AN INTRODUCTION TO STATISTICAL DATA SCIENCE

Giorgio Picci

Department of information Engineering,

University of Padova, Italy

Contents

Preface

0.1 ▪ Statistical Learning, Data Science and System Identification

In a modern technologically driven society the demand for accurate predictions and rational decision making is becoming more and more compelling. In Engineering, but also in many other sectors of the human society, predictions and rational decision making should be made based on **quantitative (i.e mathematical)** models much more than based on a single person intuition and experience as it could happen in the past. Therefore mathematical modeling of Engineering, Biological, Finance, Metheorology, and many other kind of systems is becoming of paramount importance. Due to the poor knowledge or unavailability of the underlying first principles, of the physical parameters and, especially, due to uncertainty and measurement errors and imprecision (called *noise* in general terms) inherent in almost all Biologica, Economic or Engineering processes, these models need, most of the times face uncertainty and comprise a description of uncertainty margins. Hence it is natural to use a probabilistic language.

Probabilistic model building from uncertain data used to be the realm of Statistical Science. The new ingredients which have recently entered the scenery are ultrafast and essentially unlimited computing power and the need to treat *Big Data*. We are facing an ever increasing amount of data in electronic form. Just think of Computer Vision and Image processing. Most systems are now susceptible to on-line measurement and data acquisition at speeds and with storage capacity which were unthinkable just a couple of decades ago. All of this data need to be interpreted and processed for the purpose of probing or classification and especially for prediction and rational decision making. This has led to a revisitation of Statistics with a new emphasis on algorithms, predictivity and decision making, much more than data explanation and parameter estimation as it was in the past. This is nowadays called *Statistical Learning* or even *Machine Learning*, the term "machine" being of course a captivating buzzword for "algorithmic".

In a typical scenario we have observable outcomes of a measurement process which may be quantitative (such as voltage or stock prices) but may also be categorical; i.e. alternatives in finite set such as presence or absence of signal in a communication channel. These are called effects or *outputs*. There is a set of variables which are called *inputs* which are also observable but play the role of a *cause* originating the output. For example the angle of the steering wheel and the pressure on the gas pedal as originators of the trajectory followed by a car.

There is a *Training Set* consisting of a data-base of measured input-output pairs from which one should be able to *learn* a mathematical model of the system. In very rough terms, this model can be seen just as a function mapping decision inputs into corresponding predicted outputs. The key issue being extracting (i.e. learning) such a function from the measured data,

Naturally to make the problem solvable one should have some a priori class of reasonable candidate models to choose from. These may sometimes be generically called "functions" but in some circles models are called “concepts”. One should then learn “concepts” from the data collected in the training set. Wording in this field has unfortunately become a kind of advertising game which can be misleading at times. Nowadays for example, fashionable families of functions are the so-called *Neural Nets*. These Neural Nets have nothing to do with brain or intelligence or whatever biological apparatus one may try to associate to them.They are just extremely simple mathematical functions very easy to fit to observed data, which can be combined together to arbitrary complexity. With very complicated Neural Nets one is supposed to achieve "Deep Learning".

In general, one should find a best mathematical model in the given class to describe the observed training data. This is a purely mathematical problem and of course the basic step once the variables have been coded mathematically in a proper way. The scope of learning is then to *predict* a next output from the knowledge of a current observed input (which is not in the training set) and *using the model just "learned" from processing the data in the training set*. Naturally the key point here is that this modeling (i.e. learning) process should be the result of **automatic procedures or algorithms** which are implemented in a computer. This seems to be the main outgrowth of traditional Statistics.

Time and modeling of Dynamic phenomena

Consider a physical/economic/biological system, say a paper machine, an electrical power plant or the stock exchange market in Bulgaria. By a “model” of the system we may often mean a mathematical description linking the *temporal behaviour* of certain observed variables of the system. In general the models that are needed to describe the temporal behaviour of systems must involve time and describe time variations, that is to say be *dynamical models*. In a deterministic setting dynamical models could be, say, difference or differential equations. For example Newton’s law is really a differential equation describing the temporal evolution of a mechanical system. Specifically, aimed at predicting the trajectory of a heavy body caused by the action of external forces.

In a dynamic setting both the input and output variables, once sampled and collected in the training data set, are called *time series*. These are just sequences of real or vector-valued measurements and the problem is to infer from the training data set a probabilistic model of the dynamic relation linking the two variables. This automatic model building from time series data is called *(Dynamic) System Identification*.

There may be different reasons to build models. In Data Science and in particular in System Identification one does not pay much attention to models which *explain* phenomena as in Physics (whatever this word may mean); one is chiefly interested in model building for the practical purpose of prediction, decision and control of a specific system. This is a basic difference from Physics where one instead looks for universal laws which apply to a large class of sys-

tems. In Physics one wants to *understand* the basic principles governing the behaviour of the world. The attitude in statistical learning is that *a "true model" of reality does not exist*. Somehow this seems to be the distinction between *instrumentalism* (just use the model for prediction or decision) and *realism* (pretend to describe the truth) that some scientists like Vapnik [**?**] are talking about.

The importance of 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 engineering 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 have led 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. Some devices based on these ideas are now part of commercially available communication systems (cellular phones for example).

0.2 ▪ About Modeling

Mathematical modeling should ideally be based on first principles, say like differential equations derived from the laws of physics, but often the physics of many systems is not known or too complicated and one has to resort to *empirical* models based just on inference from observed data. In this course we will only concentrate on the construction and validation of empirical models.
The distinction may seem rather crude; and is actually made only for the sake of clarity. In practice there always is some knowledge of underlying physical (or economic or biological) laws which helps in choosing suitable model classes.

Variables of the system which are accessible to measurement, can be are classified as "inputs" or *exogenous or externally imposed* variables, normally denoted by the symbols $\mathbf{x}$ or $\mathbf{u}$, and "outputs" or *explained* variables which will generally be denoted by the symbol $\mathbf{y}$. Normally these variables can only be measured at discrete instants of time t and collected in a string of data which in econometric applications are called *time series* or *discrete-time signals* in communication and control engineering.

In real systems, there are always many other variables besides the preselected inputs and outputs which influence the time evolution of the system.

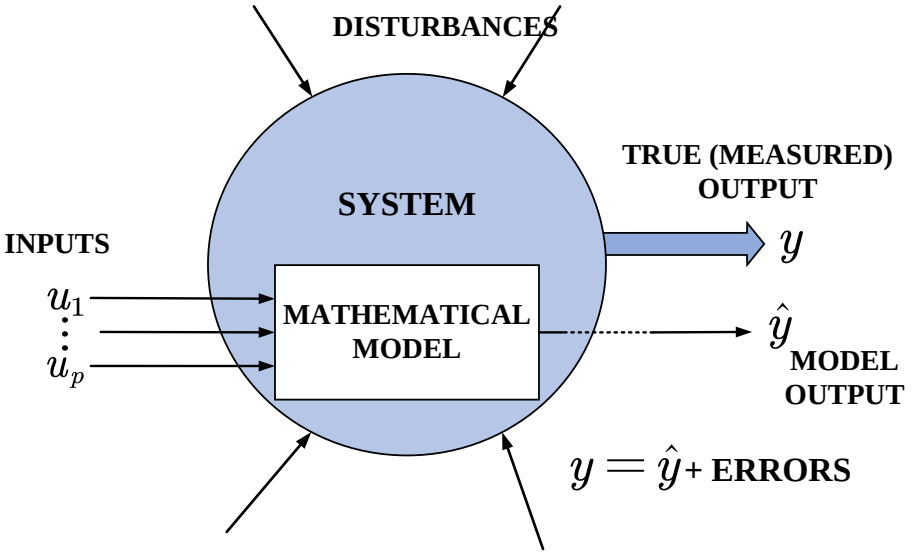


Figure 0.2.1. *System vs Model*

These other 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 $\mathbf{y}$ 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 and unknown parameters. We need to work with models of small complexity and treat the unpredictable fluctuations in some simple *aggregate* manner.

A realistic formulation of the modeling problem requires a satisfactory notion of *non-rigid*, i.e. *flexible or approximate*, notion of mathematical model of the observed data.
A model should be able to accept as legitimate, data sets (time series) which may possibly differ slightly from each another. The rationale for probabilistic models is that they fulfill precisely this request.

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 his early astronomical work *Theoria motus corporum caelestium* sect. III, p. 236 (1809).

Figure 0.2.2. *Carl Friedrich Gauss*

Example: fitting a straight line to N experimental data $\{u(t),\ y(t); t = 1, 2, \ldots, N\}$. Exact modeling means trying to match *exactly* a linear equation by solving N linear equations in the parameters a, b,

$$y(t) = au(t) + b, \qquad t = 1, 2, \ldots, N\,.$$

Can we exactly match N data points with the linear model ?
Obviously no; even if possible, the results would be extremely sensitive even to small perturbations in the data. New incoming data may change the model drastically, which means that a model determined in this way has very poor predictive capabilities.
One would say that real data obey exactly rigid relations of this kind "with probability zero". If in addition the model class is restricted to be linear finite-dimensional depending only on two parameters. A severe restriction that no real measured data can obey. In general insisting on "exact" modeling on real data leads to disastrous results. This is by now very well-known and documented in the early literature. In the language of numerical analysis, fitting rigid models to measured data invariably leads to *ill-posed problems*.

A critique to the probabilistic approach

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

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

Although in large sectors of the literature the statistical framework is often imposed dogmatically, in our opinion however, the critique originates from a tendency to confuse physical reality with mathematical models. The urn model (i.e. the underlying probability space) is just a mathematical device which describes many possible alternatives (yet statistically similar according to some probability distribution) sequences of data but is **not required to have any physical interpretation**. It could in principle be used to model systems which, may be described deterministically but would require extremely complicated mathematical models with myriads of variables.
On the same grounds it could be questioned if there are in nature objects like differential or difference equations.

Names

There is an enormous literature of books and published articles in technical journals on statistical data modeling and classification. Many such writings are sold under a variety of titles such as: Machine Learning, Data Science, Statistical Learning Theory, Neural Networks, Deep Learning, etc. Many are just centered on static Classification problems, called Pattern Recognition, and techniques called Support Vector Machines, to name just the most widespread and widely quoted nicknames. Honestly one should say that most of these names seem just

invented to make audience. For example all books titled on Machine Learning are just books building on Statistics and statistical concepts. Not to talk about *Artificial Intelligence* which nobody has ever been able to define in a reasonably precise way and seems to be a loosely defined empty box inside which one can put anything. In our opinion, it may be much better to leave the word "intelligence" to psichologists and to the literary language..

References

Without even trying to quote the enormous literature of published articles in technical journals, we shall just limit to say that there are many recent books on statistical data modeling and classification which are sold under a variety of titles. There is ample space for deepening the issues which are just touched on in these notes; for this we may refer the reader for example to [75, 44, 95, 100, 61, 65, 23].

Chapter 1

A REVIEW OF CLASSICAL STATISTICAL INFERENCE

1.1 ▪ Introduction

Modern Probability Theory is *axiomatic*. It assumes an abstract model of reality consisting of a space of elementary events Ω (for example the set of all possible outcomes of a dice throwing experiment or of a measurement process), a "σ-algebra" $\mathcal{A}$ of observable *events* (the subsets of Ω 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 Ω and the class of interesting events by a σ-algebra of subsets of Ω (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, specifying a rational 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 matter 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 assignement 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 $\mathcal{P}_i$, $i = 1, \ldots, k$ (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) .

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 $\bar{\omega} \to \mathcal{P}$, or $\bar{\omega} \to \{1, 2, \cdots, k\}$. 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 instead the toss will show tail; 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 Ω contains all sequences made of N symbols T and C in any possible order. Let $\mathcal{A}$ be the family of all subsets of Ω. 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

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

where $n(T)$ is the number of symbols T in the sequence ω. 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.

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

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

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

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

Parametric problems

The family of possible probability measures $\mathcal{P}$ (or the k classes$\mathcal{P}_i$, $i = 1, \ldots, k$) 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

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

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

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

The coin tossing problem above is parametric. Here Θ 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 $\mathbf{y} = [\mathbf{y}_1 \cdots \mathbf{y}_m]^\top$ be an m-dimensional random vector defined on the space $\{\Omega, \mathcal{A}\}$ that is, a measurable function from Ω into $\mathbb{R}^m$. The **sample space** of $\mathbf{y}$ is just the space of possible values of $\mathbf{y}$, that is some subset of $\mathbb{R}^m$, together with its Borel σ-algebra $\mathcal{B}^m$ (the smallest σ-algebra of subsets of $\mathbb{R}^m$ containing all open intervals). If P is any probability measure defined on $\mathcal{A}$, there is a corresponding *probability induced by* $\mathbf{y}$, $P_\mathbf{y}$, on its sample space $\{\mathbb{R}^m, \mathcal{B}^m\}$ which is defined in Appendix A.4. There is a "canonical" representation of a random variable on its sample space as the identity function see (A.1.3). This representation is very handy since it permits to identify $\mathbf{y}$ **just by assigning its** PDF. This is actually well-known, as one commonly speaks say about a "Gaussian random variable" of mean μ and variance σ^2, implicitly meaning that the random variable is the identity function on $\mathbb{R}$

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

defined on the sample (probability) space $\{\mathbb{R}, \mathcal{B}, P_\mathbf{y}\}$ with $P_\mathbf{y}$ defined by

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

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

Note that the sample space representation of a random variable is "sewn up" about $\mathbf{y}$ and every random variable defined on the sample space of $\mathbf{y}$, being a function of the independent variable y *is necessarily a function of* $\mathbf{y}$. Note that on the sample space of $\mathbf{y}$ there cannot exist random variables independent of $\mathbf{y}$. 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 $P \in \mathcal{P}$ ($\mathcal{P}_k$) 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

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

where the "functional form" (i.e. the analytic dependence on the independent variable y) of each F_θ 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 θ. The set Θ is the set of *admissible values* of the parameter.

The underlying conceptual scheme is that the experimental data $y = \{y_1, \ldots, y_m\}$ come from m measurement devices which are modeled as the (in general correlated) components $\mathbf{y}_1 \ldots, \mathbf{y}_m$ of an m-dimensional random variable $\mathbf{y}$ 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

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

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 , $t = 1, 2, \ldots, y_N$ all coming form experiments governed by the same PDF F. This scheme can equivalently be described as the observation of a *sequence* of random variables

$$\{\mathbf{y}_1, \ldots, \mathbf{y}_N\} \quad ,$$

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

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

where p is the fixed density of F. A result going back to Shannon [82, 83] then states that the optimal p must be the product

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

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 $\mathbf{y}_1, \ldots, \mathbf{y}_N$ 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 the (sample values of) $\mathbf{y}_1, \ldots, \mathbf{y}_N$ 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 or any $\theta \in \Theta$, as:

$$F_\theta^N(y_1, \ldots, y_N) = F_\theta(y_1), \ldots, F_\theta(y_N) \quad , \qquad y_t \in \mathbb{R}^m \quad . \tag{1.1.3}$$

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. [18].

However the situation of main interest for us is when the data are correlated, that is, the past history at time t, say $(y_1, \ldots, y_t)$ 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 fundational background for attacking this more general setup.

Definition 1.2. *Let $(\mathbf{y}_1, \ldots, \mathbf{y}_N)$ 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 ϕ, of $(\mathbf{y}_1, \ldots, \mathbf{y}_N)$, say*

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

which does not depend on the parameter θ.

Being a function of random variables, a statistic is itself a random variable, $\phi(\mathbf{y}_1, \ldots, \mathbf{y}_N)$, whose PDF can, at least in simple cases, be computed from the

joint distribution F_θ^N of the sample. Some simple examples will be presented below [1]

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

$$\bar{\mathbf{y}}_N = \frac{1}{N} \sum_{t=1}^{N} \mathbf{y}_t \quad ; \tag{1.1.4}$$

is an m-dimensonal statistics. When $\{\mathbf{y}_t\}$ is a random sample drawn from an unknown "true" PDF F_{θ_0} where $F_{\theta_0} \in \{F_\theta \ ; \ \theta \in \Theta\}$, one has $\mathbb{E}_0 \bar{\mathbf{y}}_N = \mathbb{E}_0 \mathbf{y}$, where $\mathbb{E}_0$ denotes expectation with respect to F_{θ_0}. By the law of large numbers (which will be recalled in Chap. **??**), the limit

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

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

$$\lim_{N\to\infty} \frac{1}{N} \sum_{1}^{N} y_t$$

exists for "almost all" possible sample sequences $\{y_1, y_2, \ldots, y_t, \ldots\}$ and is actually equal to the mean $\mathbb{E}_0 \mathbf{y}$ of the true distribution F_{θ_0}. This explains the origin of the name.

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

$$\hat{\mathbf{\Sigma}}_N^2 := \frac{1}{N} \sum_{t=1}^{N} (\mathbf{y}_t - \bar{\mathbf{y}}_N) \, (\mathbf{y}_t - \bar{\mathbf{y}}_N)^\top \tag{1.1.5}$$

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_{θ_0}, the limit

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

still exists with probability one, that is for "almost all" possible strings of observations $\{y_t\}$, and is equal to $\mathbb{E}_0 (\mathbf{y} - \mathbb{E}_0 \mathbf{y}) \, (\mathbf{y} - \mathbb{E}_0 \mathbf{y})^\top$ which is the variance matrix of $\mathbf{y}$ (or of F_{θ_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* θ.

[1] Often it is of interest to study the behavior of a statistic as a function of N; in particular for $N \to \infty$. For this reason a subscript N is often attached to the symbol, e.g. using a notation like ϕ_N.

Fisher vs Bayes

It is commonly recognized that there are two main philosophical approaches to statistical inference, the *Classical or Frequentist, or Fisherian*[2] and the *Bayesian* approach.

In the Fisherian approach [35] one postulates that the parameter is a deterministic but unknown quantity which by its nature could in principle be determined exactly, ideally by performing an infinite series of experiments. This viewpoint can be acceptable when θ 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 a difference equation describing the dynamics of a random signal. One classically postulates the existence of a "true" value, θ_0, of the unknown parameter indexing a "true PDF", F_{θ_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 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 θ 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

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

which is always possible provided the map $\theta \to F_\theta$ is a measurable function of θ, 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 about this. The so-called *subjectivistic school* [31] 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

[2] Frm R.A. Fisher one of the founding fathers of statistical theory.

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*). Nowdays, ultrafast computers and efficient optimization algorithms permit to apply the Bayesian philosophy to a much wider 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 deal with inference problems regarding probabilistic 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 we have a priori informations about the parameter distribution of these models. For these reasons, in this context we shall normally follow the classical approach.

On the other hand, many signal estimation problems occurring in data processing for telecommunication, automatic control, navigation and tracking and econometric forecasting (to name just a few), are more realistically modeled and solved within a Bayesian approach. For this reason we shall also devote a good part of this book to Bayesian Statistics.

1.2 ▪ Classical Theory of Parameter estimation

Consider a sample $(\mathbf{y}_1, \ldots, \mathbf{y}_N)$ 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 θ is any statistic ϕ with values in Θ. The value taken on by the random variable $\phi(\mathbf{y}_1, \ldots, \mathbf{y}_N)$ corresponding to the sample values $(y_1, \ldots, y_N)$ of $\mathbf{y}_1, \ldots, \mathbf{y}_N$,*

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

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

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, θ_0. One would in particular like that the average estimate corresponding to a large set of experimental measurements, say, $(y'_1, \ldots, y'_N)$, $(y''_1, \ldots, y''_N)$, ..., should be equal to θ_0. This condition can be expressed as

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

where $\mathbb{E}_{\theta_0}$ is the expectation operator with respect to the true PDF of the observations, $F^N_{\theta_0}$. However, since θ_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 ϕ is said to be* **(uniformly) unbiased** *if*

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

A desirable class of estimators to consider seems at first sight to be the class of (uniformly) unbiased estimators. However for certain classes of parametric PDF's unbiased estimators may not even exist. Below are some conterexamples.

Example 1.1. Let $\mathbf{y}$ be a scalar random variable having an exponential density with parameter $p > 0$, that is, let

$$\mathbb{P}\{\mathbf{y} \leq a\} = \int_0^a pe^{-pt}dt\,.$$

Assume $\phi(\mathbf{y})$ is an unbiased estimator of some function $f(p)$ of the parameter p, then it must satisfy the unbiasedness condition, that is,

$$\mathbb{E}_p\phi(\mathbf{y}) = \int_0^{+\infty} p\phi(t)e^{-pt}dt = f(p)$$

for all $p > 0$. But this equation can have a solution only if $f(p)$ is the Laplace transform of ϕ. There are plenty of functions which are not (real) holomorphic and therefore cannot be Laplace transforms. For example $f(p) = p$ would imply that the Laplace transform of ϕ should be 1, that is ϕ should be a delta distribution at $t = 0$ which is not a random variable (not a measurable function).

Example 1.2. Here is another counterexample. Let $\mathbf{y}$ be a scalar random variable having a binomial distribution with parameter p, that is, for some natural number n, let

$$\mathbf{P}\{\mathbf{y} = k\} = \binom{n}{k}p^k q^{n-k}\,, \quad k = 0, 1, \ldots, n\,.$$

Assume $\phi(\mathbf{y})$ is an unbiased estimator of some function $f(p)$ of the parameter p, then it must satisfy the unbiasedness equation, that is,

$$\mathbb{E}_p\phi(\mathbf{y}) = \sum_{k=0}^{n} \phi(k)\binom{n}{k}p^k q^{n-k} = f(p)$$

for all $0 < p < 1$. But in general this equation has no solution $\phi(k)$. Take for example $f(p) = 1/p$. This function is unbounded for $p \to 0+$ while the sum on the left must tend to zero.

More practical examples are discussed in the following.

Example 1.3. Suppose $\mathbf{y}_1, \ldots, \mathbf{y}_N$ is a random sample from a Gaussian density with unknown mean and known variance; that is $\mathbf{y}_k \sim \mathcal{N}(\theta, \sigma^2)\,; k = 1, 2, \ldots, N$ and consider the sample mean $\bar{\mathbf{y}}_N$ as an estimator of θ. This is clearly an unbiased estimator since

$$\mathbb{E}_\theta \mathbf{y}_N = \theta\,; \qquad \text{for all} \quad \theta \in \mathbb{R}\,.$$

Suppose now that the variance is the unknown parameter and let us denote it by θ^2. We want to see if the sample variance $\hat{\sigma}_N^2$ defined in (1.1.5) is also an

unbiased estimator. To this end, consider the identity

$$\sum_{k=1}^{N} (\mathbf{y}_k - \mu)^2 = \sum_{k=1}^{N} [\,(\mathbf{y}_k - \bar{\mathbf{y}}_N) + (\bar{\mathbf{y}}_N - \mu)\,]^2 =$$
$$\sum_{k=1}^{N} (\mathbf{y}_k - \bar{\mathbf{y}}_N)^2 + 2\sum_{k=1}^{N} (\mathbf{y}_k - \bar{\mathbf{y}}_N)(\bar{\mathbf{y}}_N - \mu) + N(\bar{\mathbf{y}}_N - \mu)^2 =$$
$$\sum_{k=1}^{N} (\mathbf{y}_k - \bar{\mathbf{y}}_N)^2 + N(\bar{\mathbf{y}}_N - \mu)^2$$

which holds since the sum of the deviations from the sample mean is zero. Dividing by N we find

$$\frac{1}{N}\sum_{k=1}^{N} (\mathbf{y}_k - \mu)^2 = \hat{\sigma}_N^2 + (\bar{\mathbf{y}}_N - \mu)^2$$

and taking expectations on both sides we get $\theta^2 = \mathbb{E}_{\theta^2}\, \hat{\sigma}_N^2 + \dfrac{\theta^2}{N}$ so that,

$$\mathbb{E}_{\theta^2}\, \hat{\sigma}_N^2 = \frac{N-1}{N}\theta^2 \tag{1.2.4}$$

which shows that the sample variance is *not an unbiased estimator*. We can however easily modify the expression to

$$\frac{N}{N-1}\hat{\sigma}_N^2 = \frac{1}{N-1}\sum_{k=1}^{N} (\mathbf{y}_k - \bar{\mathbf{y}}_N)^2 \tag{1.2.5}$$

and turn it into an unbiased estimator. The discussion should clarify the origin of the mysterious division by $N-1$ which often appears without explanation in the formulas of experimental physics.

Another rather natural request is that a good estimator should provide estimates $\phi(y_1, \ldots, y_N)$ which are tightly clustered about the their average value. In other words $\boldsymbol{\phi}$ 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 ϕ has* **(uniformly) minimum variance** *in the class $\mathcal{C}$ if the variance*

$$\mathrm{var}_\theta(\phi) := \mathbb{E}_\theta(\phi - \mathbb{E}_\theta\, \phi)^\top\, (\phi - \mathbb{E}_\theta\, \phi) \tag{1.2.6}$$

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

$$\mathrm{var}_\theta(\phi) \leq \mathrm{var}_\theta(\psi) \quad , \qquad \forall \psi \in \mathcal{C} \quad , \tag{1.2.7}$$

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

The bias-variance tradeoff

We should in any case remember that an optimal estimator should be as close as possible to the (unknown) *true value* θ_0 of the parameter. If we denote by $\mathbb{E}_0$ the expectation with respect to the true probability distribution F_{θ_0} then a natural optimality criterion measuring this distance should be the **mean square error** (MSE) of an estimator, defined as $\mathbb{E}_0\|\phi - \theta_0\|^2$. This is a natural measure of closedness of an estimator ϕ to the true parameter value even if, of course, the expectation is not computable in practice. Nevertheless the considerations which follow have a great practical significance.

Let $\mu(\theta) := \mathbb{E}_\theta[\phi]$ be the mean value of the estimator ϕ. We have

$$\begin{aligned}\mathbb{E}_\theta\|\phi-\theta\|^2 &= \mathbb{E}_\theta\|(\phi-\mu(\theta))+(\mu(\theta)-\theta)\|^2 = \\ &= \mathbb{E}_\theta\|\,\phi-\mu(\theta)\,\|^2 + \mathbb{E}_\theta\|\,\mu(\theta)-\theta\|^2 = \\ &= \mathrm{var}_\theta(\phi) + \mathrm{b}ias(\theta)^2 \end{aligned} \tag{1.2.8}$$

where $\mathrm{var}_\theta(\phi)$ is the scalar variance of the random variable ϕ with respect to any probability distribution of the family $\{F_\theta\}$. Naturally, the relation holds in particular for the true PDF, i.e. for $\theta = \theta_0$ whereby

$$\mathrm{M}SE(\phi) = \mathrm{var}_0(\phi) + \mathrm{b}ias(\theta_0)^2$$

This means that the mean square error is composed of two terms. In this sense, unbiasedness and minimum variance taken separately are only **partial conditions which alone do not imply optimality of an estimator**. The two terms can often be controlled separately and it may sometimes be better to choose a biased estimator having a reduced variance rather than insisting on unbiasedness.

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 $(\mathbf{y}_1, \ldots, \mathbf{y}_N)$, 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_θ admits a density $p(\cdot, \theta)$ which is twice differentiable with respect to θ.

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

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

for $i = 1, \ldots, p$ and fo revery $\theta \in \Theta$. In particular,

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

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

for all $i,j = 1,\ldots,p$ and fo revery $\theta \in \Theta$.

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

$$I(\theta) := \left[\mathbb{E}_\theta\ \left(\frac{\partial\ \log p(\mathbf{x},\theta)}{\partial\theta_i}\cdot\frac{\partial\ \log p(\mathbf{x},\theta)}{\partial\theta_j}\right)\right]_{i,j=1,\ldots,p} \tag{1.2.9}$$

which can also be written as

$$I(\theta) = \left[-\mathbb{E}_\theta\ \frac{\partial^2\ \log p(\mathbf{x},\theta)}{\partial\theta_i\,\partial\theta_j}\right]_{i,j=1,\ldots,p} \quad . \tag{1.2.10}$$

That (1.2.10) and (1.2.9) are equivalent follows by differentiating the identity $\int p(x,\theta)\,dx = 1$ (constant with respect to θ) termwise with respect to θ getting

$$\int_{\mathbb{R}^r} \frac{\partial p(x,\theta)}{\partial\theta_i}\,dx = 0 \quad , \qquad i = 1,\ldots,p \quad , \tag{1.2.11}$$

$$\int_{\mathbb{R}^r} \frac{\partial^2 p(x,\theta)}{\partial\theta_i\,\partial\theta_j}\,dx = 0 \quad , \qquad i,j = 1,\ldots,p \quad . \tag{1.2.12}$$

Equation (1.2.10) 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 (1.2.12).

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 θ [3],

$$\mathbf{z}_\theta := \left[\frac{\partial\ \log p(\mathbf{x},\theta)}{\partial\theta_i}\right]_{i=1,\ldots,p} \tag{1.2.13}$$

and note that

$$I(\theta) = \mathbb{E}_\theta\,\mathbf{z}_\theta\,\mathbf{z}_\theta^\top \geq 0 \quad . \tag{1.2.14}$$

where ≥ 0 means that the matrix on the left is positive semidefinite. From (1.2.11) it easily follows that $\mathbb{E}_\theta\,\frac{\partial\,\log p}{\partial\theta_i} = 0$ for all i's and so

$$\mathbb{E}_\theta\,\mathbf{z}_\theta = 0 \tag{1.2.15}$$

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* Θ *to* $\mathbb{R}^q$ *and* ϕ *be an unbiased estimator of* $g(\theta)$*. Let* $V(\theta)$ *be the variance matrix of* ϕ *and* $G(\theta)$ *the Jacobian matrix of* g*,*

$$G(\theta) = \left[\frac{\partial g_i(\theta)}{\partial\theta_j}\right]_{\substack{i\,=\,1,\ldots,q\\ j\,=\,1,\ldots,p}} \quad . \tag{1.2.16}$$

[3] This is often called *score*, an horrendous denomination which we shall avoid in this book.

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

$$V(\theta) - G(\theta)\, I^{-1}(\theta)\, G(\theta)^{\top} \geq 0 \quad , \tag{1.2.17}$$

where ≥ 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

$$\mathrm{Var}_\theta\{\phi(\mathbf{x}) - \hat{\phi}(\mathbf{x})\} = \mathrm{Var}_\theta\{\phi(\mathbf{x})\} - \mathrm{Cov}_\theta\{\phi(\mathbf{x}),\, \mathbf{z}_\theta\}\, \mathrm{Var}_\theta\{\mathbf{z}_\theta\}^{-1}\, \mathrm{Cov}_\theta\{\phi(\mathbf{x}),\, \mathbf{z}_\theta\}^{\top} . \tag{1.2.18}$$

See for example (5.9.4) or [70, p. 27].

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

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

by applying property A.3) one gets

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

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

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

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

that is,

$$\mathbb{E}_\theta\, \phi\; \mathbf{z}_\theta^{\top} = G(\theta) \quad . \tag{1.2.19}$$

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. □

Remarks

When ϕ is an unbiased estimator of θ (that is if g is the identity map) one has $G(\theta) = I\ (p \times p)$ and (1.2.17) becomes

$$V(\theta) - I(\theta)^{-1} \geq 0 \quad . \tag{1.2.20}$$

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

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

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

$$\mathrm{var}_\theta(\phi) \geq \mathrm{Tr}\ \left[I(\theta)^{-1}\right] \quad , \qquad \forall \theta \quad . \tag{1.2.21}$$

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* ϕ.

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 $\mathrm{Tr}\ [I(\theta)^{-1}]$ but nevertheless have minimum variance.

Example 1.4. Let $\mathbf{y}\sim\mathcal{N}(\theta,\sigma^2)$ be a scalar random variable with a known variance σ^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

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

Hence the variance of any unbiased estimator of θ based on a sample of size one, cannot be smaller that the variance of $\mathbf{y}$. Assume now we have a random sample of size N from the same Gaussian distribution. Now we have a random vector $\mathbf{x}=(\mathbf{y}_1,\ldots,\mathbf{y}_N)$ of dimension $r=N$ and

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

and hence

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

Since the random variables $\mathbf{y}_1,\ldots,\mathbf{y}_N$ are independent, it follows that,

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

Let us consider the sample mean

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

which has distribution $\mathcal{N}(\theta,\sigma^2/N)$. Since $\bar{\mathbf{y}}_N$ is an unbiased estimator of θ with variance σ^2/N, exactly equal to the inverse of the Fisher information, we conclude that the sample mean is the best possible estimator of θ (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*. $\diamondsuit$

Example 1.5. Let $\mathbf{y} \sim \mathcal{N}(\mu, \theta^2)$, where μ is known and $(\mathbf{y}_1, \ldots, \mathbf{y}_N)$ is a random sample from $\mathcal{N}(\mu, \theta^2)$. Consider the unbiased estimator of θ^2,

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

in Appendix A we shall see that $\frac{N\mathbf{s}_N^2}{\theta^2}$ has a *chi squared* distribution with $N-1$ degrees of freedom, which has expectation $N-1$ and variance $2(N-1)$. It follows that $\bar{\mathbf{s}}^2$ hasf variance $\frac{2\theta^4}{N-1}$. The Cramèr-Rao bound in this case is $2\theta^4/N$ and hence the variance of $\mathbf{s}_N^2$ is strictly larger than $I(\theta^2)^{-1}$. One can however show [73] that an unbiased estimator of θ^2 cannot have a smaller variance than that of $\mathbf{s}_N^2$. From this example it follows that $I(\theta)^{-1}$ is not the best possible lower bound. $\diamondsuit$

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 θ.

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*

$$K(f,p) := \int_{\mathbb{R}^r} [\log f - \log p]\, f(x)\, dx \;= \int_{\mathbb{R}^r} \log f/p \; f(x)\, dx = \mathbb{E}_f \log f/p\,; \tag{1.2.22}$$

It is immediate that $K(f,p) = 0$ if and only if $f = p$. For proving positivity we need to review one result of convex analysis.

Jensen's Inequality

Theorem 1.2. *For a real* **convex function** *f, arbitrary real numbers $x_1, \ldots, x_n$ in its domain and positive weights $a_1, \ldots, a_n$, it holds that*

$$f\left(\sum a_k x_k\right) \leq \sum a_k f(x_k),.$$

For a real **concave function** *f the inequality is reversed.*

The logarithm is **concave** so for a_k and x_k positive

$$\log\left(\sum a_k x_k\right) \geq \sum a_k \log x_k$$

passing to the limit this inequality holds for integrals in place of sums. Since $\log$ is a concave function, from the inequality above we get :

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

which holds for $g(x) > 0$ and an arbitrary probability measure μ. Apply then the inequality to the negative Kullback-Leibler distance we get

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

which proves that 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 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

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

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

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

it follows that

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

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

In the same way one can verify that

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

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 θ, it holds

$$K(\theta_0,\theta) \cong \frac{1}{2}\,\Delta\theta^\top\, I(\theta_0)\, \Delta\theta \quad ; \tag{1.2.23}$$

which says that, for small deviations $\Delta\theta$ of the parameter from the reference value θ_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.

Identifiability

There are situations in which the observations are structurally incapable of providing enough information to uniquely locate the value of the parameter θ which has generated them. A rather trivial example could be the following. Let θ be a two-dimensional parameter $[\theta_1,\theta_2]^\top$, ranging on $\Theta = \mathbb{R}^2$ and

let F_θ depend on (θ_1, θ_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 $\hat{\theta} = \left(\alpha\bar{\theta}_1, \frac{1}{\alpha}\,\bar{\theta}_2\right)^\top$, $\alpha \neq 0$, define the same PDF; that is $F_{\bar{\theta}}(x) = F_{\hat{\theta}}(x), \forall x$. 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 $\theta^\top$ and $\theta^{''}$ in Θ are said to be* indistinguishable *if $F_{\theta_1}(x) = F_{\theta_2}(x), \forall x \in \mathbb{R}^r$. Notation: $\theta^\top \simeq \theta^{''}$.*

Evidently $\simeq$ is an equivalence relation in Θ; in fact it is symmetric, reflexive and transitive. Hence it induces a partition of Θ in equivalence classes $[\theta\,] := \{\theta' \mid \theta' \simeq \theta\}$ such that $F_{\theta'} = F_{\theta''}$ if and only if θ' and θ'' 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 θ', θ'' in Θ.*

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

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 θ_0 *if there exists an open neighborhood of θ_0 which does not contain parameter values θ which are indistinguishable from θ_0 (of course, except θ_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 [49, 72, 76, 17]. 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 1.3 (Rothenberg). *Let the parametric model $\{p_\theta\,;\,\theta \in \Theta\}$ satisfy the assumptions A.1, A.2, A.3 of Section 1.2. Then θ_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 θ about the reference value θ_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 neighborhhod of θ_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. □

In the previous trivial example one has

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

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 θ. 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.

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 $L(x_0, \cdot) : \Theta \to \mathbb{R}_+$ (the nonnegative reals) defined by*

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

The "Maximum Likelihood principle", introduced by Gauss in 1856 [37] and successively popularized by R.A. Fisher, suggests to assume as estimate of θ, 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* θ. 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 θ (assuming of course that a maximum exists $\forall x_0 \in \mathbb{R}^r$) that is

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

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 θ). The resulting function of θ

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

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

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

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.

ML for a Gaussian random sample

Let $\mathbf{x} = (\mathbf{y}_1, \ldots, \mathbf{y}_N)$ 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 $x = (y_1, \ldots, y_N)$ 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 necessary conditions (1.3.4) provide the equations

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

$$\frac{\partial \ell}{\partial \theta_2^2} = -\frac{N}{2\theta_2^2} + \frac{1}{2\theta_2^4} \; \sum_{1}^{N} \; (y_i - \theta_1)^2 = 0$$

the first of which is depending only on θ_1 which yields

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

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 θ_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 (Gauss). *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 $(\mathbf{y}_1, \ldots, \mathbf{y}_N)$ are the sample mean and the sample variance.*

As we shall see, the result holds unchanged in the multivariable case. Next we give a few simple examples.

Example 1.6. The i.i.d. sample observations $\{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$ have a common Gaussian distribution $\mathcal{N}(\theta, \sigma^2)$ where the unknown mean is **nonnegative** that is $\Theta = \{0 \leq \theta \leq +\infty\}$. Compute its maximum likelihood estimator.
Solution: Recalling that by summing and subtracting $\bar{y}_N$ inside the square you get

$$\sum_{k=1}^N (y_k - \theta)^2 = \sum_{k=1}^N (y_k - \bar{y}_N)^2 + N(\bar{y}_N - \theta)^2$$

the maximization of the likelihood reduces to solving the constrained minimization

$$\min_{\theta \in \Theta} (\bar{y}_N - \theta)^2$$

which for $\bar{y}_N \geq 0$ is solved by $\hat{\theta} = \bar{y}_N$ and in case $\bar{y}_N < 0$ is solved by $\hat{\theta} = 0$. Therefore $\hat{\theta} = \max\{0, \bar{y}_N\}$.

Example 1.7. Suppose you have a random sample $\{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N\}$ drawn from a unilateral exponential distribution:

$$p(y, \theta) = \begin{cases} (1/\theta) \exp(-y/\theta) & y \geq 0 \\ 0 & y < 0 \end{cases}$$

Compute the maximum likelihood estimator of the parameter θ. Is this an unbiased estimator?

Solution: Taking the derivative of the log-likelihood function

$$\ell(\theta, y^N) = -N \log \theta - \frac{1}{\theta} \sum_{t=1}^N y_t$$

we find it is minimized by $\hat{\theta} = \bar{y}_N$ (the sample mean). To check unbiasedness let us first compute the expected value

$$\mathbb{E}\, \mathbf{y} = \int_0^{+\infty} \frac{y}{\theta} \exp(-y/\theta)\, dy = \theta$$

so that $\mathbb{E}\,\bar{\mathbf{y}}_N = \dfrac{1}{N}\, N\theta = \theta$. The estimator is unbiased.

Example 1.8. Compute the maximum likelihood estimator of the parameter θ from a random sample $\mathbf{y} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N\}$ drawn from the uniform probability density

$$p(x\,;\,\theta) = \frac{1}{\theta}\, I_{[\,0,\,\theta\,]}(x)\,, \qquad \theta > 0$$

where $I_A(x)$ is the indicator function of the set A, equal to 1 if $x \in A$ and zero otherwise.
Solution: To maximize the likelihood

$$L_N(y,\theta) = \frac{1}{(\theta)^N} \prod_{k=1}^{N} I_{[\,0,\,\theta\,]}(y_k)$$

we need to make θ as small as possible with the constraint to keep all indicator functions $I_{[\,0,\,\theta\,]}(y_k)$ equal to one; that is, we need to have $y_k \le \theta$ for all $k = 1, 2, \ldots, N$. This constraint will be satisfied if and only if $y_{Max} := \max_k \{y_k\}$ is smaller or equal to θ. Therefore the minimizer

$$\min_{\theta \ge 0} \{\, \theta \ge y_{Max} \,\}$$

is obviously $\hat{\theta}(y) = y_{Max}$. Note that this statistic depends on all variables of the sample. Draw a picture of the likelihood function $L_N(y,\cdot)$ (obviously as a function of θ).

ML estimators for the multivariate Gaussian density

Theorem 1.4. *Let $\mathbf{y}$ be an i.i.d. sample of N random vectors from a Gaussian distribution $\mathcal{N}(\mu,\Sigma)$, where $\mu \in \mathbb{R}^p$ and $\Sigma \in \mathbb{R}^{p\times p}$ are the unknown mean and variance matrix. Assume $\Sigma > 0$ and that N is large enough so that the (matrix-valued) sample variance $\hat{\Sigma}_N$ is positive definite, then $\ell_N(\mathbf{y},\mu,\Sigma)$ is maximized by the estimator $\phi = [\,\bar{\mathbf{y}}_N,\, \hat{\Sigma}_N\,]$ where $\bar{\mathbf{y}}_N$ and $\hat{\Sigma}_N$ are the sample mean and the sample variance matrix.*

Proof. We shall use the identity

$$\sum_{k=1}^{N}(y_k-\mu)^\top\Sigma^{-1}(y_k-\mu) = \sum_{k=1}^{N}(y_k-\bar{\mathbf{y}}_N)^\top\Sigma^{-1}(y_k-\bar{\mathbf{y}}_N)+\sum_{k=1}^{N}(\bar{\mathbf{y}}_N-\mu)^\top\Sigma^{-1}(\bar{\mathbf{y}}_N-\mu)$$

which holds since $2\sum_{k=1}^{N}(y_k-\bar{\mathbf{y}}_N)^\top\ \Sigma^{-1}(\bar{\mathbf{y}}_N-\mu) = 0$. Moreover

$$\sum_{k=1}^{N}(y_k-\bar{\mathbf{y}}_N)^\top\Sigma^{-1}(y_k-\bar{\mathbf{y}}_N) = \mathrm{Trace}\,\{\sum_{k=1}^{N}(y_k-\bar{\mathbf{y}}_N)(y_k-\bar{\mathbf{y}}_N)^\top\Sigma^{-1}\} = N\ \mathrm{Trace}\ \hat{\Sigma}_N\Sigma^{-1}$$

so that the negative log-likelihhod can be written

$$-\ell_N(\mathbf{y},\mu,\Sigma) = \frac{N}{2}\,\{\log\det\Sigma + \mathrm{Trace}\ \hat{\Sigma}_N\Sigma^{-1}\} + \sum_{k=1}^{N}(\bar{\mathbf{y}}_N-\mu)^\top\Sigma^{-1}(\bar{\mathbf{y}}_N-\mu)\,. \tag{1.3.7}$$

The last term is obviously minimized by taking $\mu = \bar{\mathbf{y}}_N$ irrespective of the value of $\Sigma > 0$. We are going to show the the first term is minimized with respect to Σ for $\Sigma = \hat{\Sigma}_N$. To this end we need the following lemma:

Lemma 1.1. *Let Y be a $p \times p$ symmetric positive definite matrix, then*

$$\text{Trace } Y - \log\det Y \geq p \tag{1.3.8}$$

Proof. In fact, from the spectral decomposition $Y = U^* \,\text{diag}\,\{\lambda_1, \ldots, \lambda_p\} U$ with $UU^* = I_p$ the inequality (1.3.8) is equivalent to

$$\sum_{k=1}^{p} (\lambda_k - \log\lambda_k - 1) \geq 0$$

which is true since for any positive number x it holds that $\log x \leq x - 1$. □

We shall use the lemma to prove that for all positive definite Σ's we have

$$\log\det\Sigma + \text{Trace } \hat{\Sigma}_N\Sigma^{-1} \geq \log\det\hat{\Sigma}_N + \text{Trace } I_p$$

which would prove the minimal property of $\hat{\Sigma}_N$. In fact after setting $Y := \hat{\Sigma}_N\Sigma^{-1}$ the above inequality can be rewritten exactly as in (1.3.8). □

Properties of ML estimators

A ML estimator is not necessarily unbiased. For example the ML estimator of θ_2^2 in example 1.3 is biased. This is a consequence of the formula

$$N\hat{\sigma}_N^2(\mathbf{y}) = \sum_1^N (\mathbf{y}_i - \theta_1)^2 - N\,(\bar{\mathbf{y}}_N - \theta_1)^2 \quad . \tag{1.3.9}$$

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 (1.3.9) and recalling that $\bar{\mathbf{y}}_N \sim \mathcal{N}(\theta_1, \theta_2^2/N)$ one finds

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

and hence

$$\mathbb{E}_\theta\,\hat{\sigma}_N^2 = \theta_2^2\,\frac{N-1}{N} \tag{1.3.10}$$

which shows that $\hat{\sigma}_N^2$ is biased with a systematic error equal to θ_2^2/N.

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

Theorem 1.5 (Invariance principle). *Let g be a function from Θ to some multidimensional interval $\Gamma \subset \mathbb{R}^k$, ($k$ finite). If $\hat{\theta}(\mathbf{x})$ is the M.L. estimator of θ, 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 [113]. Let g^{-1} be the inverse of g and define

$$\tilde{\ell}(x,\gamma) = \ell\Big(x,\, g^{-1}(\gamma)\Big) = \ell(x,\theta)\Big|_{\theta=g^{-1}(\gamma)} \tag{1.3.11}$$

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

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

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

It is then clear that if $\hat{\theta}$ is an unbiased estimator of θ, $g(\hat{\theta})$ 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.

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

unless g is linear.

Remarks 1.1. The invariance principle comes out handy when one needs to evaluate the statistical properties of parameter estimates, for example when the pdf of estimators is of interest. In fact even if an estimator of a parameter θ, say $\hat{\theta}(\mathbf{y}_1,\ldots,\mathbf{y}_N)$, is guaranteed to produce values which are close (in a probabilistic sense) to the true parameter value θ_0, since $\hat{\theta}$ is obviously a random variable, the "closedness" to the true value can in practice only be evaluated by means of statistical parameters such as the pdf or the variance of the estimate. Unfortunately the variance of $\hat{\theta}$ is very often itself a function of the unknown parameter θ, say $\sigma^2(\theta)$, where the "true " value of θ is still unknown. This fact may at the outset render the expressions of the variance provided by the statistical theory, practically useless.
When however the estimate $\hat{\theta}$ is maximum likelihood, a maximum likelihood estimate of the variance say $\hat{\sigma}^2(\theta)$ is directly provided by the invariance principle as

$$\hat{\sigma}^2(\theta) = \sigma^2(\hat{\theta})\,.$$

For example, let $\mathbf{y} \sim \mathcal{N}(\mu,\theta_2^2)$ and let $\hat{\theta}_2^2(\mathbf{y}_1,\ldots,\mathbf{y}_N)$ be the M.L. estimator of the parameter θ_2^2 based on a sample of size N. As it will be shown in the next section, the variance of $\hat{\theta}_2^2$, is

$$s_N^2 := \mathrm{var}\,\{\hat{\theta}_2^2\} \;=\; 2(\theta_2^2)^2\frac{N-1}{N^2} \quad , \tag{1.3.12}$$

which clearly still depends on θ_2^2. This formula, as such, does not look very useful as it depends on the unknown parameter θ_2^2. Using the invariance principle we can however provide an explicit expression for the ML estimate of the variance s_N^2, namely

$$\hat{s}_N^2 = 2(\hat{\theta}_2^2)^2 \frac{N-1}{N^2} \,. \tag{1.3.13}$$

Example 1.9 (Cramèr).

In many applications the variables are by nature nonnegative as for example concentrations, densities, prices, returns, etc. and cannot be modeled by Gaussian distributions. In these cases one often takes logarithms and assumes Gaussianity. Of course at the end of the calculations with Gaussian distributions one needs to go back to the original variables and compute their means, variances etc. The following definition is the key to operate in these situations.

A scalar nonnegative random variable $\mathbf{y}$ has a *log normal distribution* if $\log \mathbf{y} \sim \mathcal{N}(\mu, \sigma^2)$ or, more generally, if for some $a \geq 0$, $\log(\mathbf{y} - a) \sim \mathcal{N}(\mu, \sigma^2)$. Obviously in this last case it should be that $\mathbf{y} \geq a$ (w.p.1). Several examples of log-normal variables can be found in the classical book by Cramèr [22, pag. 219–220]). We just note that

Proposition 1.2. *The product and power with deterministic exponent of log-normal random variables is still log-normal.*

Let for example $\mathbf{x}$, $\mathbf{y}$, $\mathbf{z}$ be log-normal scalar random variables. Then the random variable $\mathbf{w} := \mathbf{x}^\alpha \mathbf{y}^\beta \mathbf{z}^\gamma$ where α, β, γ are real numbers is still log-normal. In fact the logarithm of $\mathbf{w}$ is a linear combination of the logarithms of the components and therefore has still a Gaussian distribution.

It is relatively easy to check that the density of a log-normal random variable has the expression

$$p(y) = \frac{1}{2\pi\,\sigma(y-a)} \;\exp -\frac{1}{2\sigma^2}\;\left[\log\left(y-a\right) - \mu\right]^2 \quad . \tag{1.3.14}$$

Suppose we want to find the ML estimate of the parameters of a log-normally distributed random variable $\mathbf{y}$ (with $a = 0$) from an i.i.d. sample of size N. One may take the logarithms of the random sample say, $\mathbf{x}_1 = \log \mathbf{y}_1, \ldots, \mathbf{x}_N = \log \mathbf{y}_N$ which are Gaussian and from these compute the ML estimates of the parameters θ_1 and θ_2^2 of the common (Gaussian) distribution of $\log \mathbf{y}$. Let ξ and λ^2 be the mean and variance of the corresponding log-normal distribution. Their expressions in function of the Gaussian parameters are given by the formulae

$$\begin{aligned} \xi &= \exp\left(\theta_1 + \frac{\theta_2^2}{2}\right) \quad , \\ \lambda^2 &= \xi^2 \left(e^{\theta_2^2} - 1\right) \quad , \end{aligned} \tag{1.3.15}$$

which can be derived from the characteristic function $\mathbb{E}\,\{e^{t\mathbf{x}}\}$ with $\mathbf{x} \sim \mathcal{N}(\theta_1, \theta_2^2)$. Let

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

be the ML estimator of θ_1 and

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

the ML estimator of θ_2^2. By the invariance principle of ML , the ML estimators of ξ and λ^2, denoted respectively $\hat{\xi}$ and $\hat{\lambda}^2$, are

$$\hat{\xi}(\mathbf{y}_1, \ldots, \mathbf{y}_N) = \exp\left(\hat{\theta}_1 + \frac{\hat{\theta}_2^2}{2}\right) \quad ,$$

$$\lambda^2(\mathbf{y}_1, \ldots, \mathbf{y}_N) = \hat{\xi}^2 \left[\exp(\hat{\theta}_2^2) - 1\right] \quad . \tag{1.3.16}$$

♢

The Asymptotic Properties of Max Likelihood

The explicit computation of the ML estimates are straightforward for Gaussian random variables but can turn out to be very hard or impossible for general pdf's. One should also say that the properties of ML estimates for small samples are hard to figure and are in general unknown. Precise results can instead be proved for the asymptotic behavior as $N \to \infty$ (large samples) of the estimates in quite general situations. One can show that under very general assumptions the ML estimator is *consistent*, has minimum variance and has an asymptotic distribution which is *Gaussian*. These properties are clearly very important and in this section we feel obliged to provide at least a sketch of a proof of consistency.

Assume $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$ is an i.i.d. sample from a family of pdf's $\{\, p_\theta(x)\,;\, \theta \in \Theta\}$ with $x \in \mathbb{R}^n$ and $\Theta \subset \mathbb{R}^p$. After performing an experiment you observe a sequence of sample values $x := (\, x_1, \, x_2, \ldots, x_N\,)$. The likelihood function of θ corresponding to these sample values is denoted

$$L_N(\theta) = L_N(\theta \mid x) = \prod_{k=1}^{N} p_\theta(x_k)\,.$$

so that the log-likelihood function, written $\ell_N(\theta \mid x) := \log L_N(\theta \mid x)$ is the sum $\sum_{k=1}^{N} \log p_\theta(x_k)$. Recall that a maximum likelihood estimate of θ is any function $\hat{\theta}_N(x)$ such that

$$L_N(\hat{\theta}_N(x)) = \sup_{\theta \in \Theta} L_N(\theta \mid x)$$

or equivalently $\ell_N(\hat{\theta}_N(x)) = \sup_{\theta\in\Theta} \ell_N(\theta \mid x)$. This supremum (which by definition always exists) may be $+\infty$ for all x and the MLE as a function of x may not exist. It certainly exists if Θ is a compact set and $L_N(\theta)$ is continuous as a function of θ (minimum requirement: upper semicontinuous).

Suppose the sample is generated by an unknown *true value* θ_0 of the parameter and assume the model is locally identifiable about θ_0 (in practice we need to check this condition for all θ), then the KL distance

$$K(\theta_0, \, \theta) := \int_{\mathbb{R}} \log \frac{p_{\theta_0}(x_k)}{p_\theta(x_k)} p_{\theta_0}(x_k)\, dx_k = \mathbb{E}_{\,\theta_0} \log \frac{p_{\theta_0}(\mathbf{x}_k)}{p_\theta(\mathbf{x}_k)}$$

is **positive** (independent of k) and can be zero only if $\theta = \theta_0$. Consider the difference

$$\ell_N(\theta_0) - \ell_N(\theta) = \sum_{k=1}^{N} (\log p_{\theta_0}(\mathbf{x}_k) - \log p_\theta(\mathbf{x}_k)\,) = \sum_{k=1}^{N} \log \frac{p_{\theta_0}(\mathbf{x}_k)}{p_\theta(\mathbf{x}_k)}\,;$$

which, by the law of large numbers [4] has the limit ,

$$\lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} \log \frac{p_{\theta_0}(\mathbf{x}_k)}{p_\theta(\mathbf{x}_k)} = \mathbb{E}_{\theta_0} \log \frac{p_{\theta_0}(\mathbf{x}_k)}{p_\theta(\mathbf{x}_k)} = K(\theta_0,\, \theta) \,>\, 0\,,$$

with probability 1, for every $\theta \in \Theta$. Equivalently, for $N \to \infty$,

$$\frac{1}{N}\,(\ell_N((\theta_0 \mid \mathbf{x}) - \ell_N((\theta \mid \mathbf{x})) \overset{a.s.}{\to} K(\theta_0,\, \theta) > 0$$

which means that for N large and (almost) all fixed sample x, we have $\ell_N((\theta_0 \mid x) \,>\, \ell_N((\theta \mid x)$ for all $\theta \neq \theta_0$ that is θ_0 is the unique asymptotic maximizer of the log likelihood. Next, by rearranging and taking exponentials we have the asymptotic exponential decay:

$$\frac{L_N(\theta)}{L_N(\theta_0)} = O(e^{-NK(\theta_0,\,\theta)})\,; \qquad \forall\, \theta \in \Theta$$

which implies that $L_N(\theta)$ must converge (almost surely) exponentially fast to $L(\theta_0)$. Therefore, at least when Θ is a compact set, this implies that any maximum, $\hat{\theta}_N$, of $L_N(\theta)$ must converge to θ_0. In summary we have the following:

Theorem 1.6. *If Θ is compact, $p_\theta(x)$ is continuous in θ for all x and there is $K(x)$ such that*

$$\log \frac{p_\theta(x)}{p_{\theta_0}(x)} \leq K(x)\,; \qquad \mathbb{E}_{\theta_0} K(\mathbf{x}_k) < \infty \tag{1.3.17}$$

then any maximizing $\hat{\theta}_N(x)$ converges almost surely to θ_0 as $N \to \infty$.

Condition (1.3.17) serves to exclude the possibility that a maximum could shift to the boundary of Θ see [30]. It allows to generalize the result to the case when Θ is not compact. For this however we shall have to refer the reader to the literature.

By virtue of these properties, in classical Statistics the maximum likelihood principle for parameter estimation is considered to be the preferred method to construct estimators. Unfortunately in many practical situations the ML estimator can only be computed by numerical procedures.

The method of moments

The method of moments was proposed by K. Pearson [67] in the early days of statistical theory. It provides a procedure to construct estimators of certain moments of a known probability distribution based on equating the theoretical expressions of these moments as functions of the unknown parameters of

[4] See Theorm A.8 in the appendix.

the distribution, to the corresponding *sample moments* which are just functions of the observed sample. For example, in this setting the theoretical mean and variance of a distribution are equated to the corresponding sample moments, as defined for example by the formulas (1.1.4) and (1.1.5). In case of Gaussian distributions, which are parametrized by their mean and variance (θ_1, θ_2^2), this principle stipulates that the estimators of (θ_1, θ_2^2) should be the sample mean and variance of the sample. Therefore for Gaussian distributions this method is equivalent to maximum likelihood. More interestingly, even for Gaussian distributions the moments (μ, Σ) could actually be complicated functions of other system parameters of interest. This is a typical situation with identification problems, where one needs to estimate the parameters of a dynamical model say an ARMA or state space model generating the data. In such situations the the maximum likelihood principle implemented with respect to the model parameters, may well result in an impossibly complicated maximization problem. The method of moments may instead lead to a much simpler procedure than maximum likelihood.

Example Let $\{\mathbf{y}_1, \dots, \mathbf{y}_N\}$ be a random sample from a Gamma distribution of parameter $\theta = [\,\alpha,\, \beta\,]$, say

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

and zero per $x < 0$. Elementary probability calculations yield the mean value of the distribution of $\mathbf{y}_k$ equal to

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

while the second order moment is

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

The estimators of α and β by the method of moments, are therefore obtained by equation the theoretical to the sample moments; which amounts to solving the equations:

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

yielding

$$\hat{\alpha} = \frac{\bar{\mathbf{y}}_N^2}{\bar{\mathbf{m}}_N^2 - \bar{\mathbf{y}}_N^2} \qquad \hat{\beta} = \frac{\bar{\mathbf{m}}_N^2 - \bar{\mathbf{y}}_N^2}{\bar{\mathbf{y}}_N}\,. \tag{1.3.18}$$

Obviously these estimators are (as they always should) functions of the sample moments. ♢

Later on it will be shown that estimators by the method of moments are, under very mild assumptions, always consistent but in general not asymptotically efficient.

An important application of estimation by the method of moments occurs in *Subspace Identification*. This will be very shortly addressed in Example 2.5.

1.4 ▪ Hypothesis Testing

Let $\mathbf{y}$ be an N-dimensional random vector with $\mathbf{y} \sim F_\theta$, where θ is an unknown parameter ranging on some parameter space $\Theta \subset \mathbb{R}^p$. Assume we have $m+1$ disjoint subregions $\Theta_0, \Theta_1, \ldots, \Theta_m$ of Θ, each specifying an "Hypothesis" regarding the unknown distribution F_θ. We shall denote the subset $\{F_\theta,\ \theta \in \Theta_k\}$ by the symbol H_k, or simply by the index k where $k = 0, 1, \ldots, m$.

Definition 1.12. *The problem of testing the $m+1$ hypotheses*

$$H_0,\ H_1,\ \ldots,\ H_m \tag{1.4.1}$$

is that of deciding, based on some observed data y drawn from the unknown distribution $\{F_\theta\}$, which of the classes $\{H_k\}$ F_θ belongs to. In other words, a test of the family of hypotheses (1.4.1) *is a function, called the* test statistics*,*

$$\phi : \mathbb{R}^N \to \{0,\ 1,\ \ldots,\ m\}$$

associating to all possible sample values y from the random observation vector $\mathbf{y}$, just one of the candidate classes H_k.

Since the experimental evidence on which this decision needs to be based is a random sample value $\mathbf{y} = y$, even if the decison rule declares which of the H_k's is "true", this decision is, by its very nature, uncertain. The map $y \mapsto H_k$ has always attached a certain probability of making an erroneous decision.

Note that from an abstract point of view there is no conceptual difference between estimation and hypothesis testing since the map ϕ can well be regarded as an *estimator* mapping the observed sample into the finite parameter set partition, each subset Θ_k being labeled by a natural number say k taking values in $\{0, 1, \ldots, m\}$. In fact, the principle of Maximum Likelihood can, for example, equally well be applied also to a finite parameter space. However the fact that hypothesis testing deals with *finite-valued* statistics permits (sometimes) to quantify the performance of the estimator in a much sharper way by providing *error probabilities* instead of bias and variances parameters as it was done in the estimation problems discussed in the previous and following section. The two settings differ by the techniques which can be put into play to measure the performance of the estimators.

Our definition is really in the frame of parametric Fisherian Statistics. In a Bayesian setting one assumes to have also some a priori probabilities attached to the given family of hypotheses. In this chapter we shall follow the classical parametric approach assuming that nothing is a priori known about the belonging of F_θ to a particular member of the class $\{H_k\}$. The reader should be advised that there is also a large body of theory and facts about *Nonparametric Hypothesis Testing* (which is really a section of general nonparametric Statistics) where the decision to be made is about a family of distributions which cannot be parameterized by a finite dimensional vector θ. Think for example of the problem of testing if a random variable has (or not) a Gaussian distribution.

Definition 1.13. *Hypothesis H_k is said to be* simple *or* composite *whether Θ_k is a singleton or contains more than one element of Θ.*

Going back to the elementary example at the beginning of this chapter, let-

ting $\mathbf{y}$ be the sequence of binary results of N successive independent coin tosses ($\mathbf{y} = 0$ corresponding to cross and $\mathbf{y} = 1$ to heads) and $\theta = p\,;\, p \in (0,\, 1)$; the hypothesis H_0 claiming that $p = \frac{1}{2}$ is a simple hypothesis while $H_1 : \{p;\, p > \frac{1}{2}\}$ is composite.

Two Simple Hypotheses

We shall initially study the case of deciding between two simple hypotheses, H_0 and H_1. The test statistic is in this case a binary decision rule

$$\phi : \mathbb{R}^N \to \{0,\, 1\}\,.$$

Traditionally one of the two hypotheses, viz. H_0, is given a privileged role, which is probably a fact rooted in applications to the evaluation of medical treatments and decisions about their effectiveness. As we shall see, classical hypothesis testing theory is ideologically *strongly asymmetric* as it really *requires* to privilege only one of the possible decisions.

When $\phi(y) = 0$ one is said to *accept* H_0, while if $\phi(y) = 1$ one *refuses* H_0. Naturally each of these decisions has a certain probability of being wrong and the theory, in principle at least, is searching for decision rules which minimize the probability of error. Traditionally, the set of observations leading to *refuse* H_0 is called the **critical region** of the test; this is a subset of the sample space $\mathbb{R}^N$ defined as

$$\mathcal{C} := \{y \in \mathbb{R}^N \;;\; \phi(y) = 1\}. \tag{1.4.2}$$

Clearly all decision functions ϕ having the same critical region are equivalent. One may say that the a test of a simple hypothesis H_0 is completely defined by assigning its critical region.

Discriminant Analysis, Classification and Pattern Recognition: In view of applications to modern data processing systems it has become more common and intuitive to interpret hypothesis testing as a classification problem of the *observed data*. Instead of deciding about probabilistic models, people tend to think about deciding (or classifying) different kinds of *data*; for example, determine whether a given email is "spam" or "non-spam" or classify handwritten digits in postal codes as one of the first ten natural numbers $\{0, 1, \ldots .9\}$ etc.

Naturally this decision needs to be made on the basis of models; which turn in fact out to be probabilistic models. Since in general the rough data (as an e-mail message) are qualitative and as such cannot be described by quantitative models they require a preliminary processing and a *coordinatization*, which should lead to a numerical codification of the data so that at the end one should eventually be dealing with numerical observations. This step, which is more an art than a science il called *feature extraction*. It lies at the heart of successful implementations of the theory to practical problems.

One may ask if (or when) it could ever be possible to find a statistic ϕ which *discriminates exactly* two probability measures H_0 and H_1, that is a decision function ϕ such that

$$\begin{cases} \phi(\mathbf{y}) = 0 & \text{iff } \mathbf{y} \sim H_0 \\ \phi(\mathbf{y}) = 1 & \text{iff } \mathbf{y} \sim H_1 \end{cases},$$

without errors. This can happen only in degenerate situations.
Assume for simplicity that both hypotheses H_0 and H_1 are simple and the two probability measures F_{θ_0} andF_{θ_1} admit densities $f_0(y)$ and $f_1(y)$.

Lemma 1.1. *One has perfect discrimination between f_0 and f_1 if and only if f_0 and f_1 are* orthogonal*; i.e.*

$$\int_{\mathbb{R}^N} f_0(x) f_1(x)\, dx = 0\,. \tag{1.4.3}$$

Intuitively, for continuous functions: f_1 must be strictly positive when f_0 is zero and, conversely on the region where $f_1 = 0$ then f_0 must be strictly positive. The statement is illustrated in the picture below.

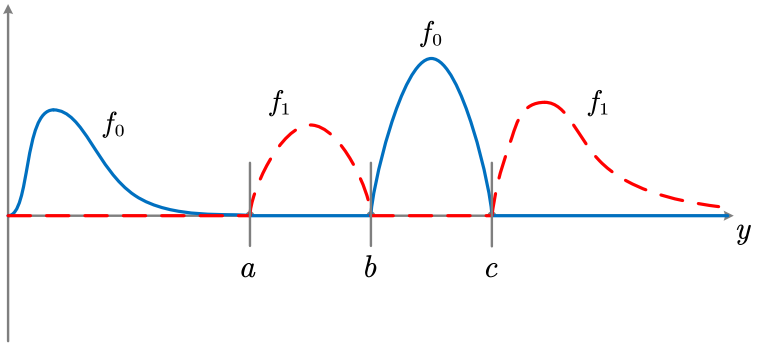


Figure 1.4.1. *Orthogonal densities*

Proof. We shall assume that f_0 and f_1 are continuous functions. Let $\mathcal{C}$ be the set on which $f_1(y) > 0$ so that $P_{\theta_1}(\mathcal{C}) = \int_{\mathcal{C}} f_1(x)\, dx = 1$. Clearly for (1.4.3) to hold, f_0 must be zero on $\mathcal{C}$; in other words, if $\mathbf{y} \sim f_0$ then y belongs to $\mathcal{C}$ with probability zero; that is

$$P_{\theta_0}(y \in \mathcal{C}) = 0\,; \qquad \text{and} \qquad P_{\theta_1}(y \in \mathcal{C}) = 1$$

Hence it is enough to pick $\mathcal{C}$ as a critical region and one has a perfect test.

For general probability measures the lemma is still valid. The general notion of orthogonality of probability measures is for example in:
Agnes Berger, On Orthogonal Probability Measures *Proceedings of the American Mathematical Society* Vol. 4, No. 5 (Oct., 1953), pp. 800-806 □

Excluding degeneracy, the situation is as follws

- If H_0 is true but $y \in \mathcal{C}$ so that $\phi(y) = 1$, one *refuses* H_0 and incurs in a so-called error of the *first kind*.

- If H_0 is false (that is $\mathbf{y} \sim H_1$) but nevertheless y belongs to the complementary of the critical set $\bar{\mathcal{C}}$; one decides for H_0 and incurs in an error of the *second kind*.

It can be pictured in tabular form as

	True Hyp.	
	H_0	H_1
Decide H_0	O.K.	II
Decide H_1	I	O.K.

(1.4.4)

These denominations may look a bit convoluted but constitute an entrenched terminology in traditional statistics. The prescribed symbol for an error of the first kind is α,

$$\alpha := \int_{\mathcal{C}} dF_0(y) = \mathbb{P}_0(\mathcal{C})$$

while β is that of incurring in an error of the second kind:

$$\beta := \int_{\overline{\mathcal{C}}} dF_1(y) = \mathbb{P}_1(\mathbb{R}^N \setminus \mathcal{C}) = 1 - \mathbb{P}_1(\mathcal{C})$$

The difference

$$1 - \beta = \mathbb{P}_1(\mathcal{C}) = \mathbb{P}(y \in \mathcal{C}) \tag{1.4.6}$$

is the probability that under H_1 the observed sample falls in the critical region (leading therefore to refuse H_0 when H_0 is actually false) is called the *power* (or discriminant power) of the test. Actually when H_1 is not a simple hypothesis so that Θ_1 contains may parameter values, the power is in general a function of θ.

The terminology used in engineering, especially in problems of signal or radar detection is more symmetric. Here H_0 and H_1 represent *absence* or *presence* of a target (in case of radar) or of a signal which may be covered by noise. One defines

$\mathbb{P}_F$: the probability of *false alarm*; deciding that the target is present when it is not present;

$\mathbb{P}_M$: probability of a *miss* that is accepting H_0, when instead the target is present;

$\mathbb{P}_D$: probability of *detection* i.e. deciding that the target is present when it is actually present.

Obviously

$$\mathbb{P}_F = \alpha, \qquad \mathbb{P}_M = \beta, \qquad \mathbb{P}_D = 1 - \beta$$

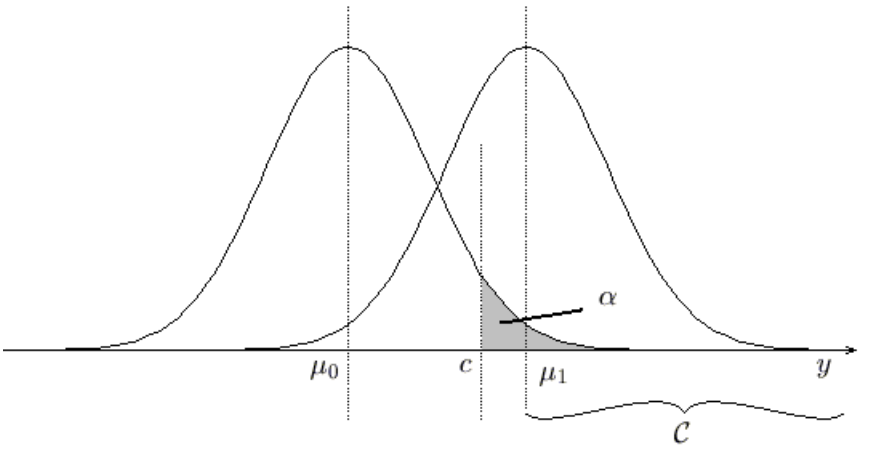
that is the probability of detection coincides with the power. Note that in certain applications (detection of hostile aircraft or diagnosis of certain diseases) a miss may be extremely costly.

In the classical setting, an *optimal test* to discriminate between H_0 and H_1 should have *both α and β as small as possibile*. Unfortunately these are conflicting objectives.

This can be seen from the simple picture below where $H_0 \sim \mathcal{N}(\mu_0, \sigma^2)$ and $H_1 \sim \mathcal{N}(\mu_1, \sigma^2)$ with $\mu_0 < \mu_1$. Suppose $N = 1$ and that the critical region $\mathcal{C}$ is an interval of the form

$$\mathcal{C} := \{y;\, y \geq c\,\}.$$

(this is actually the shape of the optimal critical region as will be proven in a few lines).



Then one clearly sees that one can make α as small as we wish by taking c very large ,

$$\alpha = \frac{1}{\sqrt{2\pi}\sigma} \int_c^{+\infty} e^{-\frac{1}{2}\frac{(y-\mu_0)^2}{\sigma^2}} dy.$$

However, the greater c, the greater will be the probability β of classifying incorrectly H_1 as H_0. In fact

$$\beta = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{c} e^{-\frac{1}{2}\frac{(y-\mu_1)^2}{\sigma^2}} dy$$

increases as c gets larger.

The classical frequentist way out of this difficulty is to *fix* α and try to find the critical region $\mathcal{C}$ *which minimizes* β, or equivalently produces the largest possible power $1-\beta$. Such a critical region is called *the best critical region* (B.C.R.) *of size* α.

In the Bayesian framework one has instead *a priori probabilities* π_0 and π_1 of the two hypotheses being true (with obviously $\pi_0 + \pi_1 = 1$) which allow to compute the *a posteriori distribution*

$$P(H_k \mid \mathbf{y}) = \frac{f_k(\mathbf{y})\pi_k}{\int_{\mathbb{R}^N} [\, f_0(y)\pi_0 + f_1(y)\pi_1 \,]\, dy}\,; \qquad k = 0, 1 \tag{1.4.7}$$

by which one can then derive a decision rule based on the standard principles of Bayesian estimation say MAP or minimum Risk, see Chapter 5.

In the classical setting the design of optimal tests for two simple hypotheses is based on the following fundamental result:

Lemma 1.2 (Neyman-Pearson). *The Best Critical Region of size* α *to discriminate between two simple hypotheses* $H_0 \equiv f_0$ *and* $H_1 \equiv f_1$ *is the set of points in the sample space* $\mathbb{R}^N$ *satisfying the inequality*

$$\mathcal{C} = \{y;\, \Lambda(y) \geq k\} \tag{1.4.8}$$

where $\Lambda(y)$ *is the likelihood ratio*

$$\Lambda(y) = \frac{f_1(y)}{f_0(y)} \tag{1.4.9}$$

the constant k being chosen in such a way that

$$\int_{\mathcal{C}} f_0(y)dy = \alpha. \tag{1.4.10}$$

The interpretation of this rule is indeed quite intuitive and in a sense follows from the maximum likelihood principle: you refuse H_0 (in fact you choose H_1) whenever the observed sample $\bar{y}$ corresponds to a probability density (actually a likelihood) $f_1(\bar{y})$ which is larger than $f_0(\bar{y})$. The value of the threshold k should hence be chosen greater or at most equal to one; $k \geq 1$.

There are versions of this Lemma dealing with the case in which H_0 and H_1 are discrete distributions or, more generally, do not admit densities. We shall refer the reader to [57] or [56] for details.

Note that by (1.4.8) $\mathcal{C}$ must contain the points y where $f_0(y) = 0$. The circumstance in which also $f_1(y)$ is zero is of no interest since these observations have probability zero under both measures and will (almost) never be observed.

Proof. Let $I_{\mathcal{C}}(y)$ be the indicatori function of the critical region $\mathcal{C}$. We want to maximize with respect to $\mathcal{C} \subseteq \mathbb{R}^N$ the quantity

$$1 - \beta = \int_{\mathcal{C}} f_1(y)dy = \int_{\mathbb{R}^N} I_{\mathcal{C}}(y)\frac{f_1(y)}{f_0(y)}f_0(y)dy = \mathbb{E}_0[I_{\mathcal{C}}\Lambda] \tag{1.4.11}$$

subject to the constraint

$$\alpha = \mathbb{E}_0[I_{\mathcal{C}}]. \tag{1.4.12}$$

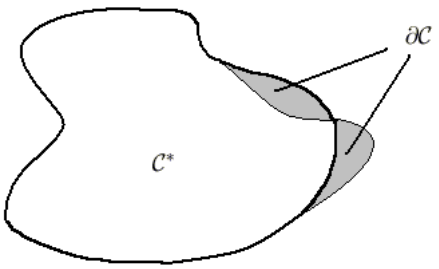
This is a so-called *free-boundary problem* in the Calculus of Variations. We shall assume that the boundary of $\mathcal{C}$ is a smooth curve.

Introducing the Lagrange multiplier λ, one needs to maximize with respect to $I_{\mathcal{C}}$

$$J(\mathcal{C}) := \mathbb{E}_0[\, I_{\mathcal{C}}\Lambda\,] - \lambda\{\mathbb{E}_0[\, I_{\mathcal{C}}\,] - \alpha\}.$$

Let $\mathcal{C}^*$ be the best critical region. Perturbing $\mathcal{C}^*$ by an infinitesimal region $\delta\mathcal{C}$ should induce a zero variation $\delta J(\mathcal{C}^*) = 0$, that is

$$\mathbb{E}_0[I_{\delta\mathcal{C}}(\Lambda - \lambda)] = \int_{\delta\mathcal{C}} [\Lambda(y) - \lambda]f_0(y)dy = 0,\ \forall\ \delta\mathcal{C},$$



from which, recalling that $f_0(y) \geq 0$, one sees that on the boundary $\mathcal{C}^*$ it should happen that

$$\Lambda(y) = \lambda \qquad\qquad y \in \partial\mathcal{C}^*.$$

so that $\Lambda(y)$ must assume a constant value $\lambda = \lambda^*$ on the boundary $\partial\mathcal{C}^*$. In particular, the multiplier λ^* must be positive, since $\Lambda(y) \geq 0$. Rewrite $J(\mathcal{C})$ as

$$J(\mathcal{C}^*) = \mathbb{E}_0[I_{\mathcal{C}^*}(\mathbf{y})(\Lambda(\mathbf{y}) - \lambda^*)] + \lambda^*\alpha$$

where the expectation can be interpreted as an inner product of a positive function (the indicator of $\mathcal{C}^*$) and the function $\Lambda(y) - \lambda$. We conclude that $\mathcal{C}^*$ can be a maximum of $J(\mathcal{C})$ only if $\mathcal{C}^*$ contains exactly the points for which $\Lambda(y) - \lambda^* \geq 0$. □

We study now the simplest example of hypothesis testing.

Example 1.10 (Comparing two means). Let $H_0 = \{\mathcal{N}(\mu_0,\, \sigma^2)\}$, $H_1 = \{\mathcal{N}(\mu_1,\, \sigma^2)\}$ both having the same variance σ^2 and known means $\mu_0 \neq \mu_1$. We observe a random sample $\mathbf{y}_1,\, \ldots,\, \mathbf{y}_N$ of size N and want to decide which of the two means is more likely to be the true one. One can rewrite the likelihood ratio

$$\Lambda(y_1,\, \ldots,\, y_N) = \exp -\frac{1}{2\sigma^2}\{\sum_1^N (y_i - \mu_1)^2 - \sum_1^N (y_i - \mu_0)^2\}.$$

by using the classical decomposition $\sum_1^N (y_t - \mu)^2 = \sum_1^N (y_t - \bar{y}_N)^2 + N(\bar{y}_N - \mu)^2$, getting

$$\begin{aligned}
\Lambda(y) &= \exp -\frac{1}{2\sigma^2}\{N(\bar{y}_N - \mu_1)^2 - N(\bar{y}_N - \mu_0)^2\} \\
&= \exp -\frac{N}{2\sigma^2}\left[(\bar{y}_N - \mu_0)^2 + 2(\bar{y}_N - \mu_0)(\mu_0 - \mu_1) + (\mu_0 - \mu_1)^2 - (\bar{y}_N - \mu_0)^2\right] \\
&= \exp -\frac{N}{2\sigma^2}[2(\bar{y}_N - \mu_0)(\mu_0 - \mu_1) + (\mu_0 - \mu_1)^2].
\end{aligned}$$

The inequality $\Lambda(y) \geq k$ is conveniently rewritten in terms of logarithms as $\log \Lambda(y) \geq \log k$, yielding

$$\frac{N}{2\sigma^2}[\, 2\, \bar{y}_N(\mu_1 - \mu_0) + \mu_1^2 - \mu_0^2] \geq \log k$$

which, in case $\mu_1 > \mu_0$ is equivalent to

$$\bar{y}_N \geq \frac{1}{2}(\mu_1 + \mu_0) + \frac{\sigma^2}{N(\mu_1 - \mu_0)} \log k$$

or, conversely, in case $\mu_0 > \mu_1$ the critical region would be defined by

$$\bar{y}_N \leq \frac{1}{2}(\mu_1 + \mu_0) - \frac{\sigma^2}{N(\mu_0 - \mu_1)} \log k\,.$$

Therefore, denoting the second member of these inequalities by c_1 or c_2; the BCR of the test is a linear half-space defined by

$$\begin{aligned}
\mathcal{C}_1 &: \{y\,;\, \bar{y}_N \geq c_1\} \quad \text{if} \quad \mu_0 < \mu_1 \\
\mathcal{C}_2 &: \{y\,;\, \bar{y}_N \leq c_2\} \quad \text{if} \quad \mu_1 < \mu_0\,.
\end{aligned}$$

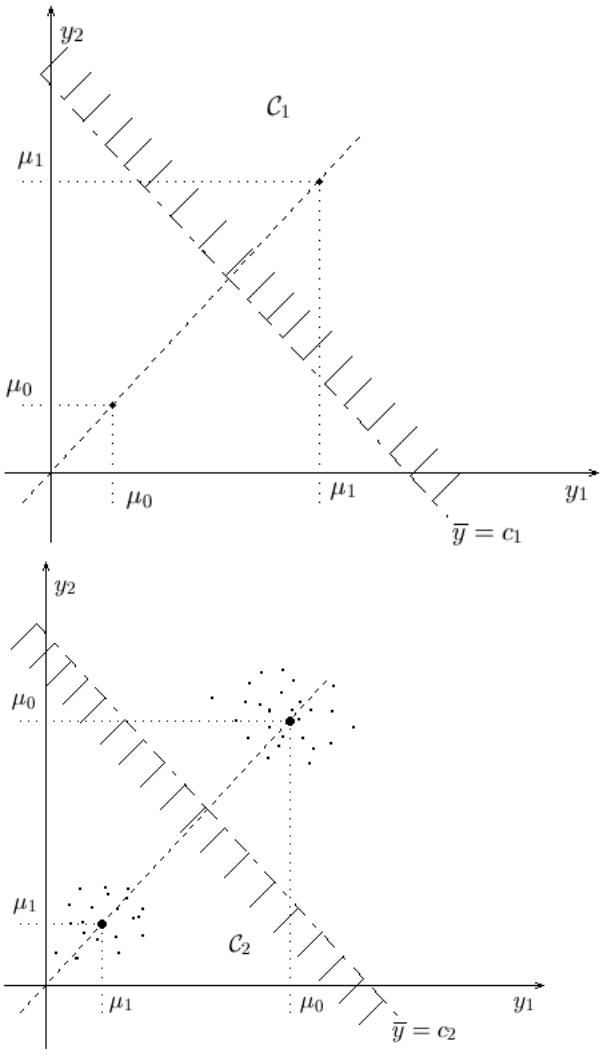


Figure 1.4.2. *Critical regions for the example* (1.10)

Note that the critical regions are defined in terms of the *sufficient statistic* $\bar{y}_N$.
Important remark: In practice, instead of using the complicated expressions for c_1 or c_2 in terms of k and the parameters of the distributions, one can argue directly to determine the boundaries as follows.
Under H_0, $\bar{\mathbf{y}}_N \sim \mathcal{N}(\mu_0,\, \frac{\sigma^2}{N})$ so that, once assigned α and supposing for example that $\mu_1 > \mu_0$, the constant c_1 can be computed from

$$\int_{c_1}^{+\infty} f_{\bar{\mathbf{y}}_N}(x)\, dx = \alpha\,.$$

Normalizing, $\frac{\sqrt{N}}{\sigma}(\bar{\mathbf{y}}_N - \mu_0) \sim \mathcal{N}(0,\, 1)$, and correspondingly changing variable

in the integral to make c_1 become

$$a_1 = \frac{\sqrt{N}}{\sigma}(c_1 - \mu_0)\,, \tag{1.4.13}$$

the critical region can conveniently be defined in terms of the normalized distribution $\mathcal{N}(0,\,1)$ by the inequality

$$\mathcal{C}_1 = \{y \mid \quad \frac{\sqrt{N}}{\sigma}(\overline{y}_N - \mu_0) \geq a_1\}\,.$$

Hence, once α is fixed, we can use the standard distribution $\mathbb{P}_0 \equiv \mathcal{N}(0,\,1)$ to find an a_1 such that

$$\mathbb{P}_0(\mathcal{C}_1) = \mathbb{P}_0(\frac{\sqrt{N}}{\sigma}(\bar{\mathbf{y}}_N - \mu_0) \geq a_1) = \alpha\,.$$

and then recover c_1 from (1.4.13) as

$$c_1 = a_1\frac{\sigma}{\sqrt{N}} + \mu_0 \tag{1.4.14}$$

In conclusion, the critical set can be computed by just resorting to the standard distribution $\mathcal{N}(0,\,1)$. ♢

There is a handful of practical examples which can be described by a similar scheme. The design of the optimal receiver in many digital communication systems implements a threshold decision of the kind just seen. In a variety of digital transmission systems the $\mathbf{y}_k$'s are samples of a continuous known waveform of period T. The transmitted signal is a series of samples of period $T_c = T/n$ which is an integer fraction of T.

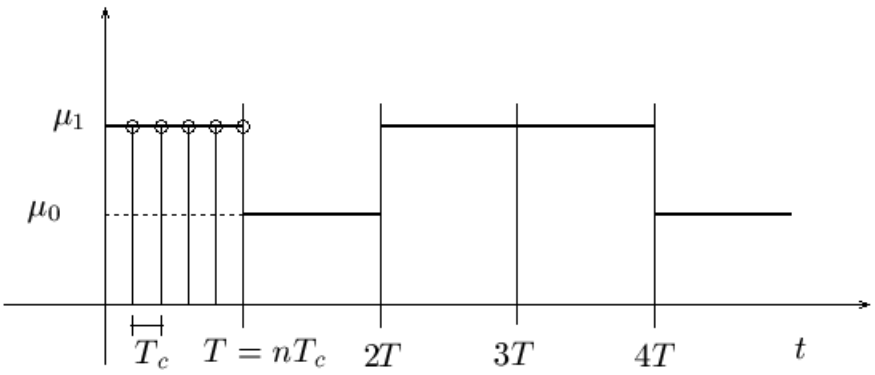


Figure 1.4.3. *Digital transmission*

After demodulation the receiver needs to process a signal $\{y_t\}$ which is the sum of the transmitted samples plus white Gaussian noise $w(t)$ of mean zero and known variance σ^2. Assuming binary waveforms and synchronous sampling, the decision problem at the receiver side can be described by two hypotheses

$$\begin{cases} y_t &= \mu_0 + w_t, \quad t = 1, \ldots, n \quad \sim \quad H_0 \\ y_t &= \mu_1 + w_t, \quad t = 1, \ldots, n \quad \sim \quad H_1 \end{cases}$$

which need to be tested on line at the end of each period T based on the last n signal samples, $y_1, \ldots, y_n$.

The receiver is designed by balancing the probability α of wrong classification with the power $1 - \beta = \mathbb{P}_D$, which is given by

$$1 - \beta = \mathbb{P}_1\{y\,;\, \overline{y} \geq c_1\}$$

where c_1 is fixed by optimizing α. In the binary transmission one can use the formula

$$\mathcal{C}_1 = \{y;\ \frac{\sqrt{n}}{\sigma}(\overline{y} - \mu_1) \geq \frac{\sqrt{n}}{\sigma}(c_1 - \mu_1)\}$$

where $\frac{\sqrt{n}}{\sigma}(\mathbf{y} - \mu_1) \sim \mathcal{N}(0,\, 1)$, under H_1.

There are standardized graphs, called *Receiver Operating Characteristic* (ROC), plotting $1 - \beta = \mathbb{P}_D$ versus $\alpha = \mathbb{P}_F$ parameterized by the ratio $d = \frac{\sqrt{n}\,|\mu_1 - \mu_0|}{\sigma}$. see e.g. [99, p 38]. They typically look like Figure 1.4.

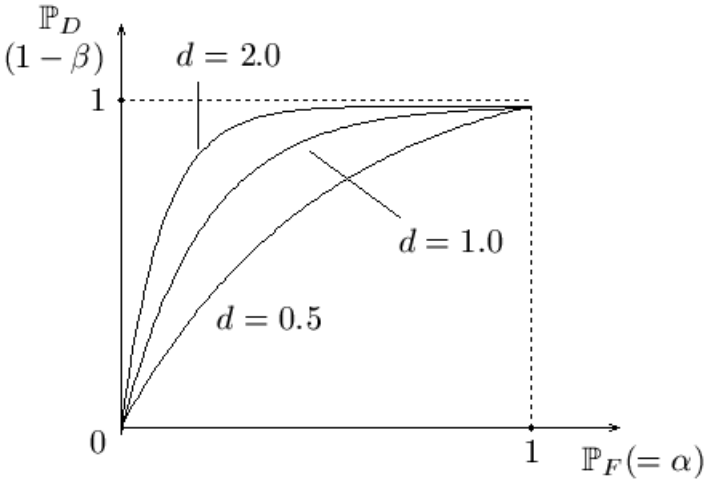


Figure 1.4.4. *Receiver Operating Characteristic (ROC).*

We have just seen that when the two hypotheses are Gaussian distributions with equal variance and the sample is i.i.d. the boundary of the critical region is a hyperplane (actually an affine variety) in the sample space. This fact holds also for

vector valued Gaussian observations and may hold also without the assumption of independent samples. The decision can always be based on checking if a certain sufficient statistics (the sample mean) of the observed sample falls on the positive or negative side of an hyperplane.

This decision rule is simple and intuitive and it has been fostered and generalized by applied statisticians to a large extent. One of the main goals being application to large amounts of data and nonlinear extensions to data which are far from being linearly separable. These classification problems are now the bulk of statistical learning (or as it is more fashionable nowadays *Machine Learning*Theory) which has generated an enormous literature, see e.g. [44, 95, 100]. We shall go back to this in Chap. 4.

Example 1.11 (Comparing variances). Consider the following decision problem: choose between two zero-mean Gaussian distributions having different variances: $H_0 = \mathcal{N}(0,\, \sigma_0^2)$, $H_1 = \mathcal{N}(0,\, \sigma_1^2)$. The observed sample consists of N i.i.d. scalar random variables, so that

$$f_i(y_1 \,\ldots\, y_N) = \frac{1}{(\sqrt{2\pi}\sigma_i)^N} \exp -\frac{1}{2}\frac{\sum_1^N y_t^2}{\sigma_i^2} \qquad i = 0,\, 1\,.$$

The likelihood ratio

$$\Lambda(y) = (\frac{\sigma_0}{\sigma_1})^N \exp\{-\frac{1}{2}(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_0^2})\sum_1^N y_t^2\}.$$

depends on the data through the sufficient statistic $\sum_1^N y_t^2$ and the critical region can be defined by the inequality $\Lambda(y) \geq k$; that is

$$\frac{1}{2}\frac{\sigma_1^2 - \sigma_0^2}{\sigma_0^2\sigma_1^2}\sum_1^N y_t^2 \geq N \log\frac{\sigma_1}{\sigma_0} + \log k\,,$$

which, assuming $\sigma_1^2 > \sigma_0^2$ turns into

$$\sum_1^N y_t^2 \geq \frac{\sigma_0^1\sigma_2^1}{\sigma_1^2 - \sigma_0^2}\left[N \log\frac{\sigma_1^2}{\sigma_0^2} + 2\log k\right] := c^2$$

The critical region is therefore the set of sample values yielding a sample second moment falling outside a circular region of radius c^2;

The threshold c^2 defining the critical region can be computed directly, without using the complicated formula above, just recalling that under H_0

$$\frac{\sum_1^N \mathbf{y}_t^2}{\sigma_0^2} \sim \chi^2(N)$$

(see the appendix Sect. A.2) so that, once α is fixed, one reads from a $\chi^2(N)$ table the value of a such that $\mathbb{P}\{\frac{\sum_1^N \mathbf{y}_t^2}{\sigma_0^2} \geq a\} = \alpha$, and obviously sets $c^2 = a\sigma_0^2$.

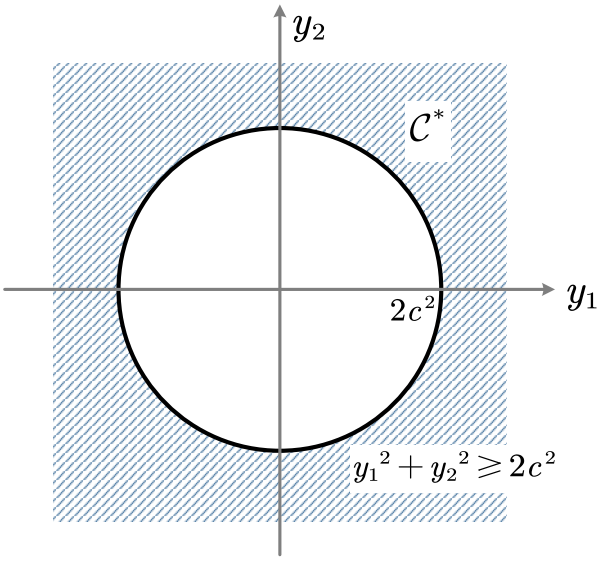


Figure 1.4.5. *Critical region for Example 1.11*

For the test power, under H_1 one has $\frac{\sum_1^N \mathbf{y}_t^2}{\sigma_1^2} \sim \chi^2(N)$ and uses an analogous procedure to get

$$\mathbb{P}_1\Big[\frac{\sum_1^N \mathbf{y}_t^2}{\sigma_1^2} \geq \frac{c^2}{\sigma_1^2}\Big] = \mathbb{P}\Big[\frac{\sum_1^N \mathbf{y}_t^2}{\sigma_1^2} \geq \frac{a\sigma_0^2}{\sigma_1^2}\Big] = 1 - \beta.$$

♢

Example 1.12 (The correlation receiver).

In this example the problem formulation is in *continuous time*. The message $\{s(t)\}$ is a train of continuous waveforms transmitted sequentially on time intervals of equal length T. It can be completely known, known up to some unknown parameter; i.e.$s(t) = s(t, \theta)$, or be a chunk of a stochastic signal whose statistics are generally assumed known. The channel distorts the signal by adding noise $w(t)$ which has zero mean and typically is modeled as a white Gaussian noise. A typical detection problem in communication engineering is to decide between the two hypotheses

$$\begin{cases} H_1 : \mathbf{y}(t) = s(t) + \mathbf{w}(t) & 0 \leq t \leq T \\ H_0 : \mathbf{y}(t) = \mathbf{w}(t) & 0 \leq t \leq T \end{cases} \tag{1.4.15}$$

which correspond to presence or absence of signal during a specific observation interval.

For simplicity assume that $s(t)$ is a known function of time and $\mathbf{w}(t)$ is white Gaussian noise of mean zero and known variance σ^2. We shall illustrate an elegant solution due to Ulf Grenander [42]. Introduce a sequence of orthonormal

functions on $L^2[0,\, T]$,

$$\phi_1(t),\, \phi_2(t),\, \ldots,\, \phi_n(t),\, \ldots, \qquad \int_0^T \phi_i(t)\phi_j(t)\, dt = \delta_{ij}.$$

where $\phi_1(t)$ is our signal normalized

$$\phi_1(t) = \frac{s(t)}{||s(\cdot)||} = \frac{s(t)}{\left[\int_0^T s^2(t)\, dt\right]^{\frac{1}{2}}} \tag{1.4.16}$$

the denominator is the square root of the energy E, of the signal.
The temporal correlation of $\mathbf{y}(t)$ and $\phi_k(t)$ is the inner product

$$\mathbf{y}_k := \langle \mathbf{y},\, \phi_k \rangle = \int_0^T \mathbf{y}(t)\phi_k(t)\, dt \qquad i = 1,\, 2,\, \ldots.$$

so that, *under* H_1 one has

$$\mathbf{y}_1 = \frac{1}{\sqrt{E}} \int_0^T s(t)s(t)\, dt + \int_0^T \mathbf{w}(t)\frac{s(t)}{E}\, dt := \sqrt{E} + \mathbf{w}_1$$

and, since $s(t)$ and $\phi_k(t)$ are orthogonal functions for $k \geq 2$

$$\mathbf{y}_k = \mathbf{w}_k := \int_0^T \mathbf{w}(t)\phi_k(t)\, dt \qquad k = 2,\, 3,\, \ldots.$$

Under H_0 one has instead

$$\begin{aligned} &\mathbf{y}_1 = \mathbf{w}_1\,, \\ &\mathbf{y}_k = \mathbf{w}_k \qquad k = 1,\, 2,\, \ldots. \end{aligned}$$

where all the variables $\{\mathbf{w}_1,\, \mathbf{w}_2\, \ldots,\, \mathbf{w}_k,\, \ldots\}$ are Gaussian independent. In fact

$$\begin{aligned} \mathbb{E}\mathbf{w}_k\mathbf{w}_j &= \mathbb{E}\int_0^T \mathbf{w}(t)\phi_k(t)\, dt \int_0^T \mathbf{w}(\tau)\phi_j(\tau)\, d\tau \\ &= \int_0^T \int_0^T \phi_k(t)\phi_j(\tau)\mathbb{E}[\mathbf{w}(t)\mathbf{w}(\tau)]\, dt\, d\tau \\ &= \int_0^T \int_0^T \phi_k(t)\phi_j(\tau)\sigma^2\delta(t-\tau)\, dt\, d\tau = \sigma^2 \int_0^T \phi_k(t)\phi_j(\tau)\, dt = \sigma^2\delta_{ij}. \end{aligned}$$

The detection problem can therefore be reformulated for a *sequence* of observations as follows

Under H_1

$$\begin{cases} \mathbf{y}_1 = \sqrt{E} + \mathbf{w}_1 \\ \mathbf{y}_k = \mathbf{w}_k \qquad k = 2,\, 3,\, \ldots \end{cases} \tag{1.4.17}$$

Under H_0

$$y_k = \mathbf{w}_k, \qquad k = 1,\, 2,\, 3,\, \ldots \tag{1.4.18}$$

where now the discrete process $\{\mathbf{w}_k\}$ is white Gaussian of mean zero and variance σ^2. Note that the random variables $\{\mathbf{y}_k\}$ are Gaussian i.i.d. under both hypotheses, with $\mathbf{y}_1$ having distributions

$$\begin{cases} \mathbf{y}_1 \sim \mathcal{N}(\sqrt{E},\, \sigma^2) & \text{under } H_1 \\ \mathbf{y}_1 \sim \mathcal{N}(0,\, \sigma^2) & \text{under } H_0 \end{cases} \tag{1.4.19}$$

while, for $k \geq 2$, under both H_1 and H_0,

$$\mathbf{y}_k \sim \mathcal{N}(0,\, \sigma^2) \tag{1.4.20}$$

It easily follows that, for all finite index n, the likelihood ratio $\Lambda(\mathbf{y}_1 \, \dots \, \mathbf{y}_n)$ is just

$$\Lambda(\mathbf{y}_1 \, \dots \, \mathbf{y}_n) = \Lambda(\mathbf{y}_1) \tag{1.4.21}$$

so that *the optimal decision is only function of the statistic* $\mathbf{y}_1$. In terms of the original data we have,

$$\begin{aligned} \Lambda(\mathbf{y}_1) &= \exp -\frac{1}{2\sigma^2}\big[(\mathbf{y}_1 - \sqrt{E})^2 - \mathbf{y}_1^2\big] \\ &= \exp -\frac{1}{2\sigma^2}\big[-2\mathbf{y}_1\sqrt{E} + E\big] \\ &= \exp \frac{1}{\sigma^2}\big[\int_0^T \mathbf{y}(t)s(t)\,dt - \frac{1}{2}\int_0^T s^2(t)\,dt\big]. \end{aligned} \tag{1.4.22}$$

This is a famous formula of continuous-time detection theory known as the *Likelihood Ratio formula*. It has been generalized to a variety of situations where $s(t)$ is unknown-deterministic or stochastic; see [81, 48][57].
The critical region (corresponding to presence of signal) is obtained by imposing

$$\log \Lambda(\mathbf{y}_1) \geq K$$

that is,

$$\mathbf{y}_1 = \int_0^T y(t)s(t)\,dt \geq \frac{1}{2}\int_0^t s^2(t)\,dt + \sigma^2 K = \frac{1}{2}E + \sigma^2 K = c.$$

Since under H_0, $\mathbf{y}_1 \sim \mathcal{N}(\sqrt{E},\, \sigma^2)$, the threshold c can be computed from $c = \frac{a}{\sigma} - \sqrt{E}$, where a is the abscissa where the graph of $\mathcal{N}(0,\, 1)$ covers an area greater than α.
The scheme of the optimal receiver is in Figure 1.4.6 below.

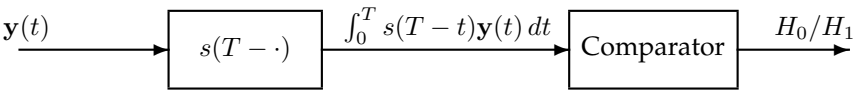


Figure 1.4.6. *Structure of the correlation receiver*

This is also called *correlation or adapted filter receiver*. The name "adapted filter" coms from the interpretation of $\mathbf{y}_1$ as the output at time T of a linear (non-causal) filter having impulse response $h(t) = s(T-t)$. ♢

1.5 ▪ Composite Hypotheses

Often H_1 (and/or H_0) is a *composite hypothesis* involving many parameter values; i.e.

$$H_i = \{f(\cdot, \theta);\, \theta \in \Theta_i\} \quad i = 0, 1$$

where Θ_i are subsets of Θ. The likelihood ratio then becomes a function of θ

$$\Lambda(y, \theta) = \frac{f(y, \theta_1)}{f(y, \theta_0)} \qquad \theta_1 \in \Theta_1,\ \theta_0 \in \Theta_0\,; \tag{1.5.1}$$

where, even if for the sake of clarity we have distinguished two parameter variables $\theta = (\theta_1, \theta_0) \in \Theta_1 \times \Theta_0$, it should be kept in mind that the parameter in our model is just one; θ. It seems natural to substitute in the expression (1.5.1) in place of θ_1, θ_0 an *estimator* $\hat{\theta}$, function of the data y, which takes into account the two admissible parameter regions defining the two hypotheses.
The natural candidate for $\hat{\theta}$ is a **maximum likelihood estimator**, this meaning that one should substitute to θ_i in the ration (1.5.1) the two ML statistics $\hat{\theta}_i$, $i = 0, 1$, which maximize $f(y, \theta_i)$ *in the respective sets* Θ_i, $i = 0, 1$. This intuitive rule can in fact be shown to have certain desirable optimality properties such as consistency, asymptotic normality etc which we shall not dig into in this book but are widely documented in the literature [56, 30].

Definition 1.14. *Let* $H_0 = \{f(\cdot, \theta);\, \theta \in \Theta_0\}$ *and* $H_1 = \{f(\cdot, \theta);\, \theta \in \Theta_1\}$. *The* Maximum Likelihood Ratio (MLR) *for the two hypotheses is the function*

$$L(y) := \frac{f(y, \hat{\theta}_1(y))}{f(y, \hat{\theta}_0(y))} \tag{1.5.2}$$

where it is assumed that the statistics $\hat{\theta}_1$, $i = 0, 1$, *are (the unique) ML parameter estimates in their respective domains* Θ_i, *that is,*

$$\hat{\theta}_1(y) := \mathrm{Arg} \max_{\{\theta \in \Theta_1\}} f(y, \theta)$$
$$\hat{\theta}_0(y) := \mathrm{Arg} \max_{\{\theta \in \Theta_0\}} f(y, \theta)\,.$$

One may then define the critical region of a MLR test exactly as in the Neyman-Pearson Lemma,

$$C := \{y;\, L(y) \geq k\}$$

where however now the constant k needs to be determined by taking into account of the fact that $\alpha = \alpha(\theta)$ is a function of $\theta \in \Theta_0$.

In general,

$$\alpha(\theta) = \int_C f(y, \theta) dy \quad \theta \in \Theta_0$$

so that, by maximizing with respect to $\theta \in \Theta_0$ and assuming that the maximum of the second member is the integral of the pointwise maximum with respect to θ, one has

$$\alpha_0 = \max_{\theta \in \Theta_0} \alpha(\theta) = \int_C f(y, \hat{\theta}_0(y)) dy\,.$$

Denoting $\hat{f}_0(y)$ the pdf $f(y, \hat{\theta}_0(y))$, if k is fixed in such a way that

$$\int_C \hat{f}_0(y)dy = \int_{\{L(y)\geq k\}} \hat{f}_0(y)dy = \int_k^\infty \hat{p}_0(l)dl = \alpha_0,$$

where $\hat{p}_0(l)$ is the pdf of the ratio $L = L(\mathbf{y})$ when $\mathbf{y} \sim \hat{f}_0(y)$, one obtains an upper bound on the error of the first kind: $\alpha(\theta) \leq \alpha_0$.

Luckily, often under H_0, $L(\mathbf{y})$ has a distribution which is independent of θ, that is, the pdf $p_0(l)$ of $L(\mathbf{y})$, when $\mathbf{y} \sim \{f(y, \theta),\, \theta \in \Theta_0\}$, does not depend on θ. In this case $p_0(\cdot)$ is just the same for *any* value of $\theta \in \Theta_0$, in particular, for the value $\hat{\theta}_0$, corresponding to the maximum of $f(y, \theta)$ in Θ_0. It follows that once fixed α, one can define the critical region $C = \{y;\, L(y) \geq k_\alpha\}$ by taking

$$\int_{k_\alpha}^\infty p_0(l)dl = \alpha\,;$$

thus getting the same probability α of incurring in a error of the first kind whatever $\theta \in \Theta_0$.

In general the pdf of $L(\mathbf{y})$ under H_1 depends on $\theta \in \Theta_1$, so that the test power

$$[1-\beta](\theta) = \int_{k_\alpha}^\infty p_1(l, \theta)dl\,,$$

$p_1(\cdot, \theta)$ being the pdf of $L(y)$ when $\mathbf{y} \sim \{f(\cdot, \theta);\, \theta \in \Theta_1\}$. This is a function of $\theta \in \Theta_1$. Note instead that in the situation discussed in the previous paragraph, when $\theta \in \Theta_0$ one has $p_1(l, \theta) \equiv p_0(l)$ and $[1-\beta](\theta)$ is independent of θ.

Example 1.13 (The "Student t" Test). Let $\mathbf{y} \sim \mathcal{N}(\mu, \sigma^2)$ where the parameters $\theta \equiv (\mu, \sigma^2)$ are unknown. Consider the hypothesis

$$H_0 :\, \mu = \mu_0\,,$$

where μ_0 is a fixed mean value. We want to verify the hypothesis H_0 *against all possible alternatives*, based on N independent observations from the parent distribution $N(\mu, \sigma^2)$. The two parameter regions are

$$\Theta_0 = \left\{\theta;\, \mu = \mu_0,\, \sigma^2 > 0\right\}$$
$$\Theta_1 = \left\{\theta;\, \mu \neq \mu_0,\, \sigma^2 > 0\right\}$$

so that Θ_1 is the open half-plane $\left\{\mu, \sigma^2 > 0\right\}$, in the parameter space, deprived of the half-line $\mu = \mu_0$.

To get the MLR we need to maximize the likelihood function

$$f(y_1, ..., y_N, \theta) = (2\pi\sigma^2)^{-\frac{N}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_1^N (y_t - \mu)^2\right\} \tag{1.5.3}$$

separately on Θ_0 and on Θ_1. On Θ_0, this just means to compute the ML estimate of σ^2 *when the mean value is known and equal to* μ_0. As we have seen,

$$\hat{\theta}_0(y) = s_N^2(y) := \frac{1}{N}\sum_1^N (y_t - \mu_0)^2$$

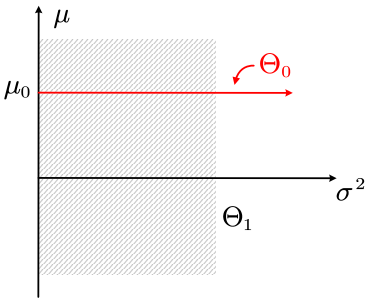


Figure 1.5.1. *Parameter spaces* Θ_0 *and* Θ_1

so that

$$f(y,\hat{\theta}_0(y)) = \left[2\pi s_N^2(y)\right]^{-\frac{N}{2}} \exp(-\frac{N}{2})\,.$$

To proceed with the maximization of $f(y,\theta)$ on Θ_1 let us first maximize on the whole set $\Theta = \left\{\mu, \sigma^2 > 0\right\}$ and then check if $\hat{\theta}_1$ may possibly lie on the line $\mu = \mu_0$. The overall maximization yields

$$\hat{\mu}(y) = \overline{y}_N = \frac{1}{N}\sum_1^N y_t$$

$$\hat{\sigma}_1^2(y) \equiv \hat{\sigma}_N^2(y) = \frac{1}{N}\sum_1^N (y_t - \overline{y}_N)^2\,.$$

Now it is obvious that $\overline{y}_N = \mu_0$ with probability zero $\forall\theta \in \Theta$, so that these estimates are also the maximizers of $f(y,\theta)$ on Θ_1. Substituting,

$$f(y,\hat{\theta}_1(y)) = \left[2\pi\hat{\sigma}_N^2(y)\right]^{-\frac{N}{2}} \exp(-\frac{N}{2})$$

and computing the MLR, one finds

$$L(y) = \left[\frac{s_N^2(y)}{\hat{\sigma}_N^2(y)}\right]^{\frac{N}{2}} = \left[\frac{\hat{\sigma}_N^2(y) + (\overline{y}_N - \mu_0)^2}{\hat{\sigma}_N^2(y)}\right]^{\frac{N}{2}}$$
$$= \left[1 + \frac{(\overline{y}_N - \mu_0)^2}{\hat{\sigma}_N^2(y)}\right]^{\frac{N}{2}}\,.$$

Note now that the random variable

$$\mathbf{t} := \frac{\overline{\mathbf{y}}_N - \mu_0}{\sqrt{\frac{\hat{\sigma}_N^2(\mathbf{y})}{N-1}}} \qquad \text{whose square is} \quad \mathbf{t}^2 := \frac{(\overline{\mathbf{y}}_N - \mu_0)^2}{\frac{\hat{\sigma}_N^2(\mathbf{y})}{N-1}} \tag{1.5.4}$$

has, under H_0 the remarkable *Student distribution* with $N-1$ degrees of freedom. Amazingly, the pdf of $\mathbf{t}$ *does not depend on the parameters* (μ,σ^2) of the parent

Gaussian distribution. See the appendix A for a definition and basic properties of the Student distribution. Since $L(\mathbf{y})$ can be expressed as

$$L(\mathbf{y}) = \left[1 + \frac{1}{N-1}\mathbf{t}^2\right]^{\frac{N}{2}} \tag{1.5.5}$$

it is clear that it depends on the data $\mathbf{y}$ only through the statistic $\mathbf{t}$. The critical region $C := \{y;\, L(y) \geq k\}$ can then equivalently be expressed as

$$C := \left\{y;\, |\mathbf{t}(y)| \geq +\sqrt{(N-1)(k^{\frac{2}{N}} - 1)}\right\} \tag{1.5.6}$$

which has the form

$$C := \{y;\, |\mathbf{t}(y)| \geq c\}\ , \quad c > 0\,.$$

Here the lucky circumstance is that under H_0, the likelihood ratio $L(\mathbf{y})$ *has a distribution which is independent of the parameters*, μ, σ^2 and hence the probability, α, of committing an error of the first kind, does in particular not depend on the unknown σ^2. For each α one can find a c_α such that

$$\int_{c_\alpha}^{\infty} p_{N-1}(t)\, dt = \frac{\alpha}{2}$$

and this value c_α defines a critical region where the probability of refusing the hypothesis $\theta = \mu_0$ (when it is true), does not depend on the unknown variance σ^2.

Under H_1, the random variable $\mathbf{t}$ has a more complicated Student distribution which now depends on μ and on σ^2 through a "non-centrality parameter"'

$$\delta = \frac{\sqrt{N}}{\sigma}(\mu - \mu_0)\,.$$

The "non-central " Student distribution $\mathcal{S}(\delta)$ is tabulated, for example in Matlab [63]. ♢

Remarks It is worth extrapolating from the simple example just discussed, some general facts. As done so far, we shall assume that the test is defined in terms of a family of pdf functions $f(y, \theta)$ depending on a p-dimensional parameter $\theta = \begin{bmatrix}\theta_1 & \dots & \theta_p\end{bmatrix}^\top$ ranging on some open set of $\mathbb{R}^p$.

The point is how to deal with an hypothesis H_0, which is defined by specifying a fixed value to some components of θ. Assume without loss of generality that one is assigning a value to the first $k(\geq 1)$ components of θ. Then, using the partitioned notation

$$\theta = \begin{bmatrix}\beta \\ \eta\end{bmatrix} \quad \beta \in \mathbb{R}^k,\, \eta \in \mathbb{R}^{p-k} \tag{1.5.7}$$

we can see H_0 as being defined by k equalities

$$H_0 := \{\theta\,;\, \beta = \beta_0\} \quad , \tag{1.5.8}$$

where b is a fixed vector in $\mathbb{R}^k$. The alternative hypothesis H_1 can then be written

$$H_1 := \{\theta\,;\, \beta \neq \beta_0\} \quad . \tag{1.5.9}$$

This parameter structure implies that either Θ_0 reduces to a point in the parameter space ($k = p$) or is an affine subspace of $\mathbb{R}^p$ having dimension k smaller than p. In any case it follows that the maximization of $f(y, \theta)$ on Θ_1 gives with probability one the same result of a maximization of $f(y, \theta)$ performed on the whole parameter space Θ. The events that $f(y, \theta)$ be maximized by functions $\hat{\theta}_i$ of the observed data which take values in a thin subspace of $\mathbb{R}^p$ have probability zero, since the maximizers inherit a continuous probability distribution function and a continuous pdf assigns probability zero to thin sets. Therefore

$$\max_{\theta \in \Theta_1} f(y, \theta) = \max_{\theta \in \Theta} f(y, \theta) \tag{1.5.10}$$

which can be read as: $\hat{\theta}_1(\mathbf{y})$ is the ordinary ML estimator $\hat{\theta}(\mathbf{y})$, of θ computed by optimization on the whole parameter space. Whence whenever H_0 and H_1 are of the form (1.5.8), (1.5.9) one can express the MLR as

$$L(y) = \frac{f(y, \hat{\theta}(y))}{f(y, \beta_0, \hat{\eta}(y))} \tag{1.5.11}$$

where $\hat{\eta}(y)$ is the "conditioned" ML estimator maximizing $f(y, \beta_0, \eta)$ with respect to η in the region Θ_0, defining the hypothesis H_0.

Clearly, since $\Theta_0 \subset \Theta$, one always has that

$$f(y, \hat{\theta}(y)) = \max_{\theta \in \Theta} f(y, \theta) \geq \max_{\theta \in \Theta_0} f(y, \theta) = f(y, \beta_0, \hat{\eta}(y))$$

and hence $L(y) \geq 1 \, \forall y \in \mathbb{R}^N$ (in (1.5.6) we implicitly assume that $k \geq 1$).
Intuitively, the larger $L(y)$, the most likely is hypothesis H_1. Ideally when $f(y, \hat{\theta}_1(y)) \gg f(y, \hat{\theta}_0(y))$, one is led to accept H_1 in the region where $\{L(y) \geq k\}$ for some prescribed k.

A Critique to the classical approach

In the classical theory of hypothesis testing, H_0 and H_1 play an unsymmetric role, H_0 describing in a sense a "priviledged" hypothesis, which is simple and very costly to refuse when it is true. The theory allows to fix a priori the probability α of refusing H_0 when H_0 is true (i.e. one can fix the probability of committing an error of the first kind). The freedom of choosing α can however lead to paradoxes. It may happen that choosing a very conservative policy (small α) to avoid a possible refusal of H_0, one may end by accepting H_0 when deciding for H_1 could be a much more reasonable choice. This is the logics behind the Neyman-Pearson Lemma whereby, although in theory one guarantees minimizing the probability β of an error of the second kind, one has no control on β. In fact, α says nothing about the probability of choosing H_0 when actually H_1 is true. In many problems in science and engineering the two hypotheses play a symmetric role and one, al least, should have a way to compare the two decisions. From this point of view, whenever a reasonable guess of probabilities $\{p_0, p_1\}$ measuring the a priori likelihood of the two hypotheses, is available, the Bayesian approach may be a sounder choice.

1.6 ▪ Problems

1-1 *Let $I(\theta)$ be the Fisher matrix relative to an arbitrary smooth 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 χ^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*

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

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.*

1-5 *Compute the Kullback-Leibler distance between the two Gaussian densities, $f \equiv \mathcal{N}(\mu\,,\, \sigma_0^2)$ and $p \equiv \mathcal{N}(\mu\,,\, \sigma^2)$. Check what happens if you invert the order of the two densities.*

1-6 *Same for the two Gaussian densities, $f \equiv \mathcal{N}(\mu_1\,,\, \sigma_0^2)$ and $p \equiv \mathcal{N}(\mu_2\,,\, \sigma^2)$.*

1-7 *Consider the linear model*

$$\mathbf{y} := \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} \mathbf{x} + \begin{bmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \end{bmatrix}$$

where $\mathbf{x}, \mathbf{e}_1, \mathbf{e}_2$ are three zero-mean mutually uncorrelated random variables of respective variances:

$$\mathrm{var}\,\{\mathbf{x}\} = 1, \quad \mathrm{var}\,\{\mathbf{e}_1\} = \lambda_1^2, \quad \mathrm{var}\,\{\mathbf{e}_2\} = \lambda_2^2\,,$$

hence the model depends on the 4-dimensional parameter

$$\theta = [a_1,\ a_2,\ \lambda_1^2,\ \lambda_2^2\,], \quad \lambda_1^2 \geq 0,\ \lambda_2^2 \geq 0\,.$$

Compute the mean and the variance matrix Σ of the vector $\mathbf{y}$ using the following notations:

$$y := \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}, \quad a = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}, \quad \Delta = \mathrm{diag}\,\{\lambda_1^2,\ \lambda_2^2\,\}$$

Study the (second-order) identifiability of the model by analyzing how Σ depends on the four parameters. Argue that identifiability must be equivalent to the map $\theta \to \Sigma$ being one-to-one. Is this the case?

1-8 *Consider again the linear model*

$$\mathbf{y} := \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} \mathbf{x} + \begin{bmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \end{bmatrix}$$

where $\mathbf{x}, \mathbf{e}_1, \mathbf{e}_2$ are three zero-mean independent Gaussian random variables of respective variances:

$$\mathrm{var}\,\{\mathbf{x}\} = 1, \quad \mathrm{var}\,\{\mathbf{e}_1\} = \lambda_1^2, \quad \mathrm{var}\,\{\mathbf{e}_2\} = \lambda_2^2\,,$$

Using the same notations write the pdf $p_\theta(y)$ of the output y of the model, depending on the 4-dimensional parameter

$$\theta = [a_1,\ a_2,\ \lambda_1^2,\ \lambda_2^2\,], \quad \lambda_1^2 \geq 0,\ \lambda_2^2 \geq 0\,.$$

Suppose you have a random sample of N (independent) measurements $\{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$. Write the log-likelihood function, and try to compute the ML estimate of θ by minimizing it.
Hint: you may use the identity

$$\sum_k y_k^\top \Sigma^{-1} y_k = \sum_k \mathrm{Trace}\,\{\Sigma^{-1} y_k y_k^\top\} = \mathrm{Trace}\,\{\Sigma^{-1} \sum_k y_k y_k^\top\} := \mathrm{Trace}\,\{\Sigma^{-1} Y\}\,.$$

where Y is a function of the data. Use the calculations in the proof of Theorem 1.4 to find the ML estimate $\hat{\Sigma}$. Then use the invariance principle to compute the estimate of θ from the theoretical expression of $\Sigma(\theta)$ by solving $\Sigma(\theta) = \hat{\Sigma}$. Could you get a unique answer?

1-9 *Suppose you have a random sample $\{\mathbf{y}_1,\ \mathbf{y}_2,\ \ldots, \mathbf{y}_N\}$ drawn from a unilateral exponential distribution:*

$$p(y,\theta) = \begin{cases} (1/\theta)\exp(-y/\theta) & y \geq 0 \\ 0 & y < 0 \end{cases}$$

Compute the maximum likelihood estimator of the parameter θ. Is this an unbiased estimator?

1-10 *Let $\mathbf{y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$ be a random (i.i.d.) sample from a Gaussian distribution having unknown mean and variance $\sigma^2 = 100$, that is $\mathbf{y}_k \sim \mathcal{N}(\theta, 100)$. We want to test the simple hypotheses*

$$\begin{aligned} H_0 &\equiv \{\theta = 2\}; \\ H_1 &\equiv \{\theta = 10\}. \end{aligned}$$

using the Neyman-Pearson Lemma. Describe the critical region $\mathcal{C}$ of the test.

Determine the number of measurements N in such a way to have $\alpha := P_0(\mathcal{C}) = 0.01$ and $P_1(\mathcal{C}) = 0.99$.
You may assume that for a Gaussian variable $\mathbf{y} \simeq \mathcal{N}(\mu, \sigma^2)$, one has $P\{\,|\mathbf{y} - \mu\,| \leq 2\sigma\,\} = 0.98$.

1-11 *The observations $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_N)$ are described by one of the two models*

$$\begin{aligned} H_0 &:\ \mathbf{y}_t = \mathbf{w}_t \\ H_1 &:\ \mathbf{y}_t = a + \mathbf{w}_t\,, \end{aligned}$$

where $a > 0$ is a known parameter and $\{\mathbf{w}_t\}$ are i.i.d. random variables described by the exponential distribution

$$p(x) = C\ \exp\{-\lambda\,|x|\} \qquad x \in \mathbb{R}\,,$$

where $\lambda > 0$ is a known parameter. Based on an observed sample of size N, we need to decide which of the two probability distributions describes the data. Find the decision rule.

1-12 *Same problem as in Example 1.8 but for a bilateral uniform density* $p(x, \theta) = \frac{1}{2\theta} I_{[-\theta, \theta]}(x)$.

1-13 *[The taxicab problem]*
You are waiting for a taxi outside of the railway station and while you wait keep notice of the number impressed on the side wall of each taxi. Assuming that that number is an enumeration of the cars owned by the taxi company, you would like to estimate the total number of taxis owned by that company. Call θ *that number and let* $\{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N\}$ *be the taxi numbers you have been taken notice of while you were waiting. Assume they are independently drawn from the uniform distribution* $U[0, \theta]$. *Find the maximum likelihood estimate of the number of taxis owned by the company.*

Chapter 2

PARAMETER ESTIMATION FOR LINEAR MODELS

A large section of classical statistics deals with the so-called *linear regression problems* which essentially include various generalizations of the antique problem of fitting a straight line to observed data points. The mathematics behind the modern formulation and solution of linear regression problems is essentially linear algebra and has little to do with probability. For this reason we shall start this chapter by discussing deterministic regression. As we shall see the mathematical apparatus brought in to solve this problem can be transferred almost verbatim to the statistical setting.

2.1 ▪ Deterministic linear least squares

The following is the simplest, yet ubiquitous problem of (parametric) model building from observed data.

Problem 2.1. *Fit, in some "reasonable " way, a parametric model of known structure to measured input-output data.*

Given: measured output data $(y_1, \dots, y_N)$, assumed real-valued for now, and "input" (or exogenous) variables $(u_1, \dots, u_N)$, both collected in N successive experiments performed on some system, and a class of candidate parametric models,

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

where the structure of the function f is assumed from a priori information and is completely determine by assigning a p-dimensional parameter vector θ. One wants to fit the observed output data in some optimal way. A reasonable and quite popular way to do so is to use a **quadratic** approximation criterion: the average squared approximation error of the observed outputs,

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

The "best" model corresponds to the value(s) $\hat{\theta}$, of θ minimizing $V(\hat{\theta})$

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

This is a simple empirical rule for constructing models from measured data. As we shall see it may come out rather naturally from statistical estimation criteria in problems where some probabilistic side information is available.
Obviously $\hat{\theta}$ depends on the data $(y_1, \ldots, y_N)$ $(u_1, \ldots, u_N)$;

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

Assuming for the moment that a unique minimizer exist, $\hat{\theta}$ is called a *Least-Squares-Estimator* of θ. No statistical significance is attached to this word.

Weighted Least Squares

It is reasonable to weight the modeling errors by some positive coefficients q_t corresponding to more or less reliable results of the experiment. This leads to *Weighted Least Squares*, criteria of the type

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

where $q_1, \ldots, q_N$ are positive numbers, which are large for reliable data and small for bad data. More generally, we may introduce a symmetric positive-definite weight matrix Q and form the criterion

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

where we have introduced vector notations

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} \qquad f(u, \theta) = \begin{bmatrix} f(u_1, \theta, 1) \\ \vdots \\ f(u_N, \theta, N) \end{bmatrix} \tag{2.1.1}$$

The minimization of $V_Q(\theta)$ can be done analytically when the model is **linear in the parameters**, that is

$$f(u_t, \theta, t) = \sum_1^p s_i(u_t, t)\, \theta_i \,, \qquad t = 1, \ldots, N \,.$$

In this problem formulation the input u_t is assumed to be a known quantity measured without errors. Therefore we can rewrite this as

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

with $s^\top(t)$ a p-dimensional row vector which is a known function of u and of the index t. Using vector notations, introducing the $N \times p$, *Signal matrix*,

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

we get the linear model class $\{\, \hat{y}_\theta = S\theta\,,\ \theta \in \Theta \,\}$ and the problem becomes to minimize with respect to θ the quadratic form

$$V_Q(\theta) = [y - S\theta]^\top\, Q[y - S\theta] = \|y - S\theta\|_Q^2\,. \tag{2.1.2}$$

The minimization can be done by elementary calculus. However it is more instructive to do this by geometric means using the *Orthogonal Projection Lemma* which in the present context is a rather intuitive condition. In Sect. 5.5 we shall provide a general statement in Hilbert space.

Make $\mathbb{R}^N$ into an inner product space by introducing the inner product $\langle x, y\rangle_Q = x^\top Q y$ and let the corresponding norm be denoted by $\|\cdot\|_Q$. Let $\mathcal{S}$ be the linear subspace of $\mathbb{R}^N$ spanned by the columns of the matrix S. Then the minimization of $\|y - S\theta\|_Q^2$ is just the minimum distance problem of finding the vector $\hat{y} \in \mathcal{S}$ of shortest distance from the data vector y. Then the minimizer of $V_Q(\theta) = \|y - S\theta\|_Q^2$ must render the error $y - S\theta$ orthogonal (according to the scalar product $\langle x, y\rangle_Q$) to the subspace $\mathcal{S}$, or, equivalently, to the columns of S, that is

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

See the picture below.

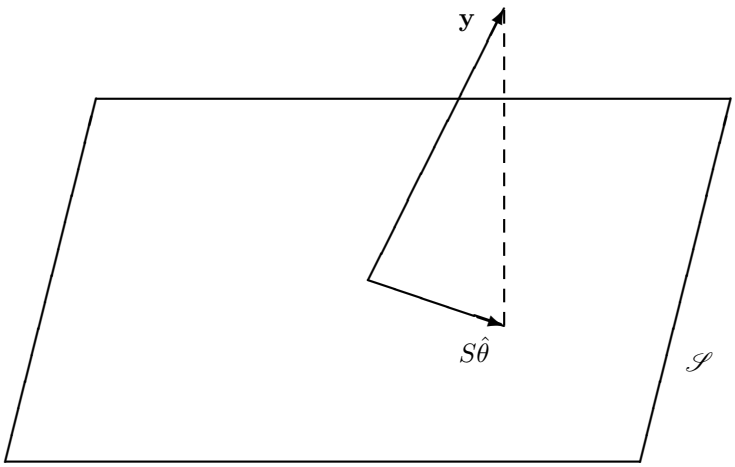


Figure 2.1.1. *Orthogonali Projectionin in* $\mathbb{R}^N$.

The orthogonality condition can be rewritten

$$S^\top Q\, S\, \theta = S^\top Q\, y\,. \tag{2.1.3}$$

These are the famous **normal equations** of the Least-Squares problem.

Let us now assume that

$$\text{rank}\ \ S = p \le N\,. \tag{2.1.4}$$

This is an *identifiability condition* of the model class. Each model corresponds $1:1$ to a unique value of the parameter. Under this condition the equation

(2.1.3) has a **unique solution** which we denote $\hat{\theta}(y)$ given by

$$\hat{\theta}(y) = [S^\top Q S]^{-1}\, S^\top Q y \quad , \tag{2.1.5}$$

which is a linear function of the observations y. For short we shall denote $\hat{\theta}(y) = Ay$. Then $S\hat{\theta}(y) := SAy$ is the orthogonal projection of y onto the subspace $\mathscr{S} = \mathrm{span}\,(S)$. In other words the matrix $P \in \mathbb{R}^{N\times N}$, defined as

$$P = SA \quad ,$$

is the orthogonal projector, with respect to the inner product $\langle\cdot,\cdot\rangle_Q$, *from* $\mathbb{R}^N$ *onto* $\mathscr{S}$. In fact P is idempotent ($P = P^2$), since

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

however P is not symmetric, as it happens with the ordinary Euclidean metric, but rather

$$P^\top = (SA)^\top = A^\top S^\top = Q\,S\,[S^\top Q\,S\,]^{-1} S^\top = Q\,S\,A\,Q^{-1} = Q\,P\,Q^{-1}\,, \tag{2.1.6}$$

so $P^\top$ is just *similar* to P. Actually, this identity just says that the projection P solving the least squares problem is a **self adjoint operator** with respect to the inner product $\langle\cdot,\cdot\rangle_Q$ as all bona-fide orthogonal projectors in general inner product spaces P should be. See the appendix B, formula (B.1.2) .

2.2 ▪ Linear Statistical Models

In the following sections we shall go back to the statistical setting where the observations are modeled as random variables. Recall that random quantities are denoted by **boldface** symbols.
Let $\mathbf{y}$ be a 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

$$\mathbf{y} = h(\theta, \mathbf{w}) \quad , \tag{2.2.1}$$

where h 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 [107] see `http://www.esat.kuleuven.be/?jwillems`. 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 the first and second order moments. In any case the *random noise* in the model (2.2.1) will be

the source of uncertainty in the relation linking the parameter θ, which is the primary object of the measurement experiment, to the observed output $\mathbf{y}$.

In principle knowledge of the model plus the PDF of the noise is equivalent to the knowledge of the distribution function $\{F_\theta \ ; \ \theta \in \Theta\}$, since one may, in principle, compute, for each fixed θ, the PDF of $\mathbf{y}$ by the well-known rules of Probability Theory. However in applied sciences and engineering it is more frequent and much more intuitive to describe the data generation mechanism by a model of the type (2.2.1). Probably the earliest example is the model used in experimental Physics called *theory of errors* which was originated by Gauss [37] while he was experimentally investigating the motion of Jupiter satellites.

Suppose one is performing measurements on a certain apparatus which can be modeled by assigning the values of a p-dimensional real parameter θ (assumed to stay constant in time) whose components are not directly accessible. In a perfectly ideal condition (or when there are no precision requirements) it should be enough to take just one measurement, since by repeating the measurement one would in principle just get the same number. It is a universally observable fact however that even if the measurements are performed with the same apparatus, one has to face the fact that the results are not the same and fluctuate slightly in an unpredictable manner. This is the main reason why it may look reasonable to take multiple measurements, performing say N successive experiments. The main question is what one should do with this bunch of numbers. What is a rational way to process these data? To answer this question one should refer to a suitable *statistical model*.

In the *theory of errors* one postulates that the individual measurements can be described as

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

where $s(\theta)$ is the ideal characteristics of the measurement instrument, which is a known function of θ and w_k is an "error" term. The question is how this quantity should be described mathematically. Gauss argues that in many measurement experiments w_k can be imagined to be a macroscopic or "aggregate" result of a large number of independent microscopic "accidental" causes which are small and their effects can be reasonably assumed to combine linearly. Gauss then argues (inventing the first known instance of a *Central Limit Theorem*) that under these conditions the possible values taken by each error variable w_k distribute according to a bell shaped probability density, which is what we now call Gaussian.

In short, the w_k's are modeled as values taken by *independent Gaussian random variables*. When there are no systematic errors the random variables $\mathbf{w}_k$ can be assumed to be zero-mean.

In this scheme the y_k are sample values of a scalar Gaussian random variable $\mathbf{y}_k$ having mean value $s(\theta)$ and variance equal to the variance of $\mathbf{w}_k$. The N observations form then a random vector $\mathbf{y} := \begin{bmatrix} \mathbf{y}_1 & \dots & \mathbf{y}_N \end{bmatrix}^\top$ which is represented in vector notation as

$$\mathbf{y} = s(\theta) + \mathbf{w} \quad , \tag{2.2.2}$$

where

$$s(\theta) = [s_1(\theta), \dots, s_N(\theta)]^\top \ , \qquad \mathbf{w} = \begin{bmatrix} \mathbf{w}_1 & \dots & \mathbf{w}_N \end{bmatrix}^\top \tag{2.2.3}$$

This model representing $\mathbf{y}$ as the sum of a deterministic "signal" plus Gaussian noise is used in a variety of applications such as for example digital communi-

cation channels. For mere notational simplicity the $\mathbf{y}_k$ have been assumed to be scalar but models describing vector valued observations are often of interest. The generalization of model (2.2.2) to describe vector valued observations is however straightforward and will be left to the reader.
The covariance matrix of the full vector $\mathbf{y}$ is clearly the same as the covariance of the noise vector

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

which also describes a possible correlation of the variables of different index and need not necessarily be diagonal. In practice, in general R may be partially unknown or poorly known. The simplest case occurs when the noise components are independent and identically distributed and R is then a scalar multiple of the identity say $R = \sigma^2 I_N$. The variance σ^2 may in general be unknown. One may then treat σ^2 (or σ) as an additional parameter to be estimated and rewrite the model as

$$\mathbf{y} = s(\theta) + \sigma\mathbf{w} \quad , \tag{2.2.4}$$

where $\mathbf{w} \sim \mathcal{N}(0, I_N)$. On the other extreme, models in which the whole noise covariance matrix is completely unknown lead to very difficult estimation problems since the whole variance needs now to be considered as an additional unknown parameter. We shall consider an intermediate situation where the noise variance is partially unknown of the form $\sigma^2 R$ with σ^2 unknown and $R = R^\top$ known and positive definite. This model can in principle be reduced to the i.i.d. noise model (2.2.4) by scaling all members of (2.2.2) multiplying from the left both members by the inverse of a square root of R; i.e.

$$R^{-1/2}\mathbf{y} = R^{-1/2}s(\theta) + R^{-1/2}\,\mathbf{w} \quad , \tag{2.2.5}$$

where $R^{1/2}(R^{1/2})^\top = R$ and $R^{-1/2}\mathbf{w} \sim \mathcal{N}(0, \sigma^2 I_N)$.

In practice however this operation is not to be recommended especially for large values of N since the explicit computation of the inverse of a square root and the scaling itself may be time consuming and numerically poorly conditioned.

In what follows we shall consider the case where $s(\theta)$ is a linear function of the parameter θ, that is

$$s(\theta) = S\theta \quad , \qquad S \in \mathbb{R}^{N\times p} \quad , \tag{2.2.6}$$

where S is a known $N \times p$ real matrix. We shall henceforth discuss parameter estimation for the *Gaussian* stochastic linear model

$$\mathbf{y} = S\theta + \mathbf{w}\,, \qquad \mathbf{w} \sim \mathcal{N}(0, \sigma^2 R) \tag{2.2.7}$$

which naturally should be compared with the deterministic least-squares model fitting of Section 2.1. Before launching into the details of statistical parameter estimation it will however be appropriate to discuss some interpretations, generalizations and limits of the model (2.2.7).

The Classical Notation

In almost all books a linear model is written

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon$$

where $\mathbf{Y}$ is the measurement (random) vector and $\mathbf{X}$ is called the *design matrix*. Both are normally written boldface which may induce the idea that both could be matrix valued of arbitrary dimensions. More dangerously, that $\mathbf{X}$ could also be a random array, a situation which in the special but frequent case of regression models (see below) leads into the setting of *Error be in Variables* problems a much more difficult setting which is so far poorly understood and cannot be dealt with by standard techniques like those discussed in this chapter.
This classical notation is, in our modest opinion, very bad, incoherent and and confusing. We shall not use it.

On regression models

In analogy to what discussed in Section 2.1, often the observations $\{y_1, \dots, y_N\}$, are the response of a physical system to external stimuli, i.e. the result of application of a sequence of *exogenous* inputs say $\{u_t \; ; \; t = 1, \dots, N \; ; \; u_t \in \mathbb{R}^q\}$ which may either be generated by some external mechanism, possibly also affected by noise, or decided by the experimenter. In this last case the u's are *known exactly* and one is really after a model relating the data $(u_t \; ; \; y_t)$ where only the outputs y are affected by measurement uncertainties. This leads to a noisy generalization of the parametric class considered earlier of the type

$$\mathbf{y}_t = f(u_t, \theta, t) + \mathbf{w}_t \,, \qquad t = 1, \dots, N \tag{2.2.8}$$

where $s(\theta, t)$ now denoted $f(u_t, \theta, t)$, is a known function, up to the assignment of a parameter value to θ, that is, a known function of $u_t, \;\; t = 1, \dots, N$ and of a p-dimensional unknown parameter θ. One says that we are *regressing y on u*. Of course when f depends linearly on θ the model is just a noisy version of the deterministic model (2.2.6) and is called a *Linear Regression model*. Here $\mathbf{w}_t$ represents the uncertainty affecting the measurement of the output variable $\mathbf{y}_t$ but may also serve as a rough description of model uncertainty at time t. Naturally this interpretation cannot have a clear probabilistic justification as was the case for the $\mathbf{w}_t$ term in the theory of errors. Note that linearity in the input variable is not required at all.

Error In Variables models

Let us observe that the structure (2.2.8) of a statistical model that we have described so far, is quite restrictive since it assumes that only the output variables y are subject to errors while the inputs in the function f or more specifically, forming the rows of the S matrix, are assumed to be known i.e. measured exactly. Often regression models are fitted routinely to the input-output data without paying much attention to this issue. In some applications this may not be realistic and may lead to questionable results.

Assuming instead that one wants to discover a model relating certain "true" output and input variables both of which are not directly accessible but only observed corrupted by additive random noise. Denoting such true variables by hatted symbols one ends up with a so-called *Error In Variables* (EIV) model

structure of the following form

$$\hat{\mathbf{y}}_t = f(\hat{\mathbf{u}}_t\,;\,\theta) \tag{2.2.9a}$$
$$\mathbf{y}_t = \hat{\mathbf{y}}_t + \tilde{\mathbf{y}}_t \tag{2.2.9b}$$
$$\mathbf{u}_t = \hat{\mathbf{u}}_t + \tilde{\mathbf{u}}_t \tag{2.2.9c}$$

where $\hat{\mathbf{y}}_t$, $\hat{\mathbf{u}}_t$ are the unobservable "true" output and input variables and $\mathbf{y}_t$, $\mathbf{u}_t$ their available observations. The subscript $_t$ is indexing repeated measurements.The terms $\tilde{\mathbf{y}}$ and $\tilde{\mathbf{u}}$ represent measurement errors which are normally assumed mutually uncorrelated or even independent and also uncorrelated or even independent of the true signals $\hat{\mathbf{y}}$, $\hat{\mathbf{u}}$. There are several variations on this model class which are discussed in the literature and we shall only briefly touch upon the linear class where the function f is just a linear map represented by some deterministic *unknown* matrix A.

In particular, assume we want to describe two scalar zero-mean random variables $\mathbf{u}$ and $\mathbf{y}$ having a positive definite covariance matrix

$$\Sigma = \begin{bmatrix} \sigma_u^2 & \sigma_{uy} \\ \sigma_{yu} & \sigma_y^2 \end{bmatrix}$$

by a scalar linear EIV model

$$\hat{\mathbf{y}}_t = a\,\hat{\mathbf{u}}_t\,; \tag{2.2.10a}$$
$$\mathbf{y}_t = \hat{\mathbf{y}}_t + \tilde{\mathbf{y}}_t \tag{2.2.10b}$$
$$\mathbf{u}_t = \hat{\mathbf{u}}_t + \tilde{\mathbf{u}}_t \tag{2.2.10c}$$

where a is an unknown parameter, $\hat{\sigma}_u^2$, $\hat{\sigma}_y^2$ are the variances of the true variables $\hat{\mathbf{u}}$ and $\hat{\mathbf{y}}$ and λ_u^2,; λ_y^2 are the variances of the corresponding additive errors. Since $\mathbb{E}\,\hat{\mathbf{y}}\,\hat{\mathbf{u}} = a\hat{\sigma}_u^2$ and $\mathbb{E}\,\hat{\mathbf{y}}^2 = a^2\hat{\sigma}_u^2$, it must hold that

$$\Sigma = \hat{\sigma}_u^2 \begin{bmatrix} 1 & a \\ a & a^2 \end{bmatrix} + \begin{bmatrix} \lambda_u^2 & 0 \\ 0 & \lambda_y^2 \end{bmatrix} \tag{2.2.11}$$

The expression on the right depends on four parameters (some of which need to be positive). A basic question is how many EIV models can represent the **same covariance matrix** Σ and therefore be indistinguishable (or equivalent) from knowledge of the joint second order statistics of the observations. It is intuitively clear that since any 2×2 covariance matrix Σ depends on three parameters there may be a multitude of EIV models representing the same random variables $\mathbf{u}$ and $\mathbf{y}$. Indeed, introduce an equivalent parameterization by setting

$$a_1 = \hat{\sigma}_u\,;\; a_2 = a\hat{\sigma}_u\,;\;\; \mathbf{a} := \begin{bmatrix} a_1 & a_2 \end{bmatrix}^\top$$

so that equation (2.2.11) is rewritten in a more symmetric form as

$$\Sigma = \mathbf{a}\mathbf{a}^\top + \begin{bmatrix} \lambda_u^2 & 0 \\ 0 & \lambda_y^2 \end{bmatrix}. \tag{2.2.12}$$

This means that by modeling with an EIV model one wants to represent the covariance $\Sigma > 0$ as the *sum of a rank deficient plus a diagonal matrix*. In particular $\mathbf{a}\mathbf{a}^\top$ being rank deficient is equivalent to the condition

$$\det \begin{bmatrix} \sigma_u^2 - \lambda_u^2 & \sigma_{uy} \\ \sigma_{yu} & \sigma_y^2 - \lambda_y^2 \end{bmatrix} = 0$$

constraining the noise variance parameters λ_u^2, λ_y^2 to satisfy the equation

$$(\sigma_u^2 - \lambda_u^2)(\sigma_y^2 - \lambda_y^2) = \sigma_{yu}^2$$

which describes an hyperbola which all feasible noise variances must belong to. Actually they need to stay only on a tract of an hyperbola

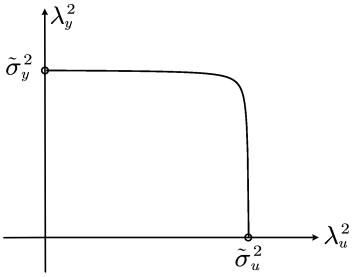


Figure 2.2.1. *EIV Hyperbola*

lying in the positive quadrant. To each feasible pair λ_u^2, λ_y^2 there corresponds a distinct EIV model with a different "true parameter" vector a obtained by factoring the matrix $\Sigma - \begin{bmatrix} \lambda_u^2 & 0 \\ 0 & \lambda_y^2 \end{bmatrix}$. The vector $\mathbf{a}$ is determined modulo a multiplicative constant of absolute value one. You may then set $\hat{\sigma}_u = a_1$ and then the unknown regression parameter a will be given by the formula

$$a = a_2/\hat{\sigma}_u \, .$$

It is clear that there are as many a's as there are feasible pairs λ_u^2, λ_y^2 lying on the hyperbola. For example one may pick $\lambda_u^2 = 0$ to get

$$\lambda_y^2 = \tilde{\sigma}_y^2 := \sigma_y^2 - \frac{\sigma_{uy}\sigma_{yu}}{\sigma_y^2} \tag{2.2.13}$$

which is the modeling error variance of a a *regression model for* $\mathbf{y}$ *in terms of* $\mathbf{u}$ *with noiseless input*[5]. The "true" a for this model is

$$a = \frac{\sigma_{yu}}{\sigma_u^2}$$

By choosing $\lambda_y^2 = 0$ one would instead get a regression model of $\mathbf{u}$ in terms of $\mathbf{y}$ and a modeling error variance $\tilde{\sigma}_u^2$ having a dual expression of (2.2.13). The so-called "total-least squares" model is obtained by picking the point $\lambda_u^2 = \lambda_y^2$ in the hyperbola and will have yet a *different regression parameter* a (which can be computed by the factorization procedure outlined above) etc. In conclusion,

[5]The expression on the right has a Bayesian modeling interpretation which will be studied in great detail in Chapter 5, in particular see Sect.5.6.

any EIV model representing the two variables $\mathbf{y}$, $\mathbf{u}$ can be fixed by, say, fixing the ratio λ_u^2/λ_y^2 of the two noise variances.

Note that this is about *stochastic modeling* and has absolutely nothing to do with statistical parameter estimation. In general an estimation method, say, least squares or total least squares or other, may be the correct approach to estimate only one of these models.

Factor Analysis (F.A.) models, to be discussed later, are related to EIV. These models involve an extra "factor " latent variable but lead to the same kind of decomposition of the observation covariance as a sum of a rank deficient plus a diagonal matrix. In certain cases and under a minimality condition, the factor variable can be eliminated leading to an EIV model. FA models belong to Bayesian modeling philosophy and we shall discuss them in Section 5.8.1. □

A distinction which is often made in the literature is between *grey box* and *black box* models. In the first category one classifies models whose structure is dictated by the physics or biology or by the economic theory etc. governing the system. Often in these models the unknown parameters have a precise physical or biological significance and the scope of the statistical procedures is just to get information about these parameters rather than discovering the system structure. Very often however the parameters enter in these models in a nonlinear fashion and standard estimation methods, say ML , can only be applied by iterative numerical algorithms.

Black-box models are instead used when the physical (or biological etc.) laws governing the phenomenon of interest are little known or uncertain and also when they maybe known but would lead to a very complicated set of equations involving a large number of unknown parameters thereby leading to identifiability issues and to very unreliable estimated models. In this case the model class is imposed by the experimenter, usually assuming the simplest parametric structure so as to allow for the application of simple estimation algorithms. Note a fundamental difference: now the model parameters are no longer the primary object of interest. The **predictive accuracy** of the model is instead the main goal of the statistical exercise. The parameter estimation criterion in this case comes out from the minimization of some measure of the *statistical prediction error*.

In Black-box estimation one may have to proceed by several trials, first trying structures which are simple and linear in the parameters. Then the theory exposed in this and following chapters becomes relevant. This "utilitaristic" philosophy of model estimation necessarily requires a successive phase of *model validation* which may well lead to increase the parametric complexity. The main principles of model validation will be discussed later in this book. Sometimes it may be necessary to resort to nonlinear models.

Examples of grey-box models are for example the state equation of a gas $pV^\gamma = kT$ where one should determine experimentally the constants γ and k from measurements of the variables p, V, T or the experimental determination of the parameters (R, L, C) of an electrical network starting from recordings of the electrical voltage at the output terminals during a discharge. While in the first example the model can be made linear in the parameters by taking logarithms, here the parameters enter nonlinearly in the time constants and in the amplitudes of the theoretical transient

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

and the estimation problem turns out to be essentially nonlinear.

2.3 ▪ Maximum Likelihood estimation of the linear model

We want to compute the ML estimates of the parameters $\theta \in \mathbb{R}^p$ and $\sigma^2 \in \mathbb{R}_+$ of the linear model (2.2.7); where $S \in \mathbb{R}^{N\times p}$ is a known matrix and $\mathbf{w}$ is a Gaussian random vector of mean zero and known variance matrix R, assumed positive definite.

Since $\mathbf{y} \sim \mathcal{N}(S\theta, \sigma^2 R)$ the log-likeliohood function is readily obtained as

$$\begin{aligned}\ell(y,\theta,\sigma^2) &= -\frac{N}{2}\,\log 2\pi - \frac{1}{2}\,\log\left[\det(\sigma^2 R)\right] - \frac{1}{2}\,(y-S\theta)^\top\,(\sigma^2 R)^{-1}\,(y-S\theta)\\ &= -\frac{N}{2}\,\log 2\pi - \frac{N}{2}\,\log\sigma^2 - \frac{1}{2}\,\log\det R - \frac{1}{2\sigma^2}\,(y-S\theta)^\top\,R^{-1}\,(y-S\theta)\,,\end{aligned} \tag{2.3.1}$$

so that, writing the gradient with respect to θ as a column vector, one gets

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

From these expressions one can compute the Fisher matrix $I(\theta,\sigma^2)$. Letting,

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

one needs to compute the entries of the matrix

$$I(\theta,\sigma) = \mathbb{E}_{\,\theta,\sigma}\,\begin{bmatrix}\mathbf{z}_\theta\mathbf{z}_\theta^\top & \mathbf{z}_\theta\mathbf{z}_\sigma\\ \mathbf{z}_\theta^\top\mathbf{z}_\sigma & \mathbf{z}_\sigma^2\end{bmatrix}\,. \tag{2.3.2}$$

which turn out to be

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

Further define the scaled variable

$$\tilde{\mathbf{y}} := R^{-1/2}\,(\mathbf{y}-S\theta) \sim \mathcal{N}(0,\sigma^2 I)$$

whereby

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

which follows since $\tilde{\mathbf{y}}$ has zero mean and the third order moments of a zero-mean Gaussian variable are zero. Note now that the quadratic form

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

has a χ^2 distribution with a number of degrees of freedom equal to N, the dimension of $\mathbf{y}$; i.e. $\frac{\|\tilde{\mathbf{y}}\|^2}{\sigma^2} \sim \chi^2(N)$. See Appendix A.4 for a definition of the chi-squared distribution and for a proof of this fact. Hence, since the expected value of $\frac{\|\tilde{\mathbf{y}}\|^2}{\sigma^2}$ is exactly N it follows from (A.2.3) that

$$\mathbb{E}_{\theta,\sigma}\, \mathbf{z}_\sigma^2 = \mathbb{E}_{\theta,\sigma} \left\{ \frac{1}{2\sigma^2} \left[\frac{\|\tilde{\mathbf{y}}\|^2}{\sigma^2} - N \right] \right\}^2 = \frac{1}{4\sigma^4} \text{ Var} \left[\frac{\|\tilde{\mathbf{y}}\|^2}{\sigma^2} \right] = \frac{N}{2\sigma^4}\,. \tag{2.3.3}$$

Putting these results together one finds a formula for the information matrix

$$I(\theta, \sigma^2) = \begin{bmatrix} \frac{1}{\sigma^2}\, S^\top R^{-1} S & 0 \\ 0 & \frac{N}{2\sigma^4} \end{bmatrix}. \tag{2.3.4}$$

Clearly $I(\theta, \sigma^2)$ is non-singular if and only $S^\top R^{-1} S$ is also non-singular which in turn happens if and only if S is full column rank. The following proposition is an immediate consequence of Rothenberg's Theorem 1.3.

Proposition 2.1. *Let (2.2.7) be a model with $N \geq p$ scalar observations. Then θ is globally identifiable if and only if*

$$\text{rank } S = p. \tag{2.3.5}$$

Whenever the nullspace of S contains a nonzero vector $\xi \neq 0$, then θ and $\theta + \xi$ would be indistinguishable.

In this section we shall assume that $\text{rank } S = p$ i.e. the p columns of S are linearly independent, which is equivalent to the existence of the inverse $I^{-1}(\theta, \sigma^2)$. Therefore the variance matrix of any unbiased estimator of θ cannot be smaller (in the matrix ordering) than $\sigma^2 [S^\top R^{-1} S]^{-1}$. Similarly, $\frac{2\sigma^4}{N}$ is a lower bound for the variance of any unbiased estimator of σ^2 although it turns out that this lower bound is not sharp.

From $\partial \ell / \partial \theta = 0$, in force of the invertibility of $S^\top R^{-1} S$, one obtains the expression for the ML estimator of θ:

$$\hat{\theta}(y) = [S^\top R^{-1} S]^{-1}\, S^\top R^{-1} y\,. \tag{2.3.6}$$

which provides indeed the absolute maximum of $\ell(y, \theta, \sigma)$ since the Hessian matrix

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

is negative definite. This expression of $\hat{\theta}$ is exactly the same as that found in Section 2.1. We shall use again the compact notation

$$\hat{\theta}(y) = Ay \quad , \qquad A := [S^\top R^{-1} S]^{-1}\, S^\top R^{-1}\,. \tag{2.3.7}$$

Theorem 2.1. *The ML estimator* (2.3.6) *of the parameter θ in the linear model* (2.2.7)

1. *is an* unbiased estimator *of the parameter θ. In fact, $\mathbb{E}_{\theta,\sigma}\, A\mathbf{y} = \theta$ for all $\theta \in \mathbb{R}^p$.*

2. *The variance matrix of $\hat{\theta}(\mathbf{y})$ is*

$$\mathrm{Var}\left\{\hat{\theta}(\mathbf{y})\right\} = \sigma^2[S^\top R^{-1}S]^{-1} \tag{2.3.8}$$

and coincides with the Cramèr-Rao lower bound. Therefore $\hat{\theta}(\mathbf{y})$ is a minimum variance estimator.

3. *The random vector $\hat{\theta}(\mathbf{y})$ is normally distributed, in fact,*

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

Proof. Property 1 follows from the fact that A is a *left-inverse* of S since

$$AS = I \quad . \tag{2.3.9}$$

Property 2 follows from

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

while property 3 is a consequence of linearity. □

Geometric interpretation and Least Squares

Just by looking at the expression of the log-likelihood (2.3.1) it is evident that $\hat{\theta}(y)$ is the function which minimizes with respect to θ the quadratic form

$$(y - S\theta)^\top R^{-1}(y - S\theta) = \|y - S\theta\|^2_{R^{-1}}\,, \tag{2.3.10}$$

which can again be interpreted as a distance in $\mathbb{R}^N$, once equipped with the inner product $\langle x, y\rangle_{R^{-1}} := x^\top R^{-1} y$. As observed already in Sec. 2.1, for any $y \in \mathbb{R}^N$ the minimizer $S\hat{\theta}(y) := S\,Ay$ of the distance (2.3.10), is just the vector $v \in \mathcal{S} := \mathrm{span}\,(S)$ (the vector space spanned by the columns of S) equal to the orthogonal projection of y onto the subspace $\mathcal{S} = \mathrm{span}\,(S)$. In other words the matrix $P \in \mathbb{R}^{N\times N}$, defined as

$$P = SA \quad , \tag{2.3.11}$$

must be the orthogonal projector (with respect to the inner product $\langle\cdot,\cdot\rangle_{R^{-1}}$) from $\mathbb{R}^N$ onto $\mathcal{S}$. In fact P is idempotent ($P = P^2$), since

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

and

$$P^\top = (SA)^\top = A^\top S^\top = R^{-1}S\left[S^\top R^{-1}S\right]^{-1}S^\top = R^{-1}SAR = R^{-1}PR\,, \tag{2.3.12}$$

which is the property of being self-adjoint with respect to the inner product $\langle\cdot,\cdot\rangle_{R^{-1}}$.

The characterization of an orthogonal projection is the orthogonality of the error, $y - S\theta$, to the subspace $\mathcal{S}$

$$S \perp (y - S\theta)\,. \tag{2.3.13}$$

which, is equivalent to

$$S^\top R^{-1} y - S^\top R^{-1}\, S\theta = 0 \tag{2.3.14}$$

and, by the invertibility of $S^\top R^{-1}\, S$, provides the expression (2.3.6). In other words, $\hat{\theta}(y)$ is a weighted least-squares estimator with weight matrix equal to the inverse of the noise covariance.

Theorem 2.2. *Under the identifiability assumption* (2.3.5) *the ML estimator of the parameter θ in the linear-Gaussian model* (2.2.7) *is the unique vector function $y \to \hat{\theta}(y)$ minimizing the distance* (2.3.10)*.*

The variance estimator

From the second log-likelihood equation $\partial\ell/\partial\sigma^2 = 0$ one gets

$$\hat{\sigma}^2(y) = \frac{1}{N}\,\left(y - S\hat{\theta}(y)\right)^\top R^{-1}\left(y - S\hat{\theta}(y)\right)^\top = \frac{1}{N}\,\|y - Py\|^2_{R^{-1}}\,, \tag{2.3.15}$$

Hence $\hat{\sigma}^2(y)$ is the average squared norm of the *residual approximation error* of the observed data y when using the value of θ equal to the ML parameter estimator, that is, when approximating y by $Py = S\hat{\theta}(y)$. We shall now need to find the pdf of the random variable $\hat{\sigma}^2(\mathbf{y})$.

Theorem 2.3. *The normalized ML estimator of the variance σ^2 in the linear model* (2.2.7) *is χ^2-distributed. More precisely,*

$$\frac{N\hat{\sigma}^2(\mathbf{y})}{\sigma^2} \sim \chi^2(N-p)\,. \tag{2.3.16}$$

In particular, the mean and variance are given by

$$\mathbb{E}_{\,\theta,\sigma^2}\;\hat{\sigma}^2(\mathbf{y}) = \sigma^2\;\frac{N-p}{N} \quad , \tag{2.3.17}$$

$$\mathrm{Var}_{\theta,\sigma^2}\;\hat{\sigma}^2(\mathbf{y}) = \sigma^4\;\frac{2(N-p)}{N^2}\,. \tag{2.3.18}$$

Proof. Notice that $\mathbf{y} - P\mathbf{y} = (\mathbf{y} - S\theta) - P(\mathbf{y} - S\theta) = \sigma(I-P)\,\mathbf{w}$. Define the random vector

$$\mathbf{z} := \; R^{-1/2}\,\mathbf{w} \;\sim \mathcal{N}(0, I)$$

and use (2.3.14) to get the representation

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

where we have used the similarity $P^\top = R^{-1} PR$ established in (2.3.12). The matrix between square brackets, say Q, is idempotent since by (2.3.12), one has

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

and has rank $N-p$. In fact $I-P$ projects onto the subspace $\mathcal{S}^\perp$, the orthogonal complement of $\mathcal{S}$ and since we have identifiability $\dim \mathcal{S} = p$. Proposition A.7 in Appendix A.4 then implies that $\mathbf{z}^\top Q \mathbf{z} \sim \chi^2(N-p)$. □

Remark 2.1. *As shown by equation (2.3.17), the estimator $\hat{\sigma}^2(\mathbf{y})$* is biased *implying a systematic bias, equal to $-\sigma^2\, p/N$, which however tends to zero as the sample size N tends to infinity. The bias can easily be compensated by modifying $\hat{\sigma}^2$ to*

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

which is clearly unbiased. The correction has however a price in terms of a larger variance. *In fact from $(N-p)\, s^2(\mathbf{y})/\sigma^2 \sim \chi^2(N-p)$ it easily follows that*

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

which is strictly larger than $2\sigma^4\,(N-p)/N^2$. Incidentally, note that the variance of $\hat{\sigma}^2(\mathbf{y})$ is actually smaller than the Cramèr-Rao bound, *which is $2\sigma^4/N$.*

In conclusion we have shown that the computation of the ML estimator of the parameter θ in the linear Gaussian model (2.2.7) can be reduced to the solution of a weighted least-squares problem. As for the general deterministic least-squares problem dealt with in Sec. 2.1 the ML estimator turns out to be a *linear function of the data*. However now this linear function is, in terms of error variance, the *best possible function of the data*, in the class of *all measurable functions of* $\mathbf{y}$ which includes arbitrary nonlinear functions. This strong optimality is a consequence of Gaussianity and would not hold for non-Gaussian noise distributions.

Statistical Least Squares with partial information

When the distribution of $\mathbf{w}$ is not Gaussian the ML estimator is no longer linear; in practice the distribution may actually be unknown or we may have only partial statistical information. Typically we may assume that only information on the first and second order moments is available. In this case it is reasonable to look for an estimator of θ which is a *linear function of the data*.

In this section we shall discuss parameter estimation on the usual linear model

$$\mathbf{y} = S\theta + \sigma\mathbf{w} \tag{2.3.19}$$

assuming only that the first and second order noise statistics

$$\mathbb{E}\,\mathbf{w} = 0 \quad , \qquad \mathrm{Var}\,(\mathbf{w}) = \mathbb{E}\,\mathbf{w}\mathbf{w}^\top = R\,. \tag{2.3.20}$$

are known, but without assuming anything on the noise distribution. As for the Gaussian model, σ^2 is the unknown parameter in the noise variance matrix, to be estimated from data.

Remarks 2.1. Note that $\mathbf{w}$ having known zero mean is not an essential assumption. For example, assuming for simplicity that $\mu := \mathbb{E}\,\mathbf{w}_t$ is the same for all $t = 1, 2, \ldots, N$, we may augment the vector θ by the additional unknown scalar parameter μ by adding an additional column to the matrix S. Setting $\tilde{\mathbf{w}}_t := \mathbf{w}_t - \mu$ one may rewrite the model as

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

where $\tilde{\mathbf{w}}$ has mean zero and the same variance matrix $\sigma^2 R$ of the original model. Introducing the new parameter $\theta_{p+1} := \mu$ and attaching a column of 1's to S one obtains the model

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

where $\tilde{\mathbf{w}}$ has mean zero. Obviously the same remark applies to the previous section.

Guided by the structure of the solution for the Gaussian case, which is a linear function of the observations depending only on the first and second order moments of $\mathbf{y}$, it is reasonable to look also in this case for an estimator of θ which is a *linear function of the data*.

Definition 2.1. *A* best linear unbiased estimator (BLUE) *is a linear function of the data* $\mathbf{y}$ *say* $\phi(\mathbf{y}) = A\mathbf{y},\ A \in R^{p\times N}$ *which,*

- *Is unbiased; i.e for all* $\theta \in \mathbb{R}^p$*,*

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

- *Has minimal variance, with respect to the semidefinite ordering of matrices; i.e.* $A \geq B$ *iff* $A - B$ *is positive semidefinite. The (matrix) variance being defined as*

$$\text{Var}\ \hat{\phi}(\mathbf{y}) := \mathbb{E}\,[\,(\phi(\mathbf{y}) - \theta)\,(\phi(\mathbf{y}) - \theta)^\top\,] .$$

It is then natural to bring in the Least-Squares estimator $\hat{\theta}$ defined in the previous Section 2.1. The question to ask is what statistical properties this estimator may have.

Proposition 2.2. *No matter what* $Q > 0$*, the weighted least squares estimator* (2.1.5) *is unbiased..*

Proof. In fact, since $\mathbb{E}\,\mathbf{w} = 0$, and

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

□

When $Q = R^{-1}$ the least squares estimator of θ is called the **Markov estimator**. Obviously when R is diagonal, i.e. $R = \text{diag}\{r_1, \ldots, r_N\}$, the diagonal weight Q with entries $q_t = \frac{1}{\text{var}\ \mathbf{y}_t} = \frac{1}{\sigma^2}\,\frac{1}{r_t}, \qquad t = 1, \ldots, N.$ is a most natural choice. Note that the unknown factor $1/\sigma^2$ does not influence the estimator.

Theorem 2.1 (Gauss-Markov). *Assume S is of full rank p. The BLUE of θ for the linear model* (2.3.19) *is the Markov estimator $\theta(\mathbf{y})$, whose variance is*

$$\text{Var } \hat{\theta}(\mathbf{y}) = \sigma^2 (S^\top R^{-1} S)^{-1} . \tag{2.3.21}$$

Proof. Since A must be a left-inverse of S it must have the form (**??**) for some positive definite Q, that is, $A = S^{-L} = [S^\top Q S]^{-1} S^\top Q$. The variance of $\hat{\phi}(\mathbf{y}) = A\mathbf{y}$ is then

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

We shall show that for all $Q = Q^\top > 0$; i.e. for all left inverses A,

$$\sigma^2 \, ARA^\top \geq \sigma^2 \, (S^\top R^{-1} S)^{-1} \; = \text{Var } \hat{\theta}(\mathbf{y}) \tag{2.3.22}$$

(a wide-sense Cramèr-Rao bound).
Proof of (2.3.22) *:*
The proof is based on the error variance formula for liner Bayesian estimation to be seen in the next Chapter. Let $\mathbf{n}$ be a random vector with orthonormal components, $\mathbf{x} := AR^{1/2}\mathbf{n}$ and $\mathbf{y} := CR^{1/2}\mathbf{n}$. The variance of the Bayesian estimation error $\tilde{\mathbf{x}} := \mathbf{x} - \mathbb{E}\,[\mathbf{x} \mid \mathbf{y}]$, is displayed in formula (5.5.13) , which is just $\Sigma_x - \Sigma_{xy}\Sigma_y^{-1}\Sigma_{xy}^\top$. Substituting the expressions for the various variance matrices as defined above and recalling that the error variance must be nonnegative we get,

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

which holds for an arbitrary full rank matrix $C \in \mathbb{R}^{p \times N}$.
Choose $C = (S^\top R^{-1} S)^{-1} S^\top R^{-1}$ and use the fact that $AS = I$ to obtain (2.3.22). □

Note that the Markov estimator of θ coincides with the ML estimator but in the present context we can only say that it is the best *linear* function of the data which is of course a much weaker statement than saying that it is the best function in the huge class of all measurable functions of the data. This holds true only in case $\mathbf{w}$ has a Gaussian distribution. When $\mathbf{w}$ is not normally distributed its variance may actually be much larger than the variance of the true ML estimator.

Estimation of σ^2 for the Markov estimator

We may still refer to the formula for the average residual estimation error (2.3.15)

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

Now $\hat{\sigma}^2(y)$ may no longer have a χ^2distribution. We may however compute its expectation. By using the properties of the projection operator P we find

$$\begin{aligned} N\mathbb{E}\,\left(\hat{\sigma}^2(\mathbf{y})\right) &= \mathbb{E}\,\left(\mathbf{y}^\top (I-P)^\top R^{-1} (I-P)\,\mathbf{y}\right) \\ &= \mathbb{E}\,\left(\mathbf{w}^\top (I-P)^\top R^{-1} (I-P)\,\mathbf{w}\right) = \mathbb{E}\,\left(\mathbf{w}^\top R^{-1} (I-P)\,\mathbf{w}\right) \\ &= \mathbb{E}\; tr\{\mathbf{w}^\top R^{-1} (I-P)\,\mathbf{w}\} = \mathbb{E}\;\; \text{Tr}\,\{R^{-1}(I-P)\;\mathbf{w}\mathbf{w}^\top\} \\ &= \mathbb{E}\;\; \text{Tr}\,\{(I-P)\;(\mathbf{w}\mathbf{w}^\top)\;R^{-1}\} = \text{Tr}\,\{(I-P)\;\mathbb{E}\,(\mathbf{w}\mathbf{w}^\top)\;R^{-1}\} \\ &= \sigma^2\;\; \text{Tr}\,(I-P) \quad ; \end{aligned} \tag{2.3.23}$$

which follows fro the identity $(I-P)\,\mathbf{y} = (I-P)\,S\theta + (I-P)\,\mathbf{w} = (I-P)\,\mathbf{w}$. Since Tr $P = \dim \mathcal{S} = p$

$$\mathbb{E}\ \hat{\sigma}^2(\mathbf{y}) = \frac{N-p}{N}\ \sigma^2\,. \tag{2.3.24}$$

which is the same formula found for the ML estimator. Of course $\frac{N}{N-p}\,\hat{\sigma}^2$ will be an unbiased estimator of σ^2.

Remarks 2.2. *Note that the BLU estimator of any linear function $c^\top\theta$ of θ ($c^\top$ is a known vector), is simply the same linear function of the Markov estimator, say $c^\top\hat{\theta}$. Of course this would be no longer true for a* non-linear *function $c(\theta)$, of θ which would instead hold for the ML estimator.*

2.4 ▪ The Case of Vector-valued Data

Suppose now that the observations and the inputs are vector-valued. We collect at each instant t, m simultaneous observations which may be the result of measurements made simultaneously on m different output channels and similarly record the p-dimensional input components fed to the system at time t. To arrange for vector valued quantities it will be convenient to change notations slightly and stack the vector data $\mathbf{y}(t)$ and the additive noise m-vectors $\mathbf{w}(t)$ into "fat" N-column matrices as follows:

$$\mathbf{Y} := \begin{bmatrix}\mathbf{y}(1) & \dots & \mathbf{y}(N)\end{bmatrix} \qquad \mathbf{W} := \begin{bmatrix}\mathbf{w}(1) & \dots & \mathbf{w}(N)\end{bmatrix}$$

Here we want to model each row $\mathbf{y}_k$ of $\mathbf{Y}$ as a linear function $\theta_k^\top S$ where S is a signal matrix with *rowspace* $\mathcal{S}$ spanned by p known N-vectors possibly depending on $\mathbf{u}$. The additive errors $\mathbf{w}(t)\,;\, t = 1, 2, \dots, N$ are in general assumed to be mutually uncorrelated each having a known variance matrix $\mathbb{E}\,\mathbf{w}(t)\mathbf{w}(t)^\top := R(t)$ which is positive definite. A common unknown scalar factor σ^2 (independent of t) can be dealt with by the techniques seen in the previous section but here we shall for simplicity assume that all $R(t)$'s are known. With the above notations the standard linear model can be rewritten compactly as

$$\mathbf{Y} = \Theta S + \mathbf{W}\,. \tag{2.4.1}$$

A matrix generalization of the linear least squares problem follows.

Matrix least-Squares Problems

In the real vector space of $m \times N$ matrices, $\mathbb{R}^{m\times N}$, we can introduce the *Frobenius* inner product defined as

$$\langle X, Z\rangle := \text{Trace}\,\{XZ^\top\}$$

which coincides with the usual Euclidean inner product of the vectors in $\mathbb{R}^{mN}$ obtained by stacking the columns of each matrix X and Z on top of each other. Using elementary properties of the trace operator you can easily show that

$$\langle X, Z\rangle := \text{Trace}\,\{X^\top Z\} = \text{Trace}\,\{Z^\top X\}\,.$$

This Inner product can be generalized introducing a positive definite $Q \in \mathbb{R}^{N\times N}$ to define

$$\langle X, Z\rangle_Q := \text{Trace}\,\{XQZ^\top\} = \text{Trace}\,\{ZQX^\top\}$$

which defines the *weighted Frobenius norm* of a matrix X as

$$\|X\|_Q^2 := \text{Trace}\,\{XQX^\top\}.$$

Let now $Y \in \mathbb{R}^{m\times N}$ and $S \in \mathbb{R}^{p\times N}$ be known real matrices. One may generalize the standard LS problem to matrix-valued data as follows: consider the problem

$$\min_{\Theta\in\mathbb{R}^{m\times p}} \|\, Y - \Theta S \,\|_F\,, \qquad \text{(Frobenius norm)} \tag{2.4.2}$$

where $\Theta \in \mathbb{R}^{m\times p}$ is an unknown matrix parameter. The Frobenius norm could actually be weighted by a positive definite weight matrix Q.

The problem can be solved for each row y_k by the orthogonality principle. Let $\mathcal{S}$ be the rowspace of S and denote a row vector in $\mathcal{S}$ by $\theta_k S\,;\ \theta_k \in \mathbb{R}^p$ (also a row vector). Then the optimality condition is

$$y_k - \theta_k S \perp \mathcal{S}\,; \quad \text{i.e.} \quad y_k QS^\top = \theta_k SQS^\top \quad k = 1, 2, \ldots, m$$

so that, assuming S of rank m, by the orthogonality principle we obtain the following solution:

$$\hat{\Theta} = YQS^\top\,[\,SQS^\top]^{-1}\,. \tag{2.4.3}$$

Example 2.1 (A vector linear regression problem). We consider a simple time-invariant situation where we want to fit to the measurements a static linear model described by a $m \times p$-dimensional matrix parameter Θ, plus additive noise

$$\mathbf{y}(t) = \Theta\,\mathbf{u}(t) + \mathbf{w}(t)\,, \qquad t = 1, \ldots, N, \quad \Theta \in \mathbb{R}^{m\times p} \tag{2.4.4}$$

where we assume $\mathbf{u}(t)$ a random zero-mean p-vector having finite second order moments for all t and $\mathbf{w}(t)$ and $\mathbf{u}(s)$ uncorrelated for all t, s. Stacking the data into "fat" N-column matrices as follows:

$$\mathbf{Y} := \begin{bmatrix}\mathbf{y}(1) & \ldots & \mathbf{y}(N)\end{bmatrix} \quad \mathbf{U} := \begin{bmatrix}\mathbf{u}(1) & \ldots & \mathbf{u}(N)\end{bmatrix} \quad \mathbf{W} := \begin{bmatrix}\mathbf{w}(1) & \ldots & \mathbf{w}(N)\end{bmatrix}$$

the model can be rewritten compactly as

$$\mathbf{Y} = \Theta\mathbf{U} + \mathbf{W}\,. \tag{2.4.5}$$

The best fit can be defined in terms of a quadratic functional

$$V(\Theta) = \|Y - \Theta U\|_Q^2 := \langle Y - \Theta U, Y - \Theta U\rangle_Q \tag{2.4.6}$$

where Y, U are the sample input-output data and Q is a weighting $N \times N$ positive definite matrix, perhaps having a block-diagonal structure. The scalar product in the vector space of $m \times N$ matrices in this context is defined as[6]

$$\langle X, Z\rangle_Q := \text{Trace}\,\{XQZ^\top\} = \text{Trace}\,\{ZQX^\top\}$$

[6]There is also an equivalent symmetric definition, namely $\langle X, Z\rangle_R := \text{Trace}\,\{X^\top RZ\}$ which however should use a different weighting matrix R of dimension $m \times m$.

which is just the inner product defining the weighted Frobenius norm $\|\cdot\|_Q$. As for the deterministic problem seen at the beginning of this section, the estimate $\hat{\Theta}$ minimizing the Frobenius distance (2.4.6) can be computed by imposing the orthogonality of each error row vector $y_k - [\,\Theta U\,]_k \in \mathbb{R}^N$ to the subspace spanned by the rows of the U matrix. This leads to the matrix equation

$$Y - \Theta U \;\perp_Q\; U$$

where $\perp_Q$ means orthogonality with respect to the Frobenius inner product (2.4.6). Assuming that U has linearly independent rows (no superfluous input channel) we get

$$\hat{\Theta} = YQU^\top[UQU^\top]^{-1} \tag{2.4.7}$$

See (2.4.3). In case $Q = I$, this formula can be rewritten by plugging in a $\frac{1}{N}$ factor, so that

$$\hat{\Theta} = \frac{1}{N}YU^\top[\frac{1}{N}\,UU^\top]^{-1} \tag{2.4.8}$$

which is a sample covariance version of a probabilistic formula for the estimate of Θ. This probabilistic formula is derived from the equation of the theoretical model (2.4.4) taking cross correlation of both members with $\mathbf{u}(t)$:

$$\mathbb{E}\,\mathbf{y}(t)\mathbf{u}(t)^\top = \Theta\mathbb{E}\,\mathbf{u}(t)\mathbf{u}(t)^\top + \mathbb{E}\,\mathbf{w}(t)\mathbf{u}(t)^\top\,,$$

given that the last term is zero by assumption. Formula (2.4.8) can then be interpreted as an estimate obtained by the *method of moments*.

If the sample covariances converge for $N \to \infty$, the estimate $\hat{\Theta}$ will converge to the true value. This is a first example of a *consistent estimator* see Sect. 2.6.6. The result has for example applications to statistical state-space system identification. See Example 2.5 below.

2.5 ▪ Empirical Prediction Error minimization

Very often statistical model building from data is not done with the goal of estimating parameters or regression functions but rather for the very purpose of **prediction**. In the jargon of Machine Learning a predictor is said to perform a *generalization* of the training data. Suppose that we have used the N-dimensional vector $\mathbf{y}$ in a standard linear Gaussian model to compute an estimator of θ, say a ML estimator. Then one can say that

$$\hat{\mathbf{y}} = S\hat{\theta}(\mathbf{y})$$

is an optimal approximation of the full vector $\mathbf{y}$ based on the known signal matrix S. This fact does not look particularly interesting *per se*. Suppose however that an $N+1$-th row vector, s_{N+1}, possibly depending on the last observations of some regression variables, is added at the bottom of matrix S and that one would like to guess (or predict) the value of the corresponding output random variable $\mathbf{y}_{N+1}$ which is *not observed*. It is then natural to suggest as a *prediction* of the $N+1$-th component $\mathbf{y}_{N+1}$, the linear function of the past data $\mathbf{y}$ given by

$$\hat{\mathbf{y}}_{N+1} = s_{N+1}\hat{\theta}(\mathbf{y})\,. \tag{2.5.1}$$

One rational of this formula is that, by the invariance principle of ML, one could generalize the linear estimator $s_{N+1}\hat{\theta}(\mathbf{y})$ of $s_{N+1}\theta$ by considering instead an arbitrary non linear function $g(\theta)$ of which $g(\hat{\theta}(\mathbf{y}))$ could then be interpreted as the ML estimate based on the N past output measurements.

Proposition 2.3. *The* Prediction Error *incurred by the ML (or by the Markov) predictor* (2.5.1)*, namely*

$$\mathbf{y}_{N+1} - \hat{\mathbf{y}}_{N+1} = s_{N+1}\left[\theta - \hat{\theta}(\mathbf{y})\right] + \sigma\mathbf{w}_{N+1} \tag{2.5.2}$$

has the smallest variance among all (linear) predictors, which are functions of the past data $\mathbf{y}$*.*

Proof. By assumption $\mathbf{w}_{N+1}$ is uncorrelated with the previous noise vector $\mathbf{w}$ and hence with the previous observations $\mathbf{y}$. Just compute the variance of the expression on the right and recall that $\hat{\theta}(\mathbf{y})$ has minimal variance.
By the invariance principle of ML, one could generalize the linear estimator $s_{N+1}^{\top}\hat{\theta}(\mathbf{y})$ of $s_{N+1}^{\top}\theta$ by considering instead an arbitrary non linear function $g(\theta)$ of which $g(\hat{\theta}(\mathbf{y}))$ could then be interpreted as the ML estimate based on the N past output measurements. □

2.6 ▪ Recursive Estimators

In many applications, especially those involving time series, the data are acquired sequentially in time. Suppose the time epoch is indexed by an integer variable $t = 0, 1, 2 \dots$ and that at each time t one needs to produce an estimate, which we denote $\hat{\theta}(t)$, of the unknown parameter θ of the linear model (2.3.19) based on measurements up to the current time t, which we denote by $\mathbf{y}^t$. Each scalar measurement being described by the linear relation

$$\mathbf{y}(t) = s(t)^{\top}\theta + \sigma\mathbf{w}(t)\,, \qquad t = 1, 2, \dots\,,$$

where we shall assume that $\{\mathbf{w}(t)\}$ is an uncorrelated sequence of unit variance (that is assume R to be the identity). The vectors $s(t)$ could in particular depend on past data $\mathbf{y}^t$ and the model could well encompass the structure described in the previous section. Since the dimension of the matrix to be inverted in the expression (2.3.6) grows with N (wich we shall now denote by the current time epoch t) the computation complexity grows with t. In fact grows approximately as $O(t\,p^2 + p^3)$ that is linearly in the dimension t. Consequently one must look for computational schemes which could possibly update the current parameter estimate sequentially when new data become available, by an algorithm which requires a fixed finite number of operations at each step, which we call a *fixed memory algorithm*. In more precise terms, one would like to **express** $\hat{\theta}(t+1)$ **as a function of the previous estimate** $\hat{\theta}(t)$ **and of the new data** $\mathbf{y}(t+1)$**.**

Since $R = I$, formula (2.3.6), written assuming measured data available up to time t, involves only the row vectors $s(k)^{\top}$ of the matrix S up to instant t and

the estimate $\hat{\theta}(t)$ is expressed as:

$$\hat{\theta}(t) = \left[\sum_{k=1}^{t} s(k)s(k)^\top\right]^{-1} \sum_{k=1}^{t} s(k)\mathbf{y}(k)\,. \tag{2.6.1}$$

The matrix

$$P(t) := \left[\sum_{k=1}^{t} s(k)s(k)^\top\right]^{-1},$$

is referred to as the *Normalized Variance* of $\hat{\theta}(t)$, whereby $\hat{\theta}(t) = P(t)\sum_{k=1}^{t} s(k)\mathbf{y}(k)$. When the next datum $\mathbf{y}(t+1)$ is acquired, we bring in the new element $s(t+1)$ of S which is a known (possibly input-dependent) vector and we can form the next estimate by the same formula, written for time $t+1$:

$$\hat{\theta}(t+1) = \left[\sum_{k=1}^{t} s(k)s(k)^\top + s(t+1)s(t+1)^\top\right]^{-1} \left[\sum_{k=1}^{t} s(k)\mathbf{y}(k) + s(t+1)\mathbf{y}(t+1)\right].$$

where the first matrix on the right side is $P(t+1)$ so that its inverse can be expressed as

$$P(t+1)^{-1} = P(t)^{-1} + s(t+1)s(t+1)^\top\,. \tag{2.6.2}$$

To transform this recursion into one for the actual normalized variances we shall use the **Matrix Inversion Lemma** (see Appendix D) which (assuming all indicated inverses exist) states that :

$$[A + BCD]^{-1} = A^{-1} - A^{-1}B\,[C^{-1} + DA^{-1}B\,]^{-1}DA^{-1}$$

Apply it to (2.6.2) letting $A = P(t)^{-1}\,,\ B = s(t+1) = D^\top$, and $C = 1$. One finds

$$P(t+1) = P(t) - P(t)s(t+1)\left[1 + s(t+1)^\top P(t)s(t+1)\right]^{-1} s(t+1)^\top P(t)\,. \tag{2.6.3}$$

Note that the term between square brackets, denoted for short

$$\beta(t) := 1 + s(t+1)^\top P(t)s(t+1)$$

is a scalar so the indicated inverse is trivial. Substitute now (2.6.3) in the expression for $\hat{\theta}(t+1)$:

$$\hat{\theta}(t+1) = P(t+1)\left[\sum_{k=1}^{t} s(k)\mathbf{y}(k) + s(t+1)\mathbf{y}(t+1)\right]$$

and define the"gain vector"

$$k(t) := P(t)s(t+1)\left[1 + s(t+1)^\top P(t)s(t+1)\right]^{-1} = P(t)s(t+1)\frac{1}{\beta(t)}$$

then after some algebra one finds

$$\hat{\theta}(t+1) = \hat{\theta}(t) + k(t)\left[\mathbf{y}(t+1) - s(t+1)^\top\hat{\theta}(t)\right] \tag{2.6.4a}$$

$$P(t+1) = P(t) - P(t)s(t+1)\left[1 + s(t+1)^\top P(t)s(t+1)\right]^{-1} s(t+1)^\top P(t)\,. \tag{2.6.4b}$$

The term $s(t+1)^\top \hat{\theta}(t)$ is the one-step-ahead prediction of $\mathbf{y}(t+1)$ based on the information available up to time t and the difference $\mathbf{e}(t+1) := \mathbf{y}(t+1) - s(t+1)^\top \hat{\theta}(t)$ is the *one-step-ahead prediction error*. The algorithm updates the old estimate $\hat{\theta}(t)$ by applying a correction term $k(t)\mathbf{e}(t+1)$ which is proportional to the prediction error adjusted via the "gain" matrix $k(t)$. Equations (2.6.4a), (2.6.4b) are known as the *deterministic Kalman Filter* for the model

$$\begin{cases} \boldsymbol{\theta}(t+1) = \boldsymbol{\theta}(t)\,; & \theta(0) = \theta_0 \\ \mathbf{y}(t) = s(t)^\top \theta(t) + \mathbf{w}(t) \end{cases}$$

Note that the unknown variance σ^2 does not enter in the equations.

The "unnormalized" variance matrix

$$\Sigma(t) := \sigma^2 P(t)$$

satisfies the same equation as $P(t)$ with the only correction on the definition of $\beta(t)$ which needs to be changed to

$$\beta(t) := \sigma^2 + s(t+1)^\top \Sigma(t) s(t+1)$$

Then we have

$$\Sigma(t+1) = \Sigma(t) - \Sigma(t)s(t+1)\left[\sigma^2 + s(t+1)^\top \Sigma(t) s(t+1)\right]^{-1} s(t+1)^\top \Sigma(t)\,.$$

How do we initialize the algorithm? One should in theory wait up to an instant t_0 such that $\sum_{s=1}^{t_0} s(k)s(k)^\top$ is invertible, to compute $P(t_0)$ and $\hat{\theta}(t_0)$. As we shall see, in many situations one can start by taking $P(0) = \alpha I_p,\ \alpha > 0$ and, say, $\hat{\theta}(0) = 0$. This has to do with the asymptotic behaviour of the algorithm which we shall now try to analize.

Theorem 2.4. *Assume that $\{\mathbf{w}(t)\}$ is an iid sequence and that*

$$\lim_{t\to\infty} \frac{1}{t} \sum_{k=1}^{t} s(k)s(k)^\top = V < \infty \tag{2.6.5}$$

where V is non singular. Then both $P(t)$ and $\Sigma(t)$ tend monotonically to zero for $t \to \infty$. Moreover, the gain $k(t)$ converges to zero and $\hat{\theta}(t)$ converges almost surely to a deterministic constant vector.

Assuming that the data are generated by a "true" parameter θ_0, the least squares estimator $\hat{\theta}(t)$ is **strongly consistent** *that is*

$$\lim_{t\to\infty} \hat{\theta}(t) = \theta_0\,. \tag{2.6.6}$$

almost surely.

Proof. Assume t is large enough so that $P(t)$ is well defined. Since (2.6.5) is the same as $\lim_{t\to\infty} \frac{1}{t} P(t)^{-1} = V$ is follows that $P(t)^{-1} \sim O(t)$ and hence $P(t) \to 0$ as $\frac{1}{t}$.

To prove that $k(t) \to 0$ just rewrite equation (2.6.3) as

$$P(t) - P(t+1) = k(t)\left[1 + s(t+1)^\top P(t) s(t+1)\right] k(t)^\top \geq k(t)k(t)^\top$$

since the term between square brackets is always positive and greater than 1. Now both $P(t)$ and $P(t+1)$ tend to 0 as $t \to \infty$ so that also $k(t)k(t)^\top$ and its trace, $\|k(t)\|^2$, must tend to zero.

In equation (2.6.4a) the prediction error $\mathbf{e}(t)$ (the term between square brackets) is a process of variance $\sigma^2 + s(t+1)^\top \Sigma(t) s(t+1)$ which tends to σ^2 as $t \to \infty$ and hence is uniformly bounded. Then it is not difficult to see that the variance of $k(t)\mathbf{e}(t)$ must tend to zero as $t \to \infty$ and hence $k(t)\mathbf{e}(t) \to 0$ almost surely. Therefore $\hat{\theta}(t+1) - \hat{\theta}(t) \to 0$ almost surely as well, so that the limit must be a constant (possibly random) vector. That this constant must be the true parameter θ_0 can be shown by substituting the "true model " equation $\mathbf{y}(t) = s(t)^\top \theta_0 + \sigma \mathbf{w}(t)$ into (2.6.1) which can then be rewritten as

$$\hat{\theta}(t) = \theta_0 + \sigma \left[\frac{1}{t} \sum_{k=1}^{t} s(k) s(k)^\top \right]^{-1} \frac{1}{t} \sum_{k=1}^{t} s(k) \mathbf{w}(k) \,. \tag{2.6.7}$$

The process $\tilde{\mathbf{w}}(t) := s(t)\mathbf{w}(t)\,;\, t = 1, 2, \ldots$ has independent variables and because of assumption (2.6.5) has bounded variance $\|s(t)\|^2 = \text{Trace } s(t)s(t)^\top < \text{Trace } V$. Therefore, by the law of large numbers the last sum converges almost surely to its mean value which is zero. Hence (2.6.6) follows. □

2.7 ▪ Examples

Example 2.2. A version of the popular *Cobb-Douglas* model in macroeconomics [14, 109] describes the relation between production Y, physical capital K, labour L and natural resources H in an economy by an equation of the following form

$$Y \cong \alpha L^{\theta_1} K^{\theta_2} H^{\theta_3} \tag{2.7.1}$$

where L, K, H are intrinsically positive and the exponents are unknown real variables which we would like to estimate using a data base consisting of a series of 86 measurements. [7] The model can be recast as a linear model by logarithmic transforms defining:

$$y = \log Y\,, \quad x_1 = \log L\,, \quad x_2 = \log K\,, \quad x_3 = \log H\,, \quad \theta_0 : \log \alpha\,,$$

by which (2.7.1) can be rewritten

$$y \simeq \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3$$

that is

$$y_t = \theta_0 + \theta_1 x_{1t} + \theta_2 x_{2t} + \theta_3 x_{3t} + \epsilon_t\,, \quad t = 1, ..., 86\,. \tag{2.7.2}$$

We shall assume that the errors ϵ_t are independent zero-mean Gaussian random variables all having the same unknown variance σ^2. The estimation problem becomes then just a standard linear Gaussian regression problem. Introduce the sample means of the three variables x_i, $i = 1, 2, 3$

$$\bar{x_i} = \frac{1}{86} \sum_1^{86} x_{it}\,, \qquad i = 1, 2, 3$$

and subtract from (2.7.2) the equation for the sample means,

$$\bar{y} = \theta_0 + \theta_1 \bar{x}_1 + \theta_2 \bar{x}_2 + \theta_3 \bar{x}_3 + \bar{\epsilon}$$

we obtain

$$y_t - \bar{y} = \sum_1^3 \theta_i (x_{it} - \bar{x}_i) + (\epsilon_t - \bar{\epsilon})\,, \qquad t = 1, 2, \ldots, 86\,. \tag{2.7.3}$$

In tis way we have reduced the parameters to 3; once the estimates $(\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3)$ are computed, $\hat{\theta}_0$ can be obtained by the formula

$$\hat{\theta_0} = \bar{y} - (\hat{\theta}_1 \bar{x}_1 + \hat{\theta}_2 \bar{x}_2 + \hat{\theta}_3 \bar{x}_3)\,. \tag{2.7.4}$$

The reader should verify that when in the standard linear model $y = S\theta + \epsilon$ the first column is all made of ones, the estimator of the first component of the parameter vector θ has an expression of the form (2.7.5)). We shall rewrite (2.7.3) using a vector notation as

$$\Delta \mathbf{y} = S\theta + \sigma \mathbf{w}\,, \qquad \mathbf{w} \sim \mathcal{N}(0, I)\,.$$

[7] The data are taken from an example in Rao's 1973 book [73, pag.227] where one has to reconstruct the cranial capacity of skulls from damaged or partially recovered specimens. One wants to estimate the cranial capacity C as a function of three linear dimensions, L, B, H, using a candidate model of the form (2.7.1).

Subtracting the sample means of y and of the three linear dimensions computed from the 86 measurements:

$$\bar{y} = 3.17\ ;\ \bar{x}_1 = 2.275\ ;\ \bar{x}_2 = 2.15\ ;\ \bar{x}_3 = 2.11$$

one can form the matrix $S \in \mathbb{R}^{86\times 3}$ and compute

$$S^\top S = \begin{bmatrix} 0.0187 & 0.0085 & 0.0068 \\ 0.0085 & 0.029 & 0.0088 \\ 0.0068 & 0.0088 & 0.029 \end{bmatrix} \qquad S^\top \Delta y = \begin{bmatrix} 0.030 \\ 0.044 \\ 0.036 \end{bmatrix}$$

from which the inverse

$$\left[S^\top S\right]^{-1} = \begin{bmatrix} 64.21 & -15.57 & -10.49 \\ -15.57 & 41.71 & -9.00 \\ -10.49 & -9.00 & 39.88 \end{bmatrix}$$

and the estimate $\hat{\theta} = \left[S^\top S\right]^{-1} S^\top \Delta y$, is found to have components

$$\hat{\theta}_1 = 0.88\,, \quad \hat{\theta}_2 = 1.04\,, \quad \hat{\theta}_3 = 0.73$$

and from (2.7.5)

$$\hat{\theta_0} = -2.618$$

so that the estimated function is

$$Y = 0.00241 L^{0.88} K^{1.04} H^{0.73}\,.$$

This model can be used to make predictions from new measured data. We shall make the assumption that the factors are measured with negligible error and can therefore be considered to be deterministic. The matrix S is then also considered to be a deterministic quantity. The main sources of randomness is the approximation due to the chosen simple model structure.
To get an estimate of the variance of the additive error in the linearized model (2.7.3) we need to compute

$$R_1^2 := \|\Delta y\|^2 - \|S\hat{\theta}\|^2 = \|\Delta y\|^2 - \langle S\hat{\theta},\, \Delta y\rangle = \|\Delta y\|^2 - \hat{\theta}^\top S^\top \Delta y \tag{2.7.5}$$

which yields

$$R_1^2 = \sum_{t=1}^{86} (y_t - \bar{y})^2 - (\hat{\theta}_1 0.030 + \hat{\theta}_2 0.044 + \hat{\theta}_3 0.036) \tag{2.7.6}$$

$$= 0.127 - 0.099 = 0.028\,. \tag{2.7.7}$$

An unbiased variance estimate is then

$$\hat{\sigma}^2 = \frac{R_1^2}{N-4} = \frac{0.028}{82} = 0.00034$$

where the denominator $N-4$ is due to the fact that the number of unknown parameters θ_i is actually 4 even if we have used the trick of reducing their apparent number to three. The ML estimate of the variance matrix of $\begin{bmatrix}\theta_1 & \theta_2 & \theta_3\end{bmatrix}^\top$ is obtained as $\hat{\sigma}^2 \left[S^\top S\right]^{-1}$. For example one gets

$$\text{var}\ \hat{\theta}_1 = 64.21 \times 3.4 \cdot 10^{-4} \cong 220 \cdot 10^{-4} = 0.022\,.$$

The variance of $\hat{\theta_0}$ can be estimated based on (2.7.5). We shall leave the details to the reader.
One should recall now that the original model has been linearized by a logarithmic transform. The quantity of interest here is the variance of the prediction error incurred when predicting the variable Y with the estimated model, but using the data coming from a $86+1$-th measurement triple made now *only on the variables* L, K, H. To this purpose recall the formula for the predictor of the vector $\mathbf{y}$ in a N-dimensional standard linear Gaussian model

$$\hat{\mathbf{y}} = S\hat{\theta}(\mathbf{y})$$

which can be particularized to express the prediction of the $N+1$-th (not yet observed) component $\mathbf{y}_{N+1}$ as

$$\hat{\mathbf{y}}_{N+1}(\mathbf{y}) = s_{N+1}\hat{\theta}(\mathbf{y}) \tag{2.7.8}$$

where s_{N+1} is the $N+1$-th row vector made with the last measurements of the regression variable which would have to be added at the bottom of matrix S to make a model describing a $N+1$-dimensional vector $\mathbf{y}$. The formula is a consequence of the invariance principle. □

Example 2.3. Some observed data $\{y(t)\}$ are described by the following linear model,

$$\mathbf{y}(t) = a + bt + \mathbf{e}(t), \qquad t = 1, \dots, N$$

where a, b are unknown parameters and $\{\mathbf{e}(t)\}$ is a Gaussian i.i.d. sequence of mean zero and unknown variance σ^2. We are only interested in estimating the angular coefficient b. To this end we model the discrete derivative $\mathbf{z}(t) := \mathbf{y}(t) - \mathbf{y}(t-1)$ of the data by the following model:

$$\mathbf{z}(t) = b + \mathbf{w}(t), \qquad t = 1, \dots, N$$

where $\mathbf{w}(t) = \mathbf{e}(t) - \mathbf{e}(t-1)$.

Write the vector form of the two linear models, in particular identify their S-matrices and the nois e variances $\sigma^2 R$.
Compare the (M.L.) estimates of the two b parameters obtained by using the two models. In particular find the variances of the estimates $\hat{\mathbf{b}}$ and in case they may result different explain why.

Solution : The original model,

$$\mathbf{y}(t) = a + bt + \mathbf{e}(t), \qquad t = 1, \dots, N$$

can be written in vector form as:

$$\mathbf{y} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ \vdots & \vdots \\ 1 & N \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} + \mathbf{e} := \begin{bmatrix} s_1 & s_2 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} + \mathbf{e}\,.$$

By applying the standard formula (2.3.8) the variance of the M.L. estimate of the vector parameter $\theta := [a\ b]^\top$ is found to be

$$\mathrm{Var}\,\{\hat{\theta}_N\} = \sigma^2 \begin{bmatrix} s_1^\top s_1 & s_1^\top s_2 \\ s_2^\top s_1 & s_2^\top s_2 \end{bmatrix}^{-1}$$

From which the variance of the second component $\hat{\mathbf{b}}_N$ is found to :

$$\begin{aligned}\text{var}\,\{\hat{\mathbf{b}}_N\} &= \sigma^2 \frac{s_1^\top s_1}{s_1^\top s_1 s_2^\top s_2 - (s_1^\top s_2)^2} = \sigma^2 \frac{N}{N s_2^\top s_2 - (\sum_{k=1}^N k)^2} \\ &= \sigma^2 \frac{1}{\sum_{k=1}^N k^2 - 1/N(\sum_{k=1}^N k)^2}\,.\end{aligned}$$

The model for the discrete derivative in vector form is:

$$\mathbf{z} = \begin{bmatrix}1\\1\\\vdots\\1\end{bmatrix} b + \mathbf{w} := s_1 b + \mathbf{w}$$

where $\mathbf{w}$ can be expressed in function of $\mathbf{e} := [\mathbf{e}(1)\ \mathbf{e}(2) \dots \mathbf{e}(N)]^\top$, as

$$\mathbf{w} := \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ -1 & 1 & 0 & \dots & 0 \\ 0 & -1 & 1 & \dots & 0 \\ & & & \ddots & \\ \dots & & & & \dots \\ 0 & \dots & & -1 & 1 \end{bmatrix} \mathbf{e} := L\mathbf{e}\,.$$

the first component of $\mathbf{w}$ has been set equal to $\mathbf{e}(1)$ since $\mathbf{e}(0)$ is not available. The variance matrix of $\mathbf{w}$ is hence $\sigma^2 R = \sigma^2 L L^\top$. According to this alternative model, the variance of $\hat{\mathbf{b}}$ is found to be:

$$\text{var}\,\{\hat{b}_N\} = \left[s_1^\top (\sigma^2 R)^{-1} s_1\right]^{-1} = \sigma^2 / \|L^{-1} s_1\|^2\,.$$

Here the inverse of L can be readily computed. One sees then that $L^{-1}s_1$ has the form

$$L^{-1}s_1 = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 1 & 1 & 0 & \dots & 0 \\ 1 & 1 & 1 & \dots & 0 \\ & & & \ddots & \\ \dots & & & & \dots \\ 1 & 1 & \dots & & 1 \end{bmatrix} s_1 = \begin{bmatrix}1\\2\\3\\\vdots\\N\end{bmatrix}$$

which leads to the formula

$$\text{var}\,\{\hat{b}_N\} = \sigma^2 / \sum_{k=1}^N k^2\,.$$

Note that this variance is *smaller* than that computed using the first linear model. The reason of this fact being that the second model is parametrized more parsimoniously than the first one. □

Example 2.4. Consider again the linear model

$$\mathbf{y}(t) = a + bt + \mathbf{e}(t), \qquad t = 1, \dots, N \tag{2.7.9}$$

where a, b are unknown parameters and $\{\mathbf{e}(t)\}$ is zero-mean i.i.d. Gaussian noise of variance σ^2. We are again interested only in estimating the angular coefficient b.

It may seem logical to introduce a fake observation vector $\mathbf{z}$ constructed by centering the true observation $\mathbf{y}$ by subtracting an estimate of the offset a, say, $\bar{a}_N$, and define $\mathbf{z} := \mathbf{y} - s_1\bar{a}_N$. This vector is described by a model where there is no mean; just as

$$\mathbf{z} = s_2\, b + \mathbf{w}$$

where $\mathbf{w}$ is still white noise.
A simple way to find an estimate $\bar{a}_N$ is to take the sample mean of $\mathbf{y}$. Check that this coincides with the estimate obtained for the linear model (2.7.9) when b is set equal to zero. Then compute the ML estimate $\hat{b}_N$ of b based on the reduced model and give necessary and sufficient conditions for its unbiasedness. Of course keeping in mind that the true model has two parameters.

Solution : Rewrite the model (2.7.9) in vector form as:

$$\mathbf{y} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ \vdots & \vdots \\ 1 & N \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} + \mathbf{e} := \begin{bmatrix} s_1 & s_2 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} + \mathbf{e}$$

Without the regressor $s_2 b$, the estimate $\bar{a}_N$ is

$$\bar{a}_N = \frac{1}{s_1^\top s_1} s_1^\top \mathbf{y} := \bar{\mathbf{y}}_N$$

which is in fact the sample mean of $\mathbf{y}$. When the second regressor is present one has instead

$$\bar{a}_N = \frac{1}{s_1^\top s_1} s_1^\top \{ \begin{bmatrix} s_1 & s_2 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} + \mathbf{e} \} = a + \frac{1}{s_1^\top s_1} s_1^\top s_2 b + \frac{1}{s_1^\top s_1} s_1^\top \mathbf{e}$$

which shows that the sample mean $\bar{a}_N$ is unbiased if and only if $s_1^\top s_2 = 0$.

Now using the model $\mathbf{z} = s_2\, b + \mathbf{w}$, we obtain

$$\hat{b}_N = \frac{1}{s_2^\top s_2} s_2^\top (\mathbf{y} - s_1 \bar{a}_N)$$

and hence

$$\mathbb{E}_\theta \hat{b}_N = \frac{1}{s_2^\top s_2} s_2^\top \mathbb{E}_\theta (\mathbf{y} - s_1 \bar{a}_N) = \frac{1}{s_2^\top s_2} s_2^\top \left[(s_1 (a - \mathbb{E}_\theta \hat{a}_N) + s_2 b \right]$$

which shows that $\hat{b}_N$ is unbiased only if $\bar{a}_N$ is; i.e. only if $s_1^\top s_2 = 0$.

An alternative argument is based on the following expression

$$\hat{b}_N = \frac{1}{s_2^\top s_2} s_2^\top \left[I - s_1 \frac{1}{s_1^\top s_1} s_1^\top \right] (s_2\, b + \mathbf{w}) .$$

which again shows that $\hat{b}_N$ is unbiased if and only if

$$\left[I - s_1 \frac{1}{s_1^\top s_1} s_1^\top \right] s_2 = s_2$$

where the term between square brackets is the orthogonal projector onto $\mathrm{span}\,\{s_1\}$. Hence $\hat{b}_N$ can be unbiased if and only if s_2 belongs to the orthogonal complement of the subspace $\mathrm{span}\,\{s_1\}$; again, if and only if $s_1^\top s_2 = 0$. □

Example 2.5 (From Subspace Identification).

Assume you observe the trajectories of the state, input and output variables, of respective dimensions n, p, m, of a linear Multi-Input-Multi-Output (MIMO) stationary stochastic system

$$\begin{bmatrix}\mathbf{x}(t+1)\\ \mathbf{y}(t)\end{bmatrix} = \begin{bmatrix}A & B\\ C & D\end{bmatrix}\begin{bmatrix}\mathbf{x}(t)\\ \mathbf{u}(t)\end{bmatrix} + \begin{bmatrix}K\\ J\end{bmatrix}\mathbf{w}(t) \tag{2.7.10}$$

where $\mathbf{w}$ is normalized white noise. With the observed trajectories from some time t onwards one constructs the data matrices (all having $N+1$ columns)

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

If these data obey the linear equation (2.7.10), there must exist a corresponding white noise trajectory $W_t := [\,w_t,\, w_{t+1},\, w_{t+2}, \ldots, w_{t+N}]$ such that

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

From this model one can now attempt to estimate the matrix parameter $\Theta := \begin{bmatrix}A & B\\ C & D\end{bmatrix}$ based on the observed data. This leads to a matrix LS problem of the kind formulated above. Assuming that the noise is Gaussian, a Maximum Likelihood estimate would need to use a weighting matrix Q constructed form the noise covariance. Unfortunately the covariance of the last (noise) term in this regression model depends on the unknown parameters $K,\ J$ and an explicit ML estimate of (A, B, C, D) looks very hard to get. The Frobenius LS solution with say $Q = I$ will generally not be ML but will provide a consistent estimator anyway.

The procedure sketched in this example is by no means the whole story since in practice the state trajectory is not observable and must be previously estimated from input-output data. See [98], [58, Chap. 13] □

2.8 ▪ Problems

2-0 *Consider the dual LS problem*

$$\min_{\theta} \|y - \theta^\top S\|_Q$$

where y is an N-dimensional row vector and $S \in \mathbb{R}^{p\times N}$. Assume that rank $S = p$ *and describe its solution.*

2-1 *Suppose that the noise sequence $\{\mathbf{w}(t)\,;\, t = 1, 2, \dots, N\}$ is Gaussian i.i.d. each vector having a positive definite covariance matrix R. In the model* (2.4.4) *the variables $u(t)$ are known deterministic vectors. Show that* (2.4.7) *is a ML estimate of Θ if Q has a special structure. Describe the structure of this matrix.*

2-2 *Show that if in the usual linear model*

$$\mathbf{y} = S\theta + \mathbf{w} \qquad \mathrm{Var}\,\{\mathbf{w}\} = \sigma^2 I \qquad \mathbb{E}\,\mathbf{w} = 0$$

the columns of $S = \begin{bmatrix} s_1 & \dots & s_p \end{bmatrix}$ are orthogonal vectors, i.e. $s_i^\top s_j = \|s_i\|^2 \delta_{i,j}$, then the components $\hat{\boldsymbol{\theta}}_i\,, i = 1, \dots p$ of the least squares estimator of θ, are mutually uncorrelated. In fact they can be computed independently of each other.

2-3 *You want to estimate the parameters μ and $\theta \in \mathbb{R}^p$ of the linear regression model*

$$\mathbf{y}_k = \mu + u_k^\top \theta + \mathbf{w}_k\,, \qquad k = 1, \dots, N$$

where $\{u_k \in \mathbb{R}^p\,;\, k = 1, \dots, N\}$ is an observed signal and $\{\mathbf{w}_k\,;\, k = 1, \dots, N\}$ is a sequence of independent zero-mean Gaussian variables of variance σ^2.

Consider the following two procedures:

- *Rewrite the model using an augmented parameter $\beta = \begin{bmatrix}\mu & \theta\end{bmatrix}^\top$, define a suitable enlarged matrix S and apply the standard estimation procedure.*
- *Let $\bar{\mathbf{y}}_N$ and $\bar{u}_N$ be the sample means of the sequences $\{\mathbf{y}_k\}$ and $\{u_k\}$ so that*

  $$\bar{\mathbf{y}}_N = \mu + \bar{u}_N^\top \theta + \bar{\mathbf{w}}_N$$

  where $\bar{\mathbf{w}}_N$ is still Gaussian. Consider subtracting this from the original regression model to eliminate μ and then compute the ML estimate $\hat{\boldsymbol{\theta}}_N$ of θ on the resulting model. Define *then the estimate $\hat{\mu} := \bar{\mathbf{y}}_N - \bar{u}_N^\top \hat{\boldsymbol{\theta}}_N$. Would you get the same result as before?*

2-4 *Consider the recursive least squares algorithm* (2.6.4a), (2.6.4b). *Show that the variance estimate can also be updated by a recursive equation of the form*

$$\hat{\sigma}^2(t{+}1) = \frac{t}{t+1}\hat{\sigma}^2(t) + \frac{1}{t+1}\,[\,\mathbf{y}(t{+}1) - s(t{+}1)\hat{\theta}(t{+}1)\,]^2\,.$$

To be really useful this recursion should be rewritten in terms of the (square of the) prediction error *involving $\hat{\theta}(t)$ instead of $\hat{\theta}(t{+}1)$. Find the corrected version.*

Chapter 3

CONDITIONING AND REGULARIZATION

3.1 ▪ Numerical Conditioning

Solving the normal equations

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

could be problematic for large dimensional datasets as numerical errors (and noise) in the data could be dramatically amplified in the solution. Need to be aware of when/why problems may arise and of possible solutions.

Most computational problems can be formalized in the following way: one has a function say $f : \mathbb{R}^k \to \mathbb{R}^p$ defined mathematically and a k-dimensional vector of "data" α. One wants to compute $x = f(\alpha)$. For example one may want to solve numerically a linear system

$$Ax = b \quad , \tag{3.1.1}$$

Here the data are $\alpha = (A, b)$ and the function f is defined mathematically by the expression $f(\alpha) = A^{-1}\,b$.

Now there are two main aspects of the problem to be taken into account.

A) The data, α, may be affected by errors. For example, numerical analysts say that real-valued data must always be represented in the computer by finite arithmetics and hence are affected by *rounding errors*. In the computer you can only store $\alpha + \delta\alpha$, where $\delta\alpha$ is the rounding error, *not* α.

B) In general there are no numerical procedures which implement *exactly* the function f or even if exact procedures are available , it may be inconvenient or uneconomical to use them. In practice f is computed approximately by some *algorithm* which implements an approximation, say, $g(\cdot)$, of $f(\cdot)$.

These are of course two distinct causes of errors which, however, always tend to sum up. Nevertheless it is convenient to discuss them separately.

Definition 3.1. *The numerical problem $x = f(\alpha)$ is* **ill-conditioned** *if small percentage errors on α generate large percentage error on the solution x. In other terms, letting*

$x = f(\alpha)$ and $x + \delta x = f(\alpha + \delta\alpha)$ one has

$$\frac{\|\delta x\|}{\|x\|} \gg \frac{\|\delta\alpha\|}{\|\alpha\|} \,. \tag{3.1.2}$$

Example 3.1. Consider the linear equation

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

whose (exact) solution is $x = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$. Introducing a small perturbation on b, say

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

the solution x becomes

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

In thi scase $\|\delta b\|/\|b\| \cong 10^{-4}$, while $\|\delta x\|/\|x\| = 1/\sqrt{2}$. Clearly, the error in the data δb is amplified by many orders of magnitude in the (exact) solution of the system.

Example 3.2. Consider now inverting the matrix

$$A = \begin{bmatrix} 100 & 100 \\ 100.2 & 100 \end{bmatrix}$$

A quick calculation shows that

$$A^{-1} = \begin{bmatrix} -5 & 5 \\ 5.01 & -5 \end{bmatrix}$$

Now suppose we want to invert the perturbed matrix

$$A + \delta A = \begin{bmatrix} 100 & 100 \\ 100.1 & 100 \end{bmatrix}$$

The inverse now is

$$(A + \delta A)^{-1} = \begin{bmatrix} -10 & 10 \\ 10.1 & -10 \end{bmatrix}$$

Evidently a 0.1% change in one entry of A has resulted in a 100% change in the entries of A^{-1}. This obviously affects in the same way the solution of the linear equation $Ax = b$.

J.H. Wilkinson, in his book *The Algebraic Eigenvalue Problem* (Oxford U.P. 1963), shows that the amplification factor in the solution of

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

is of the order of 10^{22} !

N.B. Ill-conditioning is an *intrinsic characteristic* of a numerical problem which cannot be modified by the use of special or "specially smart" algorithms. Errors due to ill-conditioning cannot be reduced or modified by the algorithm used to implement the computation of $x = f(\alpha)$. Nevertheless a well-conditioned problem can be "ruined" by a poor algorithm. Intuitively a "good" algorithm should perturb the theoretical f so little that the perturbation could well be attributed to rounding errors in the data.

Definition 3.2. *An algorithm g fo rthe numerical problem $x = f(\alpha)$, is* **numerically stable** *if for every $\alpha \in \mathbb{R}^k$ there is a perturbation $\delta\alpha$, of the same order of magnitude of the underlying rounding errors, such that $f(\alpha + \delta\alpha)$ differs from $g(\alpha)$ percentagewise of a quantity of the same order of $f(\alpha + \delta\alpha) - f(\alpha)$.*

In other words, the errors introduced by a numerically stable algorithm can always be attributed to errors due to the finite precision arithmetics. In other words, g is numerically stable if the computed solution $y = g(\alpha)$ can in principle be obtained by an "exact solver" using perturbed data, namely $y = f(\alpha + \delta\alpha)$ where $\|\delta\alpha\|/\|\alpha\|$ is of the same order of the underlying rounding errors.

Clearly no algorithm, no matter how numerically stable, can provide accurate solutions to an ill-conditioned problem. An unstable algorithm can however easily destroy a well-conditioned problem.

Remarks 3.1. *In Numerical Linear Algebra the perturbations considered are due to finite precision arithmetics (rounding errors) however the theory which follows does not depend at all on this interpretation and the perturbations on the data may have in fact any origin, say measurement noise or approximation errors of various kinds.*

Numerical Conditioning and the Condition Number

The normal equations are a special case of the ubiquitous linear system $Ax = b$. So we shall first discuss this problem assuming for the moment that $A \in \mathbb{R}^{n\times n}$ is nonsingular so that the solution of this problem is well-defined.

Assume for the moment that A has no perturbations ($\delta A = 0$); say can be stored exactly in the computer. Want to get an estimate of how much the relative error on the data $\|\delta b\|/\|b\|$ influences $\|\delta x\|/\|x\|$. For this purpose we shall use Euclidean norms

Recall that $\|A\|$ (normally denoted $\|A\|_2$ when there is a danger of confusion) is the smallest number $k > 0$ for which the inequality $\|Ax\| \leq k\,\|x\|$ holds. It can be computed as follows:

$$\|A\|^2 = \sup_{x\neq 0} \frac{x^\top A^\top A x}{x^\top x}\,. \tag{3.1.3}$$

the second member is known as a *Rayleigh quotient* and is actually equal to **maximal eigenvalue of** $A^\top A$, hence to the square of the maximal singular value of A:

$$\|A\|^2 = \max_i\ \lambda_i(A^\top A)\ = \sigma_1^2(A) \tag{3.1.4}$$

Problem 3.1. *Prove this equality.*

From the relations $x = A^{-1}\,b$ and $b = Ax$ one easily gets the estimates $\|\delta x\| \leq \|A^{-1}\|\|\delta b\|$ and $\|x\| \geq \|A\|^{-1}\|b\|$, so that

$$\frac{\|\delta x\|}{\|x\|} \leq \|A\|\,\|A^{-1}\|\,\frac{\|\delta b\|}{\|b\|} \tag{3.1.5}$$

The number $c(A) := \|A\|\,\|A^{-1}\|$ can be interpreted as an amplification gain of the errors on the right hand side of the linear system $Ax = b$. It is called **condition number** of the problem $Ax = b$ (or, of the matrix A). As we shall see in a moment, $c(A)$ has a more general meaning. First, let us observe that from $I = AA^{-1}$ it follows that

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

so that $c(A)$ is always an *amplification* coefficient.
Recalling that

$$\|A\|^2 = \lambda_{\rm MAX}\,(A^\top A)\,,$$
$$\|A^{-1}\|^2 = \lambda_{\rm MAX}\,(A^{-\top}\,A^{-1}) = \lambda_{\rm MAX}\,(AA^\top)^{-1} = \frac{1}{\lambda_{\rm MIN}(AA^\top)}$$

one immediately sees that

$$c^2(A) = \frac{\lambda_{\rm MAX}(A^\top A)}{\lambda_{\rm MIN}(A^\top A)} = \frac{\sigma_1^2(A)}{\sigma_n^2(A)} \tag{3.1.6}$$

where σ_1 and σ_n are the *maximal and minimal singular values of* A. In particular when A is symmetric,

$$c(A) = \frac{\lambda_{\rm MAX}(A)}{\lambda_{\rm MIN}(A)}\,. \tag{3.1.7}$$

From this formula one sees that when A is nearly singular, the minimum singular value is near zero and $c(A)$ may become large. However this is not always the case since for example $A = \epsilon I$ with $\epsilon \to 0$ has numerical conditioning equal to one. In any case the best conditioned matrices are those for which $A^\top A = \alpha I$. In this case one has $c(A) = 1$. These matrices are sometimes called *orthogonal* while those for which $AA^\top = I$ are *orthonormal*. Orthogonal matrices play a fundamental role in Numerical Linear Algebra.

Problem 3.2. *Compute the numerical conditioning of the* 2×2 *matrix in Example 3.1.*

Problem 3.3. *Assume that* A *is symmetric and* b *is parallel to the eigenvector of* A *corresponding to* $\lambda_{\rm MAX}$*, while* δb *is parallel to the eigenverctor of* A *corresponding to* $\lambda_{\rm MIN}$*. Show that one has* exactly*:*

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

Let's now examine the effect of rounding errors on A. Assume for the moment that $\delta b = 0$. It is immediate to see that a perturbation δx in the solution of $(A + \delta A)\,\bar{x} = b$ satisfies, up tho first order the relation

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

where of course $x = x + \delta x$ and $Ax = b$. From this it readily follows that

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

and when $c(A)\,\|\delta A\|/\|A\|$ is much smaller than 1,

$$\frac{\|\delta x\|}{\|x\|} \leq c(A)\ \frac{\|\delta A\|}{\|A\|} \quad , \tag{3.1.8}$$

which is an estimate of the same kind of (3.1.5). Hence the condition number $c(A)$ describes the effect of perturbations both on b as well as on the matrix A.

Case of A singular. This includes also the situation where A may be non-square and the solution is actually to be interpreted in the least-squares sense. We shall agree to look always for least-squares (LS) solutions of minimum norm. In this case the proper inverse to consider is the Moore-Penrose.

Problem 3.4. *Show that the formula for numerical conditioning in case of a general A (and solution to be interpreted in the LS sense) is*

$$c(A) = \|A\|\,\|A^{+}\| \tag{3.1.9}$$

where A^{+} is the Moore-Penrose pseudoinverse, see the end of Sect. B.2 for a formula defining the Moore-Penrose pseudoinverse.

Conditioning of the Least Squares Problem

I an attempt to solve an overdetermined system $Ax = b$ by multiplying both members of the equation by $A^{\top}$ one gets

$$A^{\top}Ax = A^{\top}b$$

which has the same form of the normal equations. Now the numerical conditioning of this problem is no longer the one of A but that of $A^{\top}A$. Just to get a rough estimate of what happens, let us suppose A is square. One has

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

It follows that even when the problem $Ax = b$ may be moderately well-conditioned, the normal equations may turn out to be badly ill-conditioned. Writing in exponential form $c(A) \cong 10^{c}$, c is a natural number which measures how many significant digits one looses in the numerical solution of $Ax = b$. Since $c(A)^2 = 10^{2c}$, by solving the problem (seemingly identical) $A^{\top}Ax = A^{\top}b$ one actually looses **twice as many** significant digits as in the solution of the original problem.

This means that solving the normal equations of a least squares problem $y \simeq S\theta$ is in general not a good idea. In the early 60's Gene Golub [39] has developed a different approach for attacking LS problems which is now universally used and extensively implemented e.g. in the Matlab package.

The QR Factorization

The imperative is to forget about the normal equations and work directly on the system!

Let's for the moment consider unweighted L S and a full column rank matrix S. Generalizations will be considered in the problems at the end of the section. We want to compute the LS estimate of a parameter θ by fitting N scalar observations y by the linear model

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

where ε is a vector denoting the approximation errors incurred in describing y by $S\theta$.
The p columns of $S = [s_1, \ldots, s_p]$ are linearly independent but in general not orthonormal. If they were so, $\langle s_i, s_j \rangle = s_i^\top s_j = \delta_{ij}$ and one would have $S^\top S = I$ so that the LS estimate $\hat{\theta}$ could be immediately written down as,

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

Note that in this case, $\hat{\theta}$ is just the vector of the first p coordinates of y with respect to the orthonomal basis $\{s_1, s_2, \ldots, s_p\}$ spanning the column space of S

$$\mathscr{S} := \text{span}\,\{s_1, s_2, \ldots, s_p\} = \text{Im}\,(S) \subset \mathbb{R}^N$$

The idea of the QR factorization is simply to orthonormalize the columns of S. This can be done by a well-known procedure called the *Gram-Schmidt algorithm*. This algorithm orthogonalizes sequentially the columns df $S = [s_1, \ldots, s_p]$ producing orthonormal vectors $\{q_1, \ldots, q_p\}$ defined by the relations

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

Solving with respect to $(s_1, \ldots, s_p)$ one obtains:

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

which can be written in matrix form as

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

or, more compactly,

$$S = \bar{Q}\,\bar{R} \quad , \tag{3.1.10}$$

where $\bar{Q} := [\,q_1, \ldots, q_p\,]$ is a $N \times p$ matrix with *orthonormal columns*; i.e. $\bar{Q}^\top \bar{Q} = I$ $(p \times p)$ and $\bar{R}$ is *upper triangular*.
Completing the basis $\{q_1, \ldots, q_p\}$ by adding $N - p$ vectors $\{q_{p+1}, \ldots, q_N\}$ so as to obtain an orthonormal basis for $\mathbb{R}^N$ and introducing the matrices

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

we can express S as

$$S = QR \quad , \tag{3.1.11}$$

which is the product of an *orthonormal times an upper triangular matrices*. This is the famous **QR factorization** of S.

Now we can use the QR factorization of S to solve our LS problem *without forming the normal equations*. Multiply both members of $y = S\theta + \varepsilon$ by $Q^\top$ to get

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

which we rewrite in partitoned form as

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} \bar{R} \\ 0 \end{bmatrix} \theta + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix} \quad , \tag{3.1.12}$$

Here y_1 and y_2 are the vectors of the components of y with respect to the two bases $\{\, q_1, \ldots, q_p \,\}$ and $\{\, q_{p+1}, \ldots, q_N \,\}$ spanning $\mathscr{S}$ and $\mathscr{S}^\perp$, namely

$$\begin{aligned} \operatorname{span}\{q_1 \ldots q_p\} &= \operatorname{span}\{s_1 \ldots s_p\} = \mathscr{S} \\ \operatorname{span}\{q_{p+1} \ldots q_N\} &= \mathscr{S}^\perp . \end{aligned}$$

It follows that $\begin{bmatrix} y_1 \\ 0 \end{bmatrix}$ is the orthogonal projection of y onto $\mathscr{S}$ (expressed with respect to the coordinates $\{q_i\}$) and $\begin{bmatrix} 0 \\ y_2 \end{bmatrix}$ is the projection of y on the orthogonal complement $\mathscr{S}^\perp$ and therefore coincides with the *residual estimation error* $\hat{\varepsilon} = y - Py$. The meaning of ε_1 and ε_2 will be discussed in a moment.

Recall now that solving our LS problem for θ just requires to minimize the norm of the approximation error $\varepsilon = \varepsilon(\theta) = y - S\theta$. Hence, since $Q^\top$ is an orthonormal matrix which preserves norms, this is the same as minimizing

$$\|Q^\top y - Q^\top S\theta\|^2 = \left\| \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} - \begin{bmatrix} \bar{R}\theta \\ 0 \end{bmatrix} \right\|^2 = \|y_1 - \bar{R}\theta\|^2 + \|y_2\|^2 . \tag{3.1.13}$$

Since the second term does not depend on θ, $\hat{\theta}$ can be computed by solving the p-dimensional system

$$\bar{R}\theta = y_1 \quad , \tag{3.1.14}$$

which is particularly simple since $\bar{R}$ is upper triangular and the solution can be computed by successive substitutions starting from the last(lowest) equation.
The estimation residual $\hat{\varepsilon} = \varepsilon(\hat{\theta})$ has norm equal to

$$\|\hat{\varepsilon}\|^2 = \|y_2\|^2 . \tag{3.1.15}$$

Morale: In the new coordinate system, $\varepsilon_1(\theta) := y_1 - \bar{R}\theta$ is the part of the approximation error which can be made null by chosing $\theta = \hat{\theta}$. In other words, with this choice one can describe exactly the first p components of the data y_1 with he model $Q^\top S\theta$.
One may argue that, since N is generally very large, forming Q, which is $N \times N$, could be very expensive. However in the actual solution algorithms, Q is never formed explicitly. In practice one starts with the data in a table

$$[S \mid y] \tag{3.1.16}$$

and by successive orthonormalization steps transforms it to the structure

$$\left[\begin{array}{c|c} \bar{R} & y_1 \\ \hline 0 & y_2 \end{array}\right] . \tag{3.1.17}$$

where $\bar{R}$ is $p \times p$ upper triangular. Besides the Gram-Schmidt algorithm there are several other procedures to accomplish this upper triangularization, such as the so-called Housholder algorithm or the Givens rotations. For these we shall refer to classical textbooks such as [54].

When an a priori statistical description of the error is available, ε becomes a random variable, say

$$\varepsilon \equiv \mathbf{w}\,, \qquad \mathbb{E}\,\mathbf{w} = 0\,, \qquad \text{Var}\,(\mathbf{w}) = \sigma^2 I\,.$$

In this case it is of interest to compute the variance of the estimate $\text{Var}\;\hat{\theta} = \sigma^2\,[S^\top S]^{-1}$ which, using the QR-factorization is a function of $\bar{R}$ alone

$$\text{Var}\;\hat{\theta} = \sigma^2\,(\bar{R}^\top \bar{R})^{-1}\,. \tag{3.1.18}$$

and can also be computed by the QR factorization.

The role of orthogonality

This example is from Strang's book **[?]**. Assume that we want to approximate a real function $f(x)$ on the interval $[0, 1]$ by a polynomial of fixed degree n, say $P_n(x)$. Let us choose as an approximation measure the mean square deviation which leads to solving the minimization problem

$$\min_{P_n(x)} \int_0^1 |f(x) - P_n(s)|^2\,dx\,.$$

This is also a linear Least-Squares problem on finite dimensional inner product spaces. Expressing $P_n(x)$ as

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

where $s^\top(x) = [1\;x, \ldots x^n]$ it is clear that P_n is just one element of the $n+1$-dimensional inner product space:

$$\mathscr{S} := \text{span}\,\{1\,, x \ldots,\, x^n\,\} \qquad x \in [\,0,\,1\,]$$

with the scalar product of functions on the interval $[0,1]$ defined by $\langle f,g\rangle = \int_0^1 f(x)\,g(x)\,dx$.
Imposing the orthogonality pronciple

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

one finds the normal equations for this problem

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

which have the explicit expression

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

The symmetric matrix on the left is the analog of $S^\top S$. It is the celebrated *Hilbert matrix* which is terribly ill-conditioned. For $n = 10$ the numerical conditioning of this matrix is about 10^{13}. This seems to render polynomial approximation an impossible problem!

In reality we know very well that this actually is a routine problem in numerical analysis. The key tool which makes this a standard problem is the use of *orthogonal polynomials*. If instead of $(1\; x \dots x^n)$ we start with linearly independent polynomials $p_0(x)\,p_1(x)\,\dots\,p_n(x)$ such that $\langle p_i,p_j\rangle = \delta_{ij}$, the least squares approximation

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

can simply be obtained by computing the scalar products

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

and using the parameter estimates

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

This is a universal idea which lies at the grounds for example of the Fourier series expansion.

Problem 3.5. *Show that a weighted LS problem*

$$\min_\theta \|y - S\theta\|_W$$

with weighting matrix $W = W^\top > 0$, can be solved by a QR factorization algorithm based on Gram-Schmidt with inner product $\langle\cdot,\cdot\rangle_W$. In particular what properties should the Q matrix have?

Fourier series and least squares

Considerthe following Problem: Given a continuous function $y(t)$ on the interval $[-T/2, T/2]$, find a linear combination of the functions $1,\ \sin\frac{2\pi}{T}\,t, \ldots, \sin\frac{2n\pi}{T}\,t$, $\cos\frac{2\pi}{T}\,t, \ldots, \cos\frac{2n\pi}{T}\,t$, with coefficients θ_i, $i = 0, 1, \ldots, 2n$, say

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

which approximates y. This was the original approach of Joseph Fourier [8] to Fourier series expansion. The approximation criterion is actually the average squared error

$$V(\theta) = \frac{1}{T}\int_{-T/2}^{T/2} |\, y(t) - f_n(t,\theta)\,|^2\, dt\,.$$

Consider the vector space $\mathcal{C}_T$ of continuous functions on $[-T/2,\ T/2]$ with scalar product $\langle f, g\rangle = \int_{-T/2}^{T/2} f(t)\, g(t)\, \frac{dt}{T}$. The functions $\{\sin k\frac{2\pi}{T}\,t\,;\ \cos k\frac{2\pi}{T}\,t;\ \ k = 0, 1, \ldots, n\}$ are orthonormal with respect to this inner product and hence form an orthonormal basis for a $2n+1$-dimensional subspace $\mathcal{S}$ of $\mathcal{C}_T$.Then $V(\theta)$ can be interpreted as the distance in $\mathcal{C}_T$ of the function y to a generic element $f_n(\cdot,\theta)$ of the subspace $\mathcal{S}$.

We are therefore considering a least squares problem with a linear parametric model $f_n(t,\theta) \simeq S\theta$ to approximate y. Due to orthonormality, the solution parameters can be obtained simply by computing the inner products of y with the basis functions. *The components of the Least Squares Estimate $\hat\theta$ are then exactly the first $2n+1$ Fourier coefficients of y.* Fourier went on trying to extend the expansion to an infinite sequence of sinusoids but the mathematicians of his time did not appreciate his efforts.

Use of the SVD in Linear Least Squares

The Singular Value Decomposition (SVD) is probably the most important instrument to analyze linear regression problems in depth. It has been brought into this field by G. Golub and co-workers [39] followed by a long series of contributions which are now classical, see e.g. [62, 54]. Let us consider again a standard linear model where we shall not need to assume that S is of full rank p but we shall keep the assumption that $N > p$. Let $S = U\begin{bmatrix}\Sigma\\ 0\end{bmatrix} V^\top$ be the SVD of the signal matrix where U is a $N \times N$ matrix with orthonormal columns and Σ is a $p \times p$ diagonal matrix exhibiting the (ordered) singular values of S. Change basis in the observation space $\mathbb{R}^N$ and in the parameter space $\mathbb{R}^p$ by the orthonormal matrices U and V to get

$$\hat y := U^\top y \qquad \beta := V^\top\theta\,, \qquad \bar w := U^\top w$$

so that the standard linear model is transformed into

$$\bar y = \begin{bmatrix}\Sigma\\ 0\end{bmatrix}\beta + \sigma_{\mathbf{w}}\bar w \tag{3.1.19}$$

[8] Joseph Fourier, *Mémoire sur la propagation de la chaleur dans les corps solides* (1807)

(the noise standard deviation $\sigma_{\mathbf{w}}$ is the same as that of the original model) and the unweighted Least Squares problem $\min_\theta \|y - S\theta\|^2$ is equivalent to

$$\min_\beta \| \begin{bmatrix} \bar{y}_1 \\ \bar{y}_2 \end{bmatrix} - \begin{bmatrix} \Sigma \\ 0 \end{bmatrix} \beta \|^2$$

where $\bar{y}_1 := \Sigma\beta$ are the first p components of $\bar{y}$ and $\bar{y}_2$ is made of the last $N - p$ components of $\bar{w}$. One may actually partition conformably U as

$$U = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \tag{3.1.20}$$

where U_1 is $N \times p$ and has orthonormal columns. Then $\bar{y}_1 = U_1^\top y$. As we have already seen, the LS minimization is equivalent to $\Sigma\,\beta = \bar{y}_1$. Hence, when Σ is non singular (i.e. $\mathrm{rank}\ S = p$) one gets the solution

$$\hat{\beta} = \Sigma^{-1}\,\bar{y}_1\,, \qquad \Leftrightarrow \quad \hat{\theta} = V\Sigma^{-1}U_1^\top y\,.$$

In this way the parameter estimates are found by inspection:

$$\hat{\boldsymbol{\beta}}_i(\mathbf{y}) = \frac{1}{\sigma_i}\bar{\mathbf{y}}_{1,i}; \qquad \mathrm{var}\,\{\hat{\boldsymbol{\beta}}_i\} = \frac{\sigma_{\mathbf{w}}^2}{\sigma_i^2} \qquad i = 1, \ldots, p\,. \tag{3.1.21}$$

where σ_i are the singular values of S. This relation actually holds even when $\mathrm{rank}\ S = r < p$; just by discarding the last $p - r$ equalities.
It is easy to check that the variance matrix of $\hat{\boldsymbol{\beta}}$ is proportional to Σ^{-2}; therefore the estimators $\hat{\boldsymbol{\beta}}_i(\mathbf{y})\,;\ i = 1, \ldots, p$ are uncorrelated (or independent in a Gaussian model). Since the singular values are ordered in decreasing magnitude one may say that **the variance of the parameter estimates increases with the model complexity** p. With N fixed, by increasing the number of parameters in the model one does in general worsen the quality of the parameter estimates. In particular, adding one more regressor; that is adding one more column s_{p+1}, although the σ_i will in general change, one will add a smaller singular value σ_{p+1} [93] and incur in a larger variance of the parameter estimates, in fact not just of the additional $\hat{\boldsymbol{\beta}}_{p+1}$. Note that in the orthonormalized model (3.1.19) *the estimator variance ratio is equal to the condition number of the Least Square problem*:

$$\frac{\mathrm{var}\ \hat{\boldsymbol{\beta}}_p}{\mathrm{var}\ \hat{\boldsymbol{\beta}}_1} = c^2(S) \tag{3.1.22}$$

So badly conditioned linear models will yield parameter estimates of large variance. In other words *ill-conditioning is directly related to the variance of the estimates*.

We should stress that the variance matrix of the original parameter estimator $\hat{\boldsymbol{\theta}} = V\hat{\boldsymbol{\beta}}$ is $V\,\mathrm{Var}\,\{\hat{\boldsymbol{\beta}}\}V^\top$, so that , since V is an orthonormal matrix, the trace of the variance matrices of $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\beta}}$ are the same. In other words, the scalar variances of the two estimators are the same.

In conclusion, we have learned that complicated regression models with many parameters may generally lead to ill-conditioned regression problems and to parameter estimates with a high variance. When it is necessary to attach more regressors one should try to introduce new columns which are "almost orthogonal" to the existing columns of S so as to keep the condition number of S within reasonable limits. This issue will be discussed in more detail in Chap. 7

3.2 ▪ Introduction to Linear Inverse Problems

Inverse Problems are extremely common in Engineering and applied sciences. In general they arise when one wants to recover inputs/cause from output/effect; e.g. recovering forces acting on an object from its motion (Newton), in image deconvolution (deblurring) and, quite often in statistical problems where one wants to recover a model from measured data. Typical inverse problems which are a prototype of the problem of system identification are:
Recovering initial conditions from given solutions to ODE or PDE
Recovering the differential or difference equation governing a dynamical system from (measurements of) a solution trajectory or from an input-output pair of trajectories.

The most elementary example is the ubiquitous linear algebra relation $Ax = b$. Here we can look at the equation in two ways:
1. Given A and x, compute $b = Ax$ (Direct problem). The solution only requires matrix multiplication.
2. Given A and b recover x (Inverse problem). The solution requires inversion of the matrix A.

In more general problems when there are functions involved, inverse problems require an operator inversion.

Inverse problems are *sensitive to perturbations*. For the $Ax = b$ problem this sensitivity is captured by the notion of *ill-conditioning*. We need a more general concept for physical problems where A is an operator acting on functions and x and b may be functions (often b might be the discretization of an observed function).

Ill-posed problems

We shall consider only *linear* inverse problems. There is some linear operator $\mathcal{A}$ acting on a Hilbert space $\mathcal{X}$ and taking values in some other Hilbert space $\mathcal{Y}$. The following is Hadamard's definition of an ill-posed problem.

Definition 3.3. *The problem*

$$\mathcal{A}x = y; \qquad x \in \mathcal{X}, \quad y \in \mathcal{Y}; \quad \mathcal{A} : \mathcal{X} \to \mathcal{Y}$$

is said to be **well-posed (in the sense of Hadamard)** *if the following conditions hold:*

1. *For each $y \in \mathcal{Y}$ there exists $x \in \mathcal{X}$, such that $\mathcal{A}x = y$ (existence)*
2. *For each $y \in \mathcal{Y}$ there exists a* **unique** *$x \in \mathcal{X}$, such that $\mathcal{A}x = y$ (uniqueness)*
3. *The solution depends continuously on the data y.*

When at least one of the three conditions above does not hold the problem is said to be **ill-posed** .

Examples of ill-posed problems which are often encountered in signal processing are :
Recovering a continuous function $f(t)$ from its sample values $\{f(t_k);\ k = 1, 2, \ldots\}$. Here the sampling operator does not have an inverse. Even worse:

Recovering the derivative of a function $f(t)$ from its sample values $\{f(t_k);\ k = 1, 2, \ldots\}$.
Solving an integral equation

$$y(t) = \int_T k(t-s)x(s)ds$$

possibly from sampled values $\{y(t_k);\ k = 1, 2, \ldots\}$ is generally ill-posed. An important example is *deconvolution* which is the prototype problem of Dynamic System Identification from input-output data and occurs in many other scientific areas such as Medical Imaging, Physical Chemistry, Extragalactic Astronomy etc. see [68], [9], [19].

From ill-posed to ill-conditioned

In practice all problems need to be solved by discretization. So one needs to transform $\mathcal{A}x = y$ into $Ax = b$. If the original problem was ill-posed then the discretized one is normally *ill-conditioned*. The solution need not exist and even if it does, the effect of small perturbations on b can be large variations of x.

Example 3.3. Consider the problem of recovering a signal $x(t)\,;\, t \in [0\,, T]$ from its integral, say

$$y(t) = \int_0^t x(s)\,ds \qquad \text{that is} \qquad y(t) = \int_0^T 1(t-s)x(s)\,ds$$

where $1(t)$ is the unit step function equal to 1 for $t \geq 0$ and zero elsewhere. The problem of recovering x from y is ill-posed. There are several reasons why it is so. Give at least one.

In practice you only have discrete measurements $\{y(kh)\,;\, k = 1, \ldots, N = T/h\,\}$ (assume N is an integer). Write the discretized problem as a linear system

$$y = Ax\,, \qquad y \in \mathbb{R}^N\,;\, x \in \mathbb{R}^N$$

and describe the structure of $A \in \mathbb{R}^{N\times N}$. What can you say about the condition number of A. Would it get better for $h \to 0$?

Solution: The inverse of the integral transform is obviously the derivative $x(t) = \dfrac{d}{dt}y(t)$ which we write symbolically as a linear operator $x = Dy$. Now unless y is a function in special spaces, this operator is never continuous. For example, assuming y is a continuous function, its derivative may not be continuous, could jump very wildly and there is in general no constant k such that $\|x\| \leq k\,\|y\|$. This actually happens in a large variety of situations.

The matrix A in the discretized problem has the form $A = hL$ where

$$L = \begin{bmatrix} 1 & 0 & \ldots & \ldots & \ldots & 0 \\ 1 & 1 & 0 & \ldots & \ldots & 0 \\ 1 & 1 & 1 & \ldots & \ldots & 0 \\ 1 & \ldots & \ldots & \ldots & \ldots & 1 \end{bmatrix} := \begin{bmatrix} a_1^\top \\ a_2^\top \\ \ldots \\ a_N^\top \end{bmatrix} \tag{3.2.1}$$

and for N very large the first columns (or equivalently the last rows) are nearly linearly dependent, the more so the larger is N. The condition number can be

roughly estimated as follows. Since

$$LL^\top \equiv \left[a_i^\top a_j\right]_{i,j=1,\dots,N} = \begin{bmatrix} 1 & 1 & \dots & \dots & 1 & 1 \\ 1 & 2 & 2 & \dots & \dots & 2 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & 2 & \dots & \dots & N-1 & N-1 \\ 1 & 2 & \dots & \dots & N-1 & N \end{bmatrix},$$

picking $v_1 := \begin{bmatrix} 1 & 0 & \dots & \dots & \dots & 0 \end{bmatrix}^\top$ and $v_2 := \begin{bmatrix} 0 & 0 & \dots & \dots & 0 & 1 \end{bmatrix}^\top$, by the Rayleigh quotient theorem we have

$$N = v_2^\top LL^\top v_2 \le \sigma_{max}^2(L)\,, \qquad 1 = v_1^\top LL^\top v_1 \ge \sigma_{min}^2(L)$$

and hence, since $\sigma_k(A) = h\,\sigma_k(L)\,, k = 1, 2, \dots, N$,

$$c(A)^2 = \frac{\sigma_{max}^2(L)}{\sigma_{min}^2(L)} \ge N$$

which tends to ∞ with N. Therefore, the smaller we choose h the worse is the condition number of A. This somewhat counter-intuitive argument applies to a multitude of inverse problems involving discretization and needs to be kept in mind. □

How do we treat an ill-posed problem which has no solution? The main idea is to "relax" it by turning it into an optimization problem. One looks for a best approximate solutions which is guaranteed to exist. Typically convert it into a least squares problem say:

$$\min_x \|y - \mathcal{A}x\|$$

yet the solution can still be very wild. Recall what happens with finite-dimensional linear least squares problems: Random perturbations on y are generally *amplified by ill-conditioning*. We want to *constrain the solution to be smooth*! This is done by **regularization**.

3.3 ▪ Regularized Least Squares problems

The idea of regularization is attributed to Tikhonov but Tikhonov regularization has been invented independently in many different contexts. It became widely known from its application to integral equations from the work of A.Tikhonov and D. L. Phillips [97], [68], [96].

Definition 3.4. *A* **regularization** *of the Least Squares problem* $\min_x \|y - \mathcal{A}x\|_{\mathcal{Y}}$ *is an optimization problem having the form*

$$\min_x \left\{ \|y - \mathcal{A}x\|_{\mathcal{Y}}^2 + \lambda \|x\|_{\mathcal{X}}^2 \right\}, \qquad \lambda \ge 0 \tag{3.3.1}$$

where the norm $\|x\|_{\mathcal{X}}$ *should weight large variations of the solution but may otherwise be arbitrary. The variable* λ*, called the* regularization parameter *is a design parameter which can be chosen by the user.*

The norm $\|x\|_{\mathcal{X}}^2$ can be chosen in many different ways, say as L^2 or ℓ^2 norms but the most perspicuous choice is done within the theory of Reproducing Kernel Hilbert Spaces (RKHS) which we can only touch very superficially in these notes; see Section 6.6 and e.g. [44, p. 167] for a short survey.

Theorem 3.1. *A necessary and sufficient condition for x to be a solution of the optimization problem* (3.3.1) *is that it should satisfy the Euler equation*

$$(\mathcal{A}^*\mathcal{A} + \lambda I)x = \mathcal{A}^* y \tag{3.3.2}$$

where $\mathcal{A}^*$ is the Hilbert space adjoint operator of $\mathcal{A}$.

Proof. Assume that λ has been fixed and let us denote by $\varphi(x, y)$ the quadratic functional between braces in (3.3.1). Then x is an optimal solution if for arbitrary real number α and any arbitrary element $\alpha h \in \mathcal{X}$ one has

$$\varphi(x, y) \le \varphi(x + \alpha h, y)\,. \tag{3.3.3}$$

Using linearity in the second argument of the inner product we obtain

$$\begin{aligned}\varphi(x + \alpha h, y) = \|y - \mathcal{A}x\|_{\mathcal{Y}}^2 + \lambda\,\|x\|_{\mathcal{X}}^2 + \\ - 2\alpha\{\,\langle y - \mathcal{A}x, \mathcal{A}h\rangle_{\mathcal{Y}} - \lambda\,\langle x, h\rangle_{\mathcal{X}}\,\} + \\ + \alpha^2\{\|\mathcal{A}h\|^2 + \lambda\|h\|^2\}\end{aligned}$$

Now for the inequality (3.3.3) to hold the middle term must be greater or equal to zero for arbitrary real number α and any arbitrary element $\alpha h \in \mathcal{X}$. Clearly this can be true only if the term between braces is zero, which is equivalent to

$$\langle y - \mathcal{A}x, \mathcal{A}h\rangle_{\mathcal{Y}} - \lambda\langle x, h\rangle_{\mathcal{X}} = 0$$

for all $h \in \mathcal{X}$. Now the adjoint $\mathcal{A}^*$ is a (possibly unbounded) operator mapping $\mathcal{Y}$ into $\mathcal{X}$ which can be moved to the left side of the inner product to yield

$$\langle \mathcal{A}^* y - \mathcal{A}^*\mathcal{A}x - \lambda I x, h\rangle_{\mathcal{X}} = 0$$

for all $h \in \mathcal{X}$. This is equivalent to the Euler equation (3.3.2). □

Corollary 3.1. *Assume $\mathcal{X}$, $\mathcal{Y}$ are finite dimensional inner product spaces with*

$$\langle \xi,\, \eta\rangle_{\mathcal{Y}} = \xi^\top Q \eta \qquad \langle \xi,\, \eta\rangle_{\mathcal{X}} = \xi^\top W \eta\,,$$

where Q and W are positive definite. Then the solution of the regularized least squares problem (3.3.1) *is*

$$\hat{x} = \left[A^\top Q A + \lambda W\right]^{-1} A^\top Q y \tag{3.3.4}$$

When $\lambda \to 0$

$$\lim_{\lambda\to 0}\left[A^\top Q A + \lambda W\right]^{-1} A^\top Q = A^+$$

where A^+ is the Moore-Penrose pseudoinverse of A.

We will see later an interpretation in terms of linear Bayesian estimation. For this material good references are [9],[101] and the web site
`https://en.wikipedia.org/wiki/Inverse_problem`.

One thing that is usually overlooked when fitting linear regression models to data is that often the inputs are not properly normalized and may have widely different ranges of variation resulting in some columns of the matrix S to have entries which are of several order of magnitudes smaller (or larger) than the others. The opposite then usually happens with the parameter estimates which may turn out to be of exceedingly large (or exceedingly small) magnitude. This, as we shall see below, should be avoided either by proper normalization or by changing the measurement units.

Example 3.4. *In a two-parameter linear model with i.i.d. error of variance σ^2, the two columns $[s_1,\ s_2]$ of the matrix $S \in \mathbb{R}^{N\times 2}$ are orthogonal and*

$$\|s_2\| = 10^{-6}\|s_1\|$$

Show that for any measurement y, the components of the parameter $\hat{\theta}$, least squares solutions of $y = S\theta + w$, are related by $\hat{\theta}_2 = \alpha\hat{\theta}_1$. Find α.
Solution: *By orthogonality we get the formulas*

$$\hat{\boldsymbol{\theta}}_1(\mathbf{y}) = \frac{1}{\|s_1\|^2} s_1^\top \mathbf{y} \qquad \hat{\boldsymbol{\theta}}_2(\mathbf{y}) = \frac{1}{\|s_2\|^2} s_2^\top \mathbf{y}$$

so that

$$\frac{\hat{\boldsymbol{\theta}}_2(\mathbf{y})}{\hat{\boldsymbol{\theta}}_1(\mathbf{y})} = \frac{\|s_1\|^2}{\|s_2\|^2}\frac{s_2^\top \mathbf{y}}{s_1^\top \mathbf{y}} = \frac{\theta_2 + \dfrac{s_2^\top \mathbf{w}}{\|s_2\|^2}}{\theta_1 + \dfrac{s_1^\top \mathbf{w}}{\|s_1\|^2}}$$

It is clear that the random error on $\hat{\boldsymbol{\theta}}_2$ can generally be much larger than that on $\hat{\boldsymbol{\theta}}_1$, roughly by a factor of 10^6. In fact, no matter the values of true parameters θ_1, θ_2, we have

$$\operatorname{var}\left(\hat{\boldsymbol{\theta}}_2(\mathbf{y})\right) \simeq 10^{12}\ \operatorname{var}\left(\hat{\boldsymbol{\theta}}_1(\mathbf{y})\right).$$

Hence the sample value of $\hat{\boldsymbol{\theta}}_2$ could be millions of times larger than that of $\hat{\boldsymbol{\theta}}_1$, in spite of the fact that both estimators are unbiased. □

More generally, that some components of $\hat{\theta}$ may take on very large values, happens in poorly conditioned problems when the columns of S are nearly dependent. In this case the parameter estimate $\hat{\theta}$ will have a large variance and some parameter estimates my have large size and almost cancel with other parameter components which have also large values but of opposite sign. For this reason it is a good idea to add to the least squares cost a penalty term which penalizes too large values of the components of θ. In fact the penalty term will, as we shall see, improve the conditioning of the problem.

A *Ridge Regression Problem* is a regularized Least Squares problem where the penalty term is a quadratic norm of θ. Usually one takes the plain ℓ^2-norm leading to the minimization problem

$$\min_\theta \{\|y - S\theta\|_Q^2 + \lambda\|\theta\|_2^2\} \tag{3.3.5}$$

where $\|\theta\|_2^2 = \sum_k \theta_k^2$. This is equivalent to minimization of $\|y - S\theta\|_Q^2$ subject to the constraint $\|\theta\|_2^2 \leq c$ for some c which corresponds to a spherical region

in $\mathbb{R}^p$. At the optimum the iso-cost lines of the Least Squares functional will be tangent to the surface of this sphere.

The *Ridge estimate* $\hat{\theta}_R$, solution of the minimization problem (3.3.5), is just a particularization of formula (3.3.4), namely

$$\hat{\theta}_R = \left[S^\top Q S + \lambda I_p\right]^{-1} S^\top Q\, y := A_\lambda\, y \tag{3.3.6}$$

Clearly, for $Q = R^{-1}$ and when S has full column rank, the limit for $\lambda \to 0$ of A_λ is just the matrix A of (2.3.7) and the ridge estimator will tend to the ordinary least squares estimator $\hat{\theta}$. Note that $\hat{\theta}_R$ is *biased* as are in general all regularized least squares estimators. We shall examine this property in detail in Section 5.9. However the variance matrix may actually turn out to be *smaller than the variance of the Least Squares estimate*.

Shrinkage: Assume for simplicity that $Q = I$; then from the SVD analysis of Section 3.1 one can see that $S^\top S = V\Sigma^2 V^\top$ and hence

$$\begin{aligned}\hat{\theta}_R &= \left[V\Sigma^2 V^\top + \lambda I_p\right]^{-1} V\Sigma U_1^\top\, y = \\ &= V\Sigma\, \left[\Sigma^2 + \lambda I_p\right]^{-1} \bar{y}_1 := A_\lambda\, y\end{aligned} \tag{3.3.7}$$

so that the components of the ridge estimator in the basis spanned by the columns of V are

$$\hat{\boldsymbol{\beta}}_i = \frac{\sigma_i}{\sigma_i^2 + \lambda}\, \bar{\mathbf{y}}_{1,i}\,, \qquad i = 1, 2, \ldots, p \tag{3.3.8}$$

This is called *Shrinkage*. Comparing with (3.1.21) one sees that the smaller is σ_i the more the components of $\hat{\boldsymbol{\beta}}$ are shrunken with respect to their LS counterparts.

The LASSO

LASSO stands for *least absolute shrinkage and selection operator*. It is a similar regularization problem to the Ridge Regression but wit the ℓ^2 norm of the parameter substituted by the ℓ^1 norm:

$$\min_\theta \left\{ \|y - S\theta\|^2 + \lambda \sum_{k=1}^{p} |\theta_k| \right\} \tag{3.3.9}$$

where we have taken $Q = I$ for simplicity. There are also weighted versions of the cost and more general penality function which are discussed in the literature [44, p.68], [112] but we shall not deal with. As for the ridge functional, the form (3.3.9) can be interpreted as the Lagrangian formulation of the minimization of the square norm $\|y - S\theta\|^2$ subject to the constraint

$$\sum_{k=1}^{p} |\theta_k| \le c\,.$$

One can see that the constraint region defined by the ℓ^1 norm is a rotated hypercube (in general a convex polytope), so that its corners lie on the axes while the region defined by the ℓ^2 norm is a p-sphere, which is rotationally invariant and, therefore, has no corners. As seen in the figure,

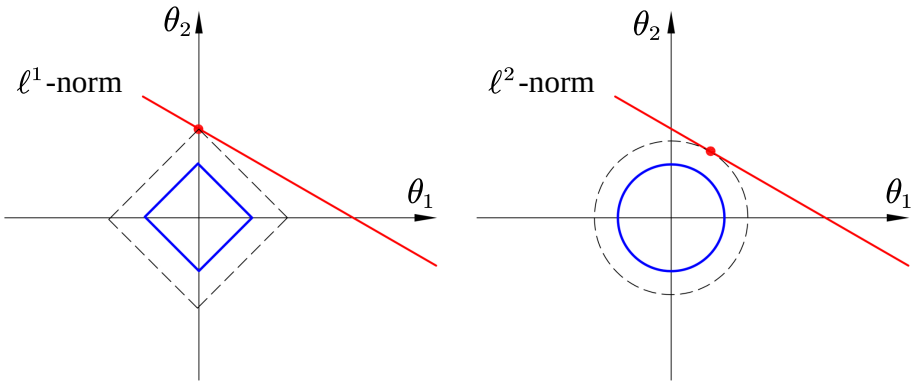


Figure 3.3.1. *Lasso vs Ridge*

a convex object that lies tangent to the boundary, such as the line shown, is likely to encounter a corner (or in higher dimensions an edge or higher-dimensional equivalent) of a hypercube, for which some components of θ are identically zero, while in the case of a p-sphere, the points on the boundary for which some of the components of θ are zero are not distinguished from the others and the convex object is no more likely to contact a point at which some components of θ are zero than one for which none of them are.

The parameter c in the constraint equation has a complicated relation with the multiplier λ although we can say that its effect is roughly like that of the reciprocal $1/\lambda$. Making c small will make some of the parameter estimates to be exactly zero. Taking c larger than the sum of the LS estimates, i.e.

$$c \geq \sum_{k=1}^{p} |\hat{\theta}_k| := c_0$$

will make the Lasso estimates coincide with the $\hat{\theta}_k$. One may guess that taking $c = c_0/2$ will cause a shrinkage of about 50%. There is however no precise rule describing the amount of shrinkage nor the number of parameters which are set to zero in the Lasso. As in the ridge regression c or λ need to be adjusted to minimize (an estimate of) the expected prediction error.

Warning: In regression problems with an unknown mean value μ of $\mathbf{y}$, it does not make sense to shrink μ. Even if μ is unknown, one should not add the mean as an extra parameter θ_0 (by introducing a column of 1's in the S matrix) but should rather center all input and output variables with respect to their sample mean as done for example in Example 2.2.

Example 3.5. *Suppose you have a linear model* $\mathbf{y} = S\theta + \sigma\mathbf{w}$ *where* S *is* $N \times p$ *but the error has an unknown systematic componente called* μ*. You augment* S *with a column*

of 1's and compute a ridge regression estimate of θ by minimizing the criterion

$$\min_{\theta,\mu}\{\|\mathbf{y} - S\theta - \mathbb{1}\mu\|^2 + \lambda\|\theta\|^2\}\,.$$

Compute the estimate of μ. What is its bias?

Solution *The minimization with respect to μ, done separately, does not involve the regularization term. Assuming for the moment θ known, we get*

$$\hat{\boldsymbol{\mu}} = (\mathbb{1}^\top\mathbb{1})^{-1}\mathbb{1}^\top(\mathbf{y} - S\theta) = \frac{1}{N}\sum_{k=1}^{N}(\mathbf{y}_k - s_k^\top\theta)$$

where $s_k^\top$ is the k-th row of S. We should really substitute the (regularized) estimate $\hat{\boldsymbol{\theta}}$ in place of θ but it is easy to see that $\sum_{k=1}^{N} s_k^\top\theta$ can be zero and the estimator unbiased no matter what value of θ, only if all columns of S are **centered**, *that is, to each element $s_{k,j}$; $k = 1,\dots,N$ of the $j-$th column one subtracts the mean*

$$\bar{s}_j = \frac{1}{N}\sum_{k=1}^{N} s_{k,j}$$

In this case the $s_{k,j}$ are deviations from their mean and their row-wise sum (with respect to the row-index k) is zero. The estimate of μ is then the sample mean $\bar{\mathbf{y}}_N$ and its variance is obviously σ^2/N.

If the columns of S are not centered the estimate of μ depends on the estimate of θ which is biased and $\hat{\boldsymbol{\mu}}$ will also be biased. The regularized estimate of θ is

$$\hat{\boldsymbol{\theta}} = [S^\top S + \lambda I_p]^{-1}S^\top(\mathbf{y} - \mathbb{1}\hat{\boldsymbol{\mu}})$$

this expression is coupled to that for $\hat{\boldsymbol{\mu}}$. In this general case one should use a joint estimator which could be obtained via Corollary 3.1 by setting $W = \operatorname{diag}\{0, I_p\}$.
You may show that the formula in fact holds even if W is not invertible. □

NB: One great advantage of regularization is that we do not need to impose the limit $p \leq N$ and in this way we may even be able to treat problems with more unknowns than data. This may well be the case in biological or bio-medical applications see [36, Sec. 5.2]. In this context, the *Variable Selection* operated by the Lasso is particularly useful and is one reason of its great success in applications.

3.4 ▪ Algorithms for the Lasso and Variable Selection

In this section we shall describe a family of techiques to compute the Lasso estimate, the solution of the optimization problem (3.3.9). We shall survey an efficient algorithm for computing the estimate discussed in [44, p. 73-78]. More details can be found in the book [94].

Let us recall the original problem formulation set as a constrained "Primal" optimization problem

$$\begin{aligned} &\underset{\theta}{\text{minimize}} && \frac{1}{2N}\|\mathbf{y} - S\theta\|_2^2 \\ &\text{subject to} && \|\theta\|_1 \leq t, \end{aligned} \tag{P}$$

where $\mathbf{y} \in \mathbb{R}^N, S \in \mathbb{R}^{N\times(p+1)}, \theta \in \mathbb{R}^{p+1}$ and we have added a normalization factor $\frac{1}{2N}$ for convenience.

The optimization problem can be written in standard Lagrangian form

$$\underset{\theta}{\text{minimize}} \quad \frac{1}{2N}\|\mathbf{y} - S\theta\|_2^2 + \lambda\|\theta\|_1 \tag{L}$$

where the regularization parameter $\lambda > 0$ has now the meaning of Lagrange multiplier. Since we want to compare the "size" of various components of θ and try to eliminate inessential variables, a preliminary *data normalization* is essential. To this end, from the $p+1$-dimensional problem ($\mathbf{1}$ is a column vector of ones) :

$$\underset{\theta}{\text{minimize}} \quad \frac{1}{2N}\|\mathbf{y} - \mathbf{1}\theta_0 - S\theta\|_2^2 + \lambda\|\theta\|_1$$

eliminate the intercept θ_0 (which does not make sense to shrink or eliminate) by centering, that is subtracting from this equation the sample averages from both sides so as to achieve the conditions

$$\frac{1}{N}\sum_{i=1}^{N} y_i = 0, \quad \frac{1}{N}\sum_{i=1}^{N} s_{ij} = 0, \quad j = 1, 2, \ldots, p.$$

and then do normalization of the remaining p columns of S (which is clearly very important for the sake of comparison of the various parameter size),

$$\frac{1}{N}\sum_{i=1}^{N} s_{ij}^2 = 1, \quad j = 1, 2, \ldots, p.$$

Keep in mind that the intercept θ_0 can be omitted after data centering, and can be recovered by $\hat{\theta}_0 = \bar{y} - \sum_{j=1}^{p} \bar{s}_j\hat{\theta}_j$ with $\hat{\theta}$ optimal.

Problem 3.6 (Relation between problems (P) and (L)).
Show that a minimizer $\hat{\theta}_\lambda$ of (L) is also a minimizer of (P) with $t = \|\hat{\theta}_\lambda\|$.
Claim: For each $t > 0$, there exists a $\lambda > 0$ such that a minimizer of (L) would also solve (P).
This can be shown by duality theory, see Appendix C in particular, the KKT conditions in terms of subdifferentials.

Coordinate descent: single variable

Assume normalization, $\frac{1}{N}\|\mathbf{z}\|^2 = 1$ and consider the one-dimensional problem

$$\underset{\theta\in\mathbb{R}}{\text{minimize}} \quad \frac{1}{2N}\|\mathbf{y} - \mathbf{z}\theta\|_2^2 + \lambda|\theta| \tag{3.4.1}$$

Lemma 3.1. *The solution to the above problem is*

$$\hat{\theta} = \begin{cases} \frac{1}{N}\langle \mathbf{z}, \mathbf{y}\rangle - \lambda & \text{if } \frac{1}{N}\langle \mathbf{z}, \mathbf{y}\rangle > \lambda \\ 0 & \text{if } \frac{1}{N}\,|\langle \mathbf{z}, \mathbf{y}\rangle| \le \lambda \\ \frac{1}{N}\langle \mathbf{z}, \mathbf{y}\rangle + \lambda & \text{if } \frac{1}{N}\langle \mathbf{z}, \mathbf{y}\rangle < -\lambda \end{cases}$$

Proof. First note that the cost function is continuous and convex and therefore has a unique minimum for arbitrary values of λ. Computing the subdifferential with respect to θ of (3.4.1) and recalling that $\|\mathbf{z}\|^2 = 1$ one finds

$$\theta = \frac{1}{N}\langle \mathbf{z}, \mathbf{y}\rangle - \lambda \ \mathrm{sign}\,(\theta)\,.$$

and if $\frac{1}{N}\langle \mathbf{z}, \mathbf{y}\rangle > \lambda$ the right hand member is certainly positive so that θ is also positive and $\mathrm{sign}\,(\theta) = 1$. A dual reasoning holds if $\frac{1}{N}\langle \mathbf{z}, \mathbf{y}\rangle < -\lambda$. Since for $\theta = 0$ the subdifferential $\lambda \ \mathrm{sign}\,(\theta)$ can take an arbitrary value between $+\lambda$ and $-\lambda$, that $\frac{1}{N}\langle \mathbf{z}, \mathbf{y}\rangle$ is between $+\lambda$ and $-\lambda$ can only mean that $\theta = 0$. □

One can write compactly the solution as $\hat{\theta} = \mathcal{S}_\lambda(\frac{1}{N}\langle \mathbf{z}, \mathbf{y}\rangle)$, where $\mathcal{S}$ is the **soft-thresholding operator:**

$$\mathcal{S}_\lambda(x) = \mathrm{sign}(x)(|x| - \lambda)_+ = \mathrm{sign}(x)\max\{|x| - \lambda, 0\}$$

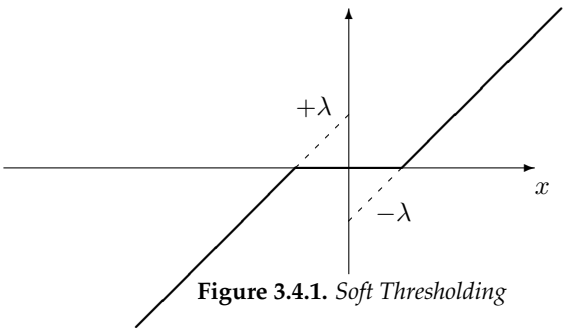


Figure 3.4.1. *Soft Thresholding*

For multivariable problems one can then do cyclic coordinatewise update according to the following scheme composed of two steps:

1. Inner loop optimization:

$$\underset{\theta_j}{\text{minimize}} \quad \frac{1}{2N}\sum_{i=1}^{N}(y_i - \sum_{k\neq j} s_{ik}\theta_k - s_{ij}\theta_j)^2 + \lambda\sum_{k\neq j}|\theta_k| + \lambda|\theta_j|$$

2. coordinate descent: First define the partial residual

$$r_i(j) = y_i - \sum_{k\neq j} s_{ik}\theta_k\,, \quad i = 1, \ldots, N,$$

Then, in each inner loop let

$$\theta_j = \mathcal{S}_\lambda(\frac{1}{N}\langle \mathbf{s}_j, \mathbf{r}(j)\rangle)\,.$$

For $j = 1 : p$ update θ_j while holding other coefficients fixed. Need to check data normalization $\frac{1}{N}\sum_{i=1}^{N} s_{ij}^2 = 1,\ j = 1, \ldots, p$ at each step.

The update can be equivalently written as

$$\theta_j \leftarrow \mathcal{S}_\lambda(\theta_j + \frac{1}{N}\langle \mathbf{s}_j, \mathbf{r}\rangle)$$

where $\mathbf{r} = \mathbf{y} - S\theta$ is the full residual.

Convergence of coordinate descent:

Suppose the objective function f has the additive decomposition

$$f(\theta_1, \ldots, \theta_p) = g(\theta_1, \ldots, \theta_p) + \sum_{j=1}^{p} h_j(\theta_j)$$

where g is differentiable and convex, and the univariate functions h_j are convex (but not necessarily differentiable), then the coordinate descent algorithm is guaranteed to converge to the global minimizer. Cf. [94, Section 5.4.1].

Proximal methods

This is is a general class of methods for minimization by a gradient-type algorithm, of a function having the structure:

$$f = g + h$$

where g is convex and differentiable and h is convex but nondifferentiable. Generalized gradient update:

$$\theta(t+1) = \text{Argmin } \theta \in \mathbb{R}^p \left\{ \underbrace{g(\theta(t)) + \langle \nabla g(\theta(t)), \theta - \theta(t)\rangle + \frac{1}{2s(t)}\|\theta - \theta(t)\|_2^2}_{\text{local approximation of } g} + h(\theta) \right\}$$

where $s(t)$ is a stepsize. Define the *proximal map* of a convex function h as

$$\mathbf{prox}_h(z) := \text{Argmin } \theta \in \mathbb{R}^p \left\{ \frac{1}{2}\|z - \theta\|_2^2 + h(\theta) \right\}$$

This is a generalized projection operator such that for $s > 0$:

$$\mathbf{prox}_{sh}(z) = \arg\min_{\theta \in \mathbb{R}^p} \left\{ \frac{1}{2s}\|z - \theta\|_2^2 + h(\theta) \right\}$$

moreover for

$$h(\theta) = \begin{cases} 0 & \text{if } \theta \in \mathcal{C} \\ +\infty & \text{otherwise} \end{cases}$$

we have $\mathbf{prox}_h(z) = \arg\min_{\theta \in \mathcal{C}} \|z - \theta\|_2^2$, the usual Euclidean projection onto $\mathcal{C}$.

Problem 3.7. *Show that the generalized gradient update is equivalent to*

$$\theta(t+1) = \boldsymbol{prox}_{s(t)h}\{\theta(t) - s(t)\nabla g(\theta(t))\}.$$

Consider now the *Proximal gradient descent for ℓ_1-penalty*. This is computationally efficient when it is easy to evaluate the proximal map.
Take $h(\theta) = \lambda\|\theta\|_1$. The t-th iteration consists of two steps:

1. First, take a gradient step $\quad z \leftarrow \theta(t) - s(t)\nabla g(\theta(t))$;

2. Second, evaluate the proximal map

$$\begin{aligned}\mathbf{prox}_{s(t)h}(z) &= \text{Argmin } \theta \in \mathbb{R}^p \left\{\frac{1}{2}\|z-\theta\|_2^2 + s(t)\lambda\|\theta\|_1\right\} \\ &= \text{Argmin } \theta \in \mathbb{R}^p \left\{\sum_{j=1}^{p}\left[\frac{1}{2}(z_j-\theta_j)^2 + s(t)\lambda|\theta_j|\right]\right\}.\end{aligned}$$

   A closed-form solution can be obtained by solving the p univariate problems separately.

Problem 3.8. *Verify that the proximal map can be evaluated by the element-wise soft thresholding operator $\mathcal{S}_\tau : \mathbb{R}^p \to \mathbb{R}^p$ with coordinates*

$$[\mathcal{S}_\tau(z)]_j = sign(z_j)(|z_j| - \tau)_+$$

with $\tau = s(t)\lambda$.

In general the algorithm has *sublinear convergence*: If g continuous differentiable with a Lipschitz gradient

$$\|\nabla g(\theta) - \nabla g(\theta')\|_2 \le L\|\theta - \theta'\|_2 \text{ for all } \theta, \theta' \in \mathbb{R}^p$$

with a constant stepsize $s(t) = s \in (0, 1/L]$, then there exists a constant C independent of the iteration number, such that

$$f(\theta(t)) - f(\theta^*) \le \frac{C}{t+1}\|\theta^0 - \theta^*\|_2 \text{ for all } t = 1, 2, \ldots,$$

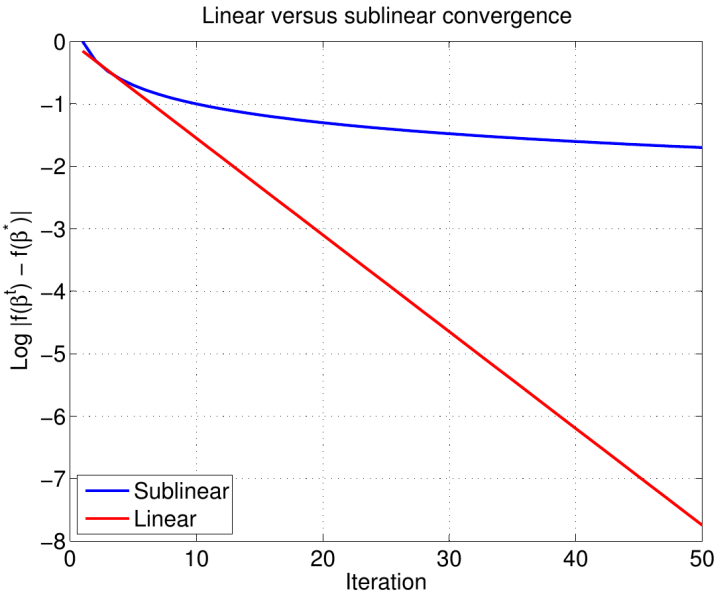
where θ^* is an optimum.
Linear convergence holds if in addition g is strongly convex, that is, there exists $\gamma > 0$ such that

$$g(\theta+\Delta) - g(\theta) - \langle\nabla g(\theta), \Delta\rangle \ge \gamma^2\|\Delta\|_2^2, \text{ for all } \theta, \Delta \in \mathbb{R}^p$$

then there exists constant $C > 0$ and $\kappa \in (0, 1)$ such that

$$f(\theta(t)) - f(\theta^*) \le C\kappa^t\|\theta^0 - \theta^*\|_2 \text{ for all } t = 1, 2, \ldots.$$

Cf. [94, Section 5.3.3] for details.



Proximal gradient for the Lasso. We have:

$$g(\theta) = \frac{1}{2}\|\mathbf{y} - S\theta\|_2^2 \text{ and } h(\theta) = \lambda\|\theta\|_1$$

Then the proximal gradient update becomes:

$$\theta(t+1) = \mathcal{S}_{s(t)\lambda}\left(\theta(t) + s(t)\frac{1}{N}S^\top(\mathbf{y} - S\theta(t))\right)$$

Remark:

1. This is a batch update version of the coordinate descent if we take $s(t) \equiv 1$ with data standardized;
2. The Lipschitz constant L here is the maximum eigenvalue of $\frac{1}{N}S^\top S$.

Some computational consideration

Strategies to improve efficiency:

1. Naive vs covariance updating

$$\sum_{i=1}^{N} s_{ij}r_i = \langle \mathbf{s}_j, \mathbf{y}\rangle - \sum_{k:\,\theta_k\neq 0}\langle \mathbf{s}_j, \mathbf{s}_k\rangle\theta_k$$

2. Warm start. When compute a sequence of lasso solutions for a decreasing sequence $\{\lambda_\ell\}_0^L$, take $\hat{\theta}(\lambda_\ell)$ as a starting point for $\hat{\theta}(\lambda_{\ell+1})$.

Ex. Verify that the largest value one needs to consider is

$$\lambda_0 = \frac{1}{N} \max_j |\langle \mathbf{s}_j, \mathbf{y} \rangle|.$$

Hint: Take initial iterate $\theta^0 = \mathbf{0}$ and check the coordinate descent.

3. Active set convergence. After one iteration through all p variables at a new value λ_ℓ starting from $\hat{\theta}(\lambda_{\ell-1})$, define the *active set* $\mathcal{A}$ to optimize only over indices of nonzero coefficients.

Concluding Remarks: Since traditional statistical methods assume many observations and a few unknown variables, they can not cope up with the situations when $p > N$. The Lasso is a powerful and general method to analyze such regression problems with many variables some of which may be redundant and could possibly be discarded. Naturally the difficult part is the final choice of a resonable value of λ. This problem can be addressed by a technique called *Cross Validation* which is briefly described in the next chapter but is still a hot area of research.

An excellent survey paper on this topic is [36].

3.5 ▪ Regression and Smoothing Splines

In a general regression problem one wants to find a continuous function $f(x)$ of the input variable $x \in \mathbb{R}^n$, which interpolates in some "optimal" way a discrete training set $\{(x_1, y_1), \ldots, (x_N, y_N)\}$ whose graph looks so far from any linear pattern to make a linear approximation unacceptable. One should note from the outset that the recovery of a continuous nonlinear function from a discrete set of values is clearly a very *ill-posed* problem since the problem can trivially have infinitely many solutions. Nevertheless a multitude of practical inference problems are naturally formulated in this way and one has to attempt some form of solution anyway. A classical approach is by approximating the unknown function by an expansion in series of basis (not necessarily orthogonal) functions $\{\varphi_k(x)\,;\, k = 0, 1, 2, \ldots\}$, say

$$f(x) \sim \sum_{k=0}^{p} \theta_k \varphi_k(x)\,. \tag{3.5.1}$$

Clearly this model is linear in the parameter $\boldsymbol{\theta} := \{\theta_k\}$ and hence the fitting of the model to data can be done by the least squares theory seen in Chapter 2. Assuming i.i.d. errors, the least squares formulation

$$\min_{\boldsymbol{\theta}} \sum_{k=1}^{N} \left[y_k - \sum_{i=0}^{p} \theta_i \varphi_i(x_k) \right]^2$$

leads to the standard formulas of Chapter 2 and if you assume that there is a "true model"

$$\mathbf{y}_k = \sum_{i=0}^{p} \theta_{0,i} \varphi_i(x_k) + \mathbf{w}_k\,, \qquad k = 1, \ldots, N \tag{3.5.2}$$

the estimates are amenable to standard statistical analysis.
The difference here is that one wants to understand the *influence of the input variables* on the statistical properties of the solution. Hence we must bring in explicitly some structural information on the regressors. For pedagogical reasons in this section we shall discuss this issue only for the case of one dimensional input data. In a sense we are considering a problem which is also called **curve fitting** in the literature.

Below we shall study an elementary case where we just take two basis functions, $\varphi_0(x) = 1$ (that is a constant) and and $\varphi_1(x) = x$.

Example 3.6 (A simple linear regression problem).
Suppose you measure scalar data pairs $\{x_k,\ y_k\ ;\ k = 1, 2, \ldots, N\}$ where the x_k's are known exactly but the y_k are affected by errors. You would like to describe approximately these data by a straight line say $y = \alpha + \beta\, x$. Suppose you model the measurement process by a statistical model

$$\mathbf{y}_k = \alpha + \beta\, x_k + \mathbf{e}_k\,, \qquad k = 1, 2, \ldots, N \tag{3.5.3}$$

where the errors $\mathbf{e}_k$ are **zero-mean independent random variables** with variances σ_k^2. In a given experimental condition ω you have observed the values $\mathbf{y}_k(\omega) = y_k\ ;\ k = k = 1, 2, \ldots, N$ corresponding to errors $\mathbf{e}_k(\omega)$ (which of course you do not know). The least squares (Markov) estimator of the parameter $(\alpha,\ \beta)$ is the solution of the minimization problem

$$\min_{(\alpha,\,\beta)} \sum_{k=1}^{N} [\, y_k - (\alpha + \beta\, x_k)\,]^2$$

Following the same procedure of Example 2.4 and Problem 2-3 we shall decouple the optimization problem for the two parameters by first introducing the sample averages on both sides of (3.5.3) to get

$$\bar{\mathbf{y}}_N = \alpha + \beta \bar{x}_N + \bar{\mathbf{e}}_N$$

and then subtracting this from the model equations. One then gets a model for the deviations $\tilde{\mathbf{y}}_k := \mathbf{y}_k - \bar{\mathbf{y}}_N$

$$\tilde{\mathbf{y}}_k = \beta(x_k - \bar{x}_N) + \tilde{\mathbf{e}}_k$$

where the intercept α is eliminated. The sequence $\{\tilde{\mathbf{e}}_k\}$ is no longer independent but approximately so for large N (check that the covariance matrix has off diagonal elements which are proportional to $\frac{1}{N}$). The two minimizers can then be computed separately:

$$\hat{\alpha}_N = \bar{y}_N - \hat{\beta}_N\, \bar{x}_N\,, \qquad \hat{\beta}_N = \frac{\sum_k (x_k - \bar{x}_N)\, y_k}{\sum_k (x_k - \bar{x}_N)^2}$$

In the numerator of the second expression we should have $y_k - \bar{y}_N$ but the sample mean $\bar{y}_N$ gives zero contribution to the product. Next, you may imagine these to be sample values of the random variables

$$\hat{\boldsymbol{\alpha}}_N = \bar{\mathbf{y}}_N - \hat{\boldsymbol{\beta}}_N\, \bar{x}_N\,, \qquad \hat{\boldsymbol{\beta}}_N = \frac{\sum_k (x_k - \bar{x}_N)\, \mathbf{y}_k}{\sum_k (x_k - \bar{x}_N)^2}\,. \tag{3.5.4}$$

We shall first directly check that these estimators are **unbiased**. This follows from the expressions

$$\mathbb{E}\,\hat{\boldsymbol{\alpha}}_N = \mathbb{E}\,(\bar{\mathbf{y}}_N - \hat{\boldsymbol{\beta}}_N\,\bar{x}_N)$$

$$\mathbb{E}\,\hat{\boldsymbol{\beta}}_N = \frac{\sum_k (x_k - \bar{x}_N)\,\mathbb{E}\,\mathbf{y}_k}{\sum_k (x_k - \bar{x}_N)^2}$$

since the errors are zero-mean $\mathbb{E}\,(\bar{\mathbf{y}}_N) = \alpha + \beta\,\bar{x}_N$ and hence

$$\mathbb{E}\,\hat{\boldsymbol{\alpha}}_N = \alpha + \mathbb{E}\,(\beta - \hat{\boldsymbol{\beta}}_N)\,\bar{x}_N\,.$$

On the other hand, $\mathbb{E}\,\mathbf{y}_k = \alpha + \beta\,x_k$, and so

$$\mathbb{E}\,(\hat{\boldsymbol{\beta}}_N - \beta) = \frac{\sum_k (x_k - \bar{x}_N)\,(\alpha + \beta\,x_k)}{\sum_k (x_k - \bar{x}_N)^2} - \beta = \beta\frac{\sum_k (x_k - \bar{x}_N)\,x_k}{\sum_k (x_k - \bar{x}_N)^2} - \beta = 0$$

since $\sum_k (x_k - \bar{x}_N)\,\alpha = 0$ and likewise $\sum_k (x_k - \bar{x}_N)\,\bar{x}_N = 0$.

We now ask the following

Question: are these consistent estimators of the parameters $(\alpha,\,\beta)$?

As we shall see the answer depends (obviously) on the error statistics but also on *how the input data are distributed on the line.*

Let us first look at the estimate

$$\hat{\boldsymbol{\beta}}_N = \frac{\sum_k (x_k - \bar{x}_N)\,\mathbf{y}_k}{\sum_k (x_k - \bar{x}_N)^2} := \sum_k w_k\,\mathbf{y}_k$$

since the $\mathbf{y}_k$ are independent (as the $\mathbf{e}_k$ are) we have

$$\text{var}\,(\hat{\boldsymbol{\beta}}_N) = \sum_k w_k^2\,\text{var}\,(\mathbf{y}_k) = \sum_{k=1}^N w_k^2\,\sigma_k^2$$

where $\sigma_k^2 = \text{var}\,(\mathbf{e}_k)$. For convergence in probability of $\hat{\boldsymbol{\beta}}_N$ to β we need

$$\lim_{N\to\infty} \sum_{k=1}^N w_k^2\,\sigma_k^2 = 0$$

which, in case $\sigma_k^2 = \sigma^2$ independent of k, implies

$$\sum_{k=1}^N w_k^2 = \frac{\sum_k (x_k - \bar{x}_N)^2}{[\sum_k (x_k - \bar{x}_N)^2\,]^2} = \frac{1}{\sum_k (x_k - \bar{x}_N)^2} \;\to\; 0\,.$$

This is the same as

$$\sum_{k=1}^{+\infty} (x_k - \bar{x}_N)^2 = \infty$$

which means that in order to have consistency of $\hat{\boldsymbol{\beta}}_N$, **the points** x_k **should not remain too close to their sample mean**. Since

$$\hat{\boldsymbol{\alpha}}_N = \alpha + (\beta - \hat{\boldsymbol{\beta}}_N)\,\bar{x}_N + \frac{1}{N}\sum_{k=1}^N \mathbf{e}_k$$

under this same condition, both the last two terms converge to zero in probability and it is easy to see that the estimator $\hat{\alpha}_N$ is also consistent in probability. □

Hence, it turns out that, in order to have consistency (and a reasonable error variance), one needs to have data points $\{x_k\}$ which do not cluster to form a too concentrated data set. On the other hand, if the approximation is meant to involve a wide range of possible input values, the result may turn out to be unsatisfactory. In particular, an unbalanced data distribution with too few data at the extremes of the regression intervals may result in a poor fit and very high variance. This is especially true with polynomials as they have unpredictable tail behavior which can be very bad for extrapolation. See for example Fig. 3.5 below One may conclude that a better policy could be just to try juxtaposition

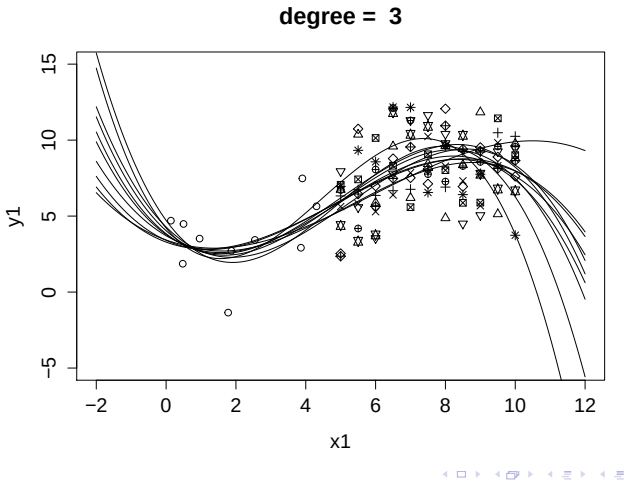


Figure 3.5.1. *Polynomial fit*

of functions which do *local approximation* of the data on a short range of input values. This is called *Local Regression* and can be done in a variety of ways. Below we shall describe a rational way to choose local approximating functions.

Splines

Let now $\{x_1 < x_2, \ldots < x_N\}$ be a sequence of N points in the interval of interest which will be called **knots**. These points will act as internal boundaries for local approximation. Instead of a single polynomial in x over the whole domain, we can rather use different polynomials of the same degree in regions defined by knots. They form a function $s(x)$ made by patching together a sequence of polynomials $s_k(x)\, k = 0, \ldots, N$, each defined on the interval $[\, x_k,\ x_{k+1}]$ with x_0 and x_{N+1} being boundaries of the approximation interval. Such functions are called **splines**, in particular linear, quadratic, cubic etc. splines depending on the degree of the polynomial pieces. We shall ask that this function should have

a maximal degree of smoothness imposing continuity of all (non constant) existing derivatives at the boundary points, that is, impose that for $k = 1, 2, \ldots, N$,

$$s_{k-1}(x_k) = s_k(x_k)\,, \qquad s'_{k-1}(x_k) = s'_k(x_k)\,, \ldots, s^{n-1}_{k-1}(x_k) = s^{n-1}_k(x_k) \quad (3.5.5)$$

where $n := \deg\{s(x)\}$ which means that the consecutive polynomial pieces join smoothly at each knot x_k together with all their (non constant) existing derivatives. In a sense they have a maximum degree of smoothness. In particular, a cubic spline is a piecewise cubic polynomial which has a continuous second derivative.
It is claimed that cubic splines are the lowest-order spline for which the knot-discontinuity is not visible to the human eye. In applications there is seldom any good reason to go beyond cubic-splines and in what follows we shall mostly discuss them.

In general splines of degree n with knots $\{x_1 < x_2, \ldots < x_N\}$ involve $N + 1$ local polynomial $s_k(x)$ (of degree n) each of which depends linearly on $n + 1$ parameters. These objects form a vector space whose dimension can be computed noting that the continuity constraints (3.5.5) impose nN linear conditions on a spline of degree n with N knots. Therefore the vector space has dimension $(n + 1)(N + 1) - nN = N + n + 1$ and hence for cubic splines there are local polynomial bases consisting of $N + 4$ elements. **B-Splines** are a particularly convenient such basis which we shall describe next.

B-Splines

The (B)-splines are a nice basis for the space of splines, in fact for splines of any order.
They are defined recursively in the following fashion:

$$B_{i,0}(x) := \{1 \,\text{if}\, x_i \le x < x_{i+1}; \quad 0 \quad \text{otherwise}\}$$
$$B_{i,k}(x) := \frac{x - x_i}{x_{i+k} - x_i} B_{i,k-1}(x) + \frac{x_{i+k+1} - x}{x_{i+k+1} - x_{i+1}} B_{i+1,k-1}(x).$$

Here each function has compact support, the B-spline of order k, $B_{i,k}(x)$ being zero for $x < x_i$ and for $x \ge x_{i+k+1}$, see Fig. 3.5 where the picture represents just one sequence of B-functions of compact support of increasing order $k = 0, 1, 2, 3$, the leftmost knot being $x_i = 0.3$ and the knot intervals have length 0.1. In the picture B_k is denoted y_{k+1}.

A cubic spline with N knots is then represented as

$$s(x) = \sum_{i=0}^{N+3} B_i(x)\theta_i \quad (3.5.6)$$

where $B_i(x)$ is a short for the cubic basis function $B_{i,3}(x)$. The first and second derivatives of $s(x)$ have a similar expansion in terms of lower order B-Splines. Both are continuously patched at the knots.

Algorithm 3.1 B-Splines Computation

function $B(i,k,knots)(e,L)$
  $e = eps(0.0)$
  $L = knots[end] - knots[1]$
  $knots = [knots..., (L + knots[2:2+k] - knots[1])...]$
    ▷ pad the knots vector for periodic case
  if $k = 0$ **then**
    $x \rightarrow float(knots[i] - e \Leftarrow x < knots[i+1] + e)$
  else
    if $k \neq 0$ **then**
      $x \rightarrow (B(i,k-1,knots)(x)*(x-knots[i])/(knots[i+k]-knots[i])+ B(i+1,k-1,knots)(x)*(knots[i+k+1]-x)/(knots[i+k+1]-knots[i+1]))$

function $Bder(i,k,knots,p)$(e, L)
  $e = eps(0.0)$
  $L = knots[end] - knots[1]$
  $knots = [knots..., (L + knots[2:2+k] - knots[1])...]$ ▷ pad the knots vector for periodic case
  if $p > k$ **then**
    $x \rightarrow zero(x)$
  else
    if $p = 0$ **then**
      $x \rightarrow B(i,k,knots)(x)$
    else
      $x \rightarrow ((Bder(i,k-1,knots,p)(x) * (x - knots[i]) + Bder(i,k-1,knots,p-1)(x))/(knots[i+k]-knots[i])+(Bder(i+1,k-1,knots,p)(x)*(knots[i+k+1]-x) - Bder(i+1,k-1,knots,p-1)(x))/(knots[i+k+1]-knots[i])$
  End

Smoothing Splines

As we have just hinted at, a basis expansion like (3.5.1), chosen without any a priori insight, irrespective of the distribution of the input data, could be a rather poor modeling choice. In this section we want to describe a special technique of local regression which turns out to be particularly successful. To understand the basics of the method we shall rephrase the problem in more general terms as a *non-parametric estimation problem*. This just means that we want to look for a function of the input data $f(x)$ which approximates in a suitable statistical sense the scattered points $\{(x_k,\, y_k)\,, k = 1, 2, \ldots, N\}$ of the training set. We shall not ask f to obey exactly the interpolation conditions $f(x_k) = y_k \;\; k = 1, 2, \ldots, N$ as this would obviously lead to an absurd overfitting but instead formulate an "approximate interpolation problem" say $f(x_k) \simeq y_k \;\; k = 1, 2, \ldots, N$ requiring that f should obey some extra smoothness constraints. Since this interpolation problem is clearly ill-posed, we shall invoke the idea of Tikhonov regularization, and, as anticipated in Section 3.3.1, reformulate it as an optimization problem. The optimization will no-longer be a parametric linear least squares but will be formulated directly *in terms of an unknown function* f. This function, besides interpolating the training data, should have a certain degree of

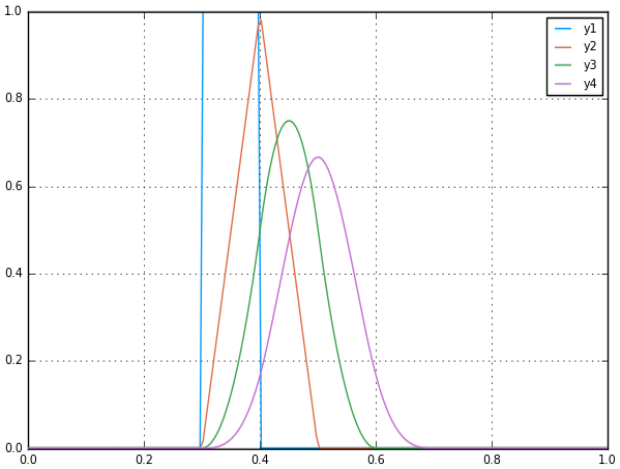


Figure 3.5.2. *B-Splines*

smoothness. To filter out the noise i.e. avoid overfitting, we shall impose that the candidate function should be smooth; in fact, we shall constrain it to be twice differentiable and add a penalty on the average amplitude of its second derivative. This turns out to be a particularly convenient choice. In formulas, we shall formulate the following variational problem:

$$\min_{f \in C^2} \left\{ \sum_{k=1}^{N} [\, y_k - f(x_k)\,]^2 + \lambda \int f''(x)^2 dx \right\}, \tag{3.5.7}$$

where the integral is extended to an interval containing the field of possible input values and λ is the regularization parameter controlling the smoothness of the solution. Obviously, for λ small we will recover the least squares interpolation problem. Although we shall not go into the details of the proof, this variational problem can be solved explicitly.

Theorem 3.2. *The solution of the problem* (3.5.7) *is a* **cubic spline**, *with knots the input points* $\{x_k, \, , k = 1, 2, \ldots, N\}$ *of the training set.*

A proof can be found in Wahba's book [101].

In general the solution $s(x)$ does not interpolate exactly the values y_k at the knots. This is why these functions are called **smoothing splines**. Note that each cubic polynomial $s_k(x)$ approximates $f(x)$ locally in the interval $[\, x_k, \; x_{k+1}]$ in a way which is "almost decoupled" from its neighbors. This is in sharp contrast with the naive approximation (3.5.1) where each basis function $\varphi_k(x)$ is supposed to approximate $f(x)$ on the whole interval of interest which is clearly a heavier task. This explains the better behaviour of spline approximation.
Spline approximation is discussed in great detail in the book [101] where the

regularization integrand is also generalized to the square of an arbitrary m-th order derivative of $f(x)$.

The numerical solution of the problem (3.5.7) can be reduced to the solution of a generalized ridge regression problem. Expressing the solution spline in a cubic B-spline basis as in (3.5.6), after introducing the matrices

$$\mathbf{B} := [B_i(x_k)]_{i=1,k=\dots,N}\,, \qquad \mathbf{\Omega} := \left[\int B_i^{''}(x) B_k^{''}(x) dx\right]_{i,k=1,\dots,N} \tag{3.5.8}$$

which are banded matrices (for example $\mathbf{\Omega}$ is symmetric tridiagonal and positive definite), the problem (3.5.7) can be recast in the regularized least squares form

$$\min_{\theta} \{\|y - \mathbf{B}\theta\|^2 + \lambda \theta^{\top} \mathbf{\Omega}\, \theta\} \tag{3.5.9}$$

which has the solution

$$\hat{\theta} = \left[\mathbf{B}^{\top}\mathbf{B} + \lambda \mathbf{\Omega}\right]^{-1} \mathbf{B}^{\top}\, y\,. \tag{3.5.10}$$

Note that both $\mathbf{B}$ and $\mathbf{\Omega}$, only depend on the input data. Although banded, these matrices are of possibly very high dimension. The smoothness of the approximant may however allow to discard a sizable subset of the original x_k's. To this end the choice of the regularization parameter is crucial; see the pictures in Fig 3.5.3. Here we shall only discuss this issue rather superficially.

The regularization parameter λ

The behaviour of the solution depends heavily on the choice of the regularization parameter λ (see e.g. Figure 3.5.3 below). In practice one has only empirical rules to choose it. One is based on a comparison with the least squares solution. Note that the solution vector $\mathbf{s} := \begin{bmatrix} s(x_1) & \dots & s(x_N) \end{bmatrix}^{\top}$ is given by the formula

$$\mathbf{s} = \mathbf{B}\left[\mathbf{B}^{\top}\mathbf{B} + \lambda \mathbf{\Omega}\right]^{-1} \mathbf{B}^{\top}\mathbf{y}$$

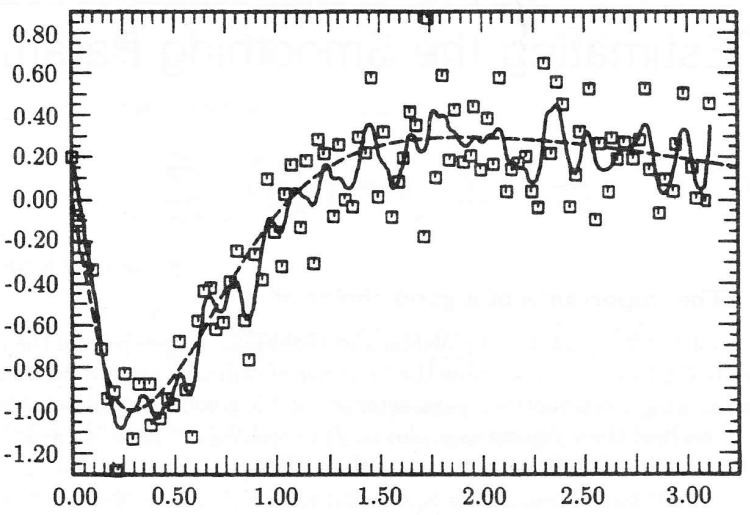
and that for $\lambda = 0$ the matrix $\mathbf{A}_{\lambda} := \mathbf{B}\left[\mathbf{B}^{\top}\mathbf{B} + \lambda \mathbf{\Omega}\right]^{-1} \mathbf{B}^{\top}$ is an orthogonal projection onto the Image space of $\mathbf{B}$. Now this matrix is square $N \times N$ but may generally have a numerical rank

$$d_0 := \text{rank } \mathbf{A}_0$$

of much smaller dimension than N which in fact can be considered as the "numerical dimension" of the solution. This dimension reduction is equivalent to eliminating some "redundant" inputs $\{x_k\}$. The rank of $\mathbf{A}_{\lambda}$ denoted

$$d(\lambda) = \text{rank } \mathbf{A}_{\lambda}$$

has a similar interpretation as numerical dimension of the solution. One can tentatively fix $d(\lambda) < N$ as a guess of the number of "important" input points and then compute numerically the corresponding value of λ using the formula for $\mathbf{A}_{\lambda}$. This value of λ can then lead to the estimate of the model by shrinking parameters to very small values or, equivalently, by weighting very little some input points.



.FIG. 4.1. *Data generated according to the model* (4.1.1). *Dashed curve is* $f(x)$. *Solid curve is fitted spline with* λ *too small.*

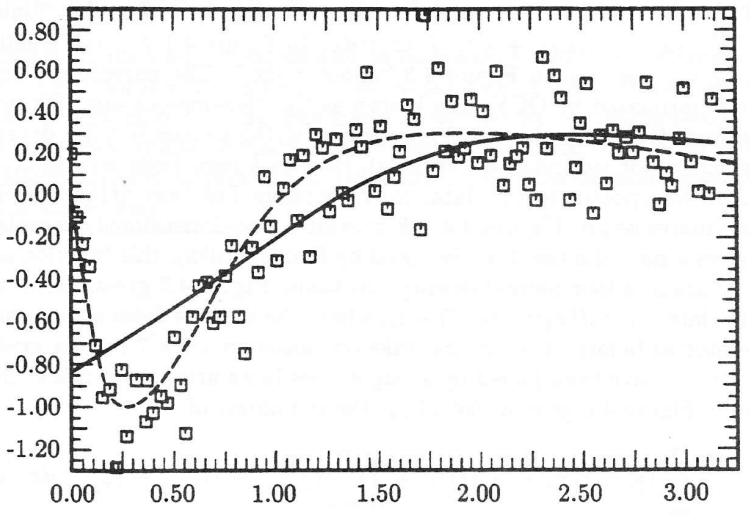


FIG. 4.2. *Same data as in Figure* 4.1. *Spline (solid curve) is fitted with* λ *too big.*

Figure 3.5.3. *From Wahba's book [101]*

The *MATLAB curve fitting toolbox* contains several functions for spline approximation. The smoothing spline algorithm is based on the function `csaps`. More details can be found in the book by Hastie et al. : `https://web.stanford.edu/~hastie/ElemStatLearn/` pages 186-189.

3.6 ▪ Problems

3-1 *Suppose the matrix* $S = \begin{bmatrix} s_1 & \dots & s_p \end{bmatrix}$ *has orthogonal columns which are ordered by decreasing norms*

$$\|s_1\| \geq \dots \geq \|s_p\|$$

and let $\mathrm{Var}\ \mathbf{w} = \sigma^2 I_N$. *Find an expression for the variance ratio* $\dfrac{\mathrm{var}\ \hat{\boldsymbol{\theta}}_1}{\mathrm{var}\ \hat{\boldsymbol{\theta}}_p}$ *in terms of these norms.*

3-2 *Consider the ridge regression estimator*

$$\hat{\theta}_R(y) = \left[S^\top Q S + \lambda I_p\right]^{-1} S^\top Q\, y := A_\lambda\, y$$

where $Q = Q^\top$ *is positive definite. Show that if* $\lambda > 0$, SA_λ *cannot be idempotent, that is, cannot be a projection matrix.*
Similarly, $A_\lambda S \neq I$ *and the ridge estimator cannot be unbiased.*

3-3 *Consider the ridge regression estimator with weight* $Q = I_N$ *and* $\lambda > 0$. *Using the SVD of* S, *in particular the fact that its squared singular values are eigenvalues of* $S^\top S$, *give an expression for the eigenvalues of* SA_λ *and for its spectral decomposition say*

$$SA_\lambda = U\Lambda U^\top, \quad UU^\top = I$$

where Λ *is the diagonal eigenvalue matrix.*

3-4 *Let* $p = 1$ *and suppose the unique column* s *of* S *is normalized so that* $\|s\|^2 = 1$. *Try to solve the Lasso problem*

$$\min_\theta \{\|\mathbf{y} - s\theta\,\|^2 + \lambda|\,\theta\,|\,\}$$

by setting the derivative with respect to θ *equal to zero, using the "almost-everywhere" derivative*

$$\frac{d|\theta|}{d\theta} = \begin{cases} 1 & \text{if} \quad \theta > 0 \\ -1 & \text{if} \quad \theta < 0 \end{cases}$$

The right hand side is called the sign *function, so that* $\dfrac{d|\theta|}{d\theta} = \mathrm{sign}\,(\theta)\ a.e.$.

3-5 *Write a Matlab program to compute the matrices* $\mathbf{B}$, $\boldsymbol{\Omega}$ *in* (3.5.6), *for a given sequence of knots.*

Chapter 4

LINEAR HYPOTHESES AND LINEAR CLASSIFICATION

4.1 ▪ Hypothesis Testing on the Linear Model

We shall initially discuss hypothesis testing on the standard N-dimensional linear model

$$\mathbf{y} = S\theta + \sigma\mathbf{w}\,, \qquad \mathbf{w} \sim \mathcal{N}(0, I_N) \tag{4.1.1}$$

where the additive noise has independent components. Recall that this is no loss of generality whenever the variance matrix of $\mathbf{w}$ is of the form $\sigma^2 R$ with a kown $R > 0$. See the normalization procedure discussed at the beginning of Chapter 2 formula (2.2.5). Later we shall generalize our discussion to linear models with arbitrary variance.

It is customary to call *linear hypotheses* those which can be expressed in terms of linear functions of the parameter θ. For example, H_0 is a linear hypothesis if it can be expressed as,

$$H_0 := \{\theta\,;\ H\theta = \beta_0\}\,. \tag{4.1.2}$$

for some matrix $H \in \mathbb{R}^{k\times p}$, $k \le p$ and β_0 some fixed vector in $\mathbb{R}^k$.

It should be quite obvious that it is no loss of generality considering only hypotheses which can be expressed in terms of a full rank matrix H; i.e. $\operatorname{rank} H = k \le p$. In other words, we shall consider only hypotheses described in a non-redundant way. Testing linear hypotheses on the linear model (4.1.2) is an important problem area which pops up in a variety of situations; typically occurring in the diagnostics of linear models estimated from the data by M.L. (or Least Squares). Some typical examples are:

1. *Adequacy of a linear model*; can be expressed as

   $$H_0 := \{\theta = 0\} \tag{4.1.3}$$

   (in which case $H = I$, $\beta_0 = 0$). The question is whether a linear model is adequate to explain the data. Accepting H_0 means that you decide that the measurements y are constituted by white noise. Refusing H_0 means that a nontrivial linear model of the form (4.1.2), should be adequate.

2. *Hypothesis on the number of significant parameters*

   $$H_0 := \{\theta_{k+1} = \theta_{k+2} = \ \ldots\ = \theta_p = 0\} \quad . \tag{4.1.4}$$

Accepting H_0 means deciding that the last $p-k$ parameters are redundant; i.e. the model is *overparameterized*. The alternatives can be several different structures compatible with the general model. It may actually make sense to compare different parametric structures. For example one may want to compare two regression models of the form

$$\mathbf{y}_t = \theta_0 + \theta_1 u_t + \mathbf{w}_t \tag{4.1.5}$$

$$\mathbf{y}_t = \theta_0 + \theta_1 u_t + \theta_2 u_t^2 + \mathbf{w}_t \tag{4.1.6}$$

from inputs $\{u_t\}$ and outputs $\{\mathbf{y}_t\}$ observed for $t = 1, ..., N$. The two alternative hypotheses to be tested could be linear versus nonlinear structure: $H_0 : \theta_2 = 0$ and $H_1 : \theta_2 \neq 0$.

3. *Analysis of Variance* Here one wants to test the equality of the means, say $\mu_1, \ldots, \mu_p$, of p mutually uncorrelated sequences of observations having Gaussian distribution and the same scalar variance. One can formalize the problem by setting up a large linear model and testing the hypothesis $\mu_1 = \mu_2 = \ldots = \mu_p$.

Let us consider the normalized linear model (4.1.1) and examine the effects of the constraint $H\theta = \beta_0$ (that is H_0) on the ML estimate of θ. Obviously, by Gaussianity of the observations, ML reduces to unweighted ($R^{-1} = I$)Least Squares.

Hence under H_0, the estimator $\hat{\theta}_0$ is found by minimizing the Euclidean distance of the vector y from the columnspace $\mathcal{S}$, of S, but by *taking into account the constraint* (4.1.2). In other words the linear combination of the columns of S which minimizes $\|y - S\theta\|^2$ canno longer use arbitrary coefficients $\theta \in \mathbb{R}^p$ but needs instead to use parameters $(\theta_1, ..., \theta_p)$ satisfying the equation $H\theta = \beta_0$.

$$\hat{\theta}_0(y) = \text{Arg} \min_{\theta \in \{\theta\,;\, H\theta = \beta_0\}} \|y - S\theta\|^2 \quad . \tag{4.1.7}$$

If $\text{rank } H = k$, the constraint $H\theta = \beta_0$ is made of k independent linear equations in θ and hence provides only $p - k$ free parameters among the p components of θ and hence $\hat{\theta}_0(y)$ will use only $p - k$ independent linear combinations of the columns of S. This means that the minimum of $\|y - S\theta\|^2$ subject to $H\theta = \beta_0$ is found by projecting y not any longer onto the whole space $\mathcal{S}$ but instead onto an affine subspace $\mathcal{H} \subset \mathcal{S}$ of dimension $p - k$, defined by

$$\mathcal{H} := \text{span } \{S\theta\,;\, H\theta = \beta_0\} \ . \tag{4.1.8}$$

All of this, obviously under the assumption that the constraint (4.1.2) is actually present, that is under the hypothesis H_0. If H_0 is not true, say the constraint (4.1.2) does not act (i.e. under H_1) the M.L. estimator say $\hat{\theta}_1(y) = \hat{\theta}(y)$ is just the usual projection of y onto $\mathcal{S}$. See tha Fig. 4.1.1

Since $\mathcal{S} \supset \mathcal{H}$, the distance of y from $\mathcal{S}$ must clearly be smaller than that from $\mathcal{H}$. It should then be clear that the sum of squared residuals in the two situations must be *differen*. Since under H_1 we can use a larger space to construct our approximation of the data y. Therefore under H_1 the norm of the approximation error of y by $S\hat{\theta}$ must be *smaller* than that of the residual error $\|y - S\hat{\theta}_0(y)\|^2$

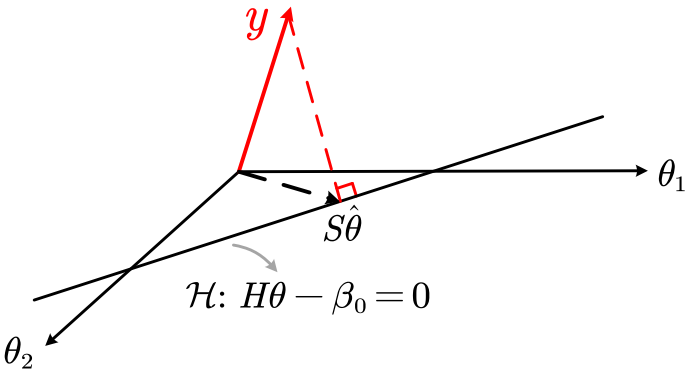


Figure 4.1.1. *Orthogonal projection on the affine subspace $\mathcal{H}$*

under H_0. Let us then introduce the sum of squared residual errors under the two hypotheses

$$\begin{array}{ll} H_0: & R_0^2(y) = \|y - S\hat{\theta}_0(y)\|^2 \\ H_1: & R_1^2(y) = \|y - S\hat{\theta}(y)\|^2 \end{array} \tag{4.1.9}$$

The following lemma provides a rather obvious formalization of the geometry of the problem.

Lemma 4.1. *The vector $S\hat{\theta}_0(y)$ is the orthogonal projection of $S\hat{\theta}(y)$ onto $\mathcal{H}$ and therefore*

$$R_0^2(y) = R_1^2(y) + \|S\hat{\theta}(y) - S\hat{\theta}_0(y)\|^2 \tag{4.1.10}$$

Proof. Introduce for convenience the symbols

$$\hat{\mu}_0 := S\hat{\theta}_0(y) \quad , \quad \hat{\mu}_1 := S\hat{\theta}(y)\,. \tag{4.1.11}$$

To prove the first statement just observe that $y - \hat{\mu}_1$ is orthogonal to $\mathcal{S}$ and hence in particular to $\mathcal{H}$. Moreover $y - \hat{\mu}_0$ is orthogonal to $\mathcal{H} \subset \mathcal{S}$ by construction. By linearity of the scalar product, $\langle y - \hat{\mu}_1 - (y - \hat{\mu}_0),\, \mathcal{H}\rangle = 0 \Rightarrow \langle \hat{\mu}_1 - \hat{\mu}_0,\, \mathcal{H}\rangle = 0$. Therefore

$$R_0^2 = \|y - \hat{\mu}_1 + \hat{\mu}_1 - \hat{\mu}_0\|^2 = \|y - \hat{\mu}_1\|^2 + \|\hat{\mu}_1 - \hat{\mu}_0\|^2 + 2\,\langle y - \hat{\mu}_1, \hat{\mu}_1 - \hat{\mu}_0\rangle \tag{4.1.12}$$

where the inner product on the right is zero since $\hat{\mu}_0$ and $\hat{\mu}_1 \in \mathcal{S}$ and likewise does their difference. Since $\hat{\mu}_1$ is the orthogonal projection onto $\mathcal{S}$, $y - \hat{\mu}_1$ is orthogonal to $\mathcal{S}$. Hence (4.1.12) reduces to

$$R_0^2 = R_1^2 + \|\hat{\mu}_1 - \hat{\mu}_0\|^2$$

(a version of Pithagora's theorem) which indeed is just (4.1.12). □

Note that if H_0 is true, the difference $\|\hat{\mu}_1 - \hat{\mu}_0\|^2$ (which is obviously random as it depends on the sample y) will in the average be *small* since the estimate,

$S\hat{\theta}(y)$, even if computed without taking into account the constraint (4.1.2) tends, by the law of large numbers, to get close to $S\theta_0$ as $N \to \infty$ since θ_0 is the true value of the parameter and by assumption $S\theta_0$ lays on $\mathcal{H}$.

Conversely, if H_0 is false, $S\hat{\theta}(y)$ keeps on lying outside of the subspace $\mathcal{H}$ even if $N \to \infty$. From this, one deduces that the ratio

$$\frac{\|\hat{\mu}_1 - \hat{\mu}_0\|^2}{R_1^2} = \frac{R_0^2 - R_1^2}{R_1^2} \tag{4.1.13}$$

should in the average be *small* under H_0 and *large* is H_1 is true.As we shall prove below this intuitive test statistic is actually what prescribes the MLR principle.

Computing the MLR

We assuma that σ^2 is *unknown* The pdf, $f(y, \theta, \sigma^2)$ of the random vector $\mathbf{y}$ described by the linear model (4.1.1), is

$$f(y, \theta, \sigma^2) = (2\pi\sigma^2)^{-\frac{N}{2}} \exp -\frac{1}{2\sigma^2}\|y - S\theta\|^2 \tag{4.1.14}$$

so that under H_1, the ML estimator $(\hat{\theta}_1, \hat{\sigma}_1^2)$ of (θ, σ^2), can be computed by maximizing over the whole parameter space. This just means that $\hat{\theta}_1$ and $\hat{\sigma}_1^2$ are the ordinary ML estimators of θ and σ^2 computed in Sect. 2,

$$\hat{\theta}_1 = \operatorname{Arg}\min_{\theta} \|y - S\theta\|^2 \qquad \hat{\sigma}_1^2 = \frac{1}{N}\|y - S\hat{\theta}_1\|^2 = \frac{1}{N}R_1^2(y)\,. \tag{4.1.15}$$

Substituting (4.1.15) into (4.1.14) one finds

$$f(y, \hat{\theta}_1(y), \hat{\sigma}_1^2(y)) = \left[2\pi\frac{R_1^2(y)}{N}\right]^{-\frac{N}{2}} \exp -\frac{N}{2}\,. \tag{4.1.16}$$

Under H_0, the estimator $\hat{\theta}_0$ solves the constrained minimization problem (4.1.7). To compute $f(y, \hat{\theta}_0(y), \hat{\sigma}_0^2(y))$ let's recall the two steps procedure seen in Sect.**??**. Assuming we have computed $\hat{\theta}_0(y)$ by solving the constrained minimization problem (4.1.7), the ML variance estimator $\hat{\sigma}_0(y)$ is found by substituting the expression of $\hat{\theta}_0(y)$ in the pdf and maximizing with respect to σ^2. This maximization does not depend on the actual expression of $\hat{\theta}_0^2(y)$ and yields the expected result

$$\hat{\sigma}_0^2(y) = \frac{1}{N}\|y - S\hat{\theta}_0(y)\| = \frac{1}{N}R_0^2(y)\,. \tag{4.1.17}$$

Substituting in the pdf (4.1.14), one finds

$$f(y, \hat{\theta}_0(y), \hat{\sigma}_0^2(y)) = \left[2\pi\frac{R_0^2(y)}{N}\right]^{-N/2} \exp(-N/2)$$

and hence

$$L(y) = \left[\frac{R_0^2(y)}{R_1^2(y)}\right]^{N/2} = \left[\frac{R_0^2(y) - R_1^2(y)}{R_1^2(y)} + 1\right]^{N/2} \tag{4.1.18}$$

which shows that the MLR $L(y)$ is a function of the ratio (4.1.13). In fact, an invertible function. Therefore (4.1.13) is equivalent to the statistic prescribed by the MLR test.

The statistical decision to choose H_0 vs H_1 requires the knowledge of the pdf of the MLR statistic $R_0^2(y) - R_1^2(y)/R_1^2(y)$ and to have a chance to see this distribution we shall need an explicit expression of $R_0^2(y) - R_1^2(y)$. This requires in turn the solution of the constrained minimization problem (4.1.7) which so far we have avoided.

To this end, let us introduce a k-dimensional parameter β by setting

$$\beta := H\theta\,. \tag{4.1.19}$$

so that H_0 can be described simply as the hypothesis $\{\beta = \beta_0 := H\theta_0\}$. Since H is full rank β is uniquely defined by the position (4.1.19). The ML estimator of β is clearly

$$\hat{\beta}(\mathbf{y}) = H\hat{\theta}(\mathbf{y}) = H\left[S^\top S\right]^{-1} S^\top \mathbf{y} \tag{4.1.20}$$

and its normalized variance matrix is

$$\frac{1}{\sigma^2}\,\mathrm{Var}\,(\hat{\beta}(\mathbf{y})) = H\left[S^\top S\right]^{-1} H^\top := D\,. \tag{4.1.21}$$

By substituting the model equation (4.1.1) we have

$$\hat{\beta}(\mathbf{y}) = H(S^\top S)^{-1}S^\top \mathbf{y} = H\theta + \sigma H(S^\top S)^{-1}S^\top \mathbf{w} := \beta + \sigma \mathbf{e}\,. \tag{4.1.22}$$

Under H_0 we have $\beta = \beta_0$ and hence,

Lemma 4.2. *Under H_0 the ML estimator of the parameter β is given by*

$$\hat{\beta}(\mathbf{y}) = H\hat{\theta}(\mathbf{y}) = \beta_0 + \sigma\mathbf{e}\,, \qquad \mathbf{e} \sim \mathcal{N}(0, D) \tag{4.1.23}$$

where the random vector $\mathbf{e} = H(S^\top S)^{-1}S^\top \mathbf{w}$, has variance matrix D defined in (4.1.21)*.*

Interpretation: We have transformed the problem (4.1.7) into one where the first variables $(\theta_1 ... \theta_k)$ are replaced by $(\beta_1 ... \beta_k)$ so that the constraint $H\theta = \beta_0$ reduces to fixing certain pre-fixed values say $(\beta_{01}, ..., \beta_{0k})$ to the new parameter. The constraint will then be acting only on the first k variables of θ leaving the remaining $p - k$ free.
This can also be seen as introducing a change of basis in the observation space. Look for a change of basis in $\mathbb{R}^N$ whereby the first k equations of the model $\mathbf{y} = S\theta + \sigma\mathbf{w}$ become like $\mathbf{z} = H\theta + \sigma\mathbf{e}$, where $\mathbf{z}$ is now a k-dimensional random vector. We are essentially looking for a matrix Q such that:

$$QS\theta = H\theta\,, \qquad \forall\,\theta \in \mathbb{R}^p$$

that is $QS = H$, where Q should be of dimension $k \times N$. This problem admits a soution (not necessarily unique) since the rows of H are $k \leq p$ linearly independent vectors in $\mathbb{R}^p$ which must then belong to the rowspace of S, which by assumption is the whole space $\mathbb{R}^p$. Let us then try a solution of the form $Q = CS^\top$, for some $C \in \mathbb{R}^{k\times p}$ which should therefore satisfy the equation $CS^\top S = H$. By our standard assumption $S^\top S$ is invertible and hence,

$$C = H(S^\top S)^{-1} \qquad Q = H(S^\top S)^{-1}S^\top\,. \tag{4.1.24}$$

Lemma 4.3. *One has*

$$\|S\hat{\theta}(\mathbf{y}) - S\hat{\theta}_0(\mathbf{y})\|^2 = \|\hat{\beta}(\mathbf{y}) - \beta_0\|^2_{D^{-1}} \tag{4.1.25}$$

where $\hat{\beta}(\mathbf{y})$ *and* D *are defined in* (4.1.23).

Proof. We need to solve the constrained minimization (4.1.7). To this end , introduce the Lagrange multiplier $\lambda \in \mathbb{R}^k$ and consider the Lagrangian

$$\min_{\theta} \{\|y - S\theta\|^2 + \lambda^\top (H\theta - \beta_0)\}\,.$$

Setting the gradient with respect to θ to zero one finds the condition

$$-2S^\top (y - S\theta) + H^\top \lambda = 0$$

which yields an expression for the extremal

$$\hat{\theta}_0(y) = [S^\top S]^{-1} S^\top y - \frac{1}{2}[S^\top S]^{-1} H^\top \lambda\,, \tag{$*$}$$

depending on λ. The multiplier is fixed by imposing the constraint $H\hat{\theta}_0(y) = \beta_0$ which is equivalent to

$$\frac{1}{2} H[S^\top S]^{-1} H^\top \lambda = H[S^\top S]^{-1} S^\top y - \beta_0$$

that is

$$\frac{1}{2} D\,\lambda = \hat{\beta}(y) - \beta_0\,.$$

Substituting into $(*)$ one finds

$$\hat{\theta}_0(y) = \hat{\theta}(y) - [S^\top S]^{-1} H^\top D^{-1}\,[\hat{\beta}(y) - \beta_0]$$

that is,

$$S\,[\hat{\theta}(y) - \hat{\theta}_0(y)\,] = S[S^\top S]^{-1} H^\top D^{-1}\,[\hat{\beta}(y) - \beta_0]$$

which immediately leads to (4.1.25). □

Theorem 4.1. *The decomposition* (4.1.10) *can be written as*

$$R^2_0(\mathbf{y}) = \|\hat{\beta}(\mathbf{y}) - \beta_0\|^2_{D^{-1}} + R^2_1(\mathbf{y})\,. \tag{4.1.26}$$

The random variables $\|\hat{\beta}(\mathbf{y}) - \beta_0\|^2_{D^{-1}}$ *and* $R^2_1(\mathbf{y})$*, are independent under both* H_0 *and* H_1*.*

Proof. Recall that $\hat{\beta}(\mathbf{y}) = H\hat{\theta}(\mathbf{y})$ e $R^2_1(\mathbf{y}) = \|\mathbf{y} - S\hat{\theta}(\mathbf{y})\|^2$*;* hence we just need to show that $\hat{\theta}(\mathbf{y})$ and $\mathbf{y} - S\hat{\theta}(\mathbf{y})$ are independent. Use now the projector $P = S(S^\top S)^{-1} S^\top$*,* to compute

$$\begin{aligned}
\mathrm{Cov}\,\left[\hat{\theta}(\mathbf{y}), (\mathbf{y} - P\mathbf{y})\right] &= \mathbb{E}\,\left[\hat{\theta}(\mathbf{y})\,(\mathbf{y} - S\hat{\theta}(\mathbf{y}))^\top\right] = \sigma(S^\top S)^{-1} S^\top \mathbb{E}\,(\mathbf{y}\mathbf{w}^\top)(I - P)^\top \\
&= \sigma^2 (S^\top S)^{-1} S^\top (I - P) \\
&= \sigma^2\left[(S^\top S)^{-1} S^\top - (S^\top S)^{-1} S^\top S (S^\top S)^{-1} S^\top\right] = 0\,.
\end{aligned}$$

□

Note that under H_0 one has $\hat{\beta}(\mathbf{y}) \sim \mathcal{N}(\beta_0, \sigma^2 D)$ (Lemma 4.2), and hence

$$\frac{R_0^2(\mathbf{y}) - R_1^2(\mathbf{y})}{\sigma^2} = \frac{1}{\sigma^2}\|\hat{\beta}(\mathbf{y}) - \beta_0\|^2_{D^{-1}} \sim \chi^2(k) \tag{4.1.27}$$

One can show that under H_1 the random variable (4.1.27) has instead a *noncentral* χ^2 distribution, $\chi^2(k, \delta)$; δ being the so-called *non-centrality parameter* [9]:

$$\delta = \frac{1}{\sigma^2}\|\hat{H}_0\theta - \beta_0\|^2_{D^{-1}} .$$

We are finally in a position to describe the pdf of the ratio (4.1.18).
Recall that the unconstrained sum of squared residuals $R_1^2(\mathbf{y})$ has a χ^2 distribution, namely:

$$\frac{R_1^2(\mathbf{y})}{\sigma^2} \approx \chi^2(N-p) . \tag{4.1.28}$$

We now just need to refer to Sec. A.2 in the appendix to conclude that:

Theorem 4.2. *Under H_0, the ratio*

$$\mathbf{z} := \frac{(N-p)}{k} \frac{\|\hat{\beta}(\mathbf{y}) - \beta_0\|^2_{D^{-1}}}{R_1^2(\mathbf{y})} \tag{4.1.29}$$

is distributed according to an F distribution, in fact as $\mathcal{F}(k, N-p)$. The critical region of the test can be expressed as

$$C := \{y\,;\, \mathbf{z}(y) \geq k_\alpha\} \tag{4.1.30}$$

where k_α is the abscissa defined by

$$\mathbb{P}_0(\mathbf{z} \geq k_\alpha) = \alpha .$$

The test based on the ratio (4.1.29) is called **F test** and is largely used in Statistics.

In some problems R_0^2 and the expression of the numerator of the statistic F can be obtained directly and there is no need of computing explicitly the estimator $\hat{\beta}$.

Example 4.1 (Known variance). Consider the following linear model with N observations

$$\begin{bmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_N \end{bmatrix} = \begin{bmatrix} s_1 & s_2 \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} + \begin{bmatrix} \mathbf{w}_1 \\ \vdots \\ \mathbf{w}_N \end{bmatrix}$$

where the $\mathbf{w}_i$ are Gaussian independent of mean zero and the variance σ^2 is known. We want to test the Hypothesis that two regressors are superfluous, choosing the null hypothesis as

$$H_0 \equiv \theta_1 = \theta_2 .$$

[9] The standard reference for this material is the classical old book by Scheffè [79]

Set $s := s_1 + s_2$ and describe the two probability densities under H_0 and under the alternative hypotheses. Use the Maximum Likelihood Ratio test to find a suitable test statistic and its distribution.

Solution:

The problem assumes that σ^2 is known; however the theory of linear hypotheses can easily be adapted to this particular case. Set $S = \begin{bmatrix} s_1 & s_2 \end{bmatrix}$ and $\theta = \begin{bmatrix} \theta_1 & \theta_2 \end{bmatrix}^\top$ and let $y \in \mathbb{R}^N$ be the observation. The two likelihood functions are

$$f_0(y;\, \theta_0) = (2\pi\sigma^2)^{-N/2}\, \exp -\frac{1}{2\sigma^2}\|y - s\theta_0\|^2$$
$$f_1(y;\, \theta_1,\, \theta_2) = (2\pi\sigma^2)^{-N/2}\, \exp -\frac{1}{2\sigma^2}\|y - S\theta\|^2$$

so that

$$L(y) = \exp \frac{N}{2\sigma^2}\left\{R_0^2(y) - R_1^2(y)\right\}$$

where $R_0^2(y) = \|y - s\hat{\theta}_0(y)\|^2$ and $R_1^2(y) = \|y - S\hat{\theta}(y)\|^2$ and the test statistic can be chosen $\varphi(y) := R_0^2(y) - R_1^2(y)$. This can be computed using the formula

$$\varphi(y) := R_0^2(y) - R_1^2(y) = \|\hat{\beta}(y) - \beta_0\|^2_{D^{-1}}\,,$$

where $\hat{\beta}(y) = \begin{bmatrix} 1 & -1 \end{bmatrix} \hat{\theta}(y)\,, \quad \beta_0 = 0$ and $D = \begin{bmatrix} 1 & -1 \end{bmatrix} \left[S^\top S\right]^{-1} \begin{bmatrix} 1 \\ -1 \end{bmatrix}$ are all scalars. By Gaussianness

$$\frac{\varphi(\mathbf{y})}{\sigma^2} \sim \chi^2(1)\,.$$

If you assume instead that σ^2 is unknown, then the procedure follows exactly the steps delineated in the above discussion and you end up with a $\mathcal{F}(1, N-2)$ distribution.

Application to the Analysis of Variance (ANOVA)

The Analysis of Variance (ANOVA) is a statistical method used to test differences between two or more means. It may seem odd that the technique is called "Analysis of Variance" rather than "Analysis of Means." As you will see, the name is appropriate because inferences about means are made by analyzing variance. It has been used widely since the seminal articles by Ronald Fisher [32, 33], especially in bio- or medical statistics for comparing medical treatments or effectiveness of drugs.

Assume we have p samples of size $N_1, ..., N_p$ extracted from p normal populations $\mathcal{N}(\mu_i, \sigma^2)$ with $i = 1, ..., p$, where the means $\mu_1, ..., \mu_p$ and the common variance σ^2 are unknown. They could for example be $N_1, ..., N_p$ measurements of a physical variable made by p different measurement devices, having however the same precision. Denote by $\mathbf{y}_i$ the i-th sample and by $\theta = [\mu_1, ..., \mu_p]^\top$ the $p-$dimensional vector of the unknown means, one may describe this set-up by a linear model of the following form

$$\mathbf{y} := \begin{bmatrix} \mathbf{y}_1 \\ . \\ . \\ \mathbf{y}_p \end{bmatrix} = \begin{bmatrix} e_{N_1} & 0 & 0 & 0 \\ 0 & . & 0 & 0 \\ 0 & 0 & . & 0 \\ 0 & 0 & 0 & e_{N_p} \end{bmatrix} \theta + \sigma\mathbf{w} \tag{4.1.31}$$

where $e_{N_i} = [1...1]^\top \in \mathbb{R}^{N_i}$ and $\mathbf{w} \sim \mathcal{N}(0, I_N)$, is $N_1+...+N_p := N-$dimensional white noise.
One wants to test the hypothesis

$$H_0 : \quad \mu_1 = \mu_2 = ... = \mu_p \tag{4.1.32}$$

against the alternative H_1 that (some) means are different.
To set up a linear model denote by $\{y_{it}\,;\, i = 1, ..., p\,;\, t = 1, ..., N_i\}$ the p strings of observations obtained in the p experiments and let us compute the residual sum of squares under H_1 and under H_0. Under H_1, one has:

$$R_1^2(y) = \min_{\mu_1,...,\mu_p} \sum_{i=1}^{p}\sum_{t=1}^{N_i}(y_{it} - \mu_i)^2 = \sum_{i=1}^{p}\min_{\mu_i}\sum_{t=1}^{N_i}(y_{it} - \mu_i)^2 \tag{4.1.33}$$

and the value of μ_i which minimizes each sum is $\hat{\mu}_i = \bar{y}_{N_i} = \frac{1}{N_i}\sum_{t=1}^{N_i} y_{it}$. Hence

$$R_1^2(y) = \sum_{i=1}^{p}\sum_{t=1}^{N_i}(y_{it} - \bar{y}_{N_i})^2 = \sum_{i=1}^{p}\hat{\sigma}^2_{N_i}\,,$$

which is distributed as $\sigma^2\chi^2(N-p)$. Under H_0, the mean is the same and therefore

$$R_0^2(y) = \min_{\mu}\sum_{i=1}^{p}\sum_{t=1}^{N_i}(y_{it} - \mu)^2 = \sum_{i=1}^{p}\sum_{t=1}^{N_i}(y_{it} - \bar{y}_N)^2\,,$$

which is distributes as $\sigma^2\chi^2(N-1)$. The difference $R_0^2 - R_1^2$ can be computed using the identity

$$\sum_{t=1}^{N_i}(y_{it} - \bar{y}_N)^2 = \sum_{t=1}^{N_i}[y_{it} - \bar{y}_{N_i} + (\bar{y}_{N_i} - \bar{y}_N)]^2 = \sum_{t=1}^{N_i}(y_{it} - \bar{y}_{N_i})^2 + N_i(\bar{y}_{N_i} - \bar{y}_N)^2$$

which leads to

$$R_0^2 - R_1^2 = \sum_{1}^{p} N_i(\bar{y}_{N_i} - \bar{y}_N)^2 \tag{4.1.34}$$

This is a weighted sum of the deviations of the sample means for each group from the overall sample mean $\bar{y}_N$. Note that in this problem we have $k = p - 1$ since (4.1.32) can be written

$$\mu_1 - \mu_2 = 0\,;\ ...\ ;\ \mu_1 - \mu_p = 0\,,$$

which are $p-1$ independent equations.

Example. Consider $p = 3$ groups with sample means as described in Table 4.1;

The hypothesis H_0 is that the three means are equal. The sum of squared differences under H_0 is $R_0^2 = 4616.64$. Using the formulas just derived one finds $R_0^2 - R_1^2 = 238.59$. Since $k = p - 1 = 2$ and $N - p = 142 - 3 = 139$ the test statistic is computed to be

$$F = \frac{139}{2}\frac{238.59}{4616.64 + 238.59} = 3.79\,.$$

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 4.1.

α	0.10	0.05	0.025	0.01
k_α	2.30	3.00	3.70	4.65

Table 4.2. *Quantiles of* $F(2, 139)$

As shown in Table 4.2, unless the probability of an error of first kind α, is chosen extremely small, we are well inside the critical region and the hypothesis that the three means are the same should be rejected.

At this point one may ask if the means μ_i of the three populations are significantly different. The answer is that they are (thus rejecting H_0), if we do not require too high "statistical certainty" to this statement. One can say that with a probability slightly higher than $97,5$ percent the means are to be considered different. There is however not enough experimental evidence to make the same statement with a statistical certainty of 99 percent of not committing a mistake by refusing the hypothesis. In this case; i.e. once fixing $\alpha = 0.01$, we would be lead to accept H_0. Unfortunately now the probability α would not tell us anything on the risk incurred in choosing H_0 when H_1 is true. To this end, we should be able to compute the probability, β, of accepting H_0 when H_1 is true, which is normally complicated since H_1 is a *compound hypothesis*. This difficulty is obviously the same as regarding the calculation of the power of the test. In any case, β normally increases when α diminishes (see for example Fig. **??**) so that the decision of accepting H_0 when α is chosen very small, reveals in general to be meaningless, as this may entail very high values for β, and, as argued in [85] possibly even close to $1 - \alpha$.

Under H_1 the ratio F defined in (4.1.29) is no longer distributed as $F(k, N-p)$. It has instead a *non central* F distribution depending on a *non-centarlity parameter* λ, defined as

$$\lambda^2 = \lambda^2(\theta, \sigma^2) = \frac{1}{\sigma^2}\|\beta - \beta_0\|^2_{D^{-1}} = \frac{1}{\sigma^2}\|H\theta - \beta_0\|^2_{D^{-1}}\,. \tag{4.1.35}$$

A non-central F distribution can however be approximated by an ordinary (central) F. In many instances it will be enough to refer to the symbolic relation (which obviously needs to be understood as a relation between random variables)

$$F(n_1, n_2, \lambda) \cong \frac{n_1 + \lambda}{n_1} F(n_1^*, n_2) \tag{4.1.36}$$

where n_1^* is given by:

$$n_1^* = (n_1 + \lambda)^2/(n_1 + 2\lambda)\,. \tag{4.1.37}$$

In this way the power of the test can be computed from $F(n_1^*, n_2)$. Since n_1^* will in general not be an integer one may use the closest integer approximation. The

formula for computing the power corresponding to $\lambda = \lambda(\theta, \sigma^2)$ is:

$$1 - \beta(\theta, \sigma^2) = \int_{\frac{n_1}{n_1+\lambda} a_\alpha}^{\infty} dF(n_1^*, n_2)\,. \tag{4.1.38}$$

This formula for $\lambda = 0$ (or $\beta = \beta_0$) returns α.

Hypothesis Testing and Confidence Regions

A *Confidence Region* for the parameter θ under some hypothesis H_0 (for H_1 there is an equivalent definition), is the complementary set to the critical region. If the critical region is defined by the error probability equal to α, then it is said that the confidence region has "size" $1-\alpha$. Let $\hat{\theta}(\bar{y})$ be an estimate of θ corresponding to the observation $\bar{y}$. One may test the goodness of (i.e. validate) the estimate by testing the hypothesis

$$H_0 : \left\{\theta = \hat{\theta}(\bar{y})\right\}\,. \tag{4.1.39}$$

The complement of the critical region of this test is a confidence region of size $1 - \alpha$.

4.2 ▪ Examples

Consider the linear model of example 2.2. You want to validate the estimates of the three exponents $\begin{bmatrix}\theta_1 & \theta_2 & \theta_3\end{bmatrix}^\top$ by running a series of statistical tests. For each problem compute the F statistic and use a table of the F distribution, for example `http://www.socr.ucla.edu/Applets.dir/F_Table.html` to compare the critical abscissa k_α for values of $\alpha = 0.01, .025, 0.05$ and .1. Discuss your decision.

4-1 Adequacy of the model
Test the Hypothesis

$$H_0 : \theta_1 = \theta_2 = \theta_3 = 0\,.$$

4-2 Equality of the exponents
Test the Hypothesis

$$H_0 = \theta_1 = \theta_2 = \theta_3\,.$$

4-3 All exponents equal to 1
Test the Hypothesis

$$H_0 = \theta_1 = \theta_2 = \theta_3 = 1\,.$$

4-4 The sum of the exponents is equal to 3
Test the Hypothesis

$$\theta_1 + \theta_2 + \theta_3 = 3$$

4-5 The exponent of H is zero
Test the Hypothesis

$$\theta_3 = 0\,.$$

◇

Solutions Recap: want to model Y as

$$Y \cong \alpha L^{\theta_1} K^{\theta_2} H^{\theta_3} \tag{4.2.1}$$

using 86 measurements of L, K, H. The logarithms satisfy a linear model: defining

$$y = \log Y\,, \quad x_1 = \log L\,, \quad x_2 = \log K\,, \quad x_3 = \log H\,, \quad \theta_0 : \log \alpha\,,$$

gives

$$y_t = \theta_0 + \theta_1 x_{1t} + \theta_2 x_{2t} + \theta_3 x_{3t} + \epsilon_t\,, \qquad t = 1, ..., 86. \tag{4.2.2}$$

The errors ϵ_t are Gaussian zero-mean and independent of unknown variance σ^2. After subtracting the sample means

$$\bar{x_i} = \frac{1}{86}\sum_1^{86} x_{it}\,, \qquad i = 1, 2, 3$$

one gets a model for the centered variables

$$\Delta y_t := y_t - \bar{y} = \sum_1^3 \theta_i (x_{it} - \bar{x}_i) + (\epsilon_t - \bar{\epsilon})\,. \tag{4.2.3}$$

The estimate of θ_0 can be obtained from

$$\bar{y} = \theta_0 + \theta_1 \bar{x}_1 + \theta_2 \bar{x}_2 + \theta_3 \bar{x}_3 + \bar{\epsilon}\,.$$

Rewrite (4.2.3) in vector form:

$$\Delta \mathbf{y} = S\theta + \sigma \mathbf{w}\,, \qquad \mathbf{w} \sim \mathcal{N}(0, I)\,.$$

Where

$$S^\top S = \begin{bmatrix} 0.0187 & 0.0085 & 0.0068 \\ 0.0085 & 0.029 & 0.0088 \\ 0.0068 & 0.0088 & 0.029 \end{bmatrix} \qquad S^\top \Delta y = \begin{bmatrix} 0.030 \\ 0.044 \\ 0.036 \end{bmatrix}$$

from which $\hat{\theta} = \left[S^\top S\right]^{-1} S^\top \Delta y$, is equal to

$$\hat{\theta}_1 = 0.88\,, \quad \hat{\theta}_2 = 1.04\,, \quad \hat{\theta}_3 = 0.73$$

and

$$\hat{\theta_0} = -2.618$$

Finally: $Y = 0.00241 L^{0.88} K^{1.04} H^{0.73}$.

The squared sum of residuals is

$$R_1^2 = \|\Delta y\|^2 - \|S\hat{\theta}\|^2 = \|\Delta y\|^2 - \langle S\hat{\theta},\, \Delta y\rangle = \|\Delta y\|^2 - \hat{\theta}^\top S^\top \Delta y \tag{4.2.4}$$

that is

$$R_1^2 = \sum_{t=1}^{86} (y_t - \bar{y})^2 - (\hat{\theta}_1 0.030 + \hat{\theta}_2 0.044 + \hat{\theta}_3 0.036) \tag{4.2.5}$$

$$= 0.127 - 0.099 = 0.028 \tag{4.2.6}$$

Will need the unbiased estimate of the variance:

$$\hat{\sigma}^2 = \frac{R_1^2}{N-4} = \frac{0.028}{82} = 0.00034\,.$$

A. Model adequacy

Let us test the hypothesis

$$H_0 : \theta_1 = \theta_2 = \theta_3 = 0$$

(Testing $\theta_0 = 0$, that is $\alpha = 1$ in (4.2.1) does not make much sense as it would correspond to $Y = 1 +'' \, noise''$). H_0 is just saying that the data would not support a model of the form (4.2.1).
Testing H_0 requires $R_0^2 - R_1^2 = \|\hat{\beta} - \beta_0\|^2_{D^{-1}} = \|\hat{\beta}\|^2_{D^{-1}}$, where $\beta = \begin{bmatrix}\theta_1 & \theta_2 & \theta_3\end{bmatrix}^\top \equiv \theta$ and $\beta_0 = 0$. Therefore

$$D = \left[S^\top S\right]^{-1}$$

and

$$\|\hat{\beta}\|^2_{D^{-1}} = \|\hat{\theta}\|^2_{D^{-1}} = \hat{\theta}^\top S^\top S\hat{\theta} = \hat{\theta}^\top S^\top \Delta y$$

the last equality follows from the orthogonality $S\hat{\theta} \perp \Delta y - S\theta$ by which $\langle S\hat{\theta}, S\hat{\theta}\rangle = \langle S\hat{\theta}, \Delta y\rangle$. It follows that $R_0^2 - R_1^2$ is simply the last term in (4.2.4). This can be checked by noting that under H_0,

$$R_0^2 = \min_{\theta_0} \|\Delta y - S\theta\|^2 = \|\Delta y\|^2$$

which is exactly the firs summand in (4.2.4). Therefore

$$F_A = \frac{N-p}{k}\frac{R_0^2 - R_1^2}{R_1^2} = \frac{86-4}{3}\frac{0.099}{0.028} = 97.4$$

Looking in the table of $F(3, 82)$ one finds the following values for k_α which lead to refuse (A) even if α is taken extremely small.

α	0.10	0.05	0.025	0.01	0.005
k_α	2.15	2.70	3.90	4.00	4.60

B. Equality of the exponents

Here

$$H_0 = \theta_1 = \theta_2 = \theta_3$$

which is equivalent to $H\theta = 0$ whereby

$$\beta := \begin{bmatrix}1 & -1 & 0\\ 1 & 0 & -1\end{bmatrix}\theta = 0\,.$$

Here we could use the formula $\|\hat{\beta}\|^2_{D^{-1}} = R_0^2 - R_1^2$, as previously done. It is however more convenient to compute R_0^2 directly. Setting $\theta_1 = \theta_2 = \theta_3 = \eta$ one has

$$\theta = \mathbf{1}\eta\,; \qquad \mathbf{1} = \begin{bmatrix}1\\1\\1\end{bmatrix}$$

since $\Delta y = (S\mathbf{1})\eta + \sigma w$, the normal equations for this problem are

$$(\mathbf{1}^\top S^\top S\mathbf{1})\eta = \mathbf{1}^\top S^\top \Delta y$$

so that $\hat{\eta}$ is the sum of the elements of $S^\top S$ and of $S^\top \Delta y$, which gives

$$0.125\eta = 0.11 \Rightarrow \hat{\eta} = 0.887$$

that is

$$R_0^2 = \|\Delta y - S\mathbf{1}\hat{\eta}\|^2 = \|\Delta y\|^2 - \hat{\eta}(\mathbf{1}^\top S^\top \Delta y)$$

and, by (4.2.4),

$$R_0^2 - R_1^2 = 0.099 - \hat{\eta} \cdot 0.11 = 0.099 - 0.098 = 0.001\,.$$

The test statistic has the sample value

$$F_B = \frac{N-p}{k}\frac{0.001}{0.028} = \frac{82}{2}\frac{0.001}{0.028} = 1.4$$

in the table of $F(2, 82)$ one finds the following values for k_α

α	0.10	0.05	0.025	0.01
k_α	2.37	3.11	3.86	4.86

For $\alpha = 0.10$, the critical abscissa of $F(2, 82)$ is 2.37 and we are already well outside the critical region. For smaller values of α, k_α will be even greater. There is no experimental evidence to refuse the hypothesis that θ_1, θ_2, θ_3 aree equal. Said in a less cryptic way we should accept the hypothesis that the three parameters are equal.

C. All exponents equal to 1

In this case

$$H_0 = \theta_1 = \theta_2 = \theta_3 = 1$$

which can be written

$$I\theta = \text{vec}\,\{1,\ 1,\ 1\}\,,$$

and yields

$$R_0^2 - R_1^2 = \|\hat{\theta} - \mathbf{1}\|^2_{[S^\top S]} = (\hat{\theta} - \mathbf{1})^\top (S^\top S)(\hat{\theta} - \mathbf{1}) = 0.0026\,.$$

In this case $k = 3$ and $F_C = 2.5$. The critical values of $F(3, 82)$ are

	0.10	0.05	0.025
	2.15	2.70	3.30

The hypothesis can be accepted with $\alpha = 0.05$ e refused with $\alpha = 0.10$. The case is dubious.

D. The sum of the exponents is 3

Want to check if

$$\theta_1 + \theta_2 + \theta_3 = 3$$

Letting $\beta := \theta_1 + \theta_2 + \theta_3$, one has $\beta_0 = 3$ and

$$\hat{\beta} = \hat{\theta}_1 + \hat{\theta}_2 + \hat{\theta}_3 = 2.65$$

while D is a scalar equal to the sum of the elements of $\left[S^\top S\right]^{-1}$; i.e. $D = 75.7$, which yields

$$\|\hat{\beta} - 3\|^2 / D = \frac{(2.65 - 3)^2}{75 \cdot 7} = \frac{(0.35)^2}{75.7}$$

and one finds that $F_D = 4.72$.

The critical abscissas of $F(1, 82)$ are

α	0.10	0.05	0.025	0.01
a_α	2.77	3.96	5.20	6.95

For $\alpha = 0.05$ F_D is in the critical region. It is reasonable to refuse H_0.

E. The exponent of H is equal to zero

H has the smalles exponent. Let us check if

$$\theta_3 = 0$$

In this case $\hat{\beta} = \hat{\theta}_3 = 0.73$, $D = 39.88$ and one immediately gets

$$\|\hat{\beta} - \beta_0\|^2_{D^{-1}} = \frac{(0.73)^2}{39.9} = 0.0135$$

from which

$$F_E = \frac{0.0135}{3.4 \cdot 10^{-4}} = 39.72$$

From the table of $F(1, 82)$ one sees that F_E falls in the critical region even for very small values of α. Therefore we reject the hypothesis that $\theta_3 = 0$. $\diamond$

4.3 ▪ Pattern Recognition and Linear Discriminant Analysis

In this section we shall address a class of classification problems which could be seen as a variant of hypothesis testing but actually have a different flavour and scope. Typically one has noisy measurements $x(t)\,;\, t = 1, 2, \ldots, N$ of certain *features* that are used to characterize a finite class of objects each of which we shall call a *pattern*. There are M possible distinct patterns to choose from, for example one would like to classify individuals as "male" or "female" based on recorded measurements of height and weight or recognize one of the first ten natural numbers $\{0,\, 1,\, \ldots, 9\}$ ($M = 10$) from features suitably extracted from a handwritten postal code. The features which characterize a pattern are a finite fixed number, say p, of real (or categorical) variables. More sophisticated applications occur in speech recognition and computer vision where the goal is to recognize objects such as words or human faces from data extracted by ad hoc signal processing techniques.

In this class of problems we are given a **training set** of N, p-dimensional observed samples $\{x(t)\,;\, t = 1, 2, \ldots, N\}$ which have already been classified; i.e. each vector of feature data $x(t)$ is already assigned to one of the M possible patterns and has attached an assignement label $y(t)$ perhaps one of the first M natural numbers. In presence of noise the data sequence will be assumed to be i.i.d. drawn from one of M possible probability distributions which in this setting are a mathematical description of the different patterns. The classification problem is then that of deciding which of the M probability distributions describes a next p-dimensional observation which does not belong to the training set.

The Automatic classification Problem: One should infer from the training set, i.e. a given stored sequence of N classified data, a *decision rule* by which classify the next incoming p-tuple of measured features; i.e. design an algorithm which decides to which pattern it belongs to. This is called *Statistical Pattern Recognition* in Engineering circles, see [75]. A key point is that here the purpose of the statistical exercise differs from the standard one in classical hypothesis testing where the decision is *after the fact* as it regards the classification or validation of a *descriptive model* (in fact, its parameters) from observed data which are *not classified*. In certain circles this is called *unsupervised learning* since the test function for deciding the partition of the sample space on which the decision should be made is designed without any a priori experience. In *supervised learning* instead one uses the past classification experience to train a model to do *prediction*.

In order to do this, one first needs to categorize different patterns in terms of a p-ple of characteristic feature values. This is the **Feature Extraction Problem**. Somehow the same as describing the patterns by means of p real or also integer-valued coordinates. Since in many applications the patterns are originally described by qualitative attributes, the coordinatization may in practice turn out to be not obvious and may require a good deal of engineering ingenuity.

In the statistical setting, one should at the end ideally be able to describe the M patterns by M distinct conditional probability distributions on the feature space, say $\mathbb{R}^p$, each distribution being conditioned by the observed training set. Then, infer a decision rule to classify new incoming feature data points as belonging to one of the M classes.

To explain this in more detail, we shall assume that the p noisy features of each pattern are described by M joint Gaussian density functions:

$$p_k(x) = \frac{1}{(2\pi)^{p/2} \det \Sigma_k^{1/2}} \exp -\frac{1}{2}(x-\mu_k)^\top \Sigma_k^{-1}(x-\mu_k)\,, \qquad k = 1, \dots, M\,. \tag{4.3.1}$$

where $x \in \mathbb{R}^p$ and the parameters $\mu_k \in \mathbb{R}^p$ and $\Sigma_k \in \mathbb{R}_+^{p\times p}$ are in general unknown. To find the decision rule we may still invoke the MLR principle. This will prescribe to substitute the unknown parameters μ_k and Σ_k with their sample estimates computed from the training set,

$$\hat\mu_k := \frac{1}{N_k}\sum_{t=1}^{N_k} x_k(t)\,, \qquad \hat\Sigma_k := \frac{1}{N_k}\sum_{t=1}^{N_k}(x_k(t)-\hat\mu_k)(x_k(t)-\hat\mu_k)^\top \tag{4.3.2}$$

where the subscript $_k$ denotes the data in the training set which are classified as belonging to the k-th class. Based on these estimates one forms a MLR to decide between classes, say class i and j:

$$L_{i,j}(x) = \frac{(2\pi)^{-p/2} \det \hat\Sigma_i^{-1/2} \exp -\frac{1}{2}(x-\hat\mu_i)^\top \hat\Sigma_i^{-1}(x-\hat\mu_i)}{(2\pi)^{-p/2} \det \hat\Sigma_j^{-1/2} \exp -\frac{1}{2}(x-\hat\mu_j)^\top \hat\Sigma_j^{-1}(x-\hat\mu_j)}\,, \tag{4.3.3}$$

the boundary between the two classes being defined by the equation $L_{i,j}(x) = 1$ while the decision (i.e. the test) statistic is defined by

$$\phi(x) = k \;\Leftrightarrow\; k = \operatorname{Arg}\max_i \hat p_i(x) \tag{4.3.4}$$

where $\hat p_i$ is the Gaussian pdf (4.3.1) *with the unknown parameters substituted by their class estimates*. This procedure could be seen as being based on the ratio of two *conditional likelihoods, given the training data*. Denoting the training set by $\mathbf{x}^N$ this means that the MLR ratio (4.3.4) is equal to the ratio of conditional pdf's

$$L_{i,j}(x) = \frac{p_i(x \mid \mathbf{x}^N = x^N)}{p_j(x \mid \mathbf{x}^N = x^N)} \tag{4.3.5}$$

where each conditional $p_i(x \mid \mathbf{x}^N = x^N)$ is actually equal to $\hat p_i(x)$ with the unknown parameters μ_i, Σ_i substituted by their ML estimates (4.3.2). This could be justified basing on the fact that $\hat\mu_i$ and $\hat\Sigma_i$ are *sufficient statistics* for the model as they contain all relevant information in the training set $\mathbf{x}^N$ which is sufficient to predict the next data [85]. The $\hat\mu_i$'s $i = 1, 2, \dots, M$, are called *centroids* of the various patterns and the decision rule to classify the observation x is based on choosing the pattern whose centroid is at smallest *Mahalanobis distance* from x, the Mahalanobis distance from the centroid of class i being defined as

$$(x-\hat\mu_i)^\top \hat\Sigma_i^{-1}(x-\hat\mu_i)\,.$$

This leads to the same decision rule defined in (4.3.4).

Linear Discriminant Analysis arises when the covariance matrices of the various class pdf's (4.3.1) are all the same and can be estimated based on the whole training data set. After canceling the multiplicative terms in (4.3.3) and taking

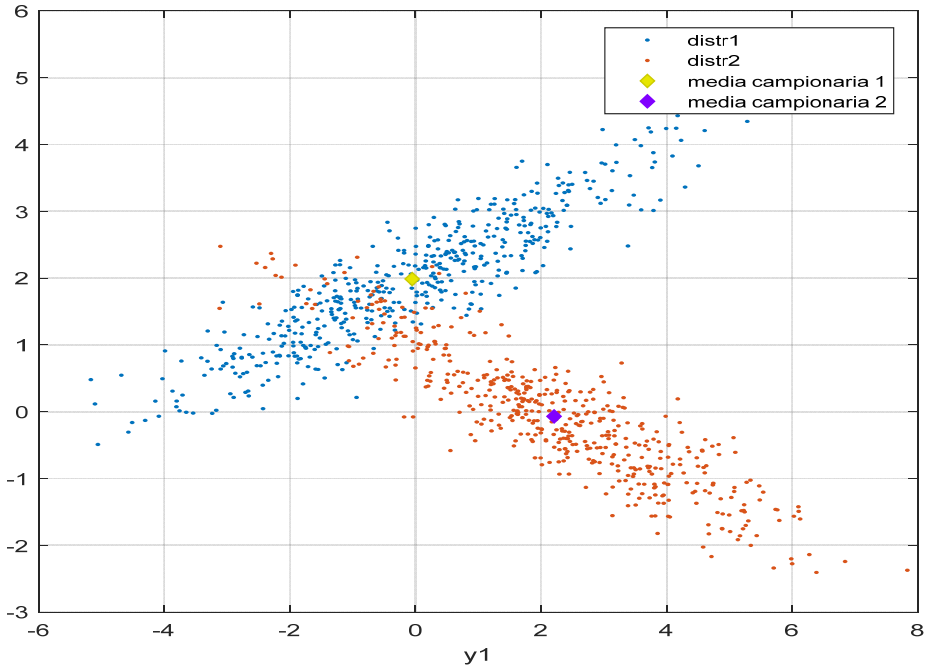


Figure 4.3.1. *Gaussian 2-dimensional feature training set*

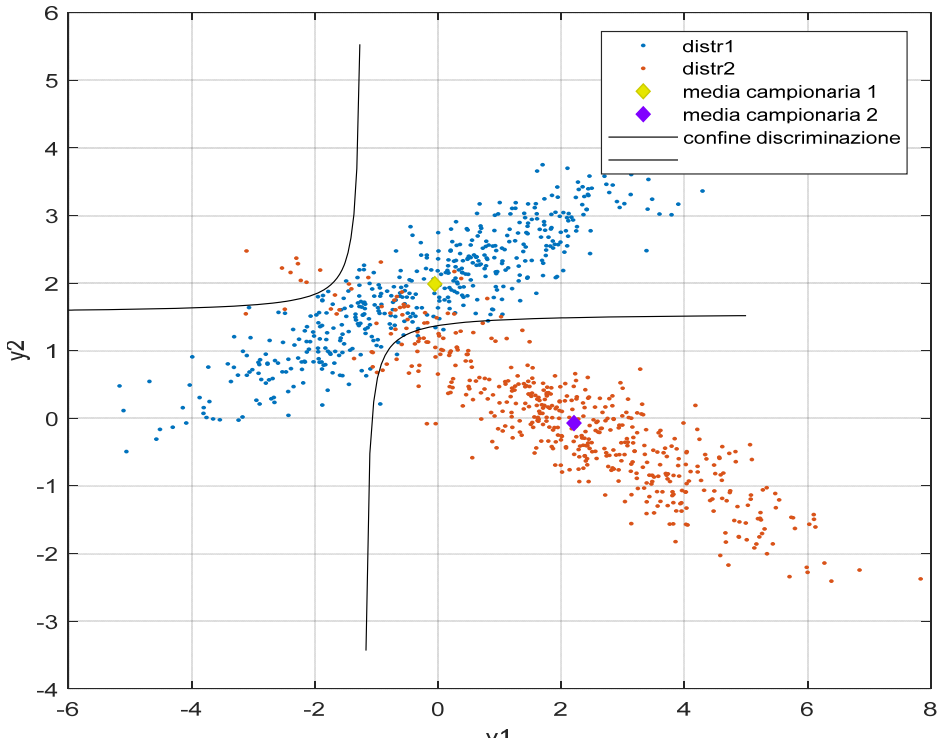


Figure 4.3.2. *Separating region*

logarithms of $L_{i,j}(x) = 1$ we find that the boundary between class i and class$j$ is defined by the equation

$$(x - \hat{\mu}_j)^\top \hat{\Sigma}^{-1}(x - \hat{\mu}_j) - (x - \hat{\mu}_i)^\top \hat{\Sigma}^{-1}(x - \hat{\mu}_i) = 0 \tag{4.3.6}$$

which after defining

$$\beta_i - \beta_j := \hat{\Sigma}^{-1}(\hat{\mu}_i - \hat{\mu}_j)\,, \qquad \gamma_i - \gamma_j := -\frac{1}{2}(\hat{\mu}_i - \hat{\mu}_j)^\top \hat{\Sigma}^{-1}(\hat{\mu}_i - \hat{\mu}_j)$$

and some algebra can be rewritten as

$$(\beta_i - \beta_j)^\top x = \gamma_i - \gamma_j \tag{4.3.7}$$

which is the equation of an *hyperplane* in $\mathbb{R}^p$. One should choose class i whenever the left hand side of (4.3.6) is positive since in this case the Mahlanobis distance of x from $\hat{\mu}_i$ is smaller than that from $\hat{\mu}_j$. In this case x falls in the region which is the intersection of the half spaces defined by the inequalities

$$(\beta_i - \beta_j)^\top x > \gamma_i - \gamma_j + k_{i,j}\,, \qquad j \neq i \tag{4.3.8}$$

where $k_{i,j}$ is to be determined by the error probability of accepting class i when instead the data are distributed according to the pdf of class j. In a Bayesian setting these numbers are much better defined as just the logs of the ratios of the a priori probabilities of the M classes, see Sect. 5.12. In any case the decision regions turn out to be convex polytopes whose boundaries are linear. Examples of partitions of $\mathbb{R}^2$ by this linear decision rule are found in [44], for example see Fig 4.5.

More examples and Figures can be generated by using the software in the website
`pmtk3support.googlecode.com`

4.4 ▪ Two Classes Separating Hyperplanes and the Perceptron

As we have seen, for Gaussian densities having the same variance matrix, the MLR decision boundaries between classes are hyperplanes. There is a large literature on classification of observed data based on separating hyperplanes, mostly without assuming anything about probability distributions. We shall here briefly discuss the binary situation ($M = 2$) following essentially the book [75].

In this context the observed feature vector $x \in \mathbb{R}^p$ is to be classified as belonging to one of two candidate patterns $\mathcal{P}_1$ and $\mathcal{P}_2$, based on a *linear discriminant function* $g : \mathbb{R}^p \to \mathbb{R}$ of the form

$$g(x) = \beta^\top x + b\,, \qquad x \in \mathbb{R}^p \tag{4.4.1}$$

which is a particular case of (4.3.7). The vector $\beta \in \mathbb{R}^p$ and the scalar b are called the *weight* vector and the *bias or threshold* of the function. The equation $g(x) = 0$ defines a linear hyperplane $\mathcal{H}$ partitioning $\mathbb{R}^p$ in two half spaces. One decides for alternative $\mathcal{P}_1$ if the scalar product $\beta^\top x$ exceeds the threshold $-b$ and for

alternative $\mathcal{P}_2$ in the opposite case. Equivalently, alternative $\mathcal{P}_1$ is chosen when $g(x) > 0$ and alternative $\mathcal{P}_2$ if $g(x) < 0$. Since for $x_1,\ x_2 \in \mathcal{H}$ one has

$$\beta^\top (x_1 - x_2) = 0$$

the vector β is normal to the hyperplane; therefore $\mathbf{n} := \beta/\|\beta\|$ is the unit vector pointing into the half space $\mathcal{H}_+$ where $g(x) > 0$. Any x can be decomposed as a sum of its orthogonal projection $\hat{x}$ onto $\mathcal{H}$ plus a vector which is normal to $\mathcal{H}$

$$x = \hat{x} + d\frac{\beta}{\|\beta\|} = \hat{x} + d\,\mathbf{n} \tag{4.4.2}$$

where $d \equiv d(x)$ is the distance (with sign) of x from $\mathcal{H}$ which can also be written as

$$d = \frac{\beta^\top x}{\|\beta\|}\,. \tag{4.4.3}$$

Moreover $g(x) = g(\hat{x}) + d\|\beta\|$ and, since $\hat{x} \in \mathcal{H}$ one has $g(x) = d\|\beta\|$ so the distance with sign of any $x \in \mathbb{R}^p$ from $\mathcal{H}$ is also given by

$$d(x) = \frac{g(x)}{\|\beta\|}\,. \tag{4.4.4}$$

In particular $b/\|\beta\|$ is the distance of the hyperplane from the origin.
A more compact description of the discriminant function is obtained pretending that the data are $p+1$-dimensional with a dummy 0-th component, say x_0, fixed equal to 1. Then, introducing an augmented weight vector $a := \begin{bmatrix} b & \beta_1 & \dots & \beta_p \end{bmatrix}^\top$ one can rewrite the discriminant function as

$$g(x) := a^\top x\,. \tag{4.4.5}$$

In this way $a^\top x = 0$ is a hyperplane through the origin in a $p+1$-dimensional space and the graph of the old discriminant function is the intersection of this plane with the shifted coordinate plane $\{x_0 = 1\}$. We shall follow this convention all through the rest of this section and *still keep the notation* x for the "augmented" vectors in $\mathbb{R}^{p+1}$.

Suppose now that we have a training set of N data $x(1), x(2), \dots, x(N)$ some of which are labeled $\mathcal{P}_1$ and the others labeled $\mathcal{P}_2$. We want to use these samples to estimate the discriminant function (4.4.5). A first naive deterministic solution could be to try to determine a exactly from the binary sequence of N given classifications $a^\top x(t),\ \ t = 1, 2, \dots, N$. To this end it will be convenient to replace all samples $x(t)$ which are classified $\mathcal{P}_2$ by their negatives $-x(t)$ so that whenever $\beta^\top x + b < 0$ we also have

$$a^\top x = \beta^\top(-x) + b(-x_0) = -g(x) > 0\,.$$

Note that the zeroth component of x has also been changed sign.

Remarks 4.1. *There is an alternative more precise notation which needs another symbol and will not be used in this section but will be convenient to use later on. Define the binary decision variable $y = y(x)$ as*

$$y = +1 \quad \textit{if } x \in \mathcal{P}_1, \quad \textit{and} \quad y = -1 \quad \textit{if } x \in \mathcal{P}_2\,. \tag{4.4.6}$$

Then $y(t)x(t)$ applies the change of sign discussed above. Note that it is based on the joint information: observed feature plus corresponding classification decision, which should actually form the content of the training set.

Now, the problem of finding a linear discriminant function for the N observed data can be solved simply by looking for a vector $a \in \mathbb{R}^{p+1}$ such that

$$a^\top x(t) > 0\,, \qquad t = 1, 2, \ldots, N\,. \tag{4.4.7}$$

This set of inequalities may have no solutions at all or multiple solutions in which case we shall say that the data set is *linearly separable*. A unique solution is exceptional. A vector $a \in \mathbb{R}^{p+1}$ will be said to *misclassify* a sample $x(t)$ if $a^\top x(t) < 0$. Clearly, if it exists, a solution of (4.4.7) cannot misclassify samples.

This naive idea can be refined in several ways, a natural one being to relax the solution of the linear inequalities(4.4.7) and transform it into an optimization problem. One should then define some optimization criterion. Probably one of the simplest is the so-called *Perceptron Criterion*:

$$J(a) := \sum_{t \in \mathbb{M}} (-a^\top x(t)) \tag{4.4.8}$$

where $\mathbb{M} = \mathbb{M}(a)$ is the *set indexing samples misclassified by* a. Since for a misclassified sample, $a^\top x(t) \leq 0$ this cost function is always nonnegative and can be zero only when a is a solution vector to (4.4.7). In fact, it is easy to check that $J(a)$ is proportional to the sum of the distances of misclassified samples to the decision boundary defined by a.

One can try to minimize the Perceptron criterion by an iterative optimization algorithm, the simplest being just the *steepest descent* algorithm which works as follows.
First compute the gradient with respect to the vector a:

$$\nabla J(a) = \sum_{t \in \mathbb{M}} (-x(t))$$

and then apply an additive update rule in the direction of the negative gradient

$$a(k+1) = a(k) + \alpha(k) \sum_{t \in \mathbb{M}} x(t) \tag{4.4.9}$$

where the sequence of stepsizes $\{\alpha(k)\}$ should be suitably chosen. The algorithm is called the *Batch Perceptron*; when it converges it can be shown to converge even if all $\alpha(k)$'s are chosen equal to a constant say 1.

Proposition 4.1. *If the training set is linearly separable, the Batch Perceptron algorithm for a suitable $\alpha(k) =$ constant converges to a solution vector of* (4.4.7).

Proof. For a formal proof we shall refer the reader to [75, p. 230]. Here we shall just provide an intuitive argument. Suppose $\alpha(k) = 1$; then if some $x(\bar{t})$ is misclassified by $a(k)$, in the next, $k+1$-th step, the inner product $a(k)^\top x(\bar{t})$, which is negative, is increased by a positive quantity since by (4.4.9),

$$a(k+1)^\top x(\bar{t}) = a(k)^\top x(\bar{t}) + \|x(\bar{t})\|^2 + B$$

where B is a sum including other misclassified samples, not depending on k. Hence $a(k)^\top x(\bar{t})$ has a positive increment so that $a(k+1)$ will tend to classify $x(\bar{t})$ correctly. That $\alpha(k)$ can be taken constant and equal to 1 is proven in [75]. □

The Perceptron algorithm has many variants which are described and discussed in great detail in the book [75]. One detail which seems to be important is that one should avoid converging to a limit point on the boundary of the feature space. This can be avoided by introducing a *margin* $b > 0$; that is requiring that $a^\top x(t) \geq b > 0$ for all $t = 1, 2, \ldots, N$. If this set of inequalities has a solution a then by (4.4.4) the corresponding separating hyperplane will be at a distance of at least $b/\|\beta\|$ from the observed feature points in $\mathbb{R}^p$.

One general limitation of the approach is that it only works if the training set is linearly separable. If it is not, the Perceptron algorithm will not converge. In this case, to stay within the linear theory, one should either accept some misclassification errors in the training set or abandon linearity and go to *nonlinear discrimination functions*.

In the first setting, one rough solution could be to estimate a separating hyperplane by using least squares. The general idea is that in the attempt of making all inner products $a^\top x(t)$ positive one may equivalently be trying to solve N simultaneous equations

$$a^\top x(t) = d\,, \qquad t = 1, 2, \ldots, N$$

for some arbitrary positive constant d (actually one might generally choose different $d = d(t) > 0$ but this is inessential).

Since $N > p$ there will be no exact solution and one should seek an approximate solution say in the Least Squares sense. Introducing the vector notations

$$X := \begin{bmatrix} x_0(1) & x_1(1) & x_2(1) & \ldots & x_p(1) \\ x_0(2) & x_1(2) & x_2(2) & \ldots & x_p(2) \\ & \ldots & \ldots & \ldots & \ldots \\ x_0(N) & x_1(N) & x_2(N) & \ldots & x_p(M) \end{bmatrix}, \qquad \mathbf{d} := \begin{bmatrix} d \\ d \\ \ldots \\ d \end{bmatrix}$$

the LS problem of finding an approximate weight vector $\hat{a}$ can be rewritten as

$$\hat{a} = \operatorname{Arg} \min_a \| \mathbf{d} - Xa \| \tag{4.4.10}$$

and, assuming $\operatorname{rank} X = p + 1$, one can use the standard formulas derived in Sect. 2.1 to obtain

$$\hat{a} = (X^\top X)^{-1} X^\top \mathbf{d}\,. \tag{4.4.11}$$

The LS estimated hyperplane $\hat{a}^\top y = d$ may in general fail to separate the two classes exactly but a wise choice of d may lead to satisfactory results.

Example 4.2. Suppose we are given a training set of four vectors in $\mathbb{R}^2$ classified as follows:

$$\begin{bmatrix} 1 \\ 2 \end{bmatrix} \text{ and } \begin{bmatrix} 2 \\ 0 \end{bmatrix} \in \mathcal{P}_1$$

and

$$\begin{bmatrix} 3 \\ 1 \end{bmatrix} \text{ and } \begin{bmatrix} 2 \\ 3 \end{bmatrix} \in \mathcal{P}_2\,.$$

The X matrix is then (recall that all training vectors are augmented by inserting a 1 on top and the vectors in $\mathcal{P}_2$ must be changed to their opposite),

$$X = \begin{bmatrix} 1 & 1 & 2 \\ 1 & 2 & 0 \\ -1 & -3 & -1 \\ -1 & -2 & -3 \end{bmatrix}$$

and by choosing $\mathbf{d} = \begin{bmatrix} 1 & 1 & 1 & 1 \end{bmatrix}^\top$ the Least Squares solution is found to be $\hat{a} = \begin{bmatrix} 11/3 & -4/3 & -2/3 \end{bmatrix}^\top$. Draw a picture of the hyperplane (actually a straight line) and check that it separates the two classes exactly. ♢

Whenever the two classes $\mathcal{P}_1$ and $\mathcal{P}_2$ are linearly separable there are in general *infinitely many* separating hyperplanes. One may attempt to find a unique optimal one in some reasonable sense. A preliminary normalization discussed in the following proposition will be instrumental to to this end.

Proposition 4.2. *Assume linear separability of the two classes and let a define a separating hyperplane. Then a can be normalized to obtain a parallel vector $\hat{a}$ such that $a^\top x(t) \geq 1$; for all $t = 1, \ldots, N$.*

Proof. Let $\{x(1), \ldots, x(N)\}$ be feature vectors in $\mathbb{R}^p$ belonging to one of two classes $\mathcal{P}_1$ and $\mathcal{P}_2$. If the two classes are linearly separable there is a minimal distance $\hat{d} = \hat{d}(a)$ of the N feature vectors from any separating hyperplane $a^\top x = 0$ so that, by (4.4.4)

$$\frac{a^\top x(t)}{\|\beta\|} \geq \hat{d}(a); \qquad t = 1, \ldots, N.$$

Now just recall that the hyperplane is defined by the homogeneous equation $a^\top x = 0$ and hence any vector proportional to a defines the same hyperplane. Therefore by picking

$$\hat{a} := \frac{a}{\|\beta\| \hat{d}(a)},$$

which is equivalent to normalize β so that $\|\beta\| \hat{d}(a) = 1$, or

$$\hat{d}(a) = 1/\|\beta\|, \tag{4.4.12}$$

linear separability (by the same hyperplane) can be written $\hat{a}^\top x(t) \geq 1$; for all $t = 1, \ldots, N$. □

In particular, feature points $x(t)$ which are at a minimal distance from the separating hyperplane will satisfy the equation $\hat{a}^\top x(t) = 1$.
In what follows we shall, without loss of generality, assume that all separating vectors a are normalized in this way.

4.5 ▪ Maximum Margin Hyperplanes and Support Vectors

We shall only provide a brief sketch of this technique. See [44, p. 132-33],[65, p. 503] and [100] for a more detailed discussion. For conciseness in this and

in the following sections we shall use a subscript to list the N features in the training set, e.g. use the symbol x_k to denote the p-dimensional vector $x(k)$. Since we will not have occasion of dealing with the components of $x(k)$, this should cause no confusion.

Assume the two classes $\mathcal{P}_1$ and $\mathcal{P}_2$ are linearly separable. We want to find an *optimal* separating hyperplane requiring that it should have **maximal distance from all feature points**. Recall that the signed distance of the feature x_k from a separating hyperplane described by the equation $\beta^\top x + b = 0$ can be expressed as

$$d(x_k) = \frac{\beta^\top x_k + b}{\|\beta\|}\,.$$

Let M be a positive number. Using the convention (4.4.6), if $y_k = 1$ the inequality

$$\frac{y_k(\beta^\top x_k + b)}{\|\beta\|} \geq M \tag{4.5.1}$$

means that x_k is at a (positive) distance greater or equal to M from the decision hyperplane, while if $y_k = -1$ it means that

$$d(x_k) = \frac{\beta^\top x_k + b}{\|\beta\|} \leq -M$$

and hence the the feature x_k is under the decision hyperplane at a distance grater or equal than M. Hence the inequality (4.5.1) covers both cases and *defines a slab centered on the decision hyperplane which contains no features*. The **Margin** M of a separating hyperplane $\mathcal{H}$ is, by definition, such that $|d(x_k)| \geq M$ for all x_k's, that is

$$|d(x_k)| = \frac{y_k(\beta^\top x_k + b)}{\|\beta\|} \geq M \quad \Leftrightarrow \quad y_k(\beta^\top x_k + b) \geq M\|\beta\|$$

Recall that the hyperplane is defined by an homogeneous equation and choosing β so that

$$\|\beta\|M = 1 \tag{4.5.2}$$

the inequality (4.5.1) defining a slab (of width $2M$) can be rewritten as

$$y_k(\beta^\top x_k + b) \geq 1\,. \tag{4.5.3}$$

Note that to define the same hyperplane the normalization, which depends on M, must of course modify also b.

Now we want to find a slab of maximal width, which means solving the problem of maximizing M subject to the conditions that all features lie outside the slab, that is, consider the

Problem 4.1 (Maximum margin hyperplane problem:). *Solve the optimization problem:*

$$\max_{\beta,b} \left\{ M\,;\, \text{subject to: } y_k(\beta^\top x_k + b) \geq 1, \quad k = 1, \ldots, N \right\} \tag{4.5.4}$$

Now, choosing β, b so that (6.6.12) is satisfied, i.e. $M = \dfrac{1}{\|\beta\|}$ the problem 4.5.4 is equivalent to **minimize** $\|\beta\|$ (equiv. its square norm) that is, solving:

$$\min_{\beta,b} \left\{ \frac{1}{2} \|\beta\|^2 \,;\, \text{subject to: } y_k(\beta^\top x_k + b) \geq 1, \quad k = 1, \ldots, N \right\} \tag{4.5.5}$$

(where the factor $\frac{1}{2}$ is added for convenience) which is a convex quadratic programming problem with linear inequality constraints. As such, since the feasible set is by assumption not empty, it must have a unique solution.

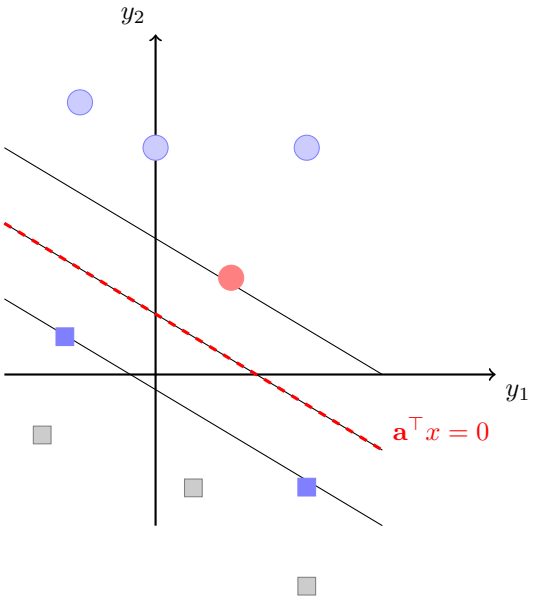


Figure 4.5.1. *Support Vectors*

Assuming the data are linearly separable, it is intuitive that the decision boundary hyperplane should depend only on a few data points, those which are *closest to the decision boundary*, which are called **support vectors**. The *margin of separation* d is the the distance between the hyperplane defined by the vector a and the closest data points (the support vectors).

To solve the optimization problem Introduce N Lagrange multipliers λ_k to form the Lagrangian function

$$L = \frac{1}{2} \|\beta\|^2 \; - \sum_{k=1}^{N} \lambda_k \; \left[y_k(\beta^\top x_k + b) - 1 \right] \; .$$

Setting the derivatives with respect to β and b to zero we obtain

$$\beta = \sum_{k=1}^{N} \lambda_k \; y_k x_k \,; \qquad 0 = \sum_{k=1}^{N} \lambda_k \; y_k \tag{4.5.6}$$

which need to be substituted in the Lagrangian to eliminate the primal variables β, b to obtain the *dual cost*

$$L_D = \sum_{k=1}^{N} \lambda_k \; - \frac{1}{2} \sum_{i=1}^{N} \sum_{k=1}^{N} \lambda_i \lambda_k \, y_i y_k x_i^\top x_k = \sum_{k=1}^{N} \lambda_k \; - \frac{1}{2} \sum_{i=1}^{N} \sum_{k=1}^{N} \lambda_i \lambda_k \, \langle y_i x_i, \, y_k x_k \rangle \tag{4.5.7}$$

which needs to be maximized with respect to the N multipliers λ_k, constrained to the positive orthant of $\mathbb{R}^N$ and by the second relation in (4.5.6). Since there is standard software to solve this problem called *maximum margin classifier*, we shall not dwell too much in the algorithmic aspects.

The optimal Lagrange multipliers λ_k^* must satisfy the Karush-Kuhn-Tucker conditions (see Appendix C), which include the equalities

$$\lambda_k^* \; \big[\, y_k(\beta^\top x_k + b) - 1 \big] \; = 0 \, , \qquad k = 1, \ldots, N \tag{4.5.8}$$

from which we can see that:
· if $\lambda_k^* > 0$ then $y_k(\beta^\top x_k + b) = 1$ which means that x_k must lie on the boundary of the optimal slab, that is be a **support vector**,
· when $y_k(\beta^\top x_k + b) > 1$ that is x_k is not on the boundary, then $\lambda_k^* = 0$. Therefore, denoting by SV the index set of the support vectors, the optimal solution can be written

$$\beta^* = \sum_{k \in SV} \lambda_k^* \; y_k x_k \, .$$

Therefore we have

Theorem 4.3. *The optimal (maximum margin) hyperplane in $\mathbb{R}^p$ is described by the affine function*

$$g^*(x) = \sum_{k \in SV} \lambda_k^* \; y_k \langle x_k, x \rangle + b^* \tag{4.5.9}$$

where the N-vector $\lambda := \begin{bmatrix} \lambda_1 & \ldots & \lambda_N \end{bmatrix}^\top$ is the unique solution of the optimization problem

$$\max_{\lambda_k \geq 0} \left\{ \sum_{k=1}^{N} \lambda_k \; - \frac{1}{2} \sum_{i=1}^{N} \sum_{k=1}^{N} \lambda_i \lambda_k \, \langle y_i x_i, \, y_k x_k \rangle \right\} \tag{4.5.10}$$

and b^* is determined by any support vector $\bar{x}_k$, by the constraint (4.5.8) *which implies $y_k(\beta^{*\top} \bar{x}_k + b) - 1 = 0$.*

For example, any support vector $\bar{x}_k$ such that, $y_k = 1$ yields

$$b^* = 1 - (\beta^*)^\top \bar{x}_k \, .$$

When the dimension of the feature space is large, one may expect a large number of nonzero components in the parameter a defining the separating boundary. Hence it makes sense to look for separating hyperplanes defined by the smallest number of parameters, which seems to imply the smallest number of support vectors.

In this section we have described an optimal linear classifier for linearly separable data. There is a far reaching generalization of this technique allowing to tackle nonlinear problems, which will be described in Section 7.9.

4.6 ▪ Problems

4-1 *(Computer simulation)*
Simulate two i.i.d. samples of sizes $N_1 = 500$ *and* $N_2 = 800$ *of Gaussian vectors having distribution* $p_1 = \mathcal{N}(\begin{bmatrix}1\\3\end{bmatrix}, \begin{bmatrix}100 & 24\\24 & 20\end{bmatrix})$ *and* $p_2 = \mathcal{N}(\begin{bmatrix}3\\-1\end{bmatrix}, \begin{bmatrix}40 & -10\\-10 & 80\end{bmatrix})$ *and find by MLR the equation defining the decision boundary. Plot your results using different colors for the two patterns.*
Generate now 50 more samples of data distributed according to p_1 *and 50 distributed according to* p_2*. Classify these samples using the above decision boundary and compute the percentage of misclassified data in the two cases. These are estimates of error probabilities. What relation do they have with* α *and* β *?*

4-2 *(Computer simulation)*
Same as above but imposing the same covariance estimate to the two populations, computed without distinguishing them. Classify the samples using two different centroids but using the new linear decision boundary.

4.7 ▪ Deciding the Complexity of a Linear Model

In many practical circumstances the number p, of parameters in the linear model $\mathbf{y} = S\theta + \sigma\mathbf{w}$ is not assigned a priori but needs rather to be chosen by the experimenter in order to satisfy certain requirements among which especially the predictive accuracy of the model. As a typical example, the order p of an autoregressive (AR) model to represent a time series $\{y(t)\,;\, t = 1, 2, \ldots, N\}$; i.e.

$$y(t) = \sum_{k=1}^{p} a_k\, y(t-k) + w(t)$$

is usually not known and needs to be estimated from the data together with the parameter vector $\theta := \begin{bmatrix} a_1 & a_2 & \ldots & a_p \end{bmatrix}^\top$. When there are enough data available, adding more lags $y(t-k)$ (i.e. increasing the order p) may look like a wise thing to do, since it will certainly correspond to a better fit of the training data set. More generally, adding more regressors, that is, increasing p in the linear model does certainly lead to a better description of the data in the sense that (with N fixed) the residual quadratic error

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

decreases as p increases and can eventually even become zero in the limit case of $p = N$, when one has as many parameters as data samples. It should however be clear that the reliability of the estimated model would seriously deteriorate when fitting models with an excessive number of parameters. Intuitively, one would not trust the prediction made with a model which fits nearly perfectly all the data points of the training set. This is called **overfitting** and is a symptom that one has been spending too much effort for modeling also the *noise* which is unavoidably superimposed to the "true" data.

Overfitting is actually a way of reducing the bias of the estimated model to the extreme. Since the nearly perfect fit obtained with very large values of p, is in general paid in terms of variance of the estimate, it will lead to models which are useless for the purpose of prediction. Here one should keep in mind that the model will have to be used with data which are *different from those forming the training set*.

In this section we shall first examine how the variance of the estimated parameters increases with increasing p in the context of classical Fisherian statistics. In this setting, the problem of choosing p can be formulated as an *hypothesis testing problem*: one should decide the complexity of the "true" model which has generated the data, based on a given fixed observed sample. We shall formulate this problem as a choice between two alternatives. First however we shall need to discuss how to update the estimates by adding regressors. This is called *Stagewise Linear Regression*.

4.8 ▪ Stagewise Linear Regression

Consider two linear Gaussian models in standard form:

$$\begin{aligned} M_1: \quad & \mathbf{y} = S_1\theta_1 + \boldsymbol{\epsilon} && \theta_1 \in \mathbb{R}^p \\ M_2: \quad & \mathbf{y} = [S_1\, S_2] \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} + \boldsymbol{\epsilon} && \theta_2 \in \mathbb{R}^k. \end{aligned} \tag{4.8.1}$$

where $\boldsymbol{\epsilon}$ is the random vector $\sigma\mathbf{w}$ which is assumed to be Gaussian, zero-mean and with variance matrix $\sigma^2 I_N$ (the $N \times N$ identity matrix). The formulas for non-standardized models can be obtained from those which will be derived below by just substituting $\mathbf{y}$ with $L^{-1}\mathbf{y}$ and S with $L^{-1}S$, where L is a left square root (say a Cholesky factor) of the variance matrix R of $\mathbf{w}$. The assumption of Gaussian distribution will be needed only later when formulating the choice of complexity as a hypothesis testing problem.

In the "simple" model M_1, we shall assume as usual that $\text{rank } S_1 = p$ while in the "complicated" model M_2, one may without loss of generality assume that the matrix $S_2 \in \mathbb{R}^{N\times k}$ is such that

$$\text{rank } [S_1\, S_2] = p + k \quad . \tag{4.8.2}$$

In case this does not happen we can eliminate the linearly dependent columns in S_2 and reparameterize the model by suitably redefining θ_2.

We shall now derive updating formulas for the parameter estimate and the relative error variance matrix of model M_2. These will be given in form of corrections to the corresponding estimates for the model M_1.

Denote by $\mathcal{S}$ the column space of the matrix $S := [S_1\, S_2]$ and let θ be the $p+k$-dimensional parameter $[\theta_1^\top\, \theta_2^\top]^\top$ in the enlarged model M_2 in (4.8.1). The Least Squares (Markov) estimate of θ and its Variance matrix are given by the well-known formulas

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

where now the matrix to be inverted has dimension $(p+k)\times(p+k)$. In force of (4.8.2) the subspace $\mathcal{S}$ can be decomposed as a direct sum

$$\text{span}\,[S] = \text{span}\,[S_1\, S_2] = \mathcal{S}_1 \oplus \mathcal{S}_2 = \text{span}\,[S_1] \oplus \text{span}\,[S_2] \tag{4.8.3}$$

and this decomposition can actually be rendered *orthogonal* by introducing the two orthogonal projectors

$$\begin{aligned} P_1 &: \mathbb{R}^N \to \mathcal{S}_1 \quad , \qquad P_1 = S_1 (S_1^\top S_1)^{-1}\, S_1^\top \quad , \\ P_1^\perp &: \mathbb{R}^N \to \mathcal{S}_1^\perp \quad , \qquad P_1^\perp = I - S_1 (S_1^\top S_1)^{-1}\, S_1^\top \quad . \end{aligned} \tag{4.8.4}$$

We shall denote by Q_1 the matrix $P_1^\perp = I - P_1$. Since $P_1 + Q_1 = I$, we have the orthogonal decomposition

$$S_2 = P_1 S_2 + Q_1 S_2\,.$$

Since the columns of $P_1 S_2$ belong by construction to $\mathcal{S}_1$, the last term in (4.8.3) can be replaced by $\text{span}\,[Q_1\, S_2]$. Therefore,

$$\text{span}\,[S] = \text{span}\,[S_1] \overset{\perp}{\oplus} \text{span}\,[Q_1\, S_2] \tag{4.8.5}$$

where the symbol $\overset{\perp}{\oplus}$ means *orthogonal* direct sum.
Let now $\hat{y}$ be the orthogonal projection of y onto the column space, $\mathcal{S}$, of the matrix S. Given that the columns of S_1 and S_2 are linearly independent, it must be possible to express $\hat{y}$ *uniquely* in the form

$$\hat{y} = S_1\, \hat{\theta}_1 + S_2\, \hat{\theta}_2 \quad , \tag{4.8.6}$$

so that $\hat{\theta}_1$ and $\hat{\theta}_2$ must obviously be the LS estimates of the parameters θ_1 and θ_2 of the augmented model M_2 with $p+k$ parameters. By the orthogonality principle of Least Squares it must be that $y - \hat{y} \perp \mathcal{S}$ and hence we must also have orthogonality separately to the two components, i.e.

$$y - \hat{y} \perp \mathcal{S}_1 \quad , \qquad y - \hat{y} \perp Q_1 \mathcal{S}_2 \quad ,$$

which can be rewritten

$$S_1^\top (y - S_1\hat{\theta}_1 - S_2\hat{\theta}_2) = 0 \quad , \tag{4.8.7}$$

$$S_2^\top Q_1 (y - S_1\hat{\theta}_1 - S_2\hat{\theta}_2) = 0 \quad . \tag{4.8.8}$$

These formulas yield

$$\hat{\theta}_1 = (S_1^\top S_1)^{-1}\, S_1^\top \left[y - S_2\hat{\theta}_2 \right] \quad , \tag{4.8.9}$$

$$\hat{\theta}_2 = (S_2^\top Q_1 S_2)^{-1}\, S_2^\top Q_1\, y \quad . \tag{4.8.10}$$

That $S_2^\top Q_1 S_2$ is invertible is readily seen by just remembering that Q_1 is a projection matrix. In fact, if for some $a \neq 0$ we could have that

$$0 = a^\top S_2^\top Q_1 S_2 a = a^\top S_2^\top Q_1^\top Q_1 S_2 a = \|Q_1 S_2 a\|^2$$

from the last equality it would follow that $S_2 a$ belongs to the nullspace of $Q_1 = P_1^\perp$. Since $\mathrm{Ker}\,(P_1^\perp) = \mathrm{Im}\,(P_1) = \mathcal{S}_1 = \mathrm{span}\,[S_1]$, it would happen that $S_2 a \in \mathrm{span}\,[S_1]$, which can happen only if $a = 0$, since the columns of S_1 and S_2 are independent.

Denoting by the symbol $\bar{\theta}_1$ the estimate of θ_1 obtained by describing the data with a p parameter model like M_1, equation (4.8.9) can be rewritten as

$$\hat{\theta}_1 = \bar{\theta}_1 - (S_1^\top S_1)^{-1}\, S_1^\top S_2 \hat{\theta}_2 \quad . \tag{4.8.11}$$

which is expressing the estimate of the first component θ_1 of the $p+k$-dimensional parameter θ of the model M_2, as the sum of $\bar{\theta}_1$ and a correction term due to the introduction of the new component θ_2.

Oblique Projections

In the decomposition (4.8.6) the two terms $S_1\,\hat{\theta}_1$ and $S_2\,\hat{\theta}_2$ have the meaning of *oblique projections* respectively, of y onto $\mathcal{S}_1$along $\mathcal{S}_2$ and of y onto S_2 along $\mathcal{S}_1$.

From (4.8.10), rcalling that $Q_1^\top Q_1 = Q_1^2 = Q_1$, one in particular sees that $\hat{\theta}_2$ can be obtained form the orthogonality relation

$$Q_1 y - Q_1 S_2 \theta_2 \perp Q_1 S_2$$

from which the oblique projection of y onto $\mathcal{S}_2$ along $\mathcal{S}_1$, can be computed by first projecting *orthogonally* the error vector $Q_1 y = y - P_1 y$ onto the subspace $(I - P_1)\mathcal{S}_2 = Q_1\mathcal{S}_2$ spanned by the columns $\tilde{s}_{2,i} := (I - P_1)\, s_{2,i}\,, i = 1, \ldots, k$ which can also be interpreted as estimation errors of estimates of the $s_{2,i}\,; i = 1, \ldots, k$, based on S_1 which can in turn be computed by solving an ordinary

least squares problem. The actual oblique projection of y is obtained by successively multiplying by S_2 the parameter $\hat{\theta}_2$ which is found by solving the above least squares problem [10].

As we shall see, in the Bayesian setting oblique projections correspond to conditional expectations.

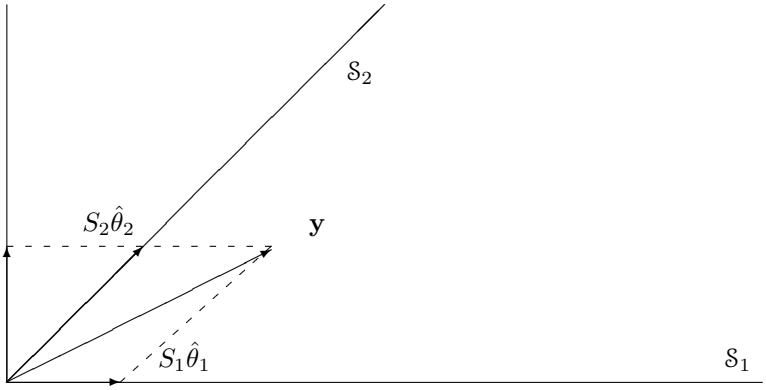


Figure 4.8.1. *Oblique Projection*

It follows that the *oblique projection operator* onto S_2 along $\mathcal{S}_1$ has the matrix representation

$$P_{2\|1} := S_2(S_2^\top Q_1 S_2)^{-1}\, S_2^\top Q_1\,. \tag{4.8.12}$$

One can in fact easily check that $P_{2\|1}^2 = P_{2\|1}$, while

$$P_{2\|1}^\top Q_1 = Q_1 P_{2\|1}\,.$$

which, in force of the fact that Q_1 is an orthogonal projector so that $Q_1 = Q_1^\top$, can be rewritten as $(Q_1 P_{2\|1})^\top = P_{2\|1}^\top Q_1^\top = Q_1 P_{2\|1}$. In other words, $Q_1 P_{2\|1}$ is symmetric idempotent and is therefore an orthogonal projection which must, by force, project onto $Q_1\mathcal{S}_2$, which is the orthogonal complement of $\mathcal{S}_1$ in $\mathcal{S}$. The following is an interpretation of the matrix $Q_1 P_{2\|1}$ which will be useful later on.

Proposition 4.3. *Let P be the orthogonal projection from $\mathbb{R}^N$ onto the subspace $\mathcal{S}$ and P_1 the orthogonal projection onto $\mathcal{S}_1 \subset \mathcal{S}$. Then $P - P_1$ is the orthogonal projection which projects onto the orthogonal complement $\mathcal{S} \cap \mathcal{S}_1^\perp$. It has the representation*

$$P - P_1 = Q_1 P_{2\|1} \tag{4.8.13}$$

where $P_{2\|1}$ is the oblique projector defined in (4.8.12)*.*

Proof. We only need to prove (4.8.13). Using (4.8.4) and (4.8.9) we obtain

$$\begin{aligned}\hat{y} = Py &= S_1(S_1^\top S_1)^{-1}\, S_1^\top\, y - S_1(S_1^\top S_1)^{-1}\, S_1^\top S_2\hat{\theta}_2(y) + S_2\hat{\theta}_2(y)\\ &= P_1 y + \left[I - S_1(S_1^\top S_1)^{-1}\, S_1^\top\right] S_2\hat{\theta}_2(y)\\ &= (P_1 + Q_1 P_{2\|1})\, y\end{aligned}$$

[10] We warn the reader that $S_2\,\hat{\theta}_2$ *cannot be the orthogonal projection of $Q_1 y = y - P_1 y$ onto $Q_1\mathcal{S}_2$*. In fact $Q_1\mathcal{S}_2$ is not even a subspace of $\mathcal{S}_2$.

which shows that indeed $P - P_1 = Q_1 P_{2\|1}$. The decomposition $P = P_1 + Q_1 P_{2\|1}$ is clearly orthogonal given that $P_1^\top (P - P_1) = P_1 Q_1 P_{2\|1} = 0$. Let us note in passing that an equivalent statement is $\mathcal{S} = P_1\mathcal{S} \oplus \mathcal{S} \cap \mathcal{S}_1^\perp$. □

Problem 4.2. Check that $P_{2\|1}$ is idempotent, its kernel is S_1 and its image is the columnspace of S_2.

One can give an analogous representation of the oblique projection of y onto $\mathcal{S}_1$ along $\mathcal{S}_2$ and prove an analogous decomposition to (4.8.6), like

$$y = P_{1\|2} y + P_{2\|1} y = S_1 (S_1^\top Q_2 S_1)^{-1}\ S_1^\top Q_2\, y + S_2 (S_2^\top Q_1 S_2)^{-1}\ S_2^\top Q_1\, y \quad (4.8.14)$$

where Q_2 has a dual meaning to Q_1. This may appear simpler than the orthogonal decomposition which we have just mentioned but is less useful since it is not orthogonal.

Comparing the Variances

In the following the subscripts $_{\theta_1}$ and $_\theta$ are to indicate the model M_1 or M_2, with respect to which the various expectations (and in particular variances) are computed.
Let us introduce the following notations:

$$\begin{aligned} \bar{\Sigma}_1 &:= [S_1^\top S_1]^{-1} \\ A_1 &:= [S_1^\top S_1]^{-1}\, S_1^\top \\ \Sigma_2 &:= [S_2^\top Q_1 S_2]^{-1} \quad ; \end{aligned}$$

wherby, $\bar{\theta}_1 = A_1 y$ and $\mathrm{Var}_{\theta_1}\ \bar{\theta}_1 = \sigma^2\, \bar{\Sigma}_1$.

Proposition 4.4. *Let $\hat{\theta}_1(\mathbf{y})$ and $\hat{\theta}_2(\mathbf{y})$ be the Markov estimators defined by formulas* (4.8.9) *and* (4.8.10)*. It holds that*

$$\mathrm{Var}_\theta \begin{bmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \end{bmatrix} = \sigma^2 \begin{bmatrix} \bar{\Sigma}_1 + A_1 S_2 \Sigma_2 S_2^\top A_1^\top & -A_1 S_2 \Sigma_2 \\ -\Sigma_2 S_2^\top A_1^\top & \Sigma_2 \end{bmatrix} \quad . \quad (4.8.15)$$

Proof. Let us start by showing that $\mathrm{Var}_\theta\,[\hat{\theta}_2] = \sigma^2\, \Sigma_2$. From (4.8.10) one has

$$\mathrm{Var}_\theta\,[\hat{\theta}_2] = \Sigma_2 S_2^\top Q_1\, \mathrm{Var}_\theta[\mathbf{y}] Q_1 S_2 \Sigma_2 = \sigma^2\, \Sigma_2 S_2^\top Q_1 S_2 \Sigma_2 = \sigma^2\, \Sigma_2 \quad ,$$

since $\mathrm{Var}_\theta\,[\mathbf{y}] = \sigma^2\, I$ and Q_1 is idempotent.

Let us now show that

Lemma 4.4. *The two estimators $\bar{\theta}_1(\mathbf{y})$ and $\hat{\theta}_2(\mathbf{y})$ are uncorrelated.*

In fact:

$$\mathrm{Cov}_\theta\ \left[\bar{\theta}_1(\mathbf{y}), \hat{\theta}_2(\mathbf{y})\right] = \bar{\Sigma}_1 S_1^\top\ \mathrm{Var}_\theta\,[\mathbf{y}]\ Q_1 S_2 \Sigma_2 = \sigma^2\, \bar{\Sigma}_1 S_1^\top Q_1 S_2 \Sigma_2 = 0 \quad ,$$

since $S_1^\top Q_1 = Q_1 S_1 = 0$.

Using now (4.8.11) we find

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

Finally we take care of $\mathrm{Var}_\theta[\hat{\theta}_1(\mathbf{y})]$. Since $\bar{\theta}_1(\mathbf{y})$ and $\hat{\theta}_2(\mathbf{y})$ are uncorrelated, one has

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

which concludes the proof of (4.8.15). □

Remark 4.1. Formula (4.8.15) describes the effect of increasing the number of parameters on the variance of the estimates. In particular the variance of the estimate $\bar{\theta}_1$ of θ_1 in a p parameter model is increased when adding more parameters. In fact is clear that Σ_1 is generally larger than the variance of $\bar{\theta}_1$, since

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

the additional term being in general at least positive semidefinite.

It should be realized however that the variance of the parameter estimates is not an objective criterion to be used for deciding on the complexity of a model. In general the purpose of the statistical exercise is not just estimating the model parameters but faithfully describing the output data, in particular for the purpose of prediction. One may as well change basis in the parameter space at wish. If it happens that in doing this the columns of S_2 become orthogonal to $\mathcal{S}_1$, that is

$$S_1^\top S_2 = 0 \qquad (\text{or } S_2^\top S_1 = 0)$$

the formulas simplify; in particular $Q_1 S_2 = S_2$ and the two estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ can be computed independently with the usual formulas

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

In particular one sees that $\hat{\theta}_1 = \bar{\theta}_1$ and therefore also $\Sigma_1 = \bar{\Sigma}_1$.

This disconcerting phenomenon can be explained whenever one is willing to accept the fact that the parameterization of a model $S\theta$ can be changed at will provided it provides the same description of the observed data y. One may for esample do a QR factorization of S and express it as the product of a $N \times (p+k)$ matrix with orthogonal columns times an upper triangular factor with a non-singular upper square sub matrix $R \in \mathbb{R}^{(p+k)\times(p+k)}$ which can be used to perform a basis change in the parameter space. Define a new parameter $\beta := R\theta$ and reparameterize the model in function of $\beta = \begin{bmatrix}\beta_1 & \beta_2\end{bmatrix}^\top$. Clearly in this new basis the two submatrices S_1 and S_2 have orthogonal columns and hence the variance of $\hat{\beta}_1$ cannot increase by adding k more (orthogonal) regressors in the model.

The moral of the story is that the variance of the parameter estimates is not invariant and depends on the basis chosen in the parameter space. Therefore comparison should be made among variables which are *invariant with respect to change of basis in the parameter space*. Quantities of this kind are for example the predictors and the residual prediction errors. □

Solution by the F test

A classical paradigm to approach the problem of choosing which of the two models (4.8.1) should be selected to describe the data is via the theory of hypothesis testing: Based on the observation $\mathbf{y} = y$ decide which of the two parametric models M_1 or M_2 has generated the data. Note that both hypotheses are composite.

We shall look for a decision rule based on the MLR test statistic. Some of the calculations will be similar to those done in Sect. 4.1.

We need to compare the quadratic residual errors using estimates computed according to a model of class M_1 (hypothesis H_0) or class M_2 (hypothesis H_1) defined in (4.8.1).

Let $\bar{\boldsymbol{\varepsilon}}(\mathbf{y}) := y - S_1\bar{\theta}_1(\mathbf{y}) = (I - P_1)\mathbf{y}$ be the residual error vector using the p parameter model M_1 and let $\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - S_1\hat{\theta}_1(\mathbf{y}) - S_2\hat{\theta}_2(\mathbf{y}) = (I - P)\mathbf{y}$ be the residual error vector of the augmented model M_2. Recall that $(I - P)$ and $(P - P_1)$ project onto orthogonal subspaces (Proposition 4.3); in fact, $I - P$ projects onto the complement $\mathcal{S}^\perp$ while $(P - P_1)$ projects onto the subspace $\mathcal{S} \cap \mathcal{S}_1^\perp$. We can therefore write,

$$\begin{aligned} R_0(\mathbf{y})^2 = \|\bar{\boldsymbol{\varepsilon}}\|^2 &= \| \left[(I - P) + (P - P_1)\right]\mathbf{y}\|^2 = \|\hat{\boldsymbol{\varepsilon}}\|^2 + \|(P - P_1)\mathbf{y}\|^2 \\ &= \|\hat{\boldsymbol{\varepsilon}}\|^2 + \|Q_1 P_{2\|1}\mathbf{y}\|^2 = R_1(\mathbf{y})^2 + \|Q_1 P_{2\|1}\mathbf{y}\|^2 . \end{aligned} \tag{4.8.16}$$

Using (4.8.12), the last term $\|Q_1 P_{2\|1}\mathbf{y}\|^2$ can be expressed in function of the estimator $\hat{\theta}_2(\mathbf{y})$ as

$$\|Q_1 P_{2\|1}\mathbf{y}\|^2 = \hat{\theta}_2(\mathbf{y})^\top \Sigma_2^{-1}\, \hat{\theta}_2(\mathbf{y})$$

so $R_0(\mathbf{y})^2$ can be written in the form

$$R_0(\mathbf{y})^2 = R_1(\mathbf{y})^2 + \hat{\theta}_2(\mathbf{y})^\top \Sigma_2^{-1}\, \hat{\theta}_2(\mathbf{y}) = R_1(\mathbf{y})^2 + \|\hat{\theta}_2(\mathbf{y})\|^2_{\Sigma_2^{-1}} \tag{4.8.17}$$

where $\sigma^2\,\Sigma_2$ is the variance of the estimator $\hat{\theta}_2(\mathbf{y})$.

How large is the difference term $\|\hat{\theta}_2(\mathbf{y})\|^2_{\Sigma_2^{-1}}$ in this expression clearly depends on which model has generated the data. In case the true model generating the data was M_1 the additional regressors $S_2\theta_2$ and the associated estimator $\hat{\theta}_2(\mathbf{y})$ would only describe the superimposed noise. One may guess that in this circumstance $S_2\,\hat{\theta}_2$ and, in fact, the term $\|\hat{\theta}_2(\mathbf{y})\|^2_{\Sigma_2^{-1}}$ would result small when compared to $R_1(\mathbf{y})^2$.

Theorem 4.4. *Under H_0, the two terms on the right side of* (4.8.17) *are independent and are both χ^2-distributed. Respectively, we have,*

$$\frac{\|\hat{\boldsymbol{\varepsilon}}\|^2}{\sigma^2} \sim \chi^2(N - p - k) \tag{4.8.18}$$

and

$$\frac{\hat{\theta}_2(\mathbf{y})^\top \Sigma_2^{-1}\, \hat{\theta}_2(\mathbf{y})}{\sigma^2} \sim \chi^2(k)\,. \tag{4.8.19}$$

Therefore the ratio

$$\varphi(\mathbf{y}) := \frac{N - p - k}{k}\, \frac{\|\hat{\theta}_2(\mathbf{y})\|^2_{\Sigma_2^{-1}}}{R_1(\mathbf{y})^2} \tag{4.8.20}$$

has pdf $\mathcal{F}(k,\ N-p-k)$.

Proof. Assume that the true model is M_1 (hypothesis H_0). The sum of squared residual errors, $\|\bar{\varepsilon}\|^2 = \|(I-P_1)\mathbf{y}\|^2$ is equal to $\|Q_1\boldsymbol{\epsilon}\|^2$. Substitute in the last term of (4.8.16) the M_1 model and use the equality $P_{2\|1}^\top Q_1 = Q_1^\top P_{2\|1} = Q_1 P_{2\|1}$, to find that $Q_1 P_{2\|1}\mathbf{y} = P_{2\|1}^\top Q_1(S_1\theta_1+\boldsymbol{\epsilon}) = P_{2\|1}^\top Q_1\boldsymbol{\epsilon}$.
On the other hand, $\hat{\varepsilon} = (I-P)\boldsymbol{\epsilon}$ and hence $\boldsymbol{\epsilon}^\top(I-P)^\top Q_1 P_{2\|1}\boldsymbol{\epsilon} = 0$, since $(I-P)$ projects onto the orthogonal complement of $\mathcal{S}$. This fact implies the incorrelation of the random vectors $\hat{\varepsilon}$ and $Q_1 P_{2\|1}\boldsymbol{\epsilon}$ which by Gaussianity implies the independence of the two terms in the right member of (4.8.16).

Finally, by Proposition A.7), $\frac{1}{\sigma^2}\|\hat{\varepsilon}\|^2 \sim \chi^2(N-(p+k))$, when the augmented model is generating the data, and under H_0, by Proposition A.6), $\frac{1}{\sigma^2}\hat{\theta}_2(\mathbf{y})^\top\Sigma_2^{-1}\,\hat{\theta}_2(\mathbf{y}) \sim \chi^2(k)$ since the mean value of $\hat{\theta}_2(\mathbf{y})$ is zero. In fact one has

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

where the two arguments in the function $\mathcal{F}$ denote the degrees of freedom of the F-distribution. □

An alternative proof can be found in [90, p. 73].

Problem 4.3. *Use the independence of the two terms in* (4.8.17) *to show, in a different way, that under* H_0,

$$\frac{1}{\sigma^2}\|\hat{\varepsilon}\|^2 \sim \chi^2(N-(p+k))\,.$$

Solution: It is clear that under H_0, $\frac{1}{\sigma^2}\|\bar{\varepsilon}\|^2 \sim \chi^2(N-p)$. As we have seen, under the same hypothesis $\frac{1}{\sigma^2}\hat{\theta}_2(\mathbf{y})^\top\Sigma_2^{-1}\,\hat{\theta}_2(\mathbf{y}) \sim \chi^2(k)$. Since the two random variables are independent, $\frac{1}{\sigma^2}\|\hat{\varepsilon}\|^2$ must necessarily have a χ^2 distribution (TheoremA.2) and the number of degrees of freedom must be $N-p-k$. ♢

Normally $N-p$ is much greater than k and the F distribution can be approximated by a χ^2 with k degrees of freedom, in the sense that for $N\to\infty$, we have the symbolic equality

$$k\,\varphi(\mathbf{y}) \ \sim\ \chi^2(k) \qquad \text{under } M_1 \tag{4.8.21}$$

Then, once fixed the probability α of deciding M_2 when M_1 is the true model, one reads in the tables of the F distribution the abscissa x_α for which

$$\mathbb{P}_{\chi^2(k)}\{k\,\varphi(\mathbf{y}) > x_\alpha\} = \alpha$$

and whenever $\varphi(y) > x_\alpha$ rejects the hypothesis that the data are generated by the "simple" model M_1, with probability α of committing an error. If the true model is M_2, the pdf of the statistic (4.8.20) is complicated. In practice the probability of an error of second kind

$$\beta := \mathbb{P}\{\text{ decide } M_1 \text{ when } \ M_2 \text{ is true }\}$$

can be estimated by Monte Carlo simulations or by the approximation of the non-central F distribution described at the end of the ANOVA Section 4.1 involving formula (4.1.35).

Finally, we should mention one aspect of stagewise regression, called in the literature *regressors collinearity* [92, 62] which needs to be considered when increasing the number of regressors. This phenomenon is related to the ill-conditioning of the underlying LS problem which could seriously hamper the reliability of the solution and is not immediately visible from the formulas given above. We shall discuss it at the end of this section.

An algorithm for Stagewise Linear Regression

The updating formulas (4.8.11) suggest a "stagewise" algorithm (someimes called "stepwise least squares") based on the *sequential introduction* (one after the other) of the columns of S. The estimate based on a model with p parameters is updated at each step by adding to the model one new parameter θ_{p+1} and, of course, a new regression variable involving only one new column s_{p+1}. It is thereby possible to obtain a recursive algorithm which sequentially updates (in fact enlarges) the estimate which allows to monitor the behaviour of the estimated model when gradually increasing itscomplexity.

Suppose one has availble the estimator $\theta^k = [\bar{\theta}_1, \ldots, \bar{\theta}_k]^\top$ obtained by fitting the data with a k parameter model:

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

where as before we use the shorthand $\varepsilon = \sigma w$. Introduce a new linearly independent column s_{k+1}, in S so that the model has now $k+1$ parameters,

$$y = S_{k+1}\,\theta + \varepsilon \quad , \tag{4.8.22}$$

where

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

Using the previous updating formula for the estimator one finds

$$\hat{\theta}_{k+1} = \frac{1}{s_{k+1}^\top\, Q_k\, s_{k+1}}\; s_{k+1}^\top Q_k y = \frac{1}{s_{k+1}^\top\, Q_k\, s_{k+1}}\; s_{k+1}^\top [y - S_k^\top \bar{\theta}^k] \tag{4.8.23}$$

$$\hat{\theta}^k = (S_k^\top S_k)^{-1}\, S_k^\top [y - s_{k+1}\hat{\theta}_{k+1}] = \bar{\theta}^k - (S_k^\top S_k)^{-1}\, S_k^\top s_{k+1}\hat{\theta}_{k+1} \quad , \tag{4.8.24}$$

where $\hat{\theta}_{k+1}$ and $\hat{\theta}^k$ represent the estimators of θ_{k+1} and $[\theta_1, \ldots, \theta_k]^\top$ relative to the augmented model (4.8.22) and Q_k is the orthogonal projector onto the orthogonal complement of the columnspace of S_k, that is

$$Q_k = I - S_k (S_k^\top S_k)^{-1}\, S_k^\top \quad . \tag{4.8.25}$$

At the next step one adds column s_{k+2} to the model (4.8.22)), and the updates the estimate as

$$\bar{\theta}^{k+1} := \begin{bmatrix} \hat{\theta}^k \\ \hat{\theta}_{k+1} \end{bmatrix} \quad , \tag{4.8.26}$$

by formulas analog to (4.8.23)–(4.8.24). The non trivial part of the algorithm is to find updating formulas for the coefficients, in particular computing the inverse

$(S_{k+1}^\top S_{k+1})^{-1}$ starting from $(S_k^\top S_k)^{-1}$. One may think of using a recursive updating of the Cholesky factorization of $S_k^\top S_k$, but a moment of thought shows that this would just be a way to do a QR factorization of the matrix S_k as the product of an orthogonal Q (which is not explicitly saved) and an upper triangular factor (which is in fact the *right Cholesky factor* of $S_k^\top S_k$). In this way one is actually updating a QR factorization of S. The *Golub-Styan Stagewise Algorithm* [41] uses an updating procedure of Householder matrices to do this. We shall leave the details to the interested reader.

Remarks 4.2. There is an important aspect of linear regression which especially appears when studying the stagewise procedure, which is called *regressor collinearity* [92] which needs to be considered quite carefully whenever one has to decide the complexity of a linear model. As we have seen, the variance of the parameter estimates is not a good criterion for this choice and one should either compare the residual error variances when increasing the model complexity or change the model parametrization in such a way that the new added columns are always orthogonal. To monitor the occurrence of this problem one should compare the condition numbers of the successive stages; in particular compare the p singular values of the original S_1, say $\bar{\sigma}_i$ with the first p singular values $\sigma_i\,;\, i = 1, \ldots, p$, of the augmented matrix. See the remark 4.2.

To analyze this effect correctly one should do a sequential updating of the Singular Value Decomposition of the matrix S by adding one column at a time and then comparing the p original singular values $\bar{\sigma}_i$ of $\bar{S}_1$ to the new $\hat{\sigma}_i\,;\, i = 1, \ldots, p+1$, of the augmented matrix. See [62] and the last subsection of Sect. 3.1 for the SVD analysis of the Least Squares problem.

Algortihms for the sequential updating of the SVD are described in [12, 13].

4.9 ▪ The FPE criterion and Cross Validation

As is evident from the initial paragraph of the previous section, the model complexity estimation problem may be framed in a different setting than the Fisherian paradigm, that is, without necessarily postulating the existence of a *true model* with a *true dimension*, which has generated the data. In this setting one may see the class of different models like (4.8.1) just as a class of simple linear approximations of an unknown "true" data generation mechanism which may actually be non-linear or infinite dimensional. One then chooses a model in this model class according to some optimality criterion. The point of view that we have enforced so far is to selected a model which best explains the available data $\mathbf{y}$ (or, said otherwise, the model which best explains the training set of data). Actually, since the ultimate scope of model buildng is to build *predictors of future data*, that is, data which have not yet been observed, a more reasonable criterion to use should be to choose the model which provides the *best prediction of future data*. Naturally here one should have some a priori information guaranteeing that future data will still be produced by some model belonging to the class. A crucial ingredient of this philosophy is that any data-based model used for prediction necessarily uses parameter estimates which are **random**. Hence the prediction error incurred by a model estimated from the data will *depend on the statistical uncertainty of the estimated parameters*. Hence the optimization criterion should take into account both the randomness inherent in the stochastic character of future data *and* the randomness of the estimated model used to construct the predictor. This point of view leads to the modern solutions of the model order estimation problem. Here we shall discuss a simple case.

Assume then that we have partitioned our data in two subvectors say $\mathbf{y} := [\mathbf{y}_1^\top \ \mathbf{y}_2^\top]^\top$ which for simplicity we shall assume of equal length N. The two vectors may represent observations made sequentially (that is the data $\mathbf{y}_2$ are collected after $\mathbf{y}_1$) and let us imagine to use the first N data $\mathbf{y}_1$ to estimate a standard linear model of dimension p. This means that we described the data $\mathbf{y}_1$ by estimating the linear model

$$\mathbf{y}_1 = S\theta + \boldsymbol{\epsilon}_1 \,, \qquad \mathrm{Var}\,[\boldsymbol{\epsilon}_1] = \sigma^2 I_N \tag{4.9.1}$$

obtaining the classical parameter estimator

$$\hat{\theta}(\mathbf{y}_1) = [S^\top S]^{-1} S^\top \mathbf{y}_1 \,.$$

We now want to evaluate how well the estimated model, $S\hat{\theta}(\mathbf{y}_1)$ can predict the future data $\mathbf{y}_2$ which have saved. Of course, for such operation to be logically consistent we must assume that the next N samples were generated by the same mechanism which did produce the past data $\mathbf{y}_1$. This can be formalized by saying that the pdf or, at least the first and second order moments of $\mathbf{y}_1$ and $\mathbf{y}_2$ are the same. In particular, we assume that the two components of the overall vector $[\mathbf{y}_1^\top \ \mathbf{y}_2^\top]^\top$ have the same mean vector μ (which could be anything) and that the overall variance of $\mathbf{y}$ is $\sigma^2 I_{2N}$. In this way $\mathbf{y}_1$ and $\mathbf{y}_2$ result to be *uncorrelated*.

Let's then consider the so-called **final prediction error** of the future data given the past,

$$\boldsymbol{\varepsilon} := \mathbf{y}_2 - S\hat{\theta}(\mathbf{y}_1) \tag{4.9.2}$$

which has mean $\mu - S[S^\top S]^{-1}S^\top \mu$. Subtracting the mean and computing the variance of $\boldsymbol{\varepsilon}$ one finds

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

As a scalar measure of the final prediction error let us take the normalized scalar variance of $\boldsymbol{\varepsilon}$, which has the expression

$$\begin{aligned}\frac{1}{N}\,\mathrm{var}\,[\boldsymbol{\varepsilon}] &= \sigma^2\frac{1}{N}\,\mathrm{Tr}\,\left\{I_N + S[S^\top S]^{-1}S^\top\right\} = \sigma^2\left\{1 + \frac{1}{N}\,\mathrm{Tr}\,([S^\top S]^{-1}S^\top S)\right\} \\ &= \sigma^2(1+\frac{p}{N})\,. \end{aligned} \tag{4.9.3}$$

From this expression one sees that the normalized scalar variance of $\boldsymbol{\varepsilon}$ **depends linearly on the model dimension** p.

However, to be coherent, in this formula we should actually substitute the theoretical variance σ^2, which is an unknown parameter, with an estimate naturally also based on a model with p parameters. We shall use the unbiased estimator of (2.3.24)

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

where $\hat{\boldsymbol{\epsilon}}_p$ is the residual error of the p parameter model. Substituting in (4.9.3) one arrives at the following expression for the final prediction error variance [NOTATION]

$$FPE(p) := \frac{1}{N}\|\hat{\boldsymbol{\epsilon}}_p\|^2\frac{(1+\frac{p}{N})}{(1-\frac{p}{N})} := \hat\sigma_p^2\frac{(1+\frac{p}{N})}{(1-\frac{p}{N})}\,. \tag{4.9.4}$$

This expression, also called **final prediction error** for short, can be used as a criterion for model order selection. One needs to compute the FPE for a class of (finite number of) models of different dimension p and then selects the model with the minimum FPE. The computations can be organized efficiently proceeding with increasing dimension and using a sequential stagewise least square algorithm of the type seen in the previous section.

There are more general criteria for model selection based on the same general principle but not necessarily relying on a past-future splitting of the data sequence. A rather extreme case is the **leave one out** cross-validation algorithm where one does model fitting using a data string of all measurements except one, that is leaving out just one element of the sequence and then computing a global error variance averaging on all possible model errors estimated by leaving out one datum. The method is discussed very clearly in Sect. 4.2 of Wahba's book [101] for general regularized least squares problems.

Chapter 5

BAYESIAN STATISTICS

5.1 ▪ Introduction to Bayesian estimation

In this chapter we shall address the *Bayesian approach* to statistical estimation. This approach, unlike the Classical Fisherian (or Frequentist) approach, refers to situations where there is an *a priori* information of probabilistic nature about the variable to be estimated. It starts from the assumption that θ should not be regarded as a "certain but unknown " parameter which can only be described by some, deterministic but unknown, *true value*, but is rather a *random variable* which, by its very nature, cannot be assigned an exact numerical value. There are instead (in general infinitely many) possible values of x, described as determinations of a random variable (or of a finite-dimensional random vector[11]) $\mathbf{x}$ which is distributed on $\mathbb{R}$ or on $\mathbb{R}^n$, according to some probability law.

From a practical point of view one may say that very often we have some a priori information available on x which is sufficient to justify the adoption of the Bayesian approach. For example, there is in general a known "nominal value" of x and its dispersion around the nominal value (say the " tolerance " or " precision class " etc.) may also be known. This information can often be translated into probabilistic parameters like the mean and variance of a probability distribution and sometimes even into a possible probability distribution. Beyond any ideological assumption about the nature of x (should it be regarded as a certain but unknown quantity, or as a random one) which may seem more or less plausible under the circumstances, the real motivation for the use of the Bayesian approach lies in the availability of a priori probabilistic information about x. In many modern applications the data acquisition systems are automatic and work at speeds and with storage capacity which allow to collect an overwhelming amount of data. Although x is by its very nature not directly measurable, it may be possible in some circumstances to use these data to *estimate a probabilistic description* of the unknown variable and this may make a Bayesian approach a natural and convenient choice [28]. It would be a mistake to disregard it.

We shall now proceed to describe formally what is meant by a Bayesian es-

[11] To adhere to standard conventions the dimension of this random vector will now be n instead of p.

timation problem. Recall that in this book we need to distinguish carefully random variables from their sample values. For random quantities we use boldface characters while for their sample values normal body typefaces. Normally we shall assume that all random variables involved are of purely continuous type and can be described by a probability density function.

Let $\mathbf{x}$ be an n-dimensional random vector whose probability distribution for the moment we assume completely known; $\mathbf{x}$ is not directly accessible to observation, which means that we can not observe the sample values x of $\mathbf{x}$. Denote

$$p_{\mathbf{x}}(x) = p_{\mathbf{x}}(x_1, \ldots, x_n) \tag{5.1.1}$$

the probability density of $\mathbf{x}$ which is also called the *a priori* probability density. Let $\mathbf{y}$ be the m-dimensional random vector representing the observations. We assume that the conditional density $p_{\mathbf{y}|\mathbf{x}}(y \mid \mathbf{x} = x)$ of $\mathbf{y}$ given a sample value x taken by $\mathbf{x}$, is a known data of our problem and denote its values by the symbol $f(y \mid x)$, say

$$f(y \mid x) = P(y \leq \mathbf{y} \leq y + dy \mid \mathbf{x} = x)/dy\,. \tag{5.1.2}$$

This function can be regarded as the mathematical description of the measuring instrument or of the transmission channel.

The *Bayesian estimation problem* is to reconstruct the random vector $\mathbf{x}$ from the probabilistic model (5.1.1) and (5.1.2) and from the observation $\mathbf{y} = y$ of the measurement device. This problem formulation can actually be understood in two ways: either as the problem to calculate the new probability distribution of $\mathbf{x}$ determined by the observation of the sample value $\mathbf{y} = y$, or as the problem of reconstruction of the *sample value* $x = \mathbf{x}(\omega)$ which was determined by the experimental condition ω at the time when the measuring experiment was performed.

If $p_{\mathbf{x}}$ is known completely, the first problem has an obvious solution dictated by Bayes rule. In fact, from the functions (5.1.1) and (5.1.2) one can get the conditional density of $\mathbf{x}$ given the observation $\mathbf{y} = y$ as,

$$p(\mathbf{x} = x \mid \mathbf{y} = y) = \frac{f(y \mid x)\; p_{\mathbf{x}}(x)}{\int_{\mathbb{R}^n} f(y \mid x)\; p_{\mathbf{x}}(x)\; dx} = \frac{p_{\mathbf{y},\mathbf{x}}(y, x)}{p_{\mathbf{y}}(y)} \tag{5.1.3}$$

and this formula shows how the observation $\mathbf{y} = y$ improves our a priori knowledge of $\mathbf{x}$, described by $p_{\mathbf{x}}$. The function (5.1.3) is commonly called the *a posteriori* probability density of $\mathbf{x}$.

However Bayes formula (5.1.3) does not solve the problem of reconstructing the sample value x which (in a probabilistic sense) determined the specific observed sample value y of the observation. In practice one often has only access to one measurement sample and needs a *point estimate* of x which should ideally be the sample value of $\mathbf{x}$ which has been giving rise to the observation $\mathbf{y} = y$.

The problem of point estimation can be better visualized when the coupling mechanism between the measured variable $\mathbf{y}$ and the unaccessible $\mathbf{x}$ is described by a statistical model; which now we write

$$\mathbf{y} = h(\mathbf{x}, \mathbf{w}) \tag{5.1.4}$$

where the parameter θ in (2.2.1) is substituted by a random vector $\mathbf{x}$. The measurement process is affected by *measurement noise* described by the random vector $\mathbf{w}$, which represents the interaction of the physical environment with the

measuring device. The random noise vector $\mathbf{w}$ is causing uncertainty in the measurement and for $\mathbf{w} = 0$ (5.1.4) the measurement becomes a certain and predictable function of $\mathbf{x}$. Note that whenever the noise distribution and a prior density for $\mathbf{x}$ are given one can in principle determine from the relation (5.1.4) the conditional density $f(y \mid x)$ and the posterior (5.1.2) by the well-known rules of Probability Theory.

The effect of the random experimental conditions ω at the time of performing the experiment is thereby condensed into a sample value of the noise $\mathbf{w}(\omega) = w$ which makes the observation y depend on the value $x = \mathbf{x}(\omega)$ as prescribed by the model (5.1.4), namely

$$y = h(x, w) \quad , \tag{5.1.5}$$

In this scheme, the problem of Bayesian point estimation appears as that of solving the equation (5.1.5) for x in terms of a collection of observation values $\mathbf{y} = y$. This is clearly an impossible task since in virtually every situation of practical interest w is inherently impossible to be known and so x can never be recovered exactly from the model (5.1.5). One can see that point estimation should be formulated as an *approximation problem*.

Naturally all approximation problems require the choice or the definition of a criterion function establishing how good the approximation is. We shall denote by $\xi := z - x$ the approximation error incurred by approximating x by z, both variables ranging in $\mathbb{R}^n$. A reasonable class of approximation criteria is defined below.

Definition 5.1. *A* cost (or loss) function *for the Bayesian point estimation problem is any scalar function $c : \mathbb{R}^n \to R$ of the variable $\xi \in \mathbb{R}^n$, which is* strictly convex, non-negative and zero at the origin. *A cost function is* symmetric, *if $c(-\xi) = c(\xi)$.*

A simple symmetric cost function is

$$c(z - x) = \|z - x\|_Q^2 \quad , \tag{5.1.6}$$

where $\|x\|_Q^2 := x^\top Q x$ and Q is a symmetric positive definite matrix. Note in fact that if $\|\xi_1\| \geq \|\xi_2\|$, then $c(\xi_1) \geq c(\xi_2)$.

Although x is unknown the probabilistic information in this problem tells us that certain values of x are more likely than others; better, that the observation process makes certain values of x more probable than others as described by the a posteriori density function $f(x \mid y)$. It is then natural to introduce the *conditional expected risk*,

$$R(z, y) := \mathbb{E}\left[c(z - \mathbf{x}) \mid \mathbf{y} = y\right] = \int_{\mathbb{R}^n} c(z - x)\, f(x \mid y)\, dx \quad , \tag{5.1.7}$$

and for a given observed y, define the *Bayesian point estimate* of x corresponding to the observation y, the vector $z = \hat{x}$ which minimizes, with respect to z, the expected risk $R(z, y)$,

$$\hat{x} = \text{Arg} \min_z R(z, y) \tag{5.1.8}$$

the existence and uniqueness of the minimum being guaranteed by the strict convexity of the function c. Of course $\hat{x}$ depends, besides the observation y, on

the choice of the cost function c. This dependence is however rather mild at least for a large class of estimation problems.

Theorem 5.1. *If the cost function c is symmetric, strictly convex and the posterior density $f(\cdot \mid y)$ is unimodal and symmetric about its mode $\mu(y)$ (so that the mode coincides with the conditional mean), then the point estimate $\hat{x}$ defined by (5.1.8) is the conditional mean of $\mathbf{x}$ given $\mathbf{y} = y$,*

$$\mu(y) = \mathbb{E}\,(\mathbf{x} \mid \mathbf{y} = y). \tag{5.1.9}$$

and does not depend on the cost function.

Proof. Assume initially that $\mu(y) = \mathbb{E}\,(\mathbf{x} \mid \mathbf{y} = y) = 0$, so the symmetry of f is written as $f(x \mid y) = f(-x \mid y)$. From this it follows that

$$R(z, y) = \mathbb{E}\,[c(z - \mathbf{x}) \mid \mathbf{y} = y] = \mathbb{E}\,[c(z + \mathbf{x}) \mid \mathbf{y} = y] = \mathbb{E}\,[c(\mathbf{x} - z) \mid \mathbf{y} = y]$$

the last equality following from the symmetry of c. Therefore, by strict convexity it also holds that

$$R(z, y) = \mathbb{E}\left[\frac{c(\mathbf{x} - z) + c(\mathbf{x} + z)}{2} \mid \mathbf{y} = y\right] > \mathbb{E}\,[c(\mathbf{x}) \mid \mathbf{y} = y] = R(0, y), \quad z \neq 0$$

which implies , $\min_z\, R(z, y) = R(0, y)$ and $Arg\, \min_z\, R(z, y) = 0 = \mu(y)$.

If $\mu(y) \neq 0$, set $\Delta\mathbf{x} := \mathbf{x} - \mu(y),\ \Delta z := z - \mu(y)$ so that $\mathbb{E}\,[\Delta\mathbf{x} \mid \mathbf{y} = y] = 0$ and $\Delta z - \Delta\mathbf{x} = z - \mathbf{x}$ which implies

$$R(z, y) = E\,[c(\Delta z - \Delta\mathbf{x}) \mid \mathbf{y} = y] > E\,[c(\Delta\mathbf{x}) \mid \mathbf{y} = y] = R(\mu(y), y), \quad z \neq \mu(y)$$

that is $\mu(y) = \text{Arg}\, \min_z\, R(z, y)$. □

The result may be different for non-symmetric probability distributions.

Proposition 5.1. *Let $\mathbf{x}$ be a scalar random variable and let $c(z) = |z|$, then*

$$\text{Arg}\, \min_z\, \mathbb{E}\,[\,|z - \mathbf{x}| \mid \mathbf{y} = y]$$

is the **conditional median** *of the a posteriori distribution given $\mathbf{y} = y$.*

Proof. First let us do the minimization of the unconditional expectation. We shall assume that $\mathbf{x}$ has a density $p(x)$; then

$$\begin{aligned}\mathbb{E}\,|\,\mathbf{x} - z\,| &= \int_z^{+\infty} (x - z)\, p(x)\, dx + \int_{-\infty}^{z} (z - x)\, p(x)\, dx = \\ &= \int_z^{+\infty} x\, p(x)\, dx - z(1 - F(z)) + zF(z) - \int_{-\infty}^{z} x\, p(x)\, dx\end{aligned}$$

computing the derivative with respect to z and setting it equal to zero, we obtain

$$F(z) = 1/2$$

that is $F(z) = 1 - F(z) = 1/2$, which is the definition of the median. The same argument works unchanged for the conditional density (or distribution). □

One can show that for symmetric unimodal probability distributions, in particular if the distributions are Gaussian, Theorem 5.1 actually holds also for non-symmetric convex cost functions and the minimum conditional risk estimator is still the conditional mean. See the article by Sherman [84]. On the other hand, as we shall see below, for a *quadratic cost function* the statement of the theorem holds for *arbitrary* (not necessarily symmetric nor unimodal) distributions. We just notice that, choosing for c the simple Euclidean distance

$$c(z, x) = \|z - x\|^2 , \tag{5.1.10}$$

this fact easily follows from a well-known variational characterization of the mean of a probability distribution which is the content of the following proposition which should be compared with the statement of Theorem 5.2 below.

Proposition 5.2. *For a quadratic cost function like* (5.1.10) *and, more generally, such as* (5.1.6)*, one has* $\hat{x} = \mathbb{E}(\mathbf{x} \mid \mathbf{y} = y)$*, for an arbitrary conditional density* $f(x \mid y)$*, provided of course that the conditional mean makes sense.*

The estimate $\hat{x}(\mathbf{y})$ minimizing the expected quadratic cost (5.1.10) is often called a *least squares (Bayesian) estimate* of $\mathbf{x}$. This terminology is somewhat ambiguous since in the literature there is a tendency to use the attribute " least squares" for too many things. We shall not use it in this probabilistic context.

Remarks 5.1. Since under our assumptions on c there always is uniqueness of the minimum (5.1.8), the point estimate $\hat{x}$ of x can be seen as the value of a function of the observation y, say $\hat{\mathbf{x}} : \mathbb{R}^m \to \mathbb{R}^n$ which is called the (minimum conditional expected risk) *Bayesian estimator* of $\mathbf{x}$ given $\mathbf{y}$. An *estimator* is just a function of the data taking values in the same range space of $\mathbf{x}$ which evidently can be interpreted as a calculation procedure (algorithm) processing the measurement data into point estimates. This is conventionally depicted as a block diagram of the type shown in Fig. 5.1.1.

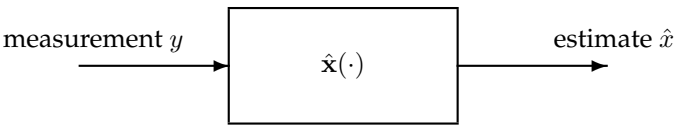


Figure 5.1.1. *Estimator.*

Notation : The symbol $\mathbb{E}(\mathbf{x} \mid \mathbf{y})$, to be read: the conditional expectation (or conditional mean) of $\mathbf{x}$ given $\mathbf{y}$ represents the function of the variable y defined on the range of $\mathbf{y}$, say $\mathbb{R}^m$, by the assignment

$$\mathbb{E}(\mathbf{x} \mid \mathbf{y}) : y \to \mathbb{E}(\mathbf{x} \mid \mathbf{y} = y) .$$

By Theorem 5.1, for convex symmetric cost functions, this function of the data is the minimal conditional risk Bayesian point estimator of $\mathbf{x}$ given the observed

value $\mathbf{y} = y$. Actually, for an arbitrary a posteriori probability distribution, the conditional mean estimator can be also characterized in the following way.

Theorem 5.2. *Consider the class of all measurable functions $g : \mathbb{R}^m \to \mathbb{R}^n$ such that $g(\mathbf{y})$ has a finite variance, then the conditional mean $\mathbb{E}(\mathbf{x} \mid \mathbf{y})$ minimizes the expected mean square deviation of $g(\mathbf{y})$ from the random vector $\mathbf{x}$; in other words,*

$$\mathbb{E}(\mathbf{x} \mid \mathbf{y}) = \operatorname{Arg} \min_{g(\cdot)} \mathbb{E}\,\|\mathbf{x} - g(\mathbf{y})\|_Q^2 \tag{5.1.11}$$

for an arbitrary symmetric positive semi definite $Q \in \mathbb{R}^{n\times n}$. If Q is positive definite then $\mathbb{E}(\mathbf{x} \mid \mathbf{y})$ is the unique minimizer.

Proof. Note first that

$$\mathbb{E}\,\|\mathbf{x}-g(\mathbf{y})\|_Q^2 = \int_{\mathbb{R}^m} \mathbb{E}\,\left[\,\|\mathbf{x} - g(\mathbf{y})\|_Q^2 \mid \mathbf{y} = y\right]\, p_{\mathbf{y}}(y)\, dy = \mathbb{E}\,\{\mathbb{E}\,\left[\|\mathbf{x} - g(\mathbf{y})\|_Q^2 \mid \mathbf{y}\,\right]\} \tag{5.1.12}$$

and then use the identity

$$\mathbb{E}\,\{\|\mathbf{x} - \mathbb{E}(\mathbf{x} \mid \mathbf{y}) + \mathbb{E}(\mathbf{x} \mid \mathbf{y}) - g(\mathbf{y})\|_Q^2 \mid \mathbf{y}\} = \mathbb{E}\,\{\|\mathbf{x} - \mathbb{E}(\mathbf{x} \mid \mathbf{y})\|_Q^2 \mid \mathbf{y}\}+$$
$$+2\mathbb{E}\,\{[\mathbf{x} - \mathbb{E}(\mathbf{x} \mid \mathbf{y})]^\top Q\,[\mathbb{E}(\mathbf{x} \mid \mathbf{y}) - g(\mathbf{y})] \mid \mathbf{y}\} + \mathbb{E}\,\{\|\mathbb{E}(\mathbf{x} \mid \mathbf{y}) - g(\mathbf{y})\|_Q^2 \mid \mathbf{y}\}$$

where the second term on the right is zero since

$$\mathbb{E}\,\{[\mathbf{x}-\mathbb{E}(\mathbf{x} \mid \mathbf{y})]^\top Q\,[\mathbb{E}(\mathbf{x} \mid \mathbf{y}) - g(\mathbf{y})] \mid \mathbf{y}\} = [\mathbb{E}(\mathbf{x} \mid \mathbf{y}) - \mathbb{E}(\mathbf{x} \mid \mathbf{y})]^\top Q\,[\mathbb{E}(\mathbf{x} \mid \mathbf{y}) - g(\mathbf{y})]$$

by a well-known property of the conditional expectation. Computing the expected value of both members of (**??**) one obtains

$$\mathbb{E}\,\|\mathbf{x} - g(\mathbf{y})\|_Q^2 = \mathbb{E}\,\|\mathbf{x} - \mathbb{E}(\mathbf{x} \mid \mathbf{y})\|_Q^2 + \mathbb{E}\,\|\mathbb{E}(\mathbf{x} \mid \mathbf{y}) - g(\mathbf{y})\|_Q^2 \quad ,$$

where both terms on right are nonnegative but the first does not depend on g. Therefore the minimum is achieved for $g(\mathbf{y}) = \mathbb{E}(\mathbf{x} \mid \mathbf{y})$. □

If we restrict the class of admissible estimators g to the class, which we shall call *mean-unbiased*[12], for which the estimation error has mean zero, namely

$$\mathbb{E}\,g(\mathbf{y}) = \mathbb{E}\,\mathbf{x}$$

then for $Q = I$ the expected value of the square norm of the estimation error $\mathbf{x}-g(\mathbf{y})$ is just its *scalar variance*. On the other hand, as indicated in the following exercise, the equality above is an obvious necessary condition for g to minimize the mean square error. Hence one may well say that the conditional mean of $\mathbf{x}$ given $\mathbf{y}$ is the estimator that has **minimum error variance** among all measurable functions g of the observations.

Problem 5.1. *Set $\tilde{\mathbf{x}} = \mathbf{x} - \mathbb{E}\,\mathbf{x}$ and $\tilde{g}(\mathbf{y}) := g(\mathbf{y}) - E\,g(\mathbf{y})$. Use the identity*

$$\mathbb{E}\,\|\mathbf{x} - g(\mathbf{y})\|_Q^2 = \mathbb{E}\,\|\tilde{\mathbf{x}} - \tilde{g}(\mathbf{y})\|_Q^2 + \|\mathbb{E}\,\mathbf{x} - \mathbb{E}\;g(\mathbf{y})\|_Q^2\,,$$

to prove that mean-unbiasedness of g is a necessary condition for optimality. ♢

[12]Note that this is a different notion than unbiasedness in the Fisherian approach.

5.2 ▪ The M.A.P estimator

It is quite obvious that the definition of point estimator as a function of the observed data that minimizes the conditional expected risk can be replaced by an equivalent one where instead one maximizes a conditional *expected gain* of the type

$$G(z,y) := \mathbb{E}\left[\gamma(z-\mathbf{x}) \mid \mathbf{y}=y\right] = \int_{\mathbb{R}^n} \gamma(z-x)\, f(x \mid y)\, dx \quad , \tag{5.2.1}$$

where now $\gamma : \mathbb{R}^n \to \mathbb{R}$ is a function of the variable $\xi = z - x$, $z \in \mathbb{R}^n$, which is *concave*, non-negative and has a maximum at the origin (the maximum being possibly infinite). Taking γ peak -shaped, centered at the origin and zero outside of a small neighborhood of $\xi = 0$, we can approximate arbitrarily well a Dirac δ function and get an expected gain function of the same form as the conditional density given $\mathbf{y} = y$; i.e.

$$G(z,y) \simeq f(z \mid y)$$

The corresponding estimator

$$\hat{x}_{MAP}(y) := \operatorname{Arg}\max_z f(z \mid y) \tag{5.2.2}$$

is inspired by the principle of choosing as a point estimate the value $z = \hat{x}$ which maximizes the a posteriori probability distribution of $\mathbf{x}$ given the observation $\mathbf{y} = y$. It is called the *Maximum a Posteriori Estimator* (MAP) and is widely used in Bayesian statistics[13]. The MAP estimator is just the *conditional mode* of the a posteriori density of $\mathbf{x}$ given the observation $\mathbf{y} = y$. Of course in the case of a unimodal and symmetric density the mode coincides with the mean and we do not find anything new.

Example 5.1 (Relation to Maximum Likelihood). *Maximum likelihood can be seen as a special case of MAP estimation which occurs when the a priori density is* uniform. *In fact, when $p(\theta)$ is a constant, the maximization of the a posteriori density of the parameter θ,*

$$p(\theta \mid x_1, \ldots, x_N) = \frac{f(x_1, \ldots, x_N \mid \theta) p(\theta)}{p(x_1, \ldots, x_N)}$$

reduces to the maximization of the likelihood function $f(x_1, \ldots, x_N \mid \theta)$ with respect to θ, since the denominator is independent of θ.

In conclusion we have seen that the Bayesian estimator is, at least in a great majority of cases of interest, a conditional mean and hence Bayesian estimation can be seen just as a chapter of probability theory without appealing to any inductive reasoning which is instead the rule in classical Statistics. Unfortunately however the conditional mean $\mathbb{E}(\mathbf{x} \mid \mathbf{y})$ can be computed explicitly only in very few cases. A particularly important one is when the joint distribution of $\mathbf{x}$ and $\mathbf{y}$ is *Gaussian*. This will be discussed in some detail below.

[13] It is actually closely related to the *maximum likelihood estimator* in classical parametric Statistics.

5.3 ▪ Conditional Expectation of Gaussian random vectors

Theorem 5.3. *Let the n- and m- dimensional random vectors $\mathbf{x}$ and $\mathbf{y}$ be jointly Gaussian, that is, let the $n+m$-dimensional vector $\mathbf{z} = [\mathbf{x}^\top, \mathbf{y}^\top]^\top$ have a Gaussian distribution with mean,*

$$\mu_{\mathbf{z}} = \begin{bmatrix} \mu_{\mathbf{x}} \\ \mu_{\mathbf{y}} \end{bmatrix} \tag{5.3.1}$$

and Covariance matrix

$$\Sigma_{\mathbf{z}} = \begin{bmatrix} \Sigma_{\mathbf{x}} & \Sigma_{\mathbf{xy}} \\ \Sigma_{\mathbf{yx}} & \Sigma_{\mathbf{y}} \end{bmatrix} \quad , \tag{5.3.2}$$

Then the conditional density of $\mathbf{x}$ given $\mathbf{y}$ is still Gaussian. If $\Sigma_{\mathbf{y}} > 0$, then its (conditional) mean and covariance matrix are:

$$\mathbb{E}\left(\mathbf{x} \mid \mathbf{y}\right) = \mu_{\mathbf{x}} + \Sigma_{\mathbf{xy}}\, \Sigma_{\mathbf{y}}^{-1}\, (\mathbf{y} - \mu_{\mathbf{y}}) \tag{5.3.3}$$

$$\mathrm{Var}\left(\mathbf{x} \mid \mathbf{y}\right) = \Sigma_{\mathbf{x}} - \Sigma_{\mathbf{xy}}\, \Sigma_{\mathbf{y}}^{-1}\, \Sigma_{\mathbf{yx}} \quad . \tag{5.3.4}$$

Proof. To simplify notations let us introduce the centered $n+m$-dimensional vector $\bar{\mathbf{z}} := \mathbf{z} - \mu_{\mathbf{z}}$, which has components

$$\bar{\mathbf{z}} = \begin{bmatrix} \bar{\mathbf{x}} \\ \bar{\mathbf{y}} \end{bmatrix} \quad .$$

Obviously $\bar{\mathbf{z}} \sim N(0, \Sigma_{\mathbf{z}})$ where $\Sigma_{\mathbf{z}}$ is displayed in (5.3.2). Introduce a linear transformation of the variables $(\bar{\mathbf{x}}, \bar{\mathbf{y}})$ of the following form

$$\begin{cases} \tilde{\mathbf{x}} = \bar{\mathbf{x}} + A\bar{\mathbf{y}} \\ \tilde{\mathbf{y}} = \bar{\mathbf{y}} \end{cases} \tag{5.3.5}$$

where $A \in \mathbb{R}^{n \times m}$ is chosen in such a way as to make $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{y}}$ uncorrelated; i.e. such that $\mathbb{E}\ \tilde{\mathbf{x}}\tilde{\mathbf{y}}^\top = 0$. Imposing this condition one finds the equation

$$0 = \Sigma_{\mathbf{xy}} + A\, \Sigma_{\mathbf{y}}$$

which, assuming $\Sigma_{\mathbf{y}} > 0$, has the unique solution

$$A = -\Sigma_{\mathbf{xy}}\, \Sigma_{\mathbf{y}}^{-1} \quad . \tag{5.3.6}$$

Clearly $\tilde{\mathbf{z}} := [\tilde{\mathbf{x}}^\top\ \tilde{\mathbf{y}}^\top]^\top$ is Gaussian zero-mean and has covariance matrix

$$\Sigma_{\tilde{\mathbf{z}}} = \begin{bmatrix} I & A \\ 0 & I \end{bmatrix} \Sigma_{\mathbf{z}} \begin{bmatrix} I & 0 \\ A^\top & I \end{bmatrix} = \begin{bmatrix} \Sigma_{\tilde{\mathbf{x}}} & 0 \\ 0 & \Sigma_{\mathbf{y}} \end{bmatrix}.$$

We want to compute $\mathbb{E}\left(\bar{\mathbf{x}} \mid \bar{\mathbf{y}}\right)$. To this end use the first equality in (5.3.5) and note that by Gaussianness $\tilde{\mathbf{x}}$ and $\bar{\mathbf{y}}$ are independent so that $\mathbb{E}\left(\tilde{\mathbf{x}} \mid \bar{\mathbf{y}}\right) = \mathbb{E}\left(\tilde{\mathbf{x}}\right) = 0$ since both $\bar{\mathbf{x}}, \bar{\mathbf{y}}$ are zero mean. Since $\bar{\mathbf{x}} = \tilde{\mathbf{x}} - A\bar{\mathbf{y}}$ and the second term is trivially a function of $\mathbf{y}$, it follows by the additivity of conditional expectation that

$$\mathbb{E}\left(\bar{\mathbf{x}} \mid \bar{\mathbf{y}}\right) = -A\bar{\mathbf{y}} = \ \Sigma_{\mathbf{xy}}\, \Sigma_{\mathbf{y}}^{-1}\, \bar{\mathbf{y}} \tag{5.3.7}$$

The same expression can be obtained computing the conditional density $p_{\bar{\mathbf{x}}|\bar{\mathbf{y}}}$ via the classical Bayes formula. By a well-known property of the Gaussian distribution, the components $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{y}}$ are independent so that

$$p_{\tilde{\mathbf{z}}}(\tilde{z}) = p_{\tilde{\mathbf{x}}}(\tilde{x})\; p_{\tilde{\mathbf{y}}}(\tilde{y}) = p_{\tilde{\mathbf{x}}}(\tilde{x})\; p_{\bar{\mathbf{y}}}(\tilde{y})\,,$$

where $\tilde{x}, \tilde{y}$ are dummy variables. From this expression it follows that the conditional density $p_{\tilde{\mathbf{x}}|\bar{\mathbf{y}}}$ is, modulo a change of variables, equal to $p_{\tilde{\mathbf{x}}}$. We need to apply the rules for computing the density of a function of random variables. Since the Jacobian of the transformation (5.3.5) is an upper triangular matrix with an identity on the main diagonal it has a determinant equal to one. Then

$$p_{\bar{\mathbf{z}}}(z) = p_{\tilde{\mathbf{z}}}(\tilde{z})\Big|_{\substack{\tilde{x}\,=\,x\,+\,Ay \\ \tilde{y}\,=\,y}} = p_{\tilde{\mathbf{x}}}(x + Ay)\; p_{\bar{\mathbf{y}}}(y) \tag{5.3.8}$$

and the conditional density of $\bar{\mathbf{x}}$ given $\bar{\mathbf{y}} = y$ is just $p_{\tilde{\mathbf{x}}}(x + Ay)$. Since $\tilde{\mathbf{x}}$ is a linear combination of Gaussian vectors, this is a Gaussian density of (conditional) mean $-Ay$ and covariance $\Sigma_{\tilde{\mathbf{x}}}$. For the mean we find again (5.3.7) and the expression for the conditional covariance is

$$\begin{aligned} \mathrm{Var}\;(\bar{\mathbf{x}} \mid \bar{\mathbf{y}}) = \Sigma_{\tilde{\mathbf{x}}} &= \mathbb{E}\,(\bar{\mathbf{x}} + A\bar{\mathbf{y}})\;(\bar{\mathbf{x}} + A\bar{\mathbf{y}})^\top \\ &= \mathbb{E}\,\bar{\mathbf{x}}\bar{\mathbf{x}}^\top + \mathbb{E}\,A\,\bar{\mathbf{y}}\bar{\mathbf{x}}^\top \\ &= \Sigma_{\mathbf{x}} - \Sigma_{\mathbf{xy}}\;\Sigma_{\mathbf{y}}^{-1}\;\Sigma_{\mathbf{yx}} \quad . \end{aligned} \tag{5.3.9}$$

Finally (5.3.3) is obtained by reintroducing the mean values in (5.3.7). □

When $\Sigma_{\mathbf{y}}$ is singular one can give similar expressions where the inverse replaced by the (Moore-Penrose) pseudoinverse of $\Sigma_{\mathbf{y}}$. See Proposition 5.6.

Now $E(\mathbf{x} \mid \mathbf{y})$ is the Bayesian estimator of $\mathbf{x}$ based on the observation vector $\mathbf{y}$, hence the difference $\tilde{\mathbf{x}}$, introduced in (5.3.5) is the *residual estimation error*

$$\tilde{\mathbf{x}} := \bar{\mathbf{x}} - \Sigma_{\mathbf{xy}}\;\Sigma_{\mathbf{y}}^{-1}\;\bar{\mathbf{y}} = \mathbf{x} - \mathbb{E}\,(\mathbf{x} \mid \mathbf{y}) \quad . \tag{5.3.10}$$

As shown in the proof of Theorem 5.3 the residual has the crucial property of being *independent of the observed data* $\mathbf{y}$. The intuition behind this fact is that $\tilde{\mathbf{x}}$ is what is left after subtracting from $\mathbf{x}$ its best approximation based on the knowledge of $\mathbf{y}$. The independence is just a manifestation of the fact that the data do not contain any more information which may be useful to change the estimation residual $\tilde{\mathbf{x}} = \mathbf{x} - \mathbb{E}\,(\mathbf{x} \mid \mathbf{y})$. In other words, this is a proof of the fact that *the data have been exploited in the best possible way*.

Incidentally the independence of the residual and the observations explains the counter-intuitive fact that the conditional covariance (5.3.4) of $\mathbf{x}$ given $\mathbf{y}$ does not depend on $\mathbf{y}$. This conditional covariance coincides in fact with the unconditional covariance of the residual estimation error, $\Sigma_{\tilde{\mathbf{x}}}$, just because of the independence of $\tilde{\mathbf{x}}$ and $\mathbf{y}$. Note also that the covariance of the residual estimation error,

$$\Sigma_{\tilde{\mathbf{x}}} = \Sigma_{\mathbf{x}} - \Sigma_{\mathbf{xy}}\;\Sigma_{\mathbf{y}}^{-1}\;\Sigma_{\mathbf{yx}} \quad ,$$

is the difference between $\Sigma_{\mathbf{x}}$, the a priori covariance of $\mathbf{x}$, and the covariance matrix of the estimator $\mathbb{E}\,(\mathbf{x} \mid \mathbf{y})$ as it easily follows from (5.3.3) or (5.3.7), since

$$\mathrm{Var}\,\mathbb{E}\,(\mathbf{x} \mid \mathbf{y}) = \Sigma_{\mathbf{xy}}\;\Sigma_{\mathbf{y}}^{-1}\;\Sigma_{\mathbf{yx}} \quad . \tag{5.3.11}$$

Therefore the difference (5.3.4) can be interpreted as *the reduction of the a priori uncertainty on* $\mathbf{x}$ *provided by the observation* $\mathbf{y}$. The smaller this difference; i.e. the closer the matrix (5.3.11)) is to $\Sigma_{\mathbf{x}}$, the more efficient is the estimator.

Bayesian inference for Gaussian pdf's.

The calculation of the conditional density in (5.3.8) can be used to discuss Bayesian inference for Gaussian random vectors.
Suppose we are *given* a Gaussian conditional density $f(y \mid \mathbf{x} = \theta)$ of an observed m-vector $\mathbf{y}$, given the value of a random vector parameter $\mathbf{x} = \theta$, together with a Gaussian a priori density of $\mathbf{x}$. We want to compute the a posteriori conditional density of $\mathbf{x}$ given $\mathbf{y} = y$. For simplicity we shall assume that both $\mathbf{x}$ and $\mathbf{y}$ have zero mean.

It follows from the second part of the proof of Theorem 5.3 that the a posteriori density $f(\theta \mid \mathbf{y} = y)$ is still Gaussian, with (conditional) mean and (conditional) variance given by formulas (5.3.3) and (5.3.4). To implement these formulas however we need the cross- and auto-covariances $\Sigma_{\mathbf{xy}}$ and $\Sigma_{\mathbf{y}}$ which are not immediately evident since from $f(y \mid \mathbf{x} = \theta)$ and $p_{\mathbf{x}}(\theta)$ we can only extract the conditional statistics of $\mathbf{y}$ given $\mathbf{x}$ besides, of course, $\mu_{\mathbf{x}}$ and $\Sigma_{\mathbf{x}}$.

We need to compute $\Sigma_{\mathbf{xy}}$ and $\Sigma_{\mathbf{y}}$ in function of the parameters of the given conditional model.
Now since $f(y \mid \mathbf{x} = \theta)$ is Gaussian, its conditional mean must be a linear function of the conditioning vector say

$$\hat{\mathbf{y}}(\mathbf{x}) = S\mathbf{x} \tag{5.3.12}$$

where S is some $m \times n$ deterministic matrix which we shall assume given and, without loss of generality, of full rank. By what we have just seen, the residual error difference $\tilde{\mathbf{y}} := \mathbf{y} - \hat{\mathbf{y}}$, which we shall re-name $\mathbf{w}$, must be uncorrelated (in fact independent) of $\hat{\mathbf{y}}(\mathbf{x}) = S\mathbf{x}$ and since S is full column rank it must be uncorrelated with (independent of) $\mathbf{x}$. In other words, given $f(y \mid \mathbf{x} = \theta)$ we discover that $\mathbf{y}$ is represented by a linear model

$$\mathbf{y} = S\mathbf{x} + \mathbf{w} \tag{5.3.13}$$

where $\mathbf{x}$ and $\mathbf{w}$ are uncorrelated and both have known variance matrices. In fact $\Sigma_{\mathbf{w}}$ is just the conditional variance of $\mathbf{y}$ given $\mathbf{x}$, which incidentally coincides with the variance of $\tilde{\mathbf{y}}$. From the model (5.3.13) it is immediate to deduce that $\Sigma_{\mathbf{yx}} = S\Sigma_{\mathbf{x}}$. Next, need to compute $\Sigma_{\mathbf{y}}$ (which is is not known) from the available densities or ,equivalently, from the linear model (5.3.13). One way to go is to exchange the role of $\mathbf{x}$ and $\mathbf{y}$ in the proof of Theorem 5.3, arriving at the dual relation

$$\Sigma_{\tilde{\mathbf{y}}} = \Sigma_{\mathbf{w}} = \Sigma_{\mathbf{y}} - \Sigma_{\mathbf{yx}}\Sigma_{\mathbf{x}}^{-1}\Sigma_{\mathbf{xy}}$$

from which

$$\Sigma_{\mathbf{y}} = \Sigma_{\mathbf{w}} + \Sigma_{\mathbf{yx}}\Sigma_{\mathbf{x}}^{-1}\Sigma_{\mathbf{xy}} = \Sigma_{\mathbf{w}} + S\Sigma_{\mathbf{x}}S^{\top}\,.$$

The last relation is anyway evident from the representation (5.3.13). Finally, assuming a zero-mean prior we get

$$\hat{\mathbf{x}}(\mathbf{y}) = \Sigma_{\mathbf{x}}S^{\top}\left[\Sigma_{\mathbf{w}} + S\Sigma_{\mathbf{x}}S^{\top}\right]^{-1}\mathbf{y}\,, \tag{5.3.14}$$

$$\text{Var}\,\{\mathbf{x} - \hat{\mathbf{x}}(\mathbf{y})\} = \Sigma_{\mathbf{x}} - \Sigma_{\mathbf{x}}S^{\top}\left[\Sigma_{\mathbf{w}} + S\Sigma_{\mathbf{x}}S^{\top}\right]^{-1}S\Sigma_{\mathbf{x}} \tag{5.3.15}$$

An alternative route could have been to compute $\Sigma_{\mathbf{y}}$ by marginalizing the joint density $p_{\mathbf{y},\mathbf{x}}(y,\theta) = f(y \mid \mathbf{x} = \theta)p_{\mathbf{x}}(\theta)$ integrating with respect to θ. We have avoided the explicit computation of this integral.
These formulas will be re-derived later in Section 5.6, in the context of linear estimation. In particular see the remark 5.2.

Example 5.2. *A scalar parameter θ is observed N-times in the presence of additive uncorrelated Gaussian noise. Letting $\mathbf{1}_N$ denote an N-vector of ones, the observation model is described by the conditional density*

$$f(y \mid \mathbf{x} = \theta) \equiv \mathcal{N}(\mathbf{1}_N\theta, \sigma^2 I_N), \tag{5.3.16}$$

where the random parameter has a Gaussian a priori distribution $p_{\mathbf{x}}(\theta) \equiv \mathcal{N}(\mu, \tau^2)$. Compute the posterior density of $\mathbf{x}$ given a N-dimensional sequence y of measurements described by the model (5.3.16) .
Solution: The posterior is still Gaussian with conditional mean and conditional variance derived by formulas (5.3.14) and (5.3.15). We have $S = \mathbf{1}_N$ and $\Sigma_{\mathbf{w}} = \sigma^2 I_N$ so that

$$\hat{\mathbf{x}}(\mathbf{y}) = \mu + \tau^2 \mathbf{1}_N^\top \left(\sigma^2 I_N + \tau^2 \mathbf{1}_N \mathbf{1}_N^\top\right)^{-1} (\mathbf{y} - \mathbf{1}_N\,\mu) \tag{5.3.17}$$

$$\sigma^2_{\hat{\mathbf{x}}} = \tau^2 - \tau^2 \mathbf{1}_N^\top \left(\sigma^2 I_N + \tau^2 \mathbf{1}_N \mathbf{1}_N^\top\right)^{-1} \mathbf{1}_N\,\tau^2 \tag{5.3.18}$$

At first sight the calculation of the inverse needed in these expressions looks quite demanding. As we shall see it is greatly facilitated by the use of the *Matrix Inversion Lemma* which will be introduced later, see Sect. 5.6 and Example 5.4.

5.4 ▪ Linear Minimum Variance Estimators

The minimum (error) variance estimator, hereafter the M.V. estimator, of $\mathbf{x}$ based on the observation $\mathbf{y}$ has a particularly simple form when $\mathbf{x}$ and $\mathbf{y}$ are Gaussian. In this case $\mathbb{E}(\mathbf{x} \mid \mathbf{y})$ is a *linear function of the observations* which can be computed based only on the first and second order joint moments of the variables $\mathbf{x}$ and $\mathbf{y}$.

When the data are not Gaussian this simplicity disappears as $\mathbb{E}(\mathbf{x} \mid \mathbf{y})$ is in general a *non-linear function* of the observations which can actually be computed explicitly only in very few cases. It is then natural to ask wether restricting a priori the candidate function of the data, g, to minimize the expected squared error $E\,\|\mathbf{x} - g(\mathbf{y})\|^2$, one could get estimators which are easier to compute. In fact the obvious first choice is to look for functions which are *linear (or affine)* in the data.

Definition 5.2. *The* linear minimum (error) variance estimator (LMV) *of the random vector $\mathbf{x}$, based on the observation vector $\mathbf{y}$ is the affine function*

$$g(\mathbf{y}) = A\mathbf{y} + b \quad , \qquad A \in \mathbb{R}^{n\times m}\ ,\ b \in \mathbb{R}^n$$

which minimizes the expected squared estimation error $\mathbb{E}\,\|\mathbf{x} - g(\mathbf{y})\|^2$.

This minimum variance linear estimator of $\mathbf{x}$ must therefore be a solution of the optimization problem

$$\min_{A\ ,\ b} \left\{\, \mathbb{E}\,\|A\mathbf{y} + b - \mathbf{x}\|^2 \ \mid A \in \mathbb{R}^{n\times m}\ ,\ b \in \mathbb{R}^n \,\right\}. \tag{5.4.1}$$

which we shall now proceed to solve. Note first that when $\mathbf{x}$ and $\mathbf{y}$ have zero mean, the value of b for which the minimum is attained is zero. In fact, assume that

$$g_*(\mathbf{y}) = A_*\mathbf{y} + b_*$$

is the optimal m.v. estimator. Since $\mathbb{E}\,\mathbf{x} = 0$, $\mathbb{E}\,\mathbf{y} = 0$, one has

$$\begin{aligned}\mathbb{E}\,\|A_*\mathbf{y} + b_* - \mathbf{x}\|^2 &= \mathbb{E}\,\|A_*\mathbf{y} - \mathbf{x}\|^2 + 2\mathbb{E}\,(A_*\mathbf{y} - \mathbf{x})^\top\, b_* + \|b_*\|^2 \\ &= \mathbb{E}\,\|A_*\mathbf{y} - \mathbf{x}\|^2 + \|b_*\|^2 \geq \mathbb{E}\,\|A_*\mathbf{y} - \mathbf{x}\|^2\end{aligned}$$

with strict inequality unless $b_* = 0$ which implies that unless $b_* = 0$, $A_*\mathbf{y}$ would be a strictly better estimator than $g_*(\mathbf{y})$.

Hence it will be enough to look for the LMV estimator of $\bar{\mathbf{x}} = \mathbf{x} - \mu_\mathbf{x}$ based on the centered data $\bar{\mathbf{y}} = \mathbf{y} - \mu_\mathbf{y}$. Once found the optimal linear function $g(\bar{\mathbf{y}}) = A\bar{\mathbf{y}}$ for the centered variables, we may just add the mean value of $\mathbf{x}$ to get $\hat{\mathbf{x}} = \hat{\bar{\mathbf{x}}} + \mu_\mathbf{x}$ and substitute back $\bar{\mathbf{y}} = \mathbf{y} - \mu_\mathbf{y}$ to obtain the formula valid for non zero mean values (see Problem 5.2 below for a formal justification). We are henceforth to consider the minimization problem with zero-mean variables,

$$\min_A\ \mathbb{E}\,\|A\bar{\mathbf{y}} - \bar{\mathbf{x}}\|^2 \quad . \tag{5.4.2}$$

Problem 5.2. *Note that*

$$\min_{A,b}\ \mathbb{E}\,\|A\mathbf{y} + b - \mathbf{x}\|^2 = \min_{A,b}\ \mathbb{E}\,\|A\bar{\mathbf{y}} - \bar{\mathbf{x}} + c\|^2$$

where $c := A\,\mu_\mathbf{y} + b - \mu_\mathbf{x}$. *Using the relation* $\mathbb{E}\,(A\bar{\mathbf{y}} - \bar{\mathbf{x}}) = 0$ *show that the minimization reduces to*

$$\min_{A,b}\ \left(\mathbb{E}\,\|A\bar{\mathbf{y}} - \bar{\mathbf{x}}\|^2 + \|\,A\,\mu_\mathbf{y} + b - \mu_\mathbf{x}\|^2\right) \quad .$$

Show that the minimum is achieved when A_* *minimizes the first term and the vector* b *is taken equal to*

$$b_* := \mu_\mathbf{x} - A_*\,\mu_\mathbf{y}\,. \tag{5.4.3}$$

5.5 ▪ Geometric formulation and the Orthogonal Projection Lemma

The minimization problem (5.4.2) can be solved by elementary means. Nevertheless it motivates the introduction of a geometric setting which, although at this point may look a bit unnatural, will become a fundamental tool when dealing with linear estimation problems for stochastic processes.

Let us consider the family of real-valued random variables defined in the same probability space $\{\Omega, \mathcal{A}, P\}$ which in our setting is to be interpreted as the space of all possible experimental conditions under which our measurement experiment could be performed.
We shall just consider random variable which have *finite variance* and (for convenience) zero mean. This set has obviously the structure of a real vector space, which we shall denote by the symbol $\mathbf{H}$. On thi svector space one can introduce a natural inner product, $\langle\cdot,\cdot\rangle_\mathbf{H}$ defined as

$$\langle\mathbf{x},\mathbf{y}\rangle_\mathbf{H} := \mathbb{E}\,\mathbf{x}\mathbf{y} = \int_\Omega \mathbf{x}(\omega)\ \mathbf{y}(\omega)\ P(d\omega) = \int_{\mathbb{R}^2} xy\ p_{\mathbf{xy}}(x,y)\ dx\ dy \tag{5.5.1}$$

kolmogorov.jpg (JPEG Image, 406 × 426 pixels)

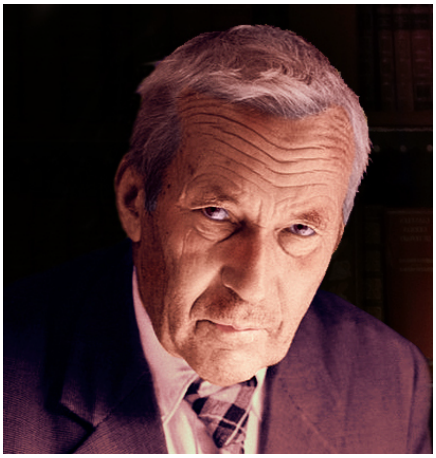


Figure 5.5.1. *A. N. Kolmogorov.*

This inner product satisfies the usual axioms of an inner product, in particular the triangular inequality and induces the *variance norm*

$$\|\mathbf{x}\|_{\mathbf{H}}^2 = \mathbb{E}\,\mathbf{x}^2 = \text{var}\,(\mathbf{x})\,, \tag{5.5.2}$$

where the equivalence between $\|\mathbf{x}\|_{\mathbf{H}} = 0$ and $\mathbf{x} = 0$ holds once we agree to declare a random variable equal to zero iff it is zero with probability one. For zero-mean random variables this is in turn equivalent to their scalar variance being equal to zero.

In this way we end up with an inner product space which is actually a real *Hilbert space*[14], as it follows from the definition of the inner product (5.5.1) which makes it an L^2 space of real measurable functions defined on the abstract set Ω. This space will still be denoted with the symbol $\mathbf{H}$. The convergence of random variables in $\mathbf{H}$ with respect to the norm $\|\cdot\|_{\mathbf{H}}$ is known as *mean-square convergence*; $\mathbf{H}$ is in fact complete with respect to mean square convergence. It will be called the *Hilbert space of second order random variables on the experiment* $\{\Omega, \mathcal{A}, P\}$. Note that the orthogonality relation in $\mathbf{H}$ corresponds to uncorrelation; it will be denoted by the symbol $\perp_{\mathbf{H}}$; i.e. $\mathbf{x} \perp_{\mathbf{H}} \mathbf{y} \Leftrightarrow \sigma_{\mathbf{x},\mathbf{y}} = \mathbb{E}\,\mathbf{x}\mathbf{y} = 0$.

The idea of introducing this geometric setting is due to Andrej Nicolayevich Kolmogorov [50]; it has allowed to phrase the linear estimation theory of random processes in a geometric setting with a great gain in simplicity and conceptual clarity.

The observation subspace: Given an m-dimensional random vector $\mathbf{y}$ the vector subspace of $\mathbf{H}$ generated by the scalar components $\mathbf{y}_k$, $i = 1, \ldots, m$, is the set of

[14]Some basic notions on Hilbert spaces are reported in Appendix D.

all linear combinations with real coefficients of the scalar components $\mathbf{y}_k$; $k = 1 \dots m$. This subspace is denoted by the symbol $\mathbf{H}(\mathbf{y})$; notation:

$$\mathbf{H}(\mathbf{y}) := \text{span}\,\{\mathbf{y}_1 \dots \mathbf{y}_m\} = \left\{ \sum_{i=1}^{m} a_k \mathbf{y}_k \ ; \ a_k \in \mathbb{R} \right\} \quad . \tag{5.5.3}$$

When the $\{\mathbf{y}_k \, k = 1, \dots, m\}$, are linearly independent (as functions of the variable ω), $\mathbf{H}(\mathbf{y})$ has dimension m. It is well-known that this happens if and only if the Gramian matrix

$$G(\mathbf{y}) := [\, \langle \mathbf{y}_k, \mathbf{y}_j \rangle_{\mathbf{H}}]_{kj=1\dots m}$$

is non-singular. In fact $G(\mathbf{y})$ is just the covariance matrix $\Sigma_{\mathbf{y}}$, of the random vector $\mathbf{y}$; hence:

Proposition 5.3. *The scalar components of the random vector $\mathbf{y}$ are linearly independent (equiv. the subspace $\mathbf{H}(\mathbf{y})$ has dimension m) if and only if the covariance matrix $\Sigma_{\mathbf{y}} = \mathbb{E}\,\mathbf{y}\mathbf{y}^\top$ is non singular (i.e.* strictly positive definite*; written $\Sigma_{\mathbf{y}} > 0$).*

In this case one may say that the data are *non redundant* as none of them can be expressed as a linear function of the others.

Consider now the following basic approximation problem.

Problem 5.3. *Given a scalar random variable $\mathbf{x} \in \mathbf{H}$, find the linear function $a^\top \mathbf{y}$, of the random vector $\mathbf{y}$, which has minimal distance from $\mathbf{x}$. Precisely, find an element $\mathbf{z} \in \mathbf{H}(\mathbf{y})$ such that the squared norm*

$$\|\mathbf{x} - \mathbf{z}\|^2_{\mathbf{H}} = \mathbb{E}\,[\mathbf{x} - \mathbf{z}]^2$$

is minimal.

Here we need not assume that $\mathbf{y}$ has a finite number of components; m could be $+\infty$ (in which case we could have an infinite sequence of observations) and $\mathbf{H}(\mathbf{y})$ could well be infinite dimensional. This generalization will be important later on when dealing with random processes. The solution is provided by the classical

Theorem 5.4 (Orthogonal projection Lemma). *Let $\mathbf{H}(\mathbf{y})$ be a closed subpace of $\mathbf{H}$. There is a unique random variable $\mathbf{z}_* \in \mathbf{H}(\mathbf{y})$ which minimizes $\|\mathbf{x} - \mathbf{z}\|^2_{\mathbf{H}}$; this variable is the orthogonal projection of $\mathbf{x}$ onto $\mathbf{H}(\mathbf{y})$.*

A necessary and sufficient condition for $\mathbf{z}_*$ to be the orthogonal projection of $\mathbf{x}$ onto $\mathbf{H}(\mathbf{y})$ is that

$$\mathbf{x} - \mathbf{z}_* \perp_{\mathbf{H}} \mathbf{H}(\mathbf{y}) \tag{5.5.4}$$

which is equivalent to

$$\mathbb{E}\,(\mathbf{x} - \mathbf{z}_*)\,\mathbf{y}_k = 0 \qquad k = 1, 2, \dots \tag{5.5.5}$$

where $\mathbf{y}_k$, $k = 1, 2, \dots$ are generators of $\mathbf{H}(\mathbf{y})$. This is called the Orthogonality Principle*.*

Proof. Let $\mathbf{z}_* \in \mathbf{H}(\mathbf{y})$ satisfy (5.5.4) and let $\mathbf{z}$ be an arbitrary element of $\mathbf{H}(\mathbf{y})$. Since $\mathbf{z}_* - \mathbf{z}$ belongs to $\mathbf{H}(\mathbf{y})$, by (5.5.4), one has:

$$\begin{aligned} \|\mathbf{x}-\mathbf{z}\|_{\mathbf{H}}^2 = & \quad \|\mathbf{x}-\mathbf{z}_*+\mathbf{z}_*-\mathbf{z}\|_{\mathbf{H}}^2 = \|\mathbf{x}-\mathbf{z}_*\|_{\mathbf{H}}^2 \\ & + 2\,\langle \mathbf{x}-\mathbf{z}_*, \mathbf{z}_*-\mathbf{z}\rangle_{\mathbf{H}} + \|\mathbf{z}_*-\mathbf{z}\|_{\mathbf{H}}^2 \\ = & \quad \|\mathbf{x}-\mathbf{z}_*\|_{\mathbf{H}}^2 + \|\mathbf{z}_*-\mathbf{z}\|_{\mathbf{H}}^2 \end{aligned} \tag{5.5.6}$$

and this expression is clearly minimal when $\mathbf{z} = \mathbf{z}_*$.
Conversely, assume that $\mathbf{z}_* \in \mathbf{H}(\mathbf{y})$ minimizes $\|\mathbf{x}-\mathbf{z}\|_{\mathbf{H}}^2$. Let $\mathbf{e}_i := \mathbf{y}_i/\|\mathbf{y}_i\|$ and define the variable

$$\mathbf{z}_i := \mathbf{z}_* + \langle \mathbf{x}-\mathbf{z}_*,\, \mathbf{e}_i\rangle_{\mathbf{H}}\, \mathbf{e}_i$$

Compute $\|\mathbf{x}-\mathbf{z}_i\|_{\mathbf{H}}^2$ to find

$$\begin{aligned} \|\mathbf{z}_i-\mathbf{x}\|_{\mathbf{H}}^2 = & \quad \|\mathbf{z}_*-\mathbf{x}\|_{\mathbf{H}}^2 + 2\langle(\mathbf{z}_*-\mathbf{x}),\, \langle \mathbf{x}-\mathbf{z}_*,\, \mathbf{e}_i\rangle_{\mathbf{H}}\mathbf{e}_i\rangle_{\mathbf{H}} + |\langle \mathbf{x}-\mathbf{z}_*,\, \mathbf{e}_i\rangle_{\mathbf{H}}|^2 \\ = & \quad \|\mathbf{z}_*-\mathbf{x}\|_{\mathbf{H}}^2 - 2\langle \mathbf{z}_*-\mathbf{x},\, \mathbf{e}_i\rangle_{\mathbf{H}}\langle \mathbf{z}_*-\mathbf{x},\, \mathbf{e}_i\rangle_{\mathbf{H}} + |\langle \mathbf{x}-\mathbf{z}_*\,\mathbf{e}_i\rangle|_{\mathbf{H}}^2 \\ = & \quad \|\mathbf{x}-\mathbf{z}_*\|_{\mathbf{H}}^2 - |\langle \mathbf{x}-\mathbf{z}_*,\, \mathbf{e}_i\rangle_{\mathbf{H}}|^2 \end{aligned}$$

and note that, by assumption $\|\mathbf{z}_i-\mathbf{x}\|_{\mathbf{H}}^2 \geq \|\mathbf{z}_*-\mathbf{x}\|_{\mathbf{H}}^2$ so that $\langle \mathbf{x}-\mathbf{z}_*\,\mathbf{e}_i\rangle_{\mathbf{H}}$ must be zero; i.e.

$$\mathbf{x}-\mathbf{z}_* \perp_{\mathbf{H}} \mathbf{y}_i\,; \qquad i = 1, 2, \ldots\,.$$

which is the orthogonality condition of (5.5.5) . Note that uniqueness follows directly from (5.5.6). For letting $\mathbf{z}_1 \in \mathbf{H}(\mathbf{y})$ to be another minimum of $\|\mathbf{x}-\mathbf{z}\|_{\mathbf{H}}^2$, would imply that $\|\mathbf{x}-\mathbf{z}_1\|_{\mathbf{H}}^2 = \|\mathbf{x}-\mathbf{z}_*\|_{\mathbf{H}}^2$ and (5.5.6) with $\mathbf{z} = \mathbf{z}_1$ implies $\|\mathbf{z}_*-\mathbf{z}_1\|_{\mathbf{H}}^2 = 0$. □

Remark 5.1. *Naturally, when* $\mathbf{y}$ *has finite dimension* m, $\mathbf{H}(\mathbf{y})$ *is obviously a* closed *subspace of the Hilbert space* $\mathbf{H}$.

It is to be stressed here that the orthogonality principle continues to be valid also in the case when $\dim \mathbf{H}(\mathbf{y}) = \infty$, *when the subspace* $\mathbf{H}(\mathbf{y})$ *is generated by an infinite family of random variables* $\{\mathbf{y}_\alpha\}$, *provided the subspace* $\mathbf{H}(\mathbf{y})$ *generated by this family is a closed subspace of* $\mathbf{H}$. *For more details see Halmos book [43, Theorem 1, p. 23].*

Theorem 5.4 has the same intuitive geometric interpretation which was illustrated for the deterministic linear least squares problem in Sect. 2.1 of the previous Chapter. See Fig. 5.5.2.

By imposing the orthogonality condition (5.5.5) in our finite-dimensional setting, one sees that a randpom variable in $\mathbf{H}(\mathbf{y})$,

$$\mathbf{z} = \sum_1^m a_i\, \mathbf{y}_i = a^\top \mathbf{y}$$

is the orthogonal projection of $\mathbf{x}$ onto $\mathbf{H}(\mathbf{y})$ if and only if the vector $a = [a_1 \ldots a_m]^\top$ satisfies the linear equations

$$\mathbb{E}\,\mathbf{x}\mathbf{y}_i = \mathbb{E}\; a^\top \mathbf{y}\; \mathbf{y}_i \quad , \qquad i = 1, \ldots, m$$

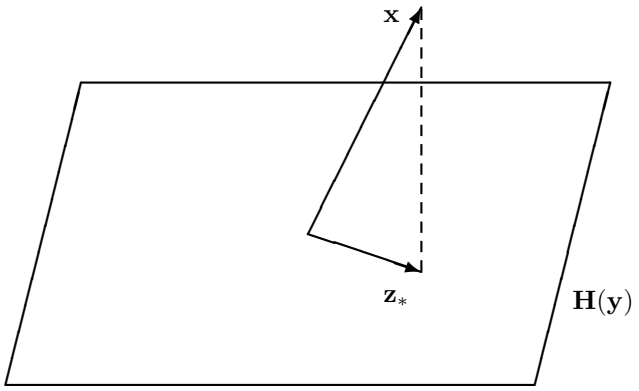


Figure 5.5.2. *The Orthogonality Principle.*

namely $a^\top \mathbb{E}\,\mathbf{y}\mathbf{y}^\top = \mathbb{E}\,\mathbf{x}\mathbf{y}^\top$ which can be written in matrix notation as

$$a^\top \Sigma_{\mathbf{y}} = \sigma_{\mathbf{xy}} \tag{5.5.7}$$

where $\sigma_{\mathbf{xy}} = [\,\sigma_{\mathbf{x}\mathbf{y}_1} \ldots \sigma_{\mathbf{x}\mathbf{y}_m}\,]$ is the row vector of cross covariances of $\mathbf{x}$ e $\mathbf{y}$. If $\Sigma_{\mathbf{y}} > 0$ equation (5.5.7) can be solved , getting

$$a_*^\top = \sigma_{\mathbf{xy}}\ \Sigma_{\mathbf{y}}^{-1} \tag{5.5.8}$$

and the minimum variance linear estimator of $\mathbf{x}$ based on the observation vector $\mathbf{y}$ (both zero-mean) is

$$\mathbf{z}_* = a_*^\top \mathbf{y} = \sigma_{\mathbf{xy}}\ \Sigma_{\mathbf{y}}^{-1}\,\mathbf{y} \quad . \tag{5.5.9}$$

When $\mathbf{x}$ is also a vector-valued, say n-dimensional, the formula above can be applied to each component $\mathbf{x}_k\,; k = 1, \ldots, n$ separately since the minimization of the sum

$$\sum_{k=1}^{n} \|\mathbf{z}_k - \mathbf{x}_k\|_{\mathbf{H}}^2 = \mathbb{E}\left(\sum_{1}^{n} (\mathbf{z}_k - \mathbf{x}_k)^2\right) = \mathbb{E}\,\|\mathbf{z} - \mathbf{x}\|^2 \tag{5.5.10}$$

is accomplished only when each term $\|\mathbf{z}_k - \mathbf{x}_k\|^2$ is minimized separately. Using a matrix notation for n-dimensional random vectors with components in $\mathbf{H}(\mathbf{y})$,

$$\mathbf{z} = A\mathbf{y} \quad , \qquad A \in \mathbb{R}^{n\times m}$$

it follows that the rows of the optimal A are obtained from (5.5.8) so that

$$A_* = \Sigma_{\mathbf{xy}}\ \Sigma_{\mathbf{y}}^{-1}$$

where $\Sigma_{\mathbf{xy}}$ is now the $n \times m$ cross covariance matrix of $\mathbf{x}$ and $\mathbf{y}$. The m.v. linear estimator is then

$$\hat{\mathbf{x}}(\mathbf{y}) = \Sigma_{\mathbf{xy}}\ \Sigma_{\mathbf{y}}^{-1}\ \mathbf{y} \quad . \tag{5.5.11}$$

Introducing the mean values in this expression, we obtain the general expression of the linear m.v. estimator.

Proposition 5.4. *In case of arbitrary mean values, the minimum variance estimator of* $\mathbf{x}$ *based on the observation* $\mathbf{y}$ *is the affine function*

$$\hat{\mathbf{x}}(\mathbf{y}) = \mu_{\mathbf{x}} + \Sigma_{\mathbf{xy}}\, \Sigma_{\mathbf{y}}^{-1}\, (\mathbf{y} - \mu_{\mathbf{y}}) \quad . \tag{5.5.12}$$

This function depends only on the joint first and second order moments of $\mathbf{x}$ *e* $\mathbf{y}$.

This expression is clearly the same that was obtained for Gaussian variables. This should come as no-surprise if only we knew a priori that in the Gaussian case the conditional expectation was a *linear function* of the data since then it could have been derived by the same argument exposed above. The highly non obvious fact is that for Gaussian random variables, among all measurable functions of the data, $g : \mathbb{R}^m \to \mathbb{R}^n$, the one which minimizes the expected mean square deviation of $g(\mathbf{y})$ from the random vector $\mathbf{x}$ happens to be *linear* (or affine).

It follows that the covariance matrix of the residual error $\tilde{\mathbf{x}} := \mathbf{x} - \hat{\mathbf{x}}(\mathbf{y})$ must have the same expression found for the Gaussian case, that is

$$\mathrm{Var}\,(\tilde{\mathbf{x}}) = \Sigma_{\mathbf{x}} - \Sigma_{\mathbf{xy}}\, \Sigma_{\mathbf{y}}^{-1}\, \Sigma_{\mathbf{yx}} \tag{5.5.13}$$

which also depends only on the joint second order moments of the two random vectors $\mathbf{x}$ and $\mathbf{y}$.

Since linear m.v. estimators only depend on the first and second order joint moments, when working in this linear setting there will be no need to ask for more detailed probabilistic information about $\mathbf{y}$ and $\mathbf{x}$. Hence, we shall normally assume that this is the only probabilistic information available for solving the inference problem we are studying. Evidently, whenever the additional information of a joint Gaussian distribution is available the estimator (5.5.12) is not only the best linear function but, indeed, the optimal one *among all possible nonlinear functions of the data*. This is so since a Gaussian distribution is completely determined by its first and second order moments.

Notations: The fact that, when the distributions are Gaussian, the linear m.v. estimator $\hat{\mathbf{x}}(\mathbf{y})$ coincides with the conditional mean $\mathbb{E}\,(\mathbf{x} \mid \mathbf{y})$, has led to a widespread use of the notation

$$\hat{\mathbf{x}}(\mathbf{y}) \equiv \hat{\mathbb{E}}\,(\mathbf{x} \mid \mathbf{y})\,, \tag{5.5.14}$$

where the hatted symbol has been named by Doob [27], the *wide sense conditional mean (or expectation) of* $\mathbf{x}$ given $\mathbf{y}$. The analogy of names is based on the fact that the properties of the two operators $\hat{\mathbb{E}}\,(\mathbf{x} \mid \mathbf{y})$ and $\mathbb{E}\,(\mathbf{x} \mid \mathbf{y})$ are formally very similar at least on a superficial ground. One just needs to substitute the word "independence" with "uncorrelation" (or orthogonality) and "general measurable function" with "linear function" of the data.

Moreover, the true conditional mean $\mathbb{E}\,(\mathbf{x} \mid \mathbf{y})$ can also be defined as an orthogonal projection on a suitable subspace of $\mathbf{H}$, see for example [58, Lemma 2.2.1]. We should however warn the reader that the two settings are quite different and the formal similarity may lead to false conclusions. When the underlying distributions are far from Gaussian it is to be expected that the linear approximation $\hat{\mathbb{E}}\,(\mathbf{x} \mid \mathbf{y})$ of the conditional expectation $\mathbb{E}\,(\mathbf{x} \mid \mathbf{y})$ could very well have no meaning. See the example 5.6 below.

We also need to introduce notations for vector valued random variables (i.e. random vectors). Although strictly speaking the Hilbert space $\mathbf{H}$ is made of

scalar (real valued) random variables, we shall nevertheless use notations like $\mathbf{z} \in \mathbf{H}(\mathbf{y})$, $\mathbf{z} \perp \mathbf{H}(\mathbf{y})$ etc., where $\mathbf{z}$ is a random *vector*, to mean that all components $\{\mathbf{z}_k\}$ of $\mathbf{z}$ belong to, or, are orthogonal to, the subspace $\mathbf{H}(\mathbf{y})$. Similarly, the symbol $\hat{\mathbb{E}}(\mathbf{x} \mid \mathbf{y})$ in case of a vector valued $\mathbf{x}$, is to be interpreted as the vector with components the orthogonal projections $\hat{\mathbb{E}}(\mathbf{x}_k \mid \mathbf{y})$ of the scalar components $\mathbf{x}_k$, $k = 1, \ldots, n$ of $\mathbf{x}$.
Finally we shall reserve the word *Covariance* to mean *cross-covariance* of two random vectors while a matrix such a s $\Sigma_{\mathbf{z}} = \mathbb{E}\,\mathbf{z}\mathbf{z}^\top$ will just be referred to as the *Variance* matrix of $\mathbf{z}$.

The following is a re-statement of the orthogonal projection lemma in vector notations. It will be useful for further reference.

Corollary 5.1. *Let $\mathbf{x}$ and $\mathbf{y}$ be second order random vectors of respective dimensions n and m; then the linear m.v. estimator $\hat{\mathbb{E}}(\mathbf{x} \mid \mathbf{y})$ is the uinque vector $\mathbf{z} \in \mathbf{H}(\mathbf{y})$ such that*

$$\mathbb{E}\,(\mathbf{x} - \mathbf{z})\,\mathbf{y}^\top = 0 \quad . \tag{5.5.15}$$

The random vector $\hat{\mathbb{E}}(\mathbf{x} \mid \mathbf{y})$ has the smallest error variance matrix among all linear functions of $\mathbf{y}$, that is

$$\mathbb{E}\,\{[\mathbf{x} - \hat{\mathbb{E}}(\mathbf{x} \mid \mathbf{y})]\,[\mathbf{x} - \hat{\mathbb{E}}(\mathbf{x} \mid \mathbf{y})]^\top\} \leq \mathbb{E}\,[(\mathbf{x} - \mathbf{z})\,(\mathbf{x} - \mathbf{z})^\top] \tag{5.5.16}$$

for every n-dimensional vector $\mathbf{z} \in \mathbf{H}(\mathbf{y})$. The inequality $A \leq B$ among symmetric matrices is to be understood to mean that $B - A$ is positive semidefinite.

Proof. Obviously (5.5.15) is just the orthogonality principle (5.5.5), written in matrix notation.

To prove (5.5.16) let $\mathbf{z} = A\mathbf{y}$ be any n-dimensional linear function of $\mathbf{y}$. The relative error variance has the expression

$$\begin{aligned}\mathrm{Var}\,(\mathbf{x} - \mathbf{z}) &= \mathrm{Var}\,(\mathbf{x} - \hat{\mathbb{E}}(\mathbf{x} \mid \mathbf{y}) + \hat{\mathbb{E}}(\mathbf{x} \mid \mathbf{y}) - A\mathbf{y}) \\ &= (\Sigma_{\mathbf{x}} - \Sigma_{\mathbf{xy}}\,\Sigma_{\mathbf{y}}^{-1}\,\Sigma_{\mathbf{yx}}) + (A - \Sigma_{\mathbf{xy}}\,\Sigma_{\mathbf{y}}^{-1})\,\Sigma_{\mathbf{y}}\,(A^\top - \Sigma_{\mathbf{y}}^{-1}\,\Sigma_{\mathbf{yx}})\end{aligned}$$

which evidently has the minimum for $A = \Sigma_{\mathbf{xy}}\,\Sigma_{\mathbf{y}}^{-1}$. □

Block-diagonalization of Symmetric Positive Definite matrices

The estimation error $\tilde{\mathbf{x}} := \mathbf{x} - \hat{\mathbb{E}}\,[\mathbf{x} \mid \mathbf{y}]$ is orthogonal to $\mathbf{y}$; hence $\mathrm{Var}\,\{\begin{bmatrix}\mathbf{y} \\ \tilde{\mathbf{x}}\end{bmatrix}\}$ is block-diagonal

$$\mathrm{Var}\,\{\begin{bmatrix}\mathbf{y} \\ \tilde{\mathbf{x}}\end{bmatrix}\} = \begin{bmatrix}\Sigma_{\mathbf{y}} & 0 \\ 0 & \Lambda\end{bmatrix}, \quad \Lambda = \Sigma_x - \Sigma_{xy}\,\Sigma_y^{-1}\,\Sigma_{yx}.$$

In matrix language the block Λ is called the **Schur Complement** of $\Sigma_{\mathbf{y}}$ in Σ. The order here is immaterial; one can exchange $\tilde{\mathbf{x}}$ with $\mathbf{y}$.
Generalization:

Lemma 5.1. *Let*

$$X = \begin{bmatrix}A & B^\top \\ B & D\end{bmatrix} \in \mathbb{R}^{(n+m)\times(n+m)}$$

be symmetric positive definite (a covariance matrix). If A is invertible, then

$$\begin{bmatrix} I & 0 \\ -BA^{-1} & I \end{bmatrix} X \begin{bmatrix} I & -A^{-1}B^\top \\ 0 & I \end{bmatrix} = \begin{bmatrix} A & 0 \\ 0 & D - BA^{-1}B^\top \end{bmatrix}$$

Proof. via Bayesian Estimation Theory. Think of A as $\Sigma_{\mathbf{y}}$, D as $\Sigma_{\mathbf{x}}$ and B as $\Sigma_{\mathbf{xy}}$. □

Therefore:
1. X is positive definite if and only if A and $S := D - BA^{-1}B^\top$ are positive definite.
2. If A is invertible, X is positive semi-definite if and only if A is positive definite and $S := D - BA^{-1}B^\top$ is positive semi-definite.

Problem 5.4. *Show, only assuming X symmetric, that,*
1. X is positive definite if and only if D and $S := A - BD^{-1}B^\top$ are positive definite.
2. If D is invertible, X is positive semi-definite if and only if D is positive definite and $S := S := A - B^\top D^{-1}B$ is positive semi-definite.

One can generalize the bock-diagonalization formulas to non-necessarily symmetric nor positive definite matrices. Consider a square block matrix

$$X = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \in \mathbb{R}^{(n+m)\times(n+m)}$$

Definition 5.3. *Assume that A is non-singular. The matrix $S_1 := D - CA^{-1}B$ is called* Schur complement *of A in X.*
Assume that D is non-singular. The matrix $S_2 := A - BD^{-1}C$ is called Schur complement *of D in X.*

Proposition 5.5. *Assume that A is invertible. Then X is invertible if and only if also the Schur complement $D - CA^{-1}B$ is invertible.*
Dually, if D is invertible X is invertible if and only if also the Schur complement $A - BD^{-1}C$ is invertible.

Proof. If A is non-singular, we have,

$$X \begin{bmatrix} I & -A^{-1}B \\ 0 & I \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} I & -A^{-1}B \\ 0 & I \end{bmatrix} = \begin{bmatrix} A & 0 \\ C & D - CA^{-1}B \end{bmatrix},$$

$$\begin{bmatrix} I & 0 \\ -CA^{-1} & I \end{bmatrix} \begin{bmatrix} A & 0 \\ C & D - CA^{-1}B \end{bmatrix} = \begin{bmatrix} A & 0 \\ 0 & D - CA^{-1}B \end{bmatrix}$$

so that

$$\begin{bmatrix} I & 0 \\ -CA^{-1} & I \end{bmatrix} X \begin{bmatrix} I & -A^{-1}B \\ 0 & I \end{bmatrix} = \begin{bmatrix} A & 0 \\ 0 & D - CA^{-1}B \end{bmatrix}$$

and the first statement follows since the two block-triangular matrices are both invertible. The second statement is proven in the same way. □

Then, if A and $D - CA^{-1}B$ are non-singular, X^{-1} is given by

$$\begin{bmatrix} I & -A^{-1}B \\ 0 & I \end{bmatrix} \begin{bmatrix} A & 0 \\ 0 & D - CA^{-1}B \end{bmatrix}^{-1} \begin{bmatrix} I & 0 \\ -CA^{-1} & I \end{bmatrix} = \\ = \begin{bmatrix} I & -A^{-1}B \\ 0 & I \end{bmatrix} \begin{bmatrix} A^{-1} & 0 \\ 0 & (D - CA^{-1}B)^{-1} \end{bmatrix} \begin{bmatrix} I & 0 \\ -CA^{-1} & I \end{bmatrix} .$$

Therefore:

$$X^{-1} = \begin{bmatrix} A^{-1} + A^{-1}B(D - CA^{-1}B)^{-1}CA^{-1} & -A^{-1}B(D - CA^{-1}B)^{-1} \\ -(D - CA^{-1}B)^{-1}CA^{-1} & (D - CA^{-1}B)^{-1} \end{bmatrix} .$$

Dually, if D and $A - BD^{-1}C$ are non-singular, then

$$X^{-1} = \begin{bmatrix} (A - BD^{-1}C)^{-1} & -(A - BD^{-1}C)^{-1}BD^{-1} \\ -D^{-1}C(A - BD^{-1}C)^{-1} & D^{-1} + D^{-1}C(A - BD^{-1}C)^{-1}BD^{-1} \end{bmatrix}$$

From these formulas the famous *Sherman-Morrison-Woodbury* formula, more commonly called the **Matrix Inversion Lemma** follows:

Lemma 5.2. *If A, D, and one of the Schur complements $(A - BD^{-1}C)$ and $(D - CA^{-1}B)$ is non-singular, then the other Schur complement is non-singular and*

$$(A - BD^{-1}C)^{-1} = A^{-1} + A^{-1}B(D - CA^{-1}B)^{-1}CA^{-1} \tag{5.5.17}$$

This formula is useful when the"perturbation" $BD^{-1}C$ of A has low rank, in particular when D is a scalar (1×1) matrix. It is used for example in recursive least squares and in Kalman filtering.

5.6 ▪ The linear model

Let us consider observation models (5.1.4) which are *linear*; that is, assume that the observation vector $\mathbf{y}$ is related to the unknown variable $\mathbf{x}$ by a linear relation of the type

$$\mathbf{y} = S\mathbf{x} + \mathbf{w} \tag{5.6.1}$$

where S is a known (deterministic) matrix, $\mathbf{x}$ and $\mathbf{w}$ are *uncorrelated random vectors* of respective variances, $P := \mathrm{Var}\,(\mathbf{x})$ and $R := \mathrm{Var}\,(\mathbf{w})$ also assumed to be known. For notational simplicity we shall initially assume that $\mathbf{x}$ and $\mathbf{w}$ have zero mean. The model (5.6.1) is widely used to represent measurements obtained by a linear sensor, which are corrupted by additive (non-observable) noise $(\mathbf{w})$. We want to compute the best linear estimator of $\mathbf{x}$ based on the measurement $\mathbf{y}$. To this end we shall rely on formula (5.5.12) which requires knowledge of the variance matrices $\Sigma_{\mathbf{xy}}$ and $\Sigma_{\mathbf{y}}$.
These matrices can be computed directly from the parameters of the model (5.6.1). Let us first note that, the orthogonality of $\mathbf{x}$ and $\mathbf{w}$, implies that $\Sigma_{\mathbf{yx}}$ is readily obtained by right multiplication of (5.6.1) by $\mathbf{x}$ and taking expectation, as

$$\Sigma_{\mathbf{yx}} = S\,\Sigma_{\mathbf{x}} = SP$$

whence

$$\Sigma_{\mathbf{xy}} = PS^{\top} \quad . \tag{5.6.2}$$

It then follows that

$$\Sigma_{\mathbf{y}} = SPS^{\top} + R \tag{5.6.3}$$

which is certainly positive definite ($\Sigma_{\mathbf{y}} > 0$) if $R > 0$, that is there are no "perfect measurements" which may be physically justifiable in most circumstances. The linear m.v. estimator of $\mathbf{x}$ given $\mathbf{y}$ is then given by

$$\hat{\mathbb{E}}\,(\mathbf{x} \mid \mathbf{y}) = PS^{\top}(SPS^{\top} + R)^{-1}\,\mathbf{y} \quad . \tag{5.6.4}$$

The residual error variance, denoted Λ, follows readily from the general formula (5.5.13) and is given by

$$\Lambda = P - PS^{\top}(SPS^{\top} + R)^{-1}\,SP \quad . \tag{5.6.5}$$

Remark 5.2. One may ask how general is the model (5.6.1) assuming knowledge of the joint second order moments of the vectors $\mathbf{x}$ and $\mathbf{y}$. It is actually not hard to see that any linear estimation problem initially formulated in terms of the joint statistics of $\mathbf{x}$ and $\mathbf{y}$ can be phrased as an estimation problem on a linear model of the form (5.6.1).

Let us just represent $\mathbf{y}$ as the sum of its orthogonal projection onto the subspace $\mathbf{H}(\mathbf{x})$ and an error term $\tilde{\mathbf{y}} = \mathbf{y} - \hat{\mathbb{E}}\,(\mathbf{y} \mid \mathbf{x})$, say

$$\mathbf{y} = \hat{\mathbb{E}}\,(\mathbf{y} \mid \mathbf{x}) + \tilde{\mathbf{y}} \tag{5.6.6}$$

Since $\hat{\mathbb{E}}\,(\mathbf{y} \mid \mathbf{x})$ is a linear function of $\mathbf{x}$ it can be written as $S\mathbf{x}$ for some matrix S. In fact, if $\Sigma_{\mathbf{x}} > 0$ S can actually be expressed by the formula

$$S = \Sigma_{\mathbf{yx}}\,\Sigma_{\mathbf{x}}^{-1} \quad .$$

Further identify $\mathbf{w}$ with the error term $\tilde{\mathbf{y}}$ which is uncorrelated with $\mathbf{x}$ by the orthogonality principle. Then (5.6.6) is formally identical to (5.6.1). When $\Sigma_{\mathbf{x}} > 0$ the variance R, of the noise term can be expressed by the formula $\Sigma_{\mathbf{y}} - \Sigma_{\mathbf{yx}}\,\Sigma_{\mathbf{x}}^{-1}\,\Sigma_{\mathbf{xy}}$.
This construction parallels that described at the end of Section 5.3 for the derivation of the posterior density in case of jointly Gaussian variables. ♢

There is an alternative expression of the formulas (5.6.4) and (5.6.5) which is more transparent and easier to compute, especially in certain special cases which are common in the applications. One such special case occurring for example when R is a diagonal matrix. The derivation of these alternative formulas uses the Matrix Inversion Lemma 5.2.

Theorem 5.5. *Assume that the a priori variance matrix P of the random vector $\mathbf{x}$ in the model* (5.6.1) *is invertible. Then the linear m.v. estimator of $\mathbf{x}$ can also be expressed in the form*

$$\hat{\mathbb{E}}\,(\mathbf{x} \mid \mathbf{y}) = \mu_{\mathbf{x}} + (P^{-1} + S^{\top}R^{-1}S)^{-1}\,S^{\top}R^{-1}(\mathbf{y} - \mu_{\mathbf{y}}) \tag{5.6.7}$$

and the relative error variance matrix as

$$\Lambda = (P^{-1} + S^{\top}R^{-1}S)^{-1} \quad . \tag{5.6.8}$$

Proof. Formula (5.6.8) is immediately obtained from (5.6.5) just by setting $A = P^{-1}$, $B = S^\top$ and $C = R^{-1}$ in the the Matrix Inversion Lemma formula (5.5.17).

The expression (5.6.7), can be obtained by the following sequence of steps

$$\begin{aligned}\hat{E}(\mathbf{x} \mid \mathbf{y}) &= PS^\top \left[R^{-1} - R^{-1}\, S(S^\top R^{-1} S + P^{-1})^{-1}\, S^\top R^{-1}\right] \mathbf{y} \\ &= \left[P - PS^\top R^{-1}\, S(S^\top R^{-1} S + P^{-1})^{-1}\right]\, S^\top R^{-1} \mathbf{y} \\ &= \left[P(S^\top R^{-1}\, S + P^{-1}) - PS^\top R^{-1} S\right](S^\top R^{-1} S + P^{-1})^{-1}\, S^\top R^{-1}\, \mathbf{y}\end{aligned}$$

and noting that the last term between square brackets is the identity. □

Formulas (5.6.7) and (5.6.8) show very clearly the influence of the a priori variance of $\mathbf{x}$ on the estimate. Roughly speaking, when the variance P is very large; i.e. the a priori knowledge of $\mathbf{x}$ is very uncertain, P^{-1} can be neglected in comparison to the other addend $S^\top R^{-1} S$ and formula (5.6.7), assuming S is of full column rank, reduces to

$$\hat{\mathbb{E}}\,(\mathbf{x} \mid \mathbf{y}) = (S^\top R^{-1} S)^{-1}\, S^\top R^{-1} \mathbf{y} \tag{5.6.9}$$

which is the weighted least-squares estimate of classical parametric Statistics (2.3.6) with weighting matrix equal to R^{-1}, which has variance

$$\Lambda = (S^\top R^{-1} S)^{-1} \quad . \tag{5.6.10}$$

Examples of use of these formulas will be seen later.

5.7 ▪ Linear Models and Marginal Gaussians

In a Bayesian setting the density of a random variable $\mathbf{y}$ having mean θ and variance σ^2 is written as a conditional density $p(y \mid \mathbf{x} = \theta)$ where $\mathbf{x}$ is another random variable, so that the **joint density** of $\mathbf{y}$ and $\mathbf{x}$ is

$$p_{\mathbf{y},\mathbf{x}}(y, \theta) = p(y \mid \mathbf{x} = \theta) p_{\mathbf{x}}(\theta) \tag{5.7.1}$$

It is clear that if $\mathbf{y} \sim \mathcal{N}(\theta, \sigma^2)$ and $\mathbf{x} \sim \mathcal{N}(\mu, \tau^2)$ then the joint density is again Gaussian. Often one needs to compute the *marginal distribution* of $\mathbf{y}$ which could formally be obtained by integrating with respect to θ (of course this distribution does not depend on θ any more). The calculation in terms of density functions even in the Gaussian case is quite complicated but for Gaussian variables there is a very simple way to do this by using linear models.

We can express both $\mathbf{y}$ and $\mathbf{x}$ by means of two linear models as

$$\mathbf{y} = \mathbf{x} + \mathbf{e}\,, \qquad \mathbf{x} = \mu + \mathbf{w}$$

where $\mathbf{e} \sim \mathcal{N}(0, \sigma^2)$ is a random variable *independent of* $\mathbf{x}$ and $\mathbf{w} \sim \mathcal{N}(\mu, \tau^2)$ is another random variable of mean μ and variance τ^2. The multiplicative relation (5.7.1) implies that $\mathbf{e}$ and $\mathbf{w}$ must be **independent** (show this). Therefore from

$$\mathbf{y} = \mu + \mathbf{w} + \mathbf{e}\,,$$

one can easily conclude that $\mathbf{y}$ has mean μ and variance $\sigma^2 + \tau^2$. In other words $\mathbf{y} \sim \mathcal{N}(\mu, \sigma^2 + \tau^2)$. Naturally here because of Gaussianness one can substitute everywhere the world "independent" with "uncorrelated".

Problem 5.5. *Compute the marginal distribution of an N-dimensoinal random vector $\mathbf{y}$ described by*

$$p(y \mid \mathbf{x} = \theta) \equiv \mathcal{N}(S\theta, \sigma^2 I_N) \,,$$

where $\mathbf{x}$ has the prior distribution $\mathbf{x} \sim \mathcal{N}(\theta_0, \tau^2 I_p)$.

5.8 ▪ Factor Analysis Models

A (static) *factor model* (or *Factor Analysis model*) is a representation

$$\mathbf{y} = A\mathbf{x} + \mathbf{w} \tag{5.8.1}$$

of m observed variables $\mathbf{y} = [\,\mathbf{y}_1 \,\dots\, \mathbf{y}_m\,]^\top$, having zero-mean and finite variance, as linear combinations of n *common factors* $\mathbf{x} = [\,\mathbf{x}_1 \,\dots\, \mathbf{x}_n\,]^\top$, plus uncorrelated "noise" or "error" terms $\mathbf{w} = [\,\mathbf{w}_1 \,\dots\, \mathbf{w}_m\,]^\top$. The m components of the error vector $\mathbf{w}$ should be zero-mean and mutually uncorrelated random variables, i.e.,

$$\Sigma_{\mathbf{xw}} := \mathbb{E}\,\{\mathbf{x}\mathbf{w}^\top\} = 0, \tag{5.8.2a}$$

$$\Delta := \mathbb{E}\,\{\mathbf{w}\mathbf{w}^\top\} = \operatorname{diag}\{\sigma_1^2, \dots, \sigma_m^2\}. \tag{5.8.2b}$$

The big difference with the standard linear model (5.6.1) is that all quantities in the right hand side of (5.8.1) are **unknown** and should be estimated from the data vector $\mathbf{y}$. This makes Factor Analysis models much harder to estimate and in fact to understand. The introduction of models of this type was initially motivated by mathematical psychology and goes back to the beginning of the twentieth century [**?**]. Their purpose is to provide an explanation of the mutual dependence of the components of an m-dimensional observed random vector $\mathbf{y}$ in terms by an (hopefully) small number n of **common factors** $\mathbf{x}$. To see that this could indeed be the scope, let $a_i^\top$ be the i-th row of the matrix A and set

$$\hat{\mathbf{y}}_i := a_i^\top \mathbf{x} = \hat{\mathbb{E}}\,[\,\mathbf{y}_i \mid \mathbf{x}\,]\,; \tag{5.8.3}$$

then one has exactly

$$\mathbb{E}\,\{\mathbf{y}_i \mathbf{y}_j\} = \mathbb{E}\,\{\hat{\mathbf{y}}_i \hat{\mathbf{y}}_j\} \tag{5.8.4}$$

for all $i \neq j$. This just means that the Bayes estimates of the components of $\mathbf{y}$ in terms of the factor vector have the same mutual correlations as the components of the observed vector $\mathbf{y}$ itself. This property is equivalent to

$$\langle \mathbf{w}_i,\, \mathbf{w}_j \rangle = \langle \mathbf{y}_i - \hat{\mathbf{y}}_i,\, \mathbf{y}_j - \hat{\mathbf{y}}_j \rangle = 0, \quad i \neq j\,.$$

The property (5.8.4) is called *conditional uncorrelation, (or conditional orthogonality)* of the random variables $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_m$ given $\mathbf{x}$ [58]. It is rather easy to see that $\mathbf{y}$ admits a representation of the type (5.8.1) if and only if $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_m$ are conditionally orthogonal given $\mathbf{x}$. Note that this property is really a property of the subspace of random variables linearly generated by the components of the vector $\mathbf{x}$, namely

$$\mathbf{X} := \{a^\top \mathbf{x} \mid a \in \mathbb{R}^n\}, \tag{5.8.5}$$

which could be called the *factor subspace* of the model. One could say that the components of $\mathbf{y}$ are conditionally orthogonal given $\mathbf{X}$. The estimates $\hat{\mathbf{y}}_i$ are then just the orthogonal projections $\hat{\mathbf{y}}_i = \hat{\mathbb{E}}\,[\,\mathbf{y}_i \mid \mathbf{X}\,]$, $i = 1, 2, \ldots, m$.

Generally $n \ll m$ and the matrix A is tall. Therefore, introducing a matrix $A^\perp$ such that $A^\perp A = 0$ one can eliminate the factors from the model (5.8.1) to obtain an *external description* of the form

$$\mathbf{y} = \hat{\mathbf{y}} + \mathbf{w}, \quad A^\perp \hat{\mathbf{y}} = 0, \tag{5.8.6}$$

in terms of "true" variables $\hat{\mathbf{y}}$ (which are still not observed) and additive errors $\mathbf{w}$. Solving the relation $A^\perp \hat{\mathbf{y}} = 0$ with respect of one of the true variables you get an explicit linear **Error-In-Variables** (EIV) model of the class (2.2.10a). Hence linear EIV models can be understood as the result of elimination of the latent variable $\mathbf{x}$ from a FA model.

A factor subspace may be unnecessarily large just because it carries redundant random variables which are uncorrelated (i.e., orthogonal) to the vector $\mathbf{y}$ to be represented. This redundancy can be eliminated by imposing a *non-redundancy condition*. Set

$$\hat{\mathbf{X}} = \text{span}\{\hat{\mathbb{E}}\,[\mathbf{y}_i \mid \mathbf{X}\,];\, i = 1, 2, \ldots, m\} := \hat{\mathbb{E}}\,[\mathbf{Y} \mid \mathbf{X}\,] \tag{5.8.7}$$

or, equivalently $\hat{\mathbf{X}} = \text{span}\{(A\mathbf{x})_i;\ i = 1, 2, \ldots, m\}$. Then $\mathbf{X}$ is non-redundant or *minimal* if $\mathbf{X} = \hat{\mathbf{X}}$. It can easily be shown that an arbitrary factor space $\mathbf{X}$ can always be substituted by its non-redundant subspace $\hat{\mathbf{X}}$ preserving the conditional orthogonality property. So, without loss of generality one can assume that the condition $\mathbf{X} = \hat{\mathbf{X}}$ is satisfied.

Any set of generating variables for $\mathbf{X}$ can serve as a common factor vector. In particular it is no loss of generality to choose the generating vector $\mathbf{x}$ as a normalized basis in $\mathbf{X}$, i.e.,

$$\mathbb{E}\,\{\mathbf{x}\mathbf{x}^\top\} = I, \tag{5.8.8}$$

which we shall do in the following. The dimension $n = \dim \mathbf{x} = \dim \mathbf{X}$ will be called the *rank* of the model. Obviously, by virtue of the non-redundancy condition $\mathbf{X} = \hat{\mathbf{X}}$, we automatically have $\text{rank}\ A = n$ for a model of rank n, i.e., A will always be left-invertible.

Two factor models for the same observable vector $\mathbf{y}$ whose factors span the same subspace $\mathbf{X}$ will be regarded as *equivalent*. Hence, with the imposed notational conventions, the factor vectors of two equivalent factor models are related by multiplication by a real orthogonal matrix.

The common factors are *nonobservable* quantities (also called **latent variables** in the literature) which, although representing the same output variables $\mathbf{y}$, could in principle be chosen in many different ways giving rise to representations (i.e., models) with different properties and of a different complexity. In applications one would like to have models with $n \ll m$ and possibly have some idea about the minimal possible number of factors necessary to represent $\mathbf{y}$. Models with a minimal number of factors correspond to factor subspaces $\mathbf{X}$ of minimal dimension. These models will be called *minimal* henceforth.

A rather disturbing fact is that there are in general many (in fact infinitely many) minimal factor subspaces for a given family of observables $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_m$. Hence there are in general many nonequivalent minimal factor models (with normalized factors) representing a fixed m-tuple of random variables $\mathbf{y}$. As a

trivial example, choose for each $k = 1, 2, \ldots, m$, the $(m-1)$-dimensional vector $\mathbf{x} := [\,\mathbf{y}_1 \ldots \mathbf{y}_{k-1}\ \mathbf{y}_{k+1} \ldots \mathbf{y}_m\,]^\top$ as a factor vector, then one obtains m "extremal" models, called *elementary regressions*, of the form

$$\begin{cases} \mathbf{y}_1 = [\,1 \ldots 0\,]\,\mathbf{x} + 0 \\ \vdots \\ \mathbf{y}_k = \hat{a}_k^\top \mathbf{x} \ + \ \mathbf{w}_k \\ \vdots \\ \mathbf{y}_m = [\,0 \ldots 1\,]\,\mathbf{x} + 0 \end{cases} \tag{5.8.9}$$

where $\hat{a}_k^\top = \mathbb{E}\{\mathbf{y}_k \mathbf{x}^\top\}\mathbb{E}\{\mathbf{x}\mathbf{x}^\top\}^{-1}$. Note that in each elementary regression model there is just one nonzero element in the error variance matrix Δ. Clearly, the elementary regression (5.8.9) corresponds to an EIV model with errors affecting only the k-th true variable.

In this example the factor subspaces are spanned by $m-1$ observable variables. A subspace $\mathbf{X}$ contained in the *data space* $\mathbf{Y} := \text{span}\{\mathbf{y}_1 \ldots \mathbf{y}_m\}$ (i.e., generated by linear functions of $\mathbf{y}$) is called *internal*. Accordingly, factor models whose factor $\mathbf{x}$ is a linear functionals of $\mathbf{y}$ are called *internal* models. Elementary regressors are internal models.

Identifiability. The inherent nonuniqueness of factor models brings up the question of which model one should use in identification. This is called *factor indeterminacy* (or *unidentifiability*) in the literature, and the term is usually referred to as parameter unidentifiability as in these models there are always "too many" parameters to be estimated. It may be argued that once a model (in essence, a factor subspace) is selected, it can always be parametrized in a one-to-one (and hence identifiable) way. The difficulty seems to be more a question of understanding the properties of the different possible models, i.e., a question of *classification*. Unfortunately, the classification of all possible (minimal) factor subspaces and an explicit characterization of minimality is, to the present time, not fully understood.

We shall address here only very superficially the question of identifiability. To this end, we need to consider the additive decomposition of the covariance matrix $\Sigma := \mathbb{E}\{\mathbf{y}\mathbf{y}^\top\}$ of the observables induced by a factor model, namely

$$\Sigma = AA^\top + \Lambda \tag{5.8.10}$$

where Λ is a diagonal matrix with positive entries and $A \in \mathbb{R}^{m\times n}$ with $n \le m$ is full column rank. This is called a *Factor Analysis decomposition* of Σ. The rank of A is also called the rank of the decomposition.

Note that for any fixed diagonal matrix Λ such that $\text{rank}\,\{\Sigma - \Lambda\} = n$, the matrix A in the decomposition (5.8.10) is just a full rank factor of $\Sigma - \Lambda$. Such a factor can be rendered unique by choosing an appropriate canonical form in the equivalence class[15] of $A \in \mathbb{R}_*^{m\times n}$ defined modulo right multiplication by $n \times n$ orthogonal matrices. Hence a canonical factor of rank $n < m$ is uniquely determined by a choice of the m positive numbers in $\Lambda = \text{diag}\,\{\lambda_1^2, \ldots, \lambda_m^2\}$ such that

$$\text{rank}\ \{\Sigma - \Lambda\} = n\,.$$

[15] We denote by $\mathbb{R}_*^{m\times n}$ the space of *full rank* $m \times n$ real matrices.

Hidden rank

The following is the main question concerning identifiability. *What is the minimal n for which a given positive definite variance Σ admits a factor analysis decomposition of rank n.* This number, $n_*(\Sigma)$, (sometimes denoted $mr(\Sigma)$ in the literature) is called the *hidden rank* of Σ. Clearly $n_*(\Sigma) \leq m-1$ for all Σ. Trivially a diagonal Σ admits a (unique) factor analysis decomposition of rank zero. Conditions for Σ to admit a factor analysis decomposition of rank one have been known since the beginning of the 20th century. In the literature a positive definite covariance matrix admitting a factor analysis decomposition of rank one is called a *Spearman matrix*.

The decompositions of rank $m-1$ are particularly simple to describe. In fact, the solutions are described in terms of the coordinates $(\lambda_1^2, \ldots, \lambda_m^2)$ in the space $\mathbb{R}_+^m$ (that is of nonnegative definite diagonal matrices $\Lambda = \operatorname{diag}\{\lambda_1^2, \ldots, \lambda_m^2\}$) by the polynomial equation $\det(\Sigma - \Lambda) = 0$. This algebraic equation, in analogy to what was found in (2.2.12), defines a smooth hypersurface (an hyperboloid with concavity facing the origin) in the positive orthant of $\mathbb{R}^m$. This hypersurface intersects the k-th coordinate axis exactly at the value λ_k^2 equal to the error variance of the k-th elementary regressor. Hence the m elementary regressors are in a sense "extremal" solutions of the FA decomposition problem of rank $m-1$.

In general, to the equation $\det(\Sigma - \Lambda) = 0$ one must couple the additional constraints that all minors of order $m-1, \ldots, m-n+1$ of the symmetric matrix $\Sigma - \Lambda$ should be zero. This also defines an algebraic surface intersecting $\mathbb{R}_+^m$ which is a subset of the hyperboloid mentioned above. If $n > n_*(\Sigma)$ there are in general many equivalent FA decompositions of Σ of rank n.

One may ask whether there may be a unique such decomposition and, in particular, if $n = n_*(\Sigma)$ implies uniqueness.

There was a popular conjecture that if

$$n \leq \frac{2m + 1 - (8m+1)^{1/2}}{2}$$

then the model would be locally identifiable; that is the set of algebraic equations above would have, for a generic Σ, a *unique solution* $\{\lambda_1^2, \ldots, \lambda_m^2\}$. In other words there would be a *unique* minimal FA model representing Σ. The upper bound is known as the *Ledermann bound*; the inequality is actually equivalent to $m + n \leq (m-n)^2$. A derivation of the bound and some relevant references can be found in [8]. However A. Shapiro in [87] and [88] shows that the bound is not universal and points out some counterexamples.

The general hidden rank question is unsolved. One may however estimate $n_*(\Sigma)$ by minimizing with respect to $\{\lambda_1^2, \ldots, \lambda_m^2\}$ the trace of $\Sigma - \Lambda$ which is interpreted as a convex surrogate of the rank, see [16] for an exhaustive bibliography.

Example 5.3. Consider the linear FA model

$$\mathbf{y} := \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} \mathbf{x} + \begin{bmatrix} \mathbf{w}_1 \\ \mathbf{w}_2 \end{bmatrix}$$

where $\mathbf{x}, \mathbf{w}_1, \mathbf{w}_2$ are zero-mean *uncorrelated* scalar random variables having variances

$$\operatorname{var}\{\mathbf{x}\} = 1, \quad \operatorname{var}\{\mathbf{w}_1\} = \sigma^2, \quad \operatorname{var}\{\mathbf{w}_2\} = \sigma^2$$

but the three scalar parameters

$$\theta = [a_1,\ a_2,\ \sigma^2\,]\,, \quad \sigma^2 > 0\,.$$

are all **unkown**. Assume that the (joint) covariance matrix Σ of $\mathbf{y}_1$, $\mathbf{y}_2$ is given and positive definite. Show that the model is identifiable and that you can uniquely compute its parameters from the spectral decomposition of Σ:

$$\Sigma = \begin{bmatrix} u_1 & u_2 \end{bmatrix} \begin{bmatrix} \lambda_1^2 & 0 \\ 0 & \lambda_2^2 \end{bmatrix} \begin{bmatrix} u_1 & u_2 \end{bmatrix}^\top$$

where u_1, u_2 are the normalized eigenvectors and $\lambda_1^2 > \lambda_2^2$.

Solution: We need to solve for a decomposition $\Sigma = aa^\top + \sigma^2 I$; which means that we must solve for a and σ^2 the two equations

$$a(a^\top u_i) = (\lambda_i^2 - \sigma^2)\, u_i\,, \qquad i = 1, 2$$

Take $\sigma^2 = \lambda_2^2$ so that $a^\top u_2 = 0$, which means that a must be parallel to u_1 (the eigenvector which is orthogonal to u_2). Then write $a = \alpha u_1$ and substitute in the other equation to get $\alpha^2 = \lambda_1^2 - \lambda_2^2$. In conclusion we have the unique solution (modulo sign of the square root)

$$a = u_1 \sqrt{(\lambda_1^2 - \lambda_2^2)} \quad \text{and} \quad \sigma^2 = \lambda_2^2\,.$$

The model is therefore identifiable.

A faster but more abstract solution is to consider $\Sigma - \lambda_2^2 I$ which must be positive semidefinite and of rank one. □

5.9 ▪ Comparison of the Bayesian and the ML estimators

It is instructive to compare the formulas of the MV estimator for the Bayesian linear model (5.6.1) with those derived in Chap. 2 for the ML (or Markov) estimators. For ease of comparison we shall assume that in the Fisherian setting the scalar parameter σ^2 is known and included in the noise variance matrix R and that the a priori covariance P is invertible. The estimators with the two approaches are compared below:

$$\text{BAYES} \quad \begin{cases} \hat{x}(\mathbf{y}) &= (S^\top R^{-1} S + P^{-1})^{-1}\, S^\top R^{-1} \mathbf{y} \\ \Lambda &= (S^\top R^{-1} S + P^{-1})^{-1} \end{cases} \tag{5.9.1}$$

$$\begin{matrix} \text{FISHER} \\ \text{(MARKOV)} \end{matrix} \quad \begin{cases} \hat{\theta}(\mathbf{y}) &= [S^\top R^{-1} S]^{-1}\, S^\top R^{-1} \mathbf{y} \\ \Sigma &= [S^\top R^{-1} S]^{-1}\,. \end{cases} \tag{5.9.2}$$

One can immediately note that when $P \to \infty$ (meaning that the prior information about $\mathbf{x}$ becomes more and more vague) the Bayesian formulas coincide with the Fisherian counterparts. It is also evident that for any P one has

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

since

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

hence (not surprisingly) the Bayesian estimator has always smaller variance than that of the Fisher estimator which is derived in the absence of any a priori information about $\mathbf{x}$. Therefore the Fisherian formulas are a limit case of the Bayesian expressions. This is however true only in a restricted sense. For *regularized* linear least squares problems we have the followiing remarkable fact:

Theorem 5.6. *The ridge estimator for a linear model $\mathbf{y} = S\theta + \mathbf{w}$, with a general quadratic penalty term on θ defined by a symmetric positive definite matrix W,*

$$\hat{\theta}_R(y) := \operatorname{Arg}\min_{\theta} \left\{ \|y - S\theta\|^2_{R^{-1}} + \lambda \|\theta\|^2_W \right\} \tag{5.9.3}$$

is the Bayes MV estimator of $\mathbf{x}$ for the model (5.6.1) *where the inverse of the a priori variance matrix is given by $P^{-1} = \lambda W$.*

This correspondence is based on formula (3.3.4). It has far reaching consequences for the interpretation and solution of many linear ill-posed problems. The interested reader may want to look at the book [101].

As we have seen the Bayesian estimator has always smaller (or equal) variance of that of the Fisher estimator. One may then wonder about the bias. It is immediate to see that the linear Bayesian estimator $\hat{x}(\mathbf{y})$ in (5.9.1) **can never be unbiased** (in the uniform sense). In fact, computing the mean with respect to the conditional density $f(y \mid x) = \mathcal{N}(Sx, R)$ (now the analog of $p_\theta(y)$ in the Fisherian setting) of $\hat{x}(\mathbf{y})$ one gets

$$\mathbb{E}\left(\hat{x}(\mathbf{y}) \mid \mathbf{x} = x\right) = \left(S^\top R^{-1} S + P^{-1}\right)^{-1} S^\top R^{-1} S x$$

which cannot be identically equal to Ix unless $P^{-1} = 0$.

There is room here for some additional remarks. The first concerns the expression of the error covariance Λ, in general given by the difference

$$\Lambda = \Sigma_x - \Sigma_{xy}\, \Sigma_y^{-1}\, \Sigma_{yx} \quad , \tag{5.9.4}$$

which is always at least positive semidefinite since it must be a covariance matrix. The term $\Sigma_{xy}\, \Sigma_y^{-1}\, \Sigma_{yx}$, which is subtracted off (Σ_x) is actually *the estimator variance* $\operatorname{Var}\ \hat{\mathbf{x}}$ so that one may say that a good estimator should have a large variance, the closer to the a priori variance of $\mathbf{x}$ the better. This may look paradoxical since it says that in the Byesian setting a good estimator should have a large variance. This is so since the sample values of $\hat{\mathbf{x}}$ should not be concentrated about the mean (as in the classical philosophy) but rather approximate as much as possible those of the variable $\mathbf{x}$ making $\mathbb{E}\ \|\mathbf{x} - \hat{x}(\mathbf{y})\|^2$ as small as possible.

It is also instructive to ask if there could be a Fisherian interpretation of the Bayesian estimation formulas. To this end, imagine that the Markov estimate could be obtained by introducing a fictitious additional measurement say

$$\mathbf{y}_0 = S_0 x + \mathbf{w}_0 \quad , \tag{5.9.5}$$

which is available prior to the actual measurements described by the standard model

$$\mathbf{y} = Sx + \mathbf{w}\,.$$

In this model $\mathbf{w}_0$ and $\mathbf{w}$ are uncorrelated and now we interpret $\mathbf{x}$ as an unknown n-dimensional deterministic *parameter*.

Let us choose S_0 and $R_0 := \mathrm{Var}\,\{\mathbf{w}_0\}$ in such a way that

$$P = (S_0^\top\, R_0^{-1}\, S_0)^{-1}\,, \tag{5.9.6}$$

thereby making P equal to the variance of the Markov estimator $\hat{x}(\mathbf{y}_0)$ for the model (5.9.5).

Write the linear Fisher estimator of x given the augmented observation vector $(\mathbf{y}_0, \mathbf{y})$. Since $\mathbf{y}_0$ and $\mathbf{y}$ are uncorrelated we obtain

$$\begin{aligned}\hat{x}(\mathbf{y}_0, \mathbf{y}) &= \hat{x}(\mathbf{y}_0) + PS^\top(R + SPS^\top)^{-1}\ [\mathbf{y} - S\hat{x}(\mathbf{y}_0)] \quad ,\\ \Sigma &= P - PS^\top(R + SPS^\top)^{-1}\, SP\,. \end{aligned} \tag{5.9.7}$$

Note that the second formula is equivalent (via the Matrix Inversion Lemma) to the expression of the Bayesian error covariance Λ in (5.6.8) but the first one coincides with the Bayes estimator if and only if the variance of $\hat{x}(\mathbf{y}_0)$ is equal to zero so that it could be interpreted as a constant mean value. But by assumption this variance is not zero. Hence there cannot be an additional measurement doing the job.

Of course this is reasonable; in the Bayesian setting the a priori information is *not an additional sample measurement* but only concerns the probability distribution of $\mathbf{x}$. As we have seen the two things are not equivalent.

5.10 ▪ Examples

Example 5.4 (Direct sensing).

This is the simplest estimation problem which may for example occur when measuring physical quantities like length, mass, resistance etc. In this case one has direct measurements of the unknown (scalar) variable $\mathbf{x}$ plus noise so that the function h in (**??**) is the identity. In other words, letting $\mathbf{y}_i$ represent the i-th measurement and assuming that $\mathbf{x}$ is real valued, one can model the observations as

$$\mathbf{y}_i = \mathbf{x} + \mathbf{w}_i \quad , \qquad i = 1, \ldots, n \tag{5.10.1}$$

where the $\mathbf{w}_i$'s are assumed to be zero-mean mutually uncorrelated random variables representing "accidental errors". We may allow the experimental apparatus to change from one trial to another so that the error (or noise) variances may be different.

To set up the experiment and actually to be able to choose the right apparatus, one should always know a nominal value, say x_0, of the variable under measurement. Quite often there is also enough information about the range of possible variations of $\mathbf{x}$ about its nominal value to guess a confidence interval about it of the form of an interval $x_0 \pm \Delta x$ where the measured value will fall with some fixed probability, say 95% . Assuming a Gaussian distribution one may estimate its standard deviation $\sigma_{\mathbf{x}}$ by using a formula of the type

$$n\,\sigma_{\mathbf{x}} = \Delta x$$

where n may be equal to 2 or 3. A similar assessment can be made for the standard deviation of the measurement errors based on the specifications of the

measurement instruments. Usually in this case the confidence interval amplitude, Δw, is a certain fixed percentage say 1 or .5 % of the maximum reading of the instrument. Hence we may well assume that in the model (5.10.1), $P \equiv \sigma_{\mathbf{x}}^2$ is known and, assuming that the measurement errors are mutually uncorrelated the overall noise variance $R_{i,j} = \sigma_i^2 \delta_{i,j}$; $i = 1, \ldots, n$ with each σ_i^2 estimated by a similar reasoning as for $\sigma_{\mathbf{x}}^2$, so that

$$R = \text{diag}\,\{\sigma_1^2, \ldots, \sigma_n^2\}\,.$$

It is normally reasonable to assume that $\mathbf{x}$ and $\mathbf{w}$ are uncorrelated. As seen already in Example 5.2, the model (5.10.1) can then be written in vector form as

$$\mathbf{y} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \mathbf{x} + \mathbf{w}$$

with $\mu_{\mathbf{x}} = x_0$, $P = \sigma_{\mathbf{x}}^2$ and R as above.

Applying formula (5.6.7) we find

$$\hat{\mathbb{E}}\,(\mathbf{x} \mid \mathbf{y}) = x_0 + \left(\frac{1}{\sigma_{\mathbf{x}}^2} + \sum_1^n \frac{1}{\sigma_i^2} \right)^{-1} \sum_1^n \frac{1}{\sigma_i^2} \,(\mathbf{y}_i - x_0) \quad .$$

When $\sigma_i^2 = \sigma^2\,;\, i = 1 \ldots n$; i.e. when the n measurements all have the same precision, this expression reduces to

$$\hat{\mathbb{E}}\,(\mathbf{x} \mid \mathbf{y}) = x_0 + \frac{1}{\frac{\sigma^2}{\sigma_{\mathbf{x}}^2} + n} \sum_1^n (\mathbf{y}_i - x_0)$$

which, when $\sigma^2/\sigma_{\mathbf{x}}^2 \ll n$ is just the sample mean. The variance of the estimate is

$$\lambda^2 = \sigma_{\mathbf{x}}^2 \left(1 + \sum_1^n \frac{\sigma_{\mathbf{x}}^2}{\sigma_i^2} \right)^{-1} \quad .$$

which, whenever $\sigma_{\mathbf{x}}^2$ is very large, reduces to

$$\lambda^2 \cong \left(\sum_1^n \frac{1}{\sigma_i^2} \right)^{-1}$$

which is just equal to σ^2/n when all measurements have the same precision.

In practice the computed value of λ is used to express the uncertainty on the estimate $\hat{x} = \hat{\mathbb{E}}\,(\mathbf{x} \mid \mathbf{y} = y)$, by reporting the result of the n measurements in the form of a confidence interval centered on $\hat{x}$; for example $x = \hat{x} \pm 2\lambda$. When the variables are Gaussian (or approximately such) one may say that x lies, with a probability of about 95%, in the interval $[\,\hat{x} - 2\lambda,\ \hat{x} + 2\lambda\,]$. ♢

Example 5.5 (Delay Estimation).

This is a common problem occurring in radar/sonar applications where the delay in receiving a reflected wave (an "echo") signal is twice the distance L of

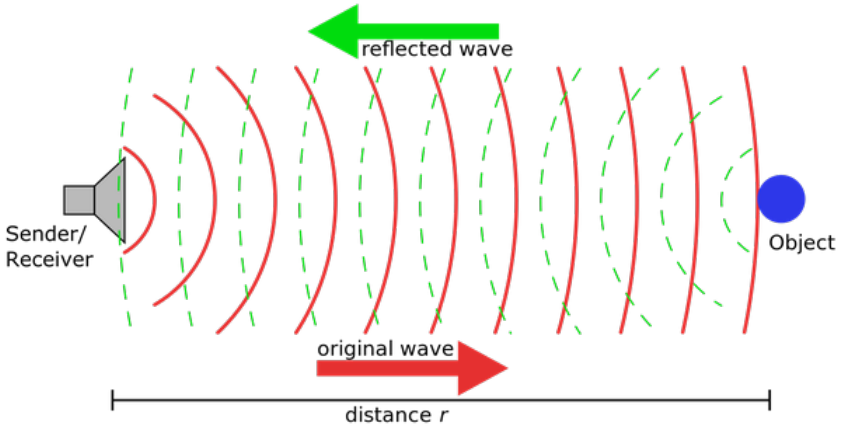


Figure 5.10.1. *Radar*

the object from the source instrument divided by the speed of light (or sound) c. A sinusoidal waveform sent at time zero, after reflexion from the target is received back with a time delay $2L/c$ so that the received signal is,

$$y(t) = A\ \sin(\Omega t - \boldsymbol{\theta}) + \mathbf{w}(t) \quad , \qquad t \in \mathbb{R} \tag{5.10.2}$$

where t is continuous time, Ω is the fixed carrier angular frequency, A is a known normalized amplitude; and

- $\boldsymbol{\theta}$ is the phase delay: $\boldsymbol{\theta} = 2\pi\dfrac{2L}{\lambda}$, where L is the distance and $\lambda = \dfrac{\Omega}{c}$ the wavelength, which we shall model as a random variable uniformly distributed on the interval $[0, 2\pi]$;[16]
- The signal at the receiver is actually sampled with period T_s and made into a finite sequence of samples of lenght N so that, setting $t = T_s k$, the actual measurement is

  $$y(k) = A\ \sin(\omega_0 k - \boldsymbol{\theta}) + \mathbf{w}(k) \quad , \qquad k = 1, \ldots, N \tag{5.10.3}$$

  where $\omega_0 := \Omega T_s$.
- Here $\mathbf{w}(k)$ is additive noise assumed i.i.d., independent of $\boldsymbol{\theta}$ and of known variance σ^2.

The problem we are facing is to estimate the phase delay $\boldsymbol{\theta}$. This problem is actually non linear but can be addressed by the linear estimation methodology which we have discussed in this chapter, by introducing a fictitious two-dimensional variable

$$\mathbf{x} = [\cos\boldsymbol{\theta}\ \ -\sin\boldsymbol{\theta}]^\top \tag{5.10.4}$$

[16] Since by periodicity we can discard the largest integer included in the wave number $\dfrac{2L}{\lambda}$.

and by rewriting (5.10.2), using the trigonometric identity

$$\sin(\omega_0 k - \boldsymbol{\theta}) = \sin\omega_0 k \;\cos\boldsymbol{\theta} - \cos\omega_0 k \;\sin\boldsymbol{\theta}\,.$$

In this way we get the standard linear model $\mathbf{y} = S\mathbf{x} + \mathbf{w}$ where

$$\mathbf{y} = \begin{bmatrix} y(1) \\ \dots \\ y(N) \end{bmatrix}, \quad S = A \begin{bmatrix} \sin\omega_0 & \cos\omega_0 \\ \dots & \dots \\ \sin\omega_0 N & \cos\omega_0 N \end{bmatrix}, \quad \mathbf{w} = \begin{bmatrix} \mathbf{w}(1) \\ \dots \\ \mathbf{w}(N) \end{bmatrix} \quad . \tag{5.10.5}$$

The first and second order moments of $\mathbf{x}$ are readily obtained from the distribution of $\boldsymbol{\theta}$. They turn out to be

$$\mu_{\mathbf{x}} = 0 \quad , \qquad P = \begin{bmatrix} 1/2 & 0 \\ 0 & 1/2 \end{bmatrix} \quad ; \tag{5.10.6}$$

while $R = \sigma^2 I_N$, where I_N is the $N \times N$ identity matrix. Substituting in the formulas (5.6.7) and (5.6.8) one gets

$$S^\top R^{-1} S = \frac{A^2}{\sigma^2} \begin{bmatrix} \sum_1^N \sin^2\omega_0 k & \sum_1^N \sin\omega_0 k\,\cos\omega_0 k \\ \sum_1^N \sin\omega_0 k\,\cos\omega_0 k & \sum_1^N \cos^2\omega_0 k \end{bmatrix}$$

$$S^\top R^{-1}\mathbf{y} = \frac{A}{\sigma^2} \begin{bmatrix} \sum_1^N \mathbf{y}(k)\,\sin\omega_0 k \\ \sum_1^N \mathbf{y}(k)\,\cos\omega_0 k \end{bmatrix}.$$

From these formulas we can obtain a simple asymptotic expression for the estimator. Rewriting (5.6.7) as

$$\hat{\mathbb{E}}\,(\mathbf{x} \mid \mathbf{y}) = \left[\frac{1}{N}\,(P^{-1} + S^\top R^{-1} S)\right]^{-1} \;\frac{1}{N}\,S^\top R^{-1}\mathbf{y} \tag{5.10.7}$$

and letting $N \to \infty$, assuming $\omega_0 \neq k\pi$ for k integer, we have

$$\frac{1}{N}\sum_1^N \sin^2\omega_0 k \to \frac{1}{2}\,, \quad \frac{1}{N}\sum_1^N \sin\omega_0 k\,\cos\omega_0 k \to 0\,, \quad \frac{1}{N}\sum_1^N \cos^2\omega_0 k \to \frac{1}{2}\,, \tag{5.10.8}$$

which lead to the asymptotic expression ($\mu_{\mathbf{x}} = [0\;0]^\top$)

$$\hat{\mathbb{E}}\,(\mathbf{x} \mid \mathbf{y}) \cong \frac{2}{A} \begin{bmatrix} \frac{1}{N}\sum_1^N \mathbf{y}(k)\,\sin\omega_0 k \\ \frac{1}{N}\sum_1^N \mathbf{y}(k)\,\cos\omega_0 k \end{bmatrix} \tag{5.10.9}$$

which can be shown to converge to the parameter (5.10.4). In fact, by independence, the law of large numbers implies that as $N \to \infty$, $\frac{1}{N}\sum_1^N \mathbf{w}(k)\sin\omega_0 k \to \mathbb{E}\,\mathbf{w}(k)\sin\omega_0 k = 0$ almost surely (and likewise for the cos function). Then by just substituting the noiseless signal $A\sin(\omega_0 k - \boldsymbol{\theta})$ in (5.10.9) and using the asymptotic expressions (5.10.8) we can conclude that

$$\lim_{N\to\infty} \frac{2}{A} \begin{bmatrix} \frac{1}{N}\sum_1^N \mathbf{y}(k)\,\sin\omega_0 k \\ \frac{1}{N}\sum_1^N \mathbf{y}(k)\,\cos\omega_0 k \end{bmatrix} = \begin{bmatrix} \cos\boldsymbol{\theta} \\ -\sin\boldsymbol{\theta} \end{bmatrix}$$

almost surely. The asymptotic error variance is

$$\Lambda \cong \frac{2\,\sigma^2}{A^2 N}\, I_2 \quad . \tag{5.10.10}$$

Using a result due to Cramèr, see e.g [30, Theorem 7, p. 45] this formula can be used to compute an asymptotic expression for the variance of the random variable $\boldsymbol{\theta}$ which is expressible as $\arctan \frac{\mathbf{x}_2}{\mathbf{x}_1}$. However we shall not insist further on this point. ♢

Example 5.6 (A case when the linear estimator is meaningless).

Consider the stochastic process $\{\mathbf{x}(t)\} := \{\sin \boldsymbol{\omega} t \ , \ t \in \mathbb{Z}\}$ where $\boldsymbol{\omega}$ is a random variable uniformly distributed on the interval $[-\pi, \pi]$ and assume we want to construct the linear estimator of a variable $\mathbf{x}(t)$ for some fixed t, based on a fixed but arbitrary number of observations $\{\mathbf{x}(s) \ , \ s < t\}$ (or even $s \neq t$).

Since, for $s \neq t$

$$\langle \mathbf{x}(t)\ \mathbf{x}(s) \rangle_{\mathbf{H}} = \frac{1}{2\pi} \int_{-\pi}^{\pi} \sin \omega t\ \sin \omega s\ d\omega = 0$$

the random variables $\mathbf{x}(t)$ and $\{\mathbf{x}(s) \ , \ s < t\}$ are uncorrelated. The linear m.v. estimator of $\mathbf{x}(t)$ given any number of variables $\{\mathbf{x}(s) \ , \ s < t\}$ is then $\mathbb{E}\,\mathbf{x}(t) = 0$ and its error variance is therefore equal to the whole a priori variance of $\mathbf{x}(t)$.

Note on the other hand that $\{\mathbf{x}(t)\}$ satisfies the linear difference equation

$$\mathbf{x}(t) - 2\cos \boldsymbol{\omega}\ \mathbf{x}(t-1) + \mathbf{x}(t-2) = 0\,, \qquad t \in \mathbb{Z}$$

from which one can obtain the angular frequency $\boldsymbol{\omega}$ as

$$\boldsymbol{\omega} = \arccos \left[\frac{\mathbf{x}(t) + \mathbf{x}(t-2)}{2\mathbf{x}(t-1)} \right]$$

and hence we can reconstruct exactly the sample value of $\boldsymbol{\omega}$ corresponding to a sample trajectory of the process $\{\mathbf{x}(t)\}$. This means that we can construct a *non linear estimator* of $\mathbf{x}(t)$ based on the variables $\mathbf{x}(t-1), \mathbf{x}(t-2), \mathbf{x}(t-3)\}$

$$\hat{\mathbf{x}}(t) = \sin \left\{ \arccos \left[\frac{\mathbf{x}(t-1) + \mathbf{x}(t-3)}{2\mathbf{x}(t-2)} \right]\ t \right\}$$

which reconstructs *exactly* the variable $\mathbf{x}(t)$ and obviously has error variance equal to zero. ♢

Example 5.7. Consider the standard linear minimum variance estimator $\hat{\mathbf{x}}(\mathbf{y})$ of the n vector $\mathbf{x}$ in the usual N-dimensional liner model where all variables are zero-mean and $\mathrm{Var}\{\mathbf{w}\} = I$,

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

Is it true that you can also write

$$\mathbf{y} = S\hat{\mathbf{x}}(\mathbf{y}) + \hat{\mathbf{w}}$$

for some other random noise vector $\hat{\mathbf{w}}$ uncorrelated with $\hat{\mathbf{x}}(\mathbf{y})$? What would be the variance matrix of this noise vector?

Solution This would clearly be true in the classical (non-Bayesian) context since in this case $\hat{\mathbf{x}}(\mathbf{y}) = (S^\top S)^{-1} S^\top \mathbf{y}$ and, recalling that $\hat{\mathbf{w}} = \mathbf{y} - S\hat{\mathbf{x}}(\mathbf{y}) = (I - \Pi)\mathbf{y}$ where $\Pi = S(S^\top S)^{-1} S^\top$ is the orthogonal projection onto $\operatorname{span} S$, it is immediate to check that

$$\mathbb{E}\,(I - \Pi)\mathbf{y}\mathbf{y}^\top S(S^\top S)^{-1} = 0\,.$$

This is not necessarily true in the Bayesian case. Recall that $\Sigma_y = SPS^\top + I$ (here we use P for the a priori variance of $\mathbf{x}$) and

$$\hat{\mathbf{w}} := \mathbf{y} - S\Sigma_{x,y}\Sigma_y^{-1}\mathbf{y} = (I - SPS^\top \Sigma_y^{-1})\mathbf{y}\,,$$

so that

$$\begin{aligned}\mathbb{E}\,\hat{\mathbf{w}}\hat{\mathbf{x}}(\mathbf{y})^\top &= (I - SPS^\top\Sigma_y^{-1})\Sigma_y\Sigma_y^{-1}SP = \\ &= SP - SPS^\top\Sigma_y^{-1}SP = SP(I - S^\top\Sigma_y^{-1}SP)\end{aligned}$$

which (assuming P non singular) could be zero only if $S^\top\Sigma_y^{-1}S = P^{-1}$. Using the matrix inversion lemma, this can be rewritten

$$S^\top S - S^\top S(S^\top S + P^{-1})^{-1}S^\top S = P^{-1}$$

which, denoting $V := S^\top S$ (an invertible $n \times n$ matrix) and using the matrix inversion lemma again, leads to

$$(V^{-1} + P)^{-1} = P^{-1}\,,$$

that is, $V^{-1} + P = P$ which can happen only if $V^{-1} = 0$, which means that $\mathbf{x}$ is a deterministic parameter.
This calculation could be done more easily in the scalar case with $N = 1$.

Example 5.8 (Comparing Bias and Variance). Consider the usual linear regression problem with data $\{S \in \mathbb{R}^{N\times p}, \mathbf{y} \in \mathbb{R}^N\}$, which is modeled under the Bayesian setting as a sample from a Gaussian conditional distribution, given the random parameter $\mathbf{x}$:

$$f(y \mid \mathbf{x} = \theta) \equiv \mathcal{N}(S\theta, \sigma^2 I_N)\,,$$

where $\mathbf{x}$ has the prior distribution $\mathbf{x} \sim \mathcal{N}(\theta_0, \tau^2 I_p)$.

We want to compute the Bayes estimate of $\mathbf{x}$ and compare its bias and variance with those of the classical parameter estimate $\hat{\boldsymbol{\theta}}$. You may assume $p = 1$ (scalar parameter and $S \equiv s$ a vector in $\mathbb{R}^N$).

Solution:

Assume $p = 1$ and for the moment $\theta_0 = 0$; then the Bayes estimator is

$$\hat{\mathbf{x}}(\mathbf{y}) = \frac{s^\top \mathbf{y}}{\|s\|^2 + \dfrac{\sigma^2}{\tau^2}}$$

which has expected value

$$\mathbb{E}_\theta \hat{\mathbf{x}}(\mathbf{y}) = \frac{\|s\|^2\theta}{\|s\|^2 + \dfrac{\sigma^2}{\tau^2}}$$

This can be equal to θ only for $\tau^2 \to \infty$. The relative bias is

$$\frac{\theta - \mathbb{E}_\theta \hat{\mathbf{x}}(\mathbf{y})}{\theta} = \frac{\dfrac{\sigma^2}{\tau^2}}{\|s\|^2 + \dfrac{\sigma^2}{\tau^2}} = \frac{1}{1 + \|s\|^2 \dfrac{\tau^2}{\sigma^2}} .$$

The ratio $\|s\|^2 \dfrac{\tau^2}{\sigma^2}$ could be called the *signal to noise (power) ratio*. The higher this ratio the smaller the error.
For $\theta_0 \neq 0$ the bias is

$$\theta - \mathbb{E}_\theta \hat{\mathbf{x}}(\mathbf{y}) = \frac{\dfrac{\sigma^2}{\tau^2}(\theta - \theta_0)}{\|s\|^2 + \dfrac{\sigma^2}{\tau^2}}$$

and we can get the same conclusion normalizing with respect to $\theta - \theta_0$ which is the deviation with respect to the a priori nominal value.

The comparison of the two variances is easily done by the theory exposed above.

5.11 ▪ Bayesian Linear Algebra

We shall start by discussing the operation of *change of basis* in a finite dimensional subspace of $\mathbf{H}$. Assume that the n components of the vector $\mathbf{y} = [\mathbf{y}_1, \ldots, \mathbf{y}_n]^\top$ are linearly independent. Consider n other linearly independent random variables in $\mathbf{H}$, $\mathbf{z}_1 , \ldots , \mathbf{z}_n$, collected in a random vector $\mathbf{z}$, which are such that

$$\mathbf{H}(\mathbf{z}) = \mathbf{H}(\mathbf{y}) \quad . \tag{5.11.1}$$

It is quite obvious that each component $\mathbf{z}$ being a linear function of $\mathbf{y}$, there must be a linear relation between $\mathbf{y}$ and $\mathbf{z}$

$$\mathbf{z} = T\mathbf{y} \tag{5.11.2}$$

where $T \in \mathbb{R}^{N \times N}$ must be non-singular since by (5.11.1) it must also be possible to write $\mathbf{y} = S\mathbf{z}$ with $S \in \mathbb{R}^{n \times n}$, which entails $\mathbf{y} = ST\mathbf{y}$, and by right-multiplying by $\mathbf{y}^\top$ and taking expectation, $\Sigma_\mathbf{y} = ST\,\Sigma_\mathbf{y}$ implies $ST = I$). The components of $\mathbf{z}$ form a new basis in $\mathbf{H}(\mathbf{y})$. It follows from (5.11.2) that T can actually be computed from the joint moments of $\mathbf{z}$ and $\mathbf{y}$, since from $\Sigma_{\mathbf{zy}} = T\,\Sigma_\mathbf{y}$ we obtain

$$T = \Sigma_{\mathbf{zy}}\,\Sigma_\mathbf{y}^{-1} \tag{5.11.3}$$

which is equivalent to saying that $\mathbf{z} = \hat{\mathbb{E}}(\mathbf{z} \mid \mathbf{y})$, that is $\mathbf{z}$ coincides with its orthogonal projection onto $\mathbf{H}(\mathbf{y})$). Likewise, we have

$$\Sigma_\mathbf{z} = T\,\Sigma_\mathbf{y}\,T^\top \quad . \tag{5.11.4}$$

Let now $\mathbf{x}$ be any random vector in $\mathbf{H}$ and let's see how the orthogonal projection operator $\hat{\mathbb{E}}(\mathbf{x} \mid \mathbf{y}) = A\mathbf{y}$ changes by introducing a new basis in $\mathbf{H}$. Let then $\hat{\mathbb{E}}(\mathbf{x} \mid \mathbf{z}) = \hat{A}\mathbf{z}$ where

$$\hat{A} = \Sigma_{\mathbf{xz}}\,\Sigma_\mathbf{z}^{-1} \quad . \tag{5.11.5}$$

From (5.11.2) we have $\Sigma_{\mathbf{xz}} = \Sigma_{\mathbf{xy}}\, T^\top$ and, using (5.11.4), it readily follows that

$$\hat{A} = \Sigma_{\mathbf{xy}}\, \Sigma_{\mathbf{y}}^{-1}\, T^{-1} = AT^{-1} \quad . \tag{5.11.6}$$

Hence, the transformation mapping $\hat{\mathbb{E}}\,(\mathbf{x} \mid \mathbf{y}) = A\mathbf{y}$ to $\hat{\mathbb{E}}\,(\mathbf{x} \mid \mathbf{z})$ induced by the change of basis, simply applies the change of basis transformatio (5.11.2) to the variable $\mathbf{y}$. In other words, the linear map A remains *invariant* and the change of basis is only applied to the argument $\mathbf{y}$.
This property is actually a "wide sense" form of a general property of conditional expectation in Probability Theory. In this setting, the conditional expectation $\mathbb{E}\,(\mathbf{x} \mid \mathbf{y})$, only depends on the *$\sigma$-algebra induced by the random vector* $\mathbf{y}$ and this σ-algebra is the same as that induced by any invertible measurable function $\mathbf{z} = \varphi\,(\mathbf{y})$ of $\mathbf{y}$. In other words, the conditional expectation given $\mathbf{z}$ is obtained by just applying the change of variable formula; i.e.

$$\mathbb{E}\,(\mathbf{x} \mid \mathbf{z}) = [\,\mathbb{E}\,(\mathbf{x} \mid \mathbf{y})]_{\mathbf{y}=\varphi^{-1}(\mathbf{z})} \quad .$$

Problem 5.6.

Let $\mathbf{y}$ be a zero-mean random vectors in $\mathbf{H}$. Again denote $\hat{\mathbb{E}}\,(\mathbf{x} \mid \mathbf{y}) := A\mathbf{y}$ but now assume that $\mathbf{z} := T\mathbf{y} + b$, with T invertible and $b \in \mathbb{R}^n$. Now since $\mathbf{z}$ is not zero mean, $\hat{\mathbb{E}}\,(\mathbf{x} \mid \mathbf{z})$ cannot, strictly speaking, be interpreted as an orthogonal projection. Show that

$$\hat{\mathbb{E}}\,(\mathbf{x} \mid \mathbf{z}) = A\,T^{-1}(\mathbf{z} - b) \quad .$$

so that the m.v. estimator still depends on the centered random variable $\mathbf{z} - b$ and is therefore an orthogonal projection. More generally, $\hat{\mathbb{E}}\,(\mathbf{x} \mid \mathbf{y} + b)$ does not depend on *b. For this reason, even when the mean of $\mathbf{y}$ is not zero, one can safely use the same symbol $\hat{E}(\mathbf{x} \mid \mathbf{y})$ to denote the linear m.v. estimator given $\mathbf{y}$ and the orthogonal projection of $\mathbf{x}$ onto $\mathbf{H}(\mathbf{y})$ whose definition requires instead $\mathbb{E}\,\mathbf{y} = 0$.* ♢

Linear estimation with linearly dependent data

In this section we want to generalize formula (5.5.9) to linearly dependent data; i.e. to the situation when $\Sigma_{\mathbf{y}}$ may be singular.
Let's then assume that $\operatorname{rank}\ \Sigma_{\mathbf{y}} = r$ with $r \leq m$, in which case there are only r components of $\mathbf{y}$ which are linearly independent random variables and let $\mathbf{z} := \bar{T}\mathbf{y},\ \bar{T} \in \mathbb{R}^{r\times m}$ be a r-dimensional basis of $\mathbf{H}(\mathbf{y})$ for example made by extracting r components of $\mathbf{y}$, so that

$$\mathbf{H}(\mathbf{z}) = \mathbf{H}(\mathbf{y})\,. \tag{5.11.7}$$

As we have just seen, since $\mathbf{z}$ and $\mathbf{y}$ span the same Hilbert subspace, it must hold that

$$\hat{\mathbb{E}}\ [\mathbf{x} \mid \mathbf{y}] = \hat{\mathbb{E}}\ [\mathbf{x} \mid \mathbf{z}] = \Sigma_{\mathbf{xz}}\, \Sigma_{\mathbf{z}}^{-1}\, \mathbf{z}. \tag{5.11.8}$$

where $\Sigma_{\mathbf{z}}$ is invertible by construction. Hence, in order to compute $\hat{\mathbb{E}}\ [\mathbf{x} \mid \mathbf{y}]$ we need to find a suitable $\mathbf{z}$ and then apply formula (5.11.8). In more concrete terms, we need a procedure to compute the matrix $\bar{T}$ which selects a basis for $\mathbf{H}(\mathbf{y})$.

Let $V \in \mathbb{R}^{m\times(m-r)}$ be a matrix whose columns form a basis for the nullspace, $\ker \Sigma_{\mathbf{y}}$, of $\Sigma_{\mathbf{y}}$; in other words, $\Sigma_{\mathbf{y}} v = 0$, $v \in \mathbb{R}^m$, if and only if v is a linear combination of the columns of V. Since $\Sigma_{\mathbf{y}}$ is symmetric, $\operatorname{Im}\ \Sigma_{\mathbf{y}} = [\ker \Sigma_{\mathbf{y}}\,]^\perp$, where

Im denotes the range (or image) space. We can therefore produce another matrix, $U \in \mathbb{R}^{m\times r}$ whose columns are a basis for the complementary image space of $\Sigma_{\mathbf{y}}$. In this way the matrix

$$S := [U \mid V]$$

is non-singular by construction. We shall now prove the following lemma.

Lemma 5.3. *The random vectors* $\mathbf{z}$ *which form a basis for* $\mathbf{H}(\mathbf{y})$ *are all expressible as*

$$\mathbf{z} = U^\top \mathbf{y} \tag{5.11.9}$$

where U *is a matrix whose columns form a basis for* $\mathrm{Im}\ \Sigma_{\mathbf{y}}$.

Given any such matrix U*, the linear m.v. estimator of* $\mathbf{x}$ *can be written in the form*

$$\hat{\mathbb{E}}\,[\mathbf{x} \mid \mathbf{y}] = \Sigma_{\mathbf{xy}} U (U^\top \Sigma_{\mathbf{y}} U)^{-1} U^\top \mathbf{y} \tag{5.11.10}$$

where the matrix $U(U^\top \Sigma_{\mathbf{y}} U)^{-1} U^\top$ *does not depend on the particular choice of* U *but only depends on the image space* $\mathrm{Im}\ \Sigma_{\mathbf{y}}$.

Proof. Let $\mathbf{y}_{_U} := U^\top \mathbf{y}$ and $\mathbf{y}_{_V} := V^\top \mathbf{y}$ where U and V are the matrices as defined before. Observe that $\mathbf{y}_{_V} := V^\top \mathbf{y} = 0$ (with probability one). In fact,

$$\Sigma_{\mathbf{y}_{_V}} = \mathbb{E}\,[V^\top \mathbf{y}\mathbf{y}^\top V] = V^\top \mathbb{E}\,[\mathbf{y}\mathbf{y}^\top] V = V^\top \Sigma_{\mathbf{y}} V = 0$$

which just says that the variance of $\mathbf{y}_{_V}$ is zero. It follows in particular, that

$$S^\top \mathbf{y} = \begin{bmatrix} U^\top \\ V^\top \end{bmatrix} \mathbf{y} = \begin{bmatrix} \mathbf{y}_{_U} \\ 0 \end{bmatrix} \tag{5.11.11}$$

Now, since $S^\top$ is non singular, $\mathbf{H}(\mathbf{y}) = \mathbf{H}(S^\top \mathbf{y})$. On the other hand, the last $m - r$ components of $S^\top \mathbf{y}$ are zero so that $\mathbf{H}(S^\top \mathbf{y}) = \mathbf{H}(\mathbf{y}_{_U})$ and hence

$$\mathbf{H}(\mathbf{y}) = \mathbf{H}(\mathbf{y}_{_U})\,.$$

Note that $\Sigma_{\mathbf{y}_{_U}} = U^\top \Sigma_{\mathbf{y}} U$ is positive definite (non singular). For if $w \in \ker \Sigma_{\mathbf{y}_{_U}}$ then $\Sigma_{\mathbf{y}} U w = 0$. But the columns of U belong to the orthogonal complement of $\ker \Sigma_{\mathbf{y}}$, and hence this implies $Uw = 0$ which, given that the columns of U are linearly independent (they are in fact a basis for $[\ker \Sigma_{\mathbf{y}}]^\perp$), implies in turn that $w = 0$. To conclude just note that

$$\Sigma_{\mathbf{xz}}\, \Sigma_{\mathbf{z}}^{-1}\, \mathbf{z} = \mathbb{E}\,\left[\mathbf{x}\mathbf{y}^\top U\right] \left(\mathbb{E}\,[U^\top \mathbf{y}\mathbf{y}^\top U]\right)^{-1} U^\top \mathbf{y}$$

which proves the equality (5.11.10). □

Observe that there are infinitely many ways to choose the matrix U and that to each such choice there corresponds a different $\mathbf{z}$; nevertheless the last member of (5.11.10) is independent of this choice.

The use of formula (5.11.10) requires a preliminary calculation of a basis matrix for $\mathrm{Im}\ \Sigma_{\mathbf{y}}$. An alternative way to compute the estimator is to use the Moore-Penrose pseudoinverse. See Appendix B.2 for a definition and a review of the main properties of this pseudoinverse.

Proposition 5.6. *For a possibly singular Variance matrix* $\Sigma_{\mathbf{y}}$*, the linear m.v. estimator can be expressed as*

$$\hat{\mathbb{E}}\,[\mathbf{x} \mid \mathbf{y}] = \Sigma_{\mathbf{xy}}\, \Sigma_{\mathbf{y}}^{+}\, \mathbf{y}. \tag{5.11.12}$$

where $\Sigma_{\mathbf{y}}^{+}$ is the Moore-Penrose pseudoinverse of $\Sigma_{\mathbf{y}}$.

Proof. By Lemma 5.3, it will be enough to show that $U(U^{\top}\Sigma_{\mathbf{y}}U)^{-1}U^{\top} = \Sigma_{\mathbf{y}}^{+}$ for whatever basis matrix U for the subspace $\mathrm{Im}\ \Sigma_{\mathbf{y}}$. Recall that, by the very definition of V, $V^{\top}\Sigma_{\mathbf{y}} = 0$ and $\Sigma_{\mathbf{y}}V = 0$ and that, without lossof generality we may choose the columns of V and of U in such a way the $S = \begin{bmatrix} V & U \end{bmatrix}$ is an orthogonal matrix. Then by known properties of the pseudoinverse, see (**??**),

$$\Sigma_{\mathbf{y}}^{+} = \left[S^{-\top} \underbrace{\begin{bmatrix} U^{\top} \\ V^{\top} \end{bmatrix}}_{S^{\top}} \Sigma_{\mathbf{y}} \underbrace{[U \mid V]}_{S} S^{-1} \right]^{+} = S \left[\begin{bmatrix} U^{\top} \\ V^{\top} \end{bmatrix} \Sigma_{\mathbf{y}} [U \mid V] \right]^{+} S^{\top} =$$

$$S \begin{bmatrix} U^{\top}\Sigma_{\mathbf{y}}U & 0 \\ 0 & 0 \end{bmatrix}^{+} S^{\top} = [U \mid V] \begin{bmatrix} (U^{\top}\Sigma_{\mathbf{y}}U)^{-1} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} U^{\top} \\ V^{\top} \end{bmatrix} =$$

$$U(U^{\top}\Sigma_{\mathbf{y}}U)^{-1}U^{\top} \tag{5.11.13}$$

which is what we wanted to show. □

Let us note that since the Moore-Penrose pseudoinverse is unique, formula (5.11.12) shows once again that the m.v. estimator does not depend on the choice of the basis $\mathbf{z}$.

When $\Sigma_{\mathbf{y}}$ is singular most numerical statistical packages automatically compute the pseudoinverse.

Problem 5.7. *Let*

$$X = \begin{bmatrix} A & B \\ B^{\top} & D \end{bmatrix} \in \mathbb{R}^{(n+m)\times(n+m)}$$

be symmetric. Prove that X is positive semi-definite if and only if $D \geq 0$, $\ker D \subseteq \ker B$ and the generalized Schur complement *$S := A - BD^{+}B^{\top}$ is positive semidefinite.*

Innovations

When the components $(\mathbf{y}_1, \ldots, \mathbf{y}_m)$ of the observation vector $\mathbf{y}$ are *orthonormal*; i.e. uncorrelated and of unit variance, the variance matrix $\Sigma_{\mathbf{y}}$ is the $m \times m$ identity and the calculation of the m.v. estimator becomes trivial. In fact, in this case formula (5.5.11) reduces to

$$\hat{\mathbb{E}}(\mathbf{x} \mid \mathbf{y}) = \Sigma_{\mathbf{xy}}\, \Sigma_{\mathbf{y}}^{-1}\, \mathbf{y} = \Sigma_{\mathbf{xy}}\, \mathbf{y} \tag{5.11.14}$$

whence the matrix A is simply the covariance $\Sigma_{\mathbf{xy}}$, of $\mathbf{x}$ and $\mathbf{y}$.

A natural idea is then to change basis in $\mathbf{H}(\mathbf{y})$, to get an orthonormal one. We shall describe a procedure, the *Gram-Schmidt orthogonalization* on $\mathbf{H}(\mathbf{y})$, which orthonormalizes the components of $\mathbf{y}$ in a *sequential* fashion. The result of this procedure depends on the ordering of the components and has important generalizations and applications to stochastic processes. In this setting one may interpret the components $\mathbf{y}_1, \ldots, \mathbf{y}_m$ as scalar observations which are made sequentially in time $\mathbf{y}_1$ being the first and $\mathbf{y}_m$ the last one. We shall assume

throughout that $\{\mathbf{y}_1, \ldots, \mathbf{y}_m\}$ are linearly independent so that the variance matrix with elements $\sigma_{ij} = \mathbb{E}(\mathbf{y}_i \mathbf{y}_j)$,

$$\Sigma_{\mathbf{y}} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \ldots & \sigma_{1m} \\ \sigma_{21} & & & \\ \sigma_{m1} & \ldots & \ldots & \sigma_{mm} \end{bmatrix} \quad ,$$

is positive definite (and of course symmetric). We shall extensively use the notation $\mathbf{y}^t := [\mathbf{y}_1, \ldots, \mathbf{y}_t]^\top$; $(t \leq m)$. Evidently $\mathbf{y}^m \equiv \mathbf{y}$.

The Gram-Schmidt Orthonormalization in $\mathbf{H}(\mathbf{y})$ works as follows: set

$$\begin{aligned} &\mathbf{e}_1 := \mathbf{y}_1 \quad ; && \varepsilon_1 := \mathbf{e}_1 / \|\mathbf{e}_1\| \\ &\mathbf{e}_2 := \mathbf{y}_2 - \langle \mathbf{y}_2, \varepsilon_1 \rangle \, \varepsilon_1 \quad ; && \varepsilon_2 := \mathbf{e}_2 / \|\mathbf{e}_2\| \\ &\qquad\qquad \ldots\ldots\ldots\ldots\ldots\ldots && \\ &\mathbf{e}_t := \mathbf{y}_t - \textstyle\sum_{k=1}^{t-1} \langle \mathbf{y}_t, \varepsilon_k \rangle \, \varepsilon_k \quad ; && \varepsilon_t := \mathbf{e}_t / \|\mathbf{e}_t\| \\ &\qquad\qquad t = 2, 3, \ldots, m \quad . && \end{aligned} \tag{5.11.15}$$

then

Theorem 5.7. *The random variables $\{\mathbf{e}_t \ , \ t = 1, 2, \ldots, m\}$ defined by the recursion* (5.11.15) *are one-step ahead linear prediction errors of $\mathbf{y}_t$ based on $\mathbf{y}^{t-1}$, in the sense that*

$$\mathbf{e}_t = \mathbf{y}_t - \hat{\mathbb{E}}(\mathbf{y}_t \mid \mathbf{y}^{t-1}), \qquad t = 1, 2, \ldots, m \tag{5.11.16}$$

and $\{\varepsilon_t \ , \ t = 1, 2, \ldots, m\}$ are the corresponding prediction errors normalized to unit variance.
There is a lower triangular non-singular matrix L_t which, for each t relates ε^t to $\mathbf{y}^t$ by a linear causal relation

$$\mathbf{y}^t = L_t \, \varepsilon^t \tag{5.11.17}$$

"causality" being defined as the subspace equality

$$\mathbf{H}(\varepsilon^t) = \mathbf{H}(\mathbf{y}^t) \quad . \tag{5.11.18}$$

holding for each $t = 1, 2, \ldots, m$. The matrix $L \equiv L_m$ is a lower triangular factor of $\Sigma_{\mathbf{y}}$,

$$\Sigma_{\mathbf{y}} = L L^\top \quad . \tag{5.11.19}$$

Proof. Orthonormality of the sequence $\{\varepsilon_t\}$ can be shown by induction, being trivially true for $t = 1$ and, based on the last relation in (5.11.15) one can argue as follows. Assume $\{\varepsilon^{t-1}\}$ has orthonormal components; then

$$\langle \mathbf{e}_t, \varepsilon_k \rangle = \langle \mathbf{y}_t, \varepsilon_k \rangle - \langle \mathbf{y}_t, \varepsilon_k \rangle = 0 \quad , \qquad \text{for all } \ k < t$$

so that $\langle \varepsilon_t, \varepsilon_k \rangle = 0$ for $k = 1, 2, \ldots, t-1$ and hence ε^t has also orthogonal components. That ε_t is also normalized is obvious.

Next, solving (5.11.15) with respect to $\mathbf{y}^t$ we find

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_t \end{bmatrix} = \begin{bmatrix} \|\mathbf{e}_1\| & 0 & 0 \\ \langle \mathbf{y}_2, \varepsilon_1 \rangle & \|\mathbf{e}_2\| 0 & 0 \\ \vdots & & \\ \langle \mathbf{y}_t, \varepsilon_1 \rangle & \ldots & \|\mathbf{e}_t\| \end{bmatrix} \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_t \end{bmatrix} \tag{5.11.20}$$

which shows the lower triangular structure of the matrix L_t in (5.11.17). Note that the diagonal terms must all be non-zero; i.e. $\|\mathbf{e}_t\| > 0$ for all t, since $\mathbf{e}_t = 0$ would imply $\mathbf{y}_t \in \mathbf{H}(\boldsymbol{\varepsilon}^{t-1}) \subset \mathbf{H}(\mathbf{y}^{t-1})$ in force of (5.11.15) and this is clearly impossibile since the components $\{\mathbf{y}_t\}$ are linearly independent. In conclusion, the transformation (5.11.20) is invertible and this proves the validity of the causality relation (5.11.18).

Since the $\{\boldsymbol{\varepsilon}_k \ , \ k \leq t\}$ are orthonormai, (5.11.14) provides

$$\hat{\mathbb{E}}\ \left[\mathbf{y}_t \mid \boldsymbol{\varepsilon}^{t-1}\right] = \sum_{k=1}^{t-1} \langle \mathbf{y}_t, \boldsymbol{\varepsilon}_k \rangle\ \boldsymbol{\varepsilon}_k$$

which is the change of variables formula (5.11.6) for the projection $\hat{\mathbb{E}}\,[\mathbf{y}_t \mid \mathbf{y}^{t-1}]$.

Finally, (5.11.19) is an immediate consequence of (5.11.17) (written for $t = m$) and of the orthonormality of $\boldsymbol{\varepsilon}^m$. □

The sequence of prediction errors $\{\mathbf{e}_t\}$ defined in (5.11.16) has been named the *innovation sequence* of the observation vector $\mathbf{y}$ while $\boldsymbol{\varepsilon}$ is called the *normalized innovation sequence* of $\mathbf{y}$. This terminology was introduced by Norbert Wiener and Pesi Masani in their famous paper [106] on prediction theory. In a sense the prediction error $\mathbf{e}_t$ or its normalized counterpart $\boldsymbol{\varepsilon}_t$ represent the "new information" brought in by the t-th observation $\mathbf{y}_t$, once the previous $t-1$ preceding observations are available and used for linear prediction. The predictable part (in fact the *linearly predictable part*) of $\mathbf{y}_t$ based on $\mathbf{y}^{t-1}$ being just the m.v. estimate $\hat{\mathbb{E}}\,[\mathbf{y}_t \mid \mathbf{y}^{t-1}]$.

A consequence of the causal equivalence relation is that by solving sequentially the equations (5.11.15), $\boldsymbol{\varepsilon}_t$ can be represented as a linear function of the past and present observations $\mathbf{y}^t$ but also, conversely, $\mathbf{y}_t$ can be expressed as a linear function of past and present normalized (or unnormalized) innovations $\boldsymbol{\varepsilon}^t$. This representation can be computed by extracting the t-th row, say g_t of L^{-1} (which is still lower triangular, see Problem **??**) thereby expressing $\boldsymbol{\varepsilon}_t$ as a causal linear function of $\mathbf{y}^t$

$$\boldsymbol{\varepsilon}_t = g_t \mathbf{y} = g_t\, \mathbf{y}^t$$

since the lower triangular structure of $L^{-1} = [\, g_{t,s}\,]_{t,s=1,\ldots,m}$ implies that $g_{t,t+1} = g_{t,t+2} = \ldots, g_{t,m} = 0$. The property of causal equivalence of the innovation $\boldsymbol{\varepsilon}$ and the obseravtions $\mathbf{y}$, expressed by (5.11.18), is of fundamental importance in prediction theory.

Problem 5.8.
a) Prove that the product of two lower triangular matrices is lower triangular.
c) Prove that the inverse of an invertible lower triangular matrix is lower triangular.
d) Prove that the inverse of a lower triangular matrix of 1's is a lower triangular matrix of 1's.

We shall now briefly look into the question of *uniqueness* of the innovation sequence.

Let then $\bar{\boldsymbol{\varepsilon}} = \{\bar{\boldsymbol{\varepsilon}}_t \ , \ t = 1, \ldots, m\}$ be another orthonormal basis in $\mathbf{H}(\mathbf{y})$. Since $\mathbf{H}(\mathbf{y}) = \mathbf{H}(\boldsymbol{\varepsilon}) = \mathbf{H}(\bar{\boldsymbol{\varepsilon}})$, the vectors $\boldsymbol{\varepsilon}$ and $\bar{\boldsymbol{\varepsilon}}$ are related by a non-singular transformation $Q \in \mathbb{R}^{m \times m}$,

$$\bar{\boldsymbol{\varepsilon}} = Q\boldsymbol{\varepsilon} \quad .$$

Since $\Sigma_{\bar{\boldsymbol{\varepsilon}}} = I$ and $\Sigma_{\boldsymbol{\varepsilon}} = I$, Q must be an orthogonal matrix; i.e.:

$$QQ^\top = I \quad . \tag{5.11.21}$$

Now to be an innovation sequence $\bar{\boldsymbol{\varepsilon}}$ must also satisfy the causal equivalence condition (5.11.18) which clearly implies that Q must have a lower triangular structure. However since Q is orthogonal; i.e. $Q^{-1} = Q^\top$, it follows that Q^{-1} must be at the same time lower and upper triangular, that is a diagonal matrix and hence Q must be a *signature matrix* that is a diagonal matrix with elements ± 1. This means that the innovation sequence is essentially unique. Therefore:

Proposition 5.7. *There is a unique (modulo sign) m-tuple of orthonormal random variables $\{\varepsilon_t\}$ which satisfies the causal equivalence condition (5.11.18).*

Cholesky factorization: As we have seen, the orthonormalization of the observation vector requires a lower triangular factorization (5.11.19) of the variance matrix $\Sigma_{\mathbf{y}}$ (which we shall assume positive definite throughout). Once calculating L the innovation vector is obtained as $\boldsymbol{\varepsilon} = L^{-1}\mathbf{y}$. We shall here address the calculation of L via a popular algorithm called *Cholesky Factorization*.

In general, finding a matrix T such that $I = T\,\Sigma_{\mathbf{y}}\,T^\top$ is the same as finding an invertible $S \in \mathbb{R}^{m\times m}$ such that $SS^\top = \Sigma_{\mathbf{y}}$ and then $T = S^{-1}$. There are many such *square roots* of $\Sigma_{\mathbf{y}}$.

Theorem 5.8. *Let $Q = Q^\top$ be positive definite. There is a* **unique** *lower triangular matrix L, with positive diagonal elements such that $Q = LL^\top$.*

Proof. By induction on the dimension k of Q. Let

$$Q_{k+1} = \begin{bmatrix} Q_k & r \\ r^\top & q \end{bmatrix} \in \mathbb{R}^{(k+1)\times(k+1)}$$

be symmetric and positive definite. Then a factorization $Q_{k+1} = L_{k+1}L_{k+1}^\top$ can hold with lower triangular factors of the form :

$$L_{k+1} = \begin{bmatrix} L_k & 0 \\ \ell^\top & \lambda \end{bmatrix} \qquad L_{k+1}^\top = \begin{bmatrix} L_k^\top & \ell \\ 0 & \lambda \end{bmatrix}$$

if and only if

$$\begin{aligned} L_k\, L_k^\top &= Q_k \\ L_k\, \ell &= r \\ \ell^\top \ell + \lambda^2 &= q\,. \end{aligned}$$

the first of which is true by the inductive hypothesis since Q_k is symmetric and positive definite. The second equation yields $\ell = L_k^{-1}\, r$ which substituted into the third yields a *unique* positive solution λ, since the difference

$$q - \ell^\top \ell = q - r^\top (L_k^\top)^{-1}\, L_k^{-1}\, r = q - r^\top Q_k^{-1}\, r$$

is the *positive* Schur complement of Q_k. In fact

$$q - r^\top Q_k^{-1}\, r = [r^\top Q_k^{-1} \;\; -1] \begin{bmatrix} Q_k & r \\ r^\top & q \end{bmatrix} \begin{bmatrix} Q_k^{-1}\, r \\ -1 \end{bmatrix} \geq 0\,.$$

An equivalent proof of positivity is obtained by interpreting Q as the variance matrix of a zero-mean random vector $\mathbf{y}$ (which is always possible) so that $\lambda^2 = q - r^\top Q_k^{-1} r$ becomes just the error variance of the m.v. estimate of the variable $\mathbf{y}_{k+1}$ based on $\mathbf{y}^k$ (see formula (5.5.13)). This error variance must be positive since the components of $\mathbf{y}^{k+1}$ are lineary independent for all k. □

The factor $L = [\ell_{ij}]$ can be computed by the following algorithm which works sequentially starting from the upper left element $\ell_{1,1} := \sqrt{q_{11}}$ descending and working from left to right.

Algorithm 5.1. *1. The diagonal elements ℓ_{ii} are computed by*

$$\ell_{ii} = \sqrt{q_{ii} - \sum_{j=1}^{i-1} \ell_{ij}^2} \qquad i = 1, \ldots, n\,. \tag{5.11.22}$$

2. Assuming the first $i-1$ rows of L have been computed, the elements of the following i-th row are given by

$$\ell_{ij} = \frac{1}{\ell_{jj}} \left(q_{ij} - \sum_{k=1}^{j-1} \ell_{ik}\, \ell_{jk} \right) \qquad j = 1, \ldots, i-1 \quad (j < i)\,. \tag{5.11.23}$$

This equation requires the elements of the i-th row $\ell_{i1}, \ldots, \ell_{i,j-1}$ and the previous elements $\ell_{j1}, \ldots, \ell_{j,j-1}$ of the j-th row which are all known since the j-th row lies above the i-th.

Note that the algorithm has a recursive structure and can be continued indefinitely as it is independent of the dimension m. It can in fact be applied to perform the orthonormalization of stochastic processes.

Solution of some algebraic equations by Cholesky factorization

Consider an algebraic linear equation

$$Q\mathbf{x} = b \tag{5.11.24}$$

where $Q \in \mathbb{R}^{n \times n}$ is symmetric and positive definite. Equations of this kind are fairly common in Statistics and frequently occur in computations of probabilistic quantities where Q is typically a Variance matrix.

A standard algorithm to solve (5.11.24) goes as follows.

1. Compute the Cholesky factorization

$$Q = LL^\top \quad .$$

2. Define $L^\top \mathbf{x} = z$ (z is an intermediate variable to be determined) and solve the triangular linear system

$$Lz = b \quad .$$

Because of the triangular structure this system can be solved easily starting from the top equation proceeding downward by successive substitutions.

3. Once z is found, solve
$$L^\top \mathbf{x} = z \quad .$$
Again, since $L^\top$ is upper triangular, the calculations can be done by successive substitutions upward starting from the solution of the bottom equation.

In particular, the algorithm can be used to compute the inverse of Q by solving the n linear equations for the columns of $Z := \begin{bmatrix} z_1 & \dots & z_n \end{bmatrix}$, one for each column of the identity matrix on the right hand side of

$$LZ = I\,.$$

This provides $Z = L^{-1}$ (lower triangular) so that

$$Q^{-1} = Z^\top Z \quad .$$

One application of the algorithm is to compute the matrix $A = \Sigma_{\mathbf{xy}}\, \Sigma_{\mathbf{y}}^{-1}$ representing the m.v. estimator of a random vector of $\mathbf{x}$ based on $\mathbf{y}$. Since A is the solution of

$$\Sigma_{\mathbf{xy}} = A\, \Sigma_{\mathbf{y}} \tag{5.11.25}$$

so the transpose of this equation is of the standard form discussed above. In particular one can see that the solution of (5.11.25) corresponding to the solution via the Cholesky factorization $\Sigma_{\mathbf{y}} = LL^\top$ has the structure

$$A = \left(\Sigma_{\mathbf{xy}}\; L^{-T}\right)\; L^{-1} \quad . \tag{5.11.26}$$

This formula has a far reaching generalization to the case where $\mathbf{x}$ and $\mathbf{y}$ are two stationary stochastic processes, as it will be seen in Chapter **??**. The reader is invited to appreciate that the estimator written in the form (5.11.26) does a preliminary "whitening" of the observations $\mathbf{y}$ by calculating the product $L^{-1}\mathbf{y}$ and then uses the cross-covariance of $\mathbf{x}$ and of the innovations process $\boldsymbol{\varepsilon}$; in fact, $\Sigma_{\mathbf{xy}}\, L^{-\top} = \Sigma_{\mathbf{x},\boldsymbol{\varepsilon}}$.

The preliminary whitening of the observation signal is a basic step in the dynamic estimation theory of stochastic processes. See e.g. Chap 4 of [58].

5.12 ▪ Bayesian Hypothesis Testing

In the Bayesian framework the partition of the parameter space is chosen at random by "nature" according to some a priori discrete probability distribution on, say, the finite set $1, 2, \ldots, M$ indexing the subsets (1.4.1). The hypothesis testing problem becomes then the *estimation* of the sample value of a *discrete random variable* which can possibly take M values on which one has an a priori distribution. We are given a complete joint probabilistic description of this random variable and of the observation vector $\mathbf{y}$ so we can compute the a posteriori probability distribution

$$f_k(y) := P\{H_k \mid \mathbf{y} = y\}\,, \qquad k = 1, 2, \ldots . M\,. \tag{5.12.1}$$

A natural estimation criterion in this finite setting is to choose as best estimate the value of k which maximizes the conditional distribution (5.12.1) for each fixed observed sample y. This is actually the same **Maximum A Posteriori** (MAP) estimate discussed in Sect. 5.2 but applied to estimation of a discrete (often just binary-valued) random variable.

Two Simple Hypotheses: Suppose that we need to decide between two hypotheses H_0 and H_1. In the Bayesian setting the two hypotheses are the two possible determination of a binary random variable (the "randomized" parameter). We assume that we know the prior probability distribution of this random variable. That is, we know $P(H_0) = p_0$ and $P(H_1) = p_1$, where of course $p_0 + p_1 = 1$. We observe a N-dimensional random vector $\mathbf{y}$ (the data) and we assume we know the conditional distribution of $\mathbf{y}$ under the two hypotheses, denoted $p(y \mid H_0)$ and $p(y \mid H_1)$.

Using Bayes' rule, we can obtain the posterior probabilities of H_0 and H_1:

$$P(H_0 \mid \mathbf{y} = y) = \frac{p(y \mid H_0)P(H_0)}{p(y)}\,, \qquad P(H_1 \mid \mathbf{y} = y) = \frac{p(y \mid H_1)P(H_1)}{p(y)}\,. \tag{5.12.2}$$

and the decision between H_0 and H_1 is just to accept the hypothesis with the highest posterior probability. As noticed before, this is exactly *maximum a posteriori* estimation and the procedure could be called *MAP test*. Note that one could equivalently say to accept H_1 if the observation y is such that

$$\frac{P(H_1 \mid \mathbf{y} = y)}{P(H_0 \mid \mathbf{y} = y)} \geq 1$$

which is the same as $\dfrac{p(y \mid H_1)P(H_1)}{p(y \mid H_0)P(H_0)} \geq 1$ or, choose H_1 if and only if the Likelihood Ratio satisfies the inequality:

$$\Lambda(y) := \frac{p(y \mid H_1)}{p(y \mid H_0)} \geq \frac{p_0}{p_1} \tag{5.12.3}$$

This inequality is very similar to the Neyman-Pearson decision rule (1.4.8), with the threshold k being now determined by the a priori distribution. It determines the region of the sample space $\mathbb{R}^N$ where one accepts H_1 (or rejects H_0). We shall still denote this region by the symbol $\mathcal{C}$. Now the probabilities of missclassification are

$$\alpha = \int_{\mathcal{C}} p(y \mid H_0)\, dy\,, \qquad \beta = \int_{\bar{\mathcal{C}}} p(y \mid H_1)\, dy$$

(here $\bar{\mathcal{C}}$ is the complement of $\mathcal{C}$) so that the the **overall error probability** is $P_e = \alpha p_0 + \beta p_1$.

Proposition 5.8. *A Maximum a Posteriori test minimizes the overall error probability.*

Proof. The expression for the overall error probability

$$P_e = \int_{\mathbb{R}^N} I_{\mathcal{C}}\, p(y \mid H_0) p_0\, dy \;+\; \int_{\mathbb{R}^N} [1 - I_{\mathcal{C}}\,]\, p(y \mid H_1) p_1\, dy \tag{5.12.4}$$

can be rewritten

$$\int_{\mathbb{R}^N} I_{\mathcal{C}}\, \{p(y \mid H_0) p_0 - p(y \mid H_1) p_1\}\, dy \;+\; \int_{\mathbb{R}^N} p(y \mid H_1)\, dy\, p_1$$

where the last term is equal to p_1 and does not depend on $\mathcal{C}$; i.e. does not depend on the decision rule. The minimum is achieved when the first integrand has the largest possible negative value; that is when $\mathcal{C}$ coincides exactly with the set where $p(y \mid H_0)p_0 - p(y \mid H_1)p_1 \leq 0$. This is precisely the decision region dictated by the MAP test (5.12.3). □

The result actually holds also when $\mathbf{y}$ is a discrete random variable. In this case one uses probability mass functions instead of the PDF.
The MAP test can be easily generalized to multiple hypotheses and the proposition continues to hold also in this case.

Exercise: Prove proposition 5.8 in the case of M simple hypotheses.

Utility and Minimum Cost Hypothesis Testing

It should be clear from the MAP decision rule that Bayesian Hypothesis testing is not privileging one particular hypothesis with respect to the other. Sometimes however one should actually consider this possibility. Suppose that you are building a sensor network to detect fires in a forest. Based on the information collected by the sensors, the system needs to decide between two opposing hypotheses:
H_0: There is no fire,
H_1: There is a fire.

There are two possible types of errors that we can make: We might accept H_0 while H_1 is true, or we might accept H_1 while H_0 is true. Note that the cost associated with these two errors are not the same. In other words, if there is a fire and we miss it, we will be making a costlier error. To address situations like this, one may associate a cost to each error type:

$c_{1|0}$: The cost of choosing H_1, when H_0 is true.
$c_{0|1}$: The cost of choosing H_0, when H_1 is true.

Of course the cost incurred by a decision will be a random variable $c(\mathbf{y})$ depending on $\mathbf{y}$. Its expected value can be written as

$$\mathbb{E}\, c(\mathbf{y}) = P(\text{choose}H_1 \mid H_0) p_0 c_{1|0} + P(\text{choose}H_0 \mid H_1) p_1 c_{0|1}.$$

where the conditional probabilities can be expressed as

$$P(\text{choose}H_1 \mid H_0) = \int_{\mathbb{R}^N} I_{\mathcal{C}}\, p(y \mid H_0)\, dy\,, \quad P(\text{choose}H_0 \mid H_1) = \int_{\mathbb{R}^N} [1 - I_{\mathcal{C}}\,]\, p(y \mid H_1)\, dy$$

and one may look for a decision rule minimizing the expected overall cost. Luckily, this can be done easily since the above expression of $\mathbb{E}\, c(\mathbf{y})$ is very similar to the overall error probability of the MAP test of (5.12.4). The only difference is that now we have $p_0 c_{1|0}$ in place of p_0, and $p_1 c_{0|1}$ instead of p_1. Therefore, we can use a decision rule similar to the MAP decision rule. More specifically, we choose H_0 if and only if

$$p(y \mid H_0) p_0 c_{1|0} > p(y \mid H_1) p_1 c_{0|1}\,. \tag{5.12.5}$$

Here is another way to interpret the above decision rule. If we divide both sides of Equation (5.12.5) by $p(y)$ and apply Bayes' rule, we conclude the following: We choose H_0 if and only if

$$P(H_0 \mid \mathbf{y} = y) c_{1|0} > P(H_1 \mid \mathbf{y} = y) c_{0|1}. \tag{5.12.6}$$

Note that $P(H_0 \mid \mathbf{y} = y) c_{1|0}$ is the *posterior conditional expected cost* of accepting H_1 when H_0 is the true probability. We call this the *posterior risk* of accepting H_1. Similarly, $P(H_1 \mid \mathbf{y} = y) c_{0|1}$ is the *posterior conditional expected cost* of accepting H_0 when H_1 is the true probability. Therefore, we can summarize the minimum cost test as follows: *We accept the hypothesis with the lowest expected posterior cost.*

Example 5.9. Suppose that a binary random variable $\mathbf{x}$ is transmitted over a communication channel. Assume that the received signal is given by

$$\mathbf{y} = \mathbf{x} + \mathbf{w},$$

where $\mathbf{w} \sim \mathcal{N}(0, \sigma^2)$ is independent of $\mathbf{x}$. Suppose that $\mathbf{x} = 1$ (H_1) with probability p, and $\mathbf{x} = 0$ (H_0) with probability $1 - p$. The goal is to decide between $\mathbf{x} = 1$ and $\mathbf{x} = 0$ by observing the random variable $\mathbf{y}$. Find the MAP test for this problem.
Generalize to a random sample $\mathbf{y}$ of size N.
Solution :
The two conditional densities are

$$p(y \mid H_1) = \mathcal{N}(1, \sigma^2)\,, \qquad p(y \mid H_0) = \mathcal{N}(0, \sigma^2)$$

and the critical region is then

$$\mathcal{C} = \{ y \mid \exp \frac{1}{2\sigma^2} \left[y^2 - (y-1)^2 \right] \geq \frac{1-p}{p} \}$$

which can be easily computed by taking logarithms and turns out to be a half line.

Composite hypotheses

Suppose now that the decision regards two (disjoint) subsets of the parameter space say

$$H_0 \equiv \{\theta \in \Theta_0\}\,; \qquad H_1 \equiv \{\theta \in \Theta_1\} \tag{5.12.7}$$

assuming that Θ_0 and Θ_1 are well behaved subsets of $\mathbb{R}^p$ and that we are given an a priori density $p(\theta)$ on the parameter space.

A *MAP* decision procedure should again be based on the rule: accept H_1 if the inequality

$$\frac{P(H_1 \mid \mathbf{y} = y)}{P(H_0 \mid \mathbf{y} = y)} > 1 \tag{5.12.8}$$

is satisfied and accept instead H_0 in the opposite case. Now however we cannot transform this into a Likelihood Ratio inequality like (5.12.3) since the conditional probabilities there are well defined only for simple hypotheses. We can instead proceed as follows: first compute the a posteriori conditional probability density $f(\theta \mid \mathbf{y} = y)$ by Bayes rule and then integrate over the two subsets (5.12.7) using the prior. The following example should clarify the procedure.

Example 5.10. *Given a random sample* $\{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N\}$ *drawn from a Gaussian density* $\mathcal{N}(\theta, 1)$ *and a Gaussian a priori distribution on the parameter*

$$\boldsymbol{\theta} \sim \mathcal{N}(0, \sigma^2)$$

where σ^2 *is known, you need to decide which of the two composite hypotheses*

$$\begin{aligned} H_0 &\equiv \{\theta \leq c,\}; \\ H_1 &\equiv \{\theta > c\,\}. \end{aligned}$$

has generated the data. Describe your decision policy using a MAP criterion.

Solution: By independence the Gaussian conditional probabilities will depend on the data only through the sample mean (a sufficient statistics). We just need to compute the two a posteriori probabilities $P\{H_0 \mid \bar{\mathbf{y}}_N\}$ and $P\{H_1 \mid \bar{\mathbf{y}}_N\}$ by integrating the conditional density

$$p(\theta \mid \bar{\mathbf{x}}_N = \bar{x}_N) = \frac{1}{p(\bar{x}_N)} \frac{N^{N/2}}{[(2\pi)^{N+1}\sigma^2]^{1/2}} exp - \{\frac{N}{2}(\bar{x}_N - \theta)^2 + \frac{\theta^2}{2\sigma^2}\}$$

on the two sets Θ_0 and Θ_1. Since the denominator cancels, the ratio (5.12.8) reduces to

$$\frac{P\{H_1 \mid \bar{\mathbf{y}}_N = \bar{x}_N\}}{P\{H_0 \mid \bar{\mathbf{y}}_N = \bar{x}_N\}} = \frac{\int_c^{+\infty} exp - \{\frac{N}{2}(\bar{x}_N - \theta)^2 + \frac{\theta^2}{2\sigma^2}\}\, d\theta}{\int_{-\infty}^{c} exp - \{\frac{N}{2}(\bar{x}_N - \theta)^2 + \frac{\theta^2}{2\sigma^2}\}\, d\theta}\,.$$

The calculations are left to the reader. One can show that the ratio depends on the data only through $\bar{\mathbf{y}}_N$. We should choose H_1 if the ratio is greater than 1 and H_0 otherwise.

Note that the priors are

$$p_0 = \int_{-\infty}^{c} \frac{1}{\sqrt{2\pi\sigma^2}} \exp -\frac{1}{2}\frac{x^2}{\sigma^2}\, dx \qquad p_1 = \int_{c}^{+\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp -\frac{1}{2}\frac{x^2}{\sigma^2}\, dx$$

so that $p_0 > p_1$ iff $c > 0$ and $p_0 < p_1$ in the other case. □

5.13 ▪ Classification by Logistic Regression

As we have just seen, in Bayesian classification, the assignment of an hypothesis (or a pattern) in a class is performed based on the posterior probabilities, $P(\cdot \mid y)$. This is in general, not an easy task since calculation of the posteriors requires the a priori distribution and the conditional pdf's of the observations given the pattern, which is lots of information often not available in practice. The idea of logistic regression is to model the posterior probabilities directly, via a simple log-linear model and estimate them directly from the data.

In this section, following [95, p. 291] we shall discuss a simple such model called *logistic regression model*. This name has been established in the statistics community, although the model refers to classification and not to regression. It is a typical example of empirical, sometimes called *discriminative*, modeling approach, where the distribution of the data is of no interest and is not estimated.

We shall only discuss *the two-class case:* Assume the feature measurements $\mathbf{y}$ are p-dimensional. The starting point is to model the log of the ratio of the posteriors as a linear function of the observation,

$$\log \frac{P(H_1 \mid \mathbf{y})}{P(H_0 \mid \mathbf{y})} = \theta^\top \mathbf{y} \tag{5.13.1}$$

where θ is a $p+1$-dimensional parameter with constant term θ_0 absorbed in the vector θ. The likelihood ratio is then an exponential function of $\theta^\top \mathbf{y}$ so that whenever $\theta^\top \mathbf{y} > 0$ one chooses H_1 and instead when $\theta^\top \mathbf{y} < 0$ one decides for H_0. The decision boundary is therefore the hyperplane $\theta^\top \mathbf{y} = 0$. This can be made into a more general affine hyperplane boundary by augmenting the dimension of the feature space introducing an artificial zeroth-index coordinate $x_0 = 1$. Since

$$P(H_1 \mid \mathbf{y}) + P(H_0 \mid \mathbf{y}) = 1$$

dividing both members by $P(H_0 \mid \mathbf{y})$ one finds

$$P(H_0 \mid \mathbf{y}) = \frac{1}{1 + e^{\theta^\top \mathbf{y}}} := \sigma(\theta^\top \mathbf{y}) \tag{5.13.2}$$

$$P(H_1 \mid \mathbf{y}) = \frac{e^{\theta^\top \mathbf{y}}}{1 + e^{\theta^\top \mathbf{y}}} = 1 - \sigma(\theta^\top \mathbf{y}) \tag{5.13.3}$$

These formulas involve the function

$$\sigma(x) := \frac{1}{1 + \exp(x)}$$

which is a symmetric version of the ubiquitous *sigmoid function* $\sigma(-x)$ widely used in Neural Network models. Although it may sound a bit mystical as to how one thought of such a model, it suffices to look more carefully at (5.13.2) to demystify it. Assuming the data in the two classes follow Gaussian distributions with $\Sigma_1 = \Sigma_0 = \Sigma$ and for simplicity that the priors are equal, i.e. $p_0 = p_1$, the log of the Gaussian likelihood ratio is written as in Section 4.3

$$\log \frac{P(H_1 \mid \mathbf{y})}{P(H_0 \mid \mathbf{y})} = (\mu_1 - \mu_0)^\top \Sigma^{-1} \mathbf{y} + \text{constant} \tag{5.13.4}$$

which is a linear function of the data. In other words *in logistic regression we adopt a Gaussian likelihood ratio irrespective of the data distribution*. However, even

if the data are distributed according to Gaussians, it may still be preferable to adopt the logistic regression formulation instead of that in (5.13.4). In the latter formula, the covariance matrix and the means have to be estimated from the training set while the logistic regression formulation only involves estimation of $p+1$ parameters (which however enter non-linearly in the model). Of course, assuming that the Gaussian assumption is valid, if one can obtain good estimates of the covariance matrix, employing this extra information can lead to more efficient estimates, in the sense of lower variance. This is natural, because more information concerning the distribution of the data is exploited. In practice, it turns out that using the logistic regression is, in general safer compared to the linear discriminant analysis (LDA).

Logistic parameter estimation

We assume an i.i.d. sample $\mathbf{y} := \{\, \mathbf{y}(t)\,;\, t = 1, \ldots, N\}$ of p-dimensional classified features (actually we make each feature vector $p+1$-dimensional letting the first component $\mathbf{y}_0(t)$ of $\mathbf{y}(t)$ equal to 1, allowing a constant intercept θ_0 in the linear function $\theta^\top \mathbf{y}$), each $\mathbf{y}(t)$ coming together with a classification variable, say $\mathbf{x}(t)$, equal to 0 or 1 according to which class, H_0 or H_1, was selected based on $\mathbf{y}(t)$. In accordance with the definitions in (5.13.2), let us introduce the notation

$$p_0(\theta; y) := P(H_0 \mid \mathbf{y} = y) = \sigma(\theta^\top y)\,; \qquad p_1(\theta; y) := P(H_1 \mid \mathbf{y} = y) = 1 - \sigma(\theta^\top y)$$

The decision sequence $\mathbf{x} := \{\, \mathbf{x}(t)\,;\, t = 1, \ldots, N\}$ has a *time-varying* Bernoulli probability distribution with random parameter $p = \sigma(\theta^\top \mathbf{y}(t)\,)$ which is a function of the current observed feature. This can be represented as

$$P(x \mid \mathbf{y} = y) = \prod_{t=1}^{N} \sigma(\theta^\top y(t)\,)^{1-x(t)}\, [\,1 - \sigma(\theta^\top y(t)\,)\,]^{x(t)} \tag{5.13.5}$$

which yields a log-likelihood

$$\ell(\theta; x, y) = \sum_{t=1}^{N} \left\{ (1 - x(t))\, \log \sigma(\theta^\top y(t)\,) + x(t)\, \log[\,1 - \sigma(\theta^\top y(t)\,)\,] \right\} \tag{5.13.6}$$

$$= \sum_{t=1}^{N} \left\{ x(t)\, \theta^\top y(t) - \log(1 + e^{\theta^\top y(t)}\,) \right\} \tag{5.13.7}$$

Unfortunately this is a non-linear function of the parameter θ which can only be maximized by numerical methods. To maximize $\ell(\theta; x, y)$ we set the gradient with respect to θ equal to zero, i.e.

$$\nabla_\theta \ell(\theta; x, y) = \sum_{t=1}^{N} y(t)\, \{\, x(t) - p_1(\theta; y(t))\} = 0\,.$$

This is a system of $p+1$ non linear equations in θ. Since whenever H_0 is chosen, $x(t)$ is equal to 0, the equation of index 0 yields

$$\frac{N_1}{N} = \frac{1}{N} \sum_{t=1}^{N} p_1(\theta; y(t))$$

where N_1 is the number of decisions for H_1 in the training set. This has an intuitive interpretation as an empirical frequency estimate. For the next developments it will be convenient to introduce vector notations. Introduce the $N \times (p+1)$ matrix Y having rows $y(1)^\top, \ldots, y(N)^\top$ and the N-vector $x - \mathbf{p}(\theta,\, y)$ having components $x(t) - p_1(\theta; y(t))\,;\; t = 1, 2, \ldots, N$ so that

$$\nabla_\theta \ell(\theta; \mathbf{x}, \mathbf{y}) = Y^\top [\, x - \mathbf{p}(\theta,\, y)\,] \,.$$

To solve the full log-likelihood equation $\nabla_\theta \ell(\theta; \mathbf{x}, \mathbf{y}) = 0$ we shall use a Newton algorithm which uses the second derivatives. Since

$$\nabla_\theta \sigma(\theta^\top y) = \sigma(\theta^\top y)\,[1 - \sigma(\theta^\top y)\,]\, y \,, \qquad y \in \mathbb{R}^{p+1}$$

the Hessian can be written

$$H(\theta, \mathbf{y}) := \frac{\partial^2}{\partial\theta\, \partial\theta^\top} \ell(\theta; \mathbf{x}, \mathbf{y}) = \sum_{t=1}^{N} \mathbf{y}(t)\mathbf{y}(t)^\top \sigma(\theta^\top \mathbf{y}(t))\,[1 - \sigma(\theta^\top \mathbf{y}(t)\,]$$

which, after introducing the $(p+1) \times (p+1)$ diagonal matrix

$$Q(\theta,\, \mathbf{y}) := \text{diag}\, \{\sigma(\theta^\top \mathbf{y}(1))\,[1 - \sigma(\theta^\top \mathbf{y}(1)\,] \ldots, \sigma(\theta^\top \mathbf{y}(N))\,[1 - \sigma(\theta^\top \mathbf{y}(N)\,]\,\}$$

can be written

$$H(\theta, \mathbf{y}) = \sum_{t=1}^{N} \mathbf{y}(t) Q(\theta,\, \mathbf{y}) \mathbf{y}(t)^\top \;= Y^\top Q(\theta,\, \mathbf{y}) Y \,. \tag{5.13.8}$$

We shall assume that for N large enough, this matrix is invertible. Then the k-th step of the Newton algorithm

$$\theta_{k+1} = \theta_k - H^{-1}(\theta_k, \mathbf{y}) \nabla_\theta \ell(\theta_k; \mathbf{x}, \mathbf{y})$$

can be rewritten

$$\begin{aligned} H(\theta_k, \mathbf{y})\theta_{k+1} &= H(\theta_k, \mathbf{y})\theta_k - Y^\top [\, x - \mathbf{p}(\theta_k,\, y)\,] \;= \\ &= Y^\top Q(\theta_k,\, \mathbf{y}) Y \theta_k - Y^\top [\, x - \mathbf{p}(\theta_k,\, y)\,] = \\ &= Y^\top Q(\theta_k,\, \mathbf{y})\, \left\{ Y\theta_k - Q(\theta_k,\, \mathbf{y})^{-1} [\, x - \mathbf{p}(\theta_k,\, y)\,] \right\} \,. \end{aligned}$$

This equation, after setting $z_k := Y\theta_k - Q(\theta_k,\, \mathbf{y})^{-1} [\, x - \mathbf{p}(\theta_k,\, y)\,]$, can be interpreted as the normal equation

$$Y^\top Q(\theta_k,\, \mathbf{y}) Y\, \theta_{k+1} = \; Y^\top Q(\theta_k,\, \mathbf{y}) z_k \tag{5.13.9}$$

which is generated as k-th step of an iterative solution of the weighted least squares problem

$$\min_{\theta_{k+1}} \| Y\, \theta_{k+1} - z_k \,\|^2_{Q(\theta_k,\, \mathbf{y})} \,, \qquad k = 1, 2, \ldots . \tag{5.13.10}$$

This algorithm is called the *iteratively reweighted least squares* (IRLS). Since the log-likelihood is concave this Newton algorithm with a suitably designed step-size sequence converges. The estimate is maximum likelihood so that, if the true model belongs to the model class, the asymptotic properties of ML such as consistency and asymptotic normality are guaranteed.

In general, when the model complexity is not assigned a regularization penalty such as shrinkage or Lasso can be added to the iterative least squares procedure. Details can be found in Sect. 4.4.4 of [44].

5.14 ▪ Problems

5-1 *Consider the standard linear model with uncorrelated Gaussian noise* $\mathbf{w}(k) \sim \mathcal{N}(0, \sigma^2)$

$$\mathbf{y}(k) = s(k)\boldsymbol{\theta} + \mathbf{w}(k)\,, \qquad k = 1, \dots, N$$

where $\boldsymbol{\theta}$ *is the scalar unknown parameter with prior the* (centered) Laplace distribution $L(0, \lambda)$ *having density*

$$p_L(\theta) = \frac{1}{2b} \exp\{-\frac{|\,\theta\,|}{b}\}\,, \qquad b > 0$$

which has mean zero and variance $2b^2$*. We would like to compute the MAP estimate of* $\boldsymbol{\theta}$ *given the* N *independent observations* $\mathbf{y}^N$*. Use Bayes rule to write the the posterior distribution and compute its logarithm disregarding the log of the denominator* $p(y^N)$ *which does not enter in the calculation of the MAP. What kind of regularized least squares problem you end up with ?*

In order to generalize your answer to a linear model with a p*-dimensional vector parameter* $\boldsymbol{\theta}$*, what kind of prior should you choose?*

5-2 *Let* $\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix}$ *be a zero-mean random vector with both* $\text{var}\,(\mathbf{y}_1)$ *and* $\text{var}\,(\mathbf{y}_2)$ *positive but with a singular joint variance matrix. Show that* $\mathbf{y}_1 = \alpha\,\mathbf{y}_2$ *for some* $\alpha \neq 0$*.*

5-3 *Let* $\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix}$ *be a zero-mean random vector with a non-singular variance matrix. Find necessary and sufficient conditions on* $\mathbf{y}_1, \mathbf{y}_2$ *for the validity of the relation*

$$\hat{\mathbb{E}}\,[\mathbf{x} \mid \mathbf{y}] = \hat{\mathbb{E}}\,[\mathbf{x} \mid \mathbf{y}_1] + \hat{E}[\mathbf{x} \mid \mathbf{y}_2]$$

for an arbitrary zero-mean random variable $\mathbf{x}$*.*

5-4 *Assume that the covariance function* $\sigma(\tau)$ *is a simple exponential*

$$\sigma(\tau) = \sigma(0)\lambda_0^{-|\tau|}\,, \qquad \tau \in \mathbb{Z}\,,$$

where $0 < \lambda_0 < 1$ *(the reader should check that this is a positive definite function). Then, assuming that* $t_1 \leq \ldots \leq t_n$*, show that the matrix*

$$\Sigma_{n+1} = \sigma(0) \begin{bmatrix} 1 & \lambda_0^{t_2-t_1} & \dots & \lambda_0^{t_n-t_1} \\ \lambda_0^{t_2-t_1} & 1 & & \dots \\ & & \ddots & \\ \dots & & & \\ \lambda_0^{t_n-t_1} & & & 1 \end{bmatrix}.$$

is always non singular but is singular for $n \geq 2$ *if* $\lambda_0 = 1$*, independently of the choice of the time instants* $t_1 \leq \ldots \leq t_n$*.*

5-5 *Consider the two-blocks linear model*

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} S_1 \\ S_2 \end{bmatrix} \mathbf{x} + \begin{bmatrix} \mathbf{w}_1 \\ \mathbf{w}_2 \end{bmatrix}$$

where the observations $\mathbf{y}_i$ *are* m_i*-dimensional,* $\text{Var}\,(\mathbf{x}) := P > 0$*,* $\text{Var}\,(\mathbf{w}_i) := R_i > 0$*,* $i = 1, 2$*, and* $\mathbf{w}_1 \perp \mathbf{x} \perp \mathbf{w}_2$ *(all random variables are zero-mean). Use the matrix*

inversion Lemma to show that the m.v. estimator $\hat{\mathbf{x}} = \hat{\mathbb{E}}[\,\mathbf{x} \mid \mathbf{y}_1, \mathbf{y}_2]$ *is a linear function of the* decentralized estimators $\hat{\mathbf{x}}_i = \hat{\mathbb{E}}[\,\mathbf{x} \mid \mathbf{y}_i]$, $i = 1, 2$, *of the form*

$$\hat{\mathbf{x}} = Q^{-1}[A\,\hat{\mathbf{x}}_1 + B\,\hat{\mathbf{x}}_2] \quad .$$

Find expressions for Q, A, B.

5-6 *(Distributed Estimation). The subvectors* $\mathbf{y}_1, \ldots, \mathbf{y}_N$ *form a partition of the observation vector* $\mathbf{y}$ *which is described by a linear model* $\mathbf{y} = S\mathbf{x} + \mathbf{w}$ *for wich all standard assumptions (zero-mean,* $R > 0$ *etc.) are assumed to hold. Let* $\hat{\mathbf{x}}_k = \hat{\mathbb{E}}[\,\mathbf{x} \mid \mathbf{y}_k]$ *be the* local estimators *of* $\mathbf{x}$ *based on knowledge of the parameters of the overall model but each using only the local observation vector* $\mathbf{y}_k$, $k = 1, \ldots, N$.

Find conditions on the model for the validity of the relation, called fusion of the estimates,

$$\hat{\mathbb{E}}[\mathbf{x} \mid \mathbf{y}_1, \ldots, \mathbf{y}_N] = \hat{\mathbb{E}}[\mathbf{x} \mid \hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_N] = \sum_{k=1}^{N} A_k\,\hat{\mathbf{x}}_k\,.$$

In other terms, when is the global estimator $\hat{\mathbf{x}} = \hat{\mathbb{E}}[\,\mathbf{x} \mid \mathbf{y}]$ *expressible as a (linear) deterministic function of the* local *estimators* $\{\hat{\mathbf{x}}_k\}$ *?*

• *Show that* $\mathbf{w}_k \perp \mathbf{x} \perp \mathbf{w}_j$, $\forall k \neq j$ *(*$\mathbf{w}_k$ *being the noise subvector of* $\mathbf{w}$ *corrupting* $\mathbf{y}_k$*) ia a sufficient condition and find the parameters of the linear relation between* $\hat{\mathbf{x}}$ *and* $\hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_N$.

• *Is this condition also necessary?*

5-7 *The three scalar random variables* $(\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3)$ *are zero-mean and have a non-singular variance matrix* Σ. *Consider the three othogonal projections* $\hat{\mathbb{E}}[\mathbf{y}_1 \mid \mathbf{y}_2\,\mathbf{y}_3]$, $\hat{\mathbb{E}}[\mathbf{y}_2 \mid \mathbf{y}_1\,\mathbf{y}_3]$, $\hat{\mathbb{E}}[\mathbf{y}_3 \mid \mathbf{y}_1\,\mathbf{y}_2]$, *each of which depends on two scalar parameters. Under what conditions do they determine* Σ *(note that* Σ*, being symmetric, depends on six parameters). Can one assign* arbitrarily *the three functions and thereby uniquely determine* Σ*?*

5-8 *Let* $\mathbf{y} := \{\mathbf{y}(\tau) \mid \tau \in \mathbb{R}\}$ *be a continuous time stochastic process of zero mean and covariance function* $\sigma(t, s)$ *which is finite at every point* $\{t,\ s\}$ *of* $\mathbb{R}^2$. *Show that*

1. $\mathbf{y}$ *is continuous in mean square (i.e.* $\lim_{t \to s} \|\mathbf{y}(t) - \mathbf{y}(s)\| = 0$ *for every* $t, s \in \mathbb{R}^2$*), if and only if* σ *is continuous at every point of the diagonal* $t = s$ *of the* $\{t,\ s\}$ *plane.*

2. $\mathbf{y}$ *is mean square differentiable if and only if the mixed second derivatives* $\partial^2\sigma/\partial t \partial s$ *exist at all points of the diagonal* $t = s$ *of the* $\{t,\ s\}$ *plane.*

3. *Let* $\mathbf{y}(t)$ *be piecewise mean square continuous on the interval* $[\,0,\ T]$. *Define the mean square integral* $\mathbf{x} := \int_0^T \mathbf{y}(s) d\,s$ *as the mean square limit of the Riemann sums. Show that* $\mathbf{x}$ *has finite variance if and only if*

$$\int_0^T \int_0^T \sigma(t,\ s)\,dt\,ds < \infty\,.$$

Second order calculus for processes in continuous time is discussed in several books, e.g. Jazswinsky [47, p. 60-70], Wong,[108, p. 77-80], [59, p.].

5-9 *Again, let* $\mathbf{y} := \{\mathbf{y}(\tau) \mid \tau \in \mathbb{R}\}$ *be a zero-mean continuous-time process with a covariance function* $\sigma(t, s)$ *continuous on* $\mathbb{R}^2$. *The process is sampled with sampling period* $h > 0$.

1. *Describe the m.v. estimate of the mean square derivative* $\frac{d\mathbf{y}}{d\tau}(t_0)$ *at a generic time* t_0*, based on the discrete sample values* $\{\mathbf{y}(t_0 - nh), \ldots, \mathbf{y}(t_0 - h), \mathbf{y}(t_0), \mathbf{y}(t_0 + h), \ldots, \mathbf{y}(t_0 + nh)\}$.

2. *Describe the m.v. estimate of the mean square integral* $\mathbf{x} := \int_0^T \mathbf{y}(s)\, ds$*, based on the discrete values*$\{\mathbf{y}(0), \mathbf{y}(h), \ldots, \mathbf{y}(nh)\}$, $nh = T$. *In other terms show how to compute the coefficients* $c(k)$ *of the mean square quadrature formula*

$$\hat{\mathbf{x}} = \sum_{k=0}^{n} c(k)\mathbf{y}(kh)$$

in such a way that the error $\mathbf{x} - \hat{\mathbf{x}}$ *has minimal variance.*

5-10 *A surveillance system is in charge of detecting intruders to a facility. There are two hypotheses to choose from:*
H_0*: No intruder is present.*
H_1*: There is an intruder.*

The system sends an alarm message if it accepts H_1. *Suppose that after processing the data, we obtain* $P(H_1 \mid y) = 0.05$. *Also, assume that the cost of missing an intruder is 10 times the cost of a false alarm. Should the system send an alarm message (accept* H_1*)?*

5-11 *We want to compute the Bayesian MAP estimate of* θ *from an i.i.d. sample* $\{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N\}$ *with* $\mathbf{y}_k \sim \mathcal{N}(S\theta, \sigma^2 I_N)$*, where* σ^2 *is known, assuming a prior density for the mean vector* θ *which is also Gaussian and has a variance* $\tau^2 I_p$*, i.e.* $\mathbf{x} \sim \mathcal{N}(\theta_0, \tau^2 I_p)$ *where* $\theta_0 \in \mathbb{R}^p$ *is the a priori mean.*

Chapter 6

PRINCIPAL COMPONENT ANALYSIS

6.1 ▪ Introduction to data compression

In this chapter we shall discuss some general techniques for statistical data compression (or noise reduction). In many practical applications, although the data seem to reside in a high-dimensional space, the true dimensionality, after subtracting the noise, known as *intrinsic dimensionality*, can be much lower. As a trivial example, in a three-dimensional space, the data may cluster around a straight line, or around a circle or a parabola, arbitrarily placed in $\mathbb{R}^3$. In these cases, the intrinsic dimensionality of the data is equal to one, as any of these curves can equivalently be described in terms of a single parameter. Learning the lower dimensional structure associated with a given set of data is gaining in importance in the context of big data processing and analysis. There is a multitude of practical examples where it is relevant, such as image processing, computer vision, medical diagnosis and especially information retrieval, where one is looking for efficient searching procedure for identifying similar patterns in large databases.

6.2 ▪ Principal Component Analysis (PCA)

Let $\mathbf{y}$ be an N-dimensional random vector whose components have finite variance. The interpretation of $\mathbf{y}$ is that of a sequence of rough redundant data from which we need to extract an hopefully small number of features. A good example to keep in mind is the description of handwritten digits in [44, p. 536-537]. Each discretized and numerically coded picture of say, a handwritten "3", could be a possible sample realization of the random vector $\mathbf{y}$. We shall look for a linear expansion of $\mathbf{y}$ in terms of a family of **deterministic** orthonormal column vectors, $\varphi_k = \begin{bmatrix}\varphi_k(1) & \dots & \varphi_k(N)\end{bmatrix}^\top$, $k = 1, 2, \dots$. Naturally the coefficients of this expansion will have to be random variables; their role parallels and in fact extends that of the Fourier coefficients in the deterministic setting.

Problem 6.1. *Express the random vector* $\mathbf{y}$ *as a linear combination of deterministic* N*-vectors* $\varphi_k \in \mathbb{R}^N$, $k = 1, 2, \dots, N$ *which are orthonormal with respect to the Euclidean inner product. The linear combination should be by means of uncorrelated*

scalar random coefficients; that is

$$\mathbf{y}(t) = \sum_k \mathbf{x}_k \varphi_k(t) \qquad t = 1, 2, \ldots, N\,, \tag{6.2.1}$$

with $\mathbb{E}\,\mathbf{x}_k \mathbf{x}_j = 0$ *for* $k \neq j$.

The expansion (6.2.1) is called *biorthogonal* if view of the fact that the $\{\mathbf{x}_k; k = 1, \ldots, N\}$ must form an orthogonal basis in $\mathbf{H}(\mathbf{y})$ and the modes φ_k are an orthonormal basis in $\mathbb{R}^N$; i.e.

$$\varphi_k^\top \varphi_j = \delta_{k,j} \qquad k, j = 1, 2, \ldots$$

$\delta_{k,j}$ being the Kronecker delta. If the $\mathbf{x}_k$ are ordered in decreasing (variance) norm the N-vectors components of $\mathbf{y}$, $\mathbf{x}_k \varphi_k$, will have the same (scalar variance) norm of the $\mathbf{x}_k$'s and can likewise be ordered assigning the first places in the list to the components with higher variance. These are called the **principal components** of $\mathbf{y}$.

Principal Component Analysis (PCA) is an important tool for feature extraction and classification in decision processes. In applications to Signal Processing, the sequence the components are indexed by the (discrete) time variable and $\mathbf{y} := \{\mathbf{y}(1), \ldots, \mathbf{y}(t), \ \ldots, \mathbf{y}(N)\}$ may represent a (time-sampled) *random signal* defined on a discrete finite time interval $[1,\ N]$. The problem addressed in this section could be seen as a generalization of the Fourier transform of a random signal by linear combinations of orthonormal deterministic functions of time which are not necessarily sinusoidal but are instead, in some sense, better "tailored" to the signal under analysis. These functions are called the *(proper) modes* of of the signal $\mathbf{y}$. They enjoy some natural properties. If N is large, an exact representation in general requires "too many" modes and the key question in applications is the optimal approximation of the signal in terms of a small number of modes. Principal Component Analysis provides the theoretical basis for such an approximation

We shall initially discuss the (exact) representation of a discrete time random signal of finite duration since it can basically be treated by linear algebra and elementary Hilbert space techniques. Principal Component Analysis of infinite duration signals (or processes) is based on the same general ideas but requires more sophisticated tools and goes under a different name. It is called *Karhunen Loève expansion* and will be discussed in the next section.

Since subtracting the means does not change the construction we shall describe below, without loss of generality we shall assume that $\mathbf{y}$ has zero mean. Denote by $\Sigma_{\mathbf{y}} = \mathbb{E}\,\mathbf{y}\mathbf{y}^\top$ the variance matrix of $\mathbf{y}$, which will be assumed to be positive definite and known.

Proposition 6.1. *The random signal* $\mathbf{y}$ *admits a biorthogonal expansion of the form* (6.2.1) *if and only if the modes* $\{\varphi_k\}$ *form a system of normalized eigenvectors for the covariance matrix,* $\Sigma_{\mathbf{y}}$. *In this case, letting* $\lambda_k > 0$ *denote the eigenvalue of* $\Sigma_{\mathbf{y}}$ *corresponding to the eigenvector* φ_k, *the random variables* $\mathbf{x}_k$ *are given by the formula*

$$\mathbf{x}_k = \varphi_k^\top \mathbf{y} = \sum_{t=1}^{N} \varphi_k(t)\, \mathbf{y}(t), \qquad k = 1, \ldots, N\,. \tag{6.2.2}$$

Proof. *Sufficiency:* Let φ_k be the k-th normalized eigenvector of $\Sigma_{\mathbf{y}}$ correspond to the (necessarily positive!) eigenvalue $\lambda_k > 0$, that is

$$\Sigma_{\mathbf{y}} \varphi_k = \lambda_k \, \varphi_k \, , \qquad k = 1, 2, \ldots, N$$

then define the random variables $\mathbf{x}_k = \varphi_k^\top \mathbf{y}$. These are uncorrelated since:

$$\mathbb{E}\, \mathbf{x}_k \mathbf{x}_j = \mathbb{E}\, \varphi_k^\top \mathbf{y}\, \varphi_j^\top \mathbf{y} \; = \; \varphi_k^\top \mathbb{E}\, [\mathbf{y}\, \mathbf{y}^\top]\, \varphi_j = \lambda_k \, \delta_{k,j}$$

We could actually make the $\{\mathbf{x}_k\}$'s into an orthonormal basis for $H(\mathbf{y})$ by normalization

$$\hat{\mathbf{x}}_k := \frac{1}{\sqrt{\lambda_k}} \, \varphi_k^\top \mathbf{y}$$

so that

$$\mathbf{y} = \sum_{k=1}^{N} \langle \mathbf{y}, \, \hat{\mathbf{x}}_k \rangle \, \hat{\mathbf{x}}_k \; = \; \sum_{k=1}^{N} \mathbb{E}\, [\mathbf{y}\, \mathbf{y}^\top] \, \frac{1}{\sqrt{\lambda_k}} \varphi_k \, \hat{\mathbf{x}}_k = \sum_{k=1}^{N} \Sigma_{\mathbf{y}} \, \frac{1}{\sqrt{\lambda_k}} \varphi_k \, \hat{\mathbf{x}}_k = \\ = \sum_{k=1}^{N} \sqrt{\lambda_k} \, \hat{\mathbf{x}}_k \, \varphi_k \, ,$$

however keeping the norms $\|\mathbf{x}_k\|^2 = \mathbb{E}\, \mathbf{x}_k^2 = \lambda_k$ will turn out to be a more convenient choice.
Necessity: Assume $\mathbf{y}$ admits a biorthogonal expansion (6.2.1). From this we obtain the following expression for the covariance matrix

$$\Sigma_{\mathbf{y}} = \sum_{k,j=1}^{N} \mathbb{E}\, [\mathbf{x}_k \mathbf{x}_j] \varphi_k \varphi_j^\top = \sum_{k=1}^{N} \varphi_k \, \mathbb{E}\, [\mathbf{x}_k]^2 \varphi_k^\top \\ = \quad \Phi \; \mathrm{diag}\, \{\|\mathbf{x}_1\|^2, \ldots, \|\mathbf{x}_N\|^2\} \, \Phi^\top \tag{6.2.3}$$

where $\Phi := [\varphi_1, \ldots, \varphi_N]$ is an orthonormal matrix (i.e. a matrix with orthonormal columns). It follows that the columns of Φ must be the normalized eigenvectors of Σ and $\|\mathbf{x}_k\|^2 \, , \;\; k = 1, 2, \ldots, N$ its eigenvalues, that is $\lambda_k = \|\mathbf{x}_k\|^2$. □

The vectors φ_k may be called the *proper modes* of the signal $\mathbf{y}$. The eigenvalues of Σ will be listed in *decreasing order*; i.e.

$$\lambda_1 \geq \ldots \geq \lambda_N > 0$$

and this ordering is transmitted to the random coefficients $\mathbf{x}_k$ and to the corresponding modes φ_k. With this convention, the biorthogonal expansion (6.2.1) is *unique*. It is commonly called the **Principal Components Analysis (PCA)** of the signal $\mathbf{y}$.

Signal approximation

It often happens that the "statistical energy" of the signal

$$\mathbb{E}\, \|\mathbf{y}\|^2 = \sum_{k=1}^{N} \mathbb{E}\, \{\mathbf{x}_k\}^2 = \sum_{k=1}^{N} \lambda_k$$

is concentrated on a few proper modes. In other words it often happens that the eigenvalues of index larger than some $n < N$ are (relatively) small, for example they may be such that

$$\lambda_1 + \ldots + \lambda_n \gg \lambda_{n+1} + \ldots + \lambda_N$$

and their contribution to the expansion (6.2.1) could therefore be neglected. The resulting approximate expansion,

$$\mathbf{y} \simeq \hat{\mathbf{y}}_n := \sum_{k=1}^{n} \mathbf{x}_k \varphi_k = \sum_{k=1}^{n} (\varphi_k^\top \mathbf{y})\, \varphi_k \tag{6.2.4}$$

is a universal tool used in data compression, source coding, data storage and especially in pattern recognition.

In practice the covariance matrix Σ is not known but can be estimated from experimental data. Assume that we have a set of M independent sample measurements of the same random signal, say $x_k := \begin{bmatrix} x_k(1) & \ldots & x_k(N) \end{bmatrix}^\top$; $k = 1, 2, \ldots, M$, one can form the sample $N \times N$ covariance estimate

$$\hat{\Sigma} = \frac{1}{M} \sum_{k=1}^{M} x_k x_k^\top$$

and use this estimate in place of the true Σ. The normalized eigenvectors, $\hat{\varphi}_k$ of $\hat{\Sigma}$ will form an orthonormal basis in $\mathbb{R}^N$ which can be used to compute the sample coefficients of a test signal y as

$$x_k := \sum_{t=1}^{N} y(t) \hat{\varphi}_k(t) = y^\top \hat{\varphi}_k\,, \qquad k = 1, 2, \ldots N$$

These yield the orthonormal expansion

$$y(t) = \sum_{k} x_k \hat{\varphi}_k(t) \qquad t = 1, 2, \ldots, N\,, \tag{6.2.5}$$

in terms of the sample modes which can compactly be written as $y = \hat{\Phi} x$ where $\hat{\Phi}$ is an orthonormal matrix and the components of the vector $x = \begin{bmatrix} x_1 & \ldots & x_N \end{bmatrix}^\top$ can be ordered in decreasing magnitude. In fact we have

$$\|y\|^2 = \|x\|^2 = x_1^2 + x_2^2 + \ldots + x_N^2$$

where each term x_k^2 is the contribution of the k-th mode $\hat{\varphi}_k$ to the "energy" $\|y\|^2$ of the signal. In practice one will retain a (hopefully small) number of coefficients which contribute to most of the energy. The most significant numbers x_k are often used as *features* for classification of a pattern. See for example [44, p. 536].

Optimality of the PCA

Note that the approximation error vector $\tilde{\mathbf{y}}_n := \mathbf{y} - \hat{\mathbf{y}}_n$ is orthogonal to $\hat{\mathbf{y}}_n$, so that

$$\Sigma = \text{Var } \mathbf{y} = \text{Var } \hat{\mathbf{y}}_n + \text{Var } \tilde{\mathbf{y}}_n := \hat{\Sigma}_n + \tilde{\Sigma}_n$$

and the variance matrix of the approximant $\hat{\mathbf{y}}_n$ can be expresses as

$$\hat{\Sigma}_n = \sum_{k=1}^{n} \lambda_k \, \varphi_k \mathbb{E}\{\mathbf{x}_k^2\} \, \varphi_k^\top = \sum_{k=1}^{n} \lambda_k \, \varphi_k \, \varphi_k^\top .$$

A very well-known property of this approximation is recalled in the following proposition.

Proposition 6.2. *The variance matrix $\hat{\Sigma}_n$, of $\hat{\mathbf{y}}_n$, is the best symmetric positive semidefinite approximant of rank n, (either in the ℓ^2 or Frobenius norm) of the original variance matrix Σ.*

Proof. The statement follows from the well-known optimal approximation property of the truncated Singular Value Decomposition of a matrix; see the appendix B.2. One just needs to apply the result to a square root (say the Cholesky factor) of Σ. □ □

It is normally claimed and often given for granted in the literature that the approximation procedure described above does provide an *optimal* approximate representation of the signal. However besides the optimality of the covariance approximation described above, we have not been able to find any clear definition nor exact statement describing what this optimality should be in terms of *random signal approximation*.

Below we shall try to understand what kind of approximation criterion should be natural and reasonable to use in this context. To begin with, let us observe that the second member of (6.2.4) can be seen as a linear transformation acting on the random vector $\mathbf{y}$, represented by a certain deterministic matrix say M, which is symmetric, positive semidefinite and of rank n.

Any M of this kind can be written in factorized form $M = WW^\top$ where W is $N \times n$ and of full column rank. That M has rank n ($\leq N$), implies that the approximation $\hat{\mathbf{y}} := M\mathbf{y}$, generates an n-dimensional subspace of $\mathbf{H}(\mathbf{y})$. In this sense we can say that (6.2.4) provides an approximation $\hat{\mathbf{y}} := M\mathbf{y}$, *of rank* n, of $\mathbf{y}$.

Motivated by the above, let us consider a problem of optimal rank n approximation of the random vector $\mathbf{y}$, having the following natural formulation.

Problem 6.2. *Find a matrix $M \in \mathbb{R}^{N\times N}$ of rank n, solving the following minimum problem*

$$\min_{\text{rank}\,(M)=n} \mathbb{E}\{\|\mathbf{y} - M\,\mathbf{y}\|^2\} \tag{6.2.6}$$

Note that an equivalent geometric formulation is to look for an optimal n-dimensional subspace of $\mathbf{H}(\mathbf{y})$ onto which $\mathbf{y}$ should be projected in order to minimize the approximation error variance. Let us stress that this is quite different from the usual minimum error variance approximation problem which amounts to projecting onto a *given subspace*.

As for (6.2.4), minimizing the square distance in (6.2.6) requires that the approximation $M\mathbf{y}$ should be uncorrelated with the approximation error; namely

$$\mathbf{y} - M\mathbf{y} \;\perp\; M\mathbf{y} \tag{6.2.7}$$

which is equivalent to

$$M\Sigma - M\Sigma M^\top = 0\,.$$

Introducing a square root $\Sigma^{1/2}$ of Σ and defining $\hat{M} := \Sigma^{-1/2} M \Sigma^{1/2}$, this condition is seen to be equivalent to

$$\hat{M} = \hat{M}\,\hat{M}^\top$$

which just says that $\hat{M}$ must be symmetric and *idempotent* (i.e. $\hat{M} = \hat{M}^2$), in other words an *orthogonal projection* from $\mathbb{R}^N$ onto some n-dimensional subspace. Hence M must have the following structure

$$M \;=\; \Sigma^{1/2}\,\Pi\,\Sigma^{-1/2}, \qquad \Pi \;=\; \Pi^2 \qquad \Pi \;=\; \Pi^\top \tag{6.2.8}$$

where $\Sigma^{1/2}$ is any square root of Σ and Π is an orthogonal projection matrix of rank n.

Theorem 6.1. *The solutions of the signal approximation problem* (6.2.6) *are of the form*

$$M = W\,W^\top, \qquad W = \Phi_n Q_n$$

where Φ_n is a $N \times n$ matrix whose columns are the first n normalized eigenvectors of Σ, ordered according to the descending magnitude ordering of the corresponding eigenvalues and Q_n is an arbitrary $n \times n$ orthogonal matrix.

Proof. Let $\Lambda := \operatorname{diag}\{\lambda_1, \ldots, \lambda_N\}$ and $\Sigma \;=\; \Phi\Lambda\Phi^\top$ the spectral decomposition of Σ in which Φ is an orthogonal matrix of eigenvectors. We can pick as a square root of Σ the matrix $\Sigma^{1/2} \;:=\; \Phi\Lambda^{1/2}$.

Now, no matter how $\Sigma^{1/2}$ is chosen, the random vector $\mathbf{e} := \Sigma^{-1/2}\,\mathbf{y}$ has orthonormal components. Hence using (6.2.8) the cost function of our minimum problem can be rewritten as

$$\begin{aligned}
\mathbb{E}\,\{\|\mathbf{y} - M\,\mathbf{y}\|^2\} \;&=\; \mathbb{E}\,\{\|\Sigma^{1/2}\mathbf{e} - \Sigma^{1/2}\,\Pi\,\Sigma^{-1/2}\mathbf{y}\|^2\} \\
&=\; E\{\|\Sigma^{1/2}(\mathbf{e} - \Pi\,\mathbf{e})\|^2\} \;=\; \mathbb{E}\,\{\|\Lambda^{1/2}(\mathbf{e} - \Pi\,\mathbf{e})\|^2\} \\
&=\; \mathbb{E}\;(\mathbf{e} - \Pi\,\mathbf{e})^\top \Lambda\,(\mathbf{e} - \Pi\,\mathbf{e}) \;=\; \operatorname{Tr}\,\left[\,\Lambda\,\mathbb{E}\,(\mathbf{e} - \Pi\,\mathbf{e})(\mathbf{e} - \Pi\,\mathbf{e})^\top\right]
\end{aligned}$$

where $\operatorname{Tr}\,A := \sum a_{kk}$ is the trace of A. Our minimum problem can therefore be rewritten as

$$\min_{\text{rank}\;(\,\Pi\,)\,=\,n} \operatorname{Tr}\,\{\Lambda\Pi^\perp\}$$

where $\Pi^\perp := I - \Pi$ is the orthogonal projection onto the orthogonal complement of the subspace $\operatorname{Im}\,\Pi$.

Since the eigenvalues are ordered in decreasing magnitude; i.e. $\{\lambda_1 \geq \ldots \geq \lambda_N\}$, one sees that the minimum of this function of Π is reached when Π projects onto the subspace spanned by the first n coordinate axes. In other words, $\Pi_{optimal} = \operatorname{diag}\,\{I_n,\, 0\}$ the minimum being $\lambda_{n+1} + \ldots + \lambda_N$. It is then evident that

$$M \;=\; \Phi\Lambda^{1/2}\,\Pi_{optimal}\Lambda^{-1/2}\Phi^\top = \Phi_n\Phi_n^\top.$$

Naturally, multiplying Φ_n by any orthogonal $n \times n$ matrix does not change the result. □ □

This result confirms in particular that the truncated expansion (6.2.4) is optimal in the sense that it provides the best M and the best approximation subspace for the criterion (6.2.6). This characterization can be exploited when dealing with subspace approximation problems; see e.g. [110].

Numerical Implementation by SVD

In practice the calculation of the eigenvectors $\hat{\varphi}_k$ can be done without forming the (sample) covariance matrix $\hat{\Sigma}$ which can be a costly operation for high dimensional data. One starts instead from the *data matrix*:

$$Y = \begin{bmatrix} x_1 & \dots & x_M \end{bmatrix} \quad \in \mathbb{R}^{N \times M}, \quad N \leq M$$

where the test signals $x_k\,; k = 1, \dots, M$ have been depurated from their sample mean, and performs a singular value decomposition

$$Y = U \Delta V^\top, \qquad \Delta = \operatorname{diag}\{\sigma_1, \dots, \sigma_N\} \tag{6.2.9}$$

It follows readily from Theorem B.2 in the appendix that the columns of U are just the normalized eigenvectors of $\hat{\Sigma}$ and the ordered sequence $\{\sigma_k^2\}$ is proportional to the sample eigenvalues $\{\hat{\lambda}_k\}$. The truncation of the SVD expansion can be seen as an optimal subspace approximation. This is the linear algebra version of the optimization problem discussed in the previous section, called *Procrustes Problem*, see [40, p. 601].

Application to Feature extraction

Handwritten digit recognition See [44, p. 536-541].

Eigenfaces : Dimensionality reduction usually becomes important when the number of features is not negligible compared to the number of training samples. As an example, suppose we would like to perform face recognition, i.e. determine the identity of the person depicted in an image, based on a training dataset of labeled face images. One approach might be to treat the brightness of each pixel of the image as a feature. If the input images are of size 32×32 pixels, this means that the feature vector contains 1024 feature values. Classifying a new face image can then be done by calculating the Euclidean distance between this 1024-dimensional vector, and the feature vectors of the people in our training dataset. The smallest distance then tells us which person we are looking at.

However, operating in a 1024-dimensional space becomes problematic if we only have a few hundred training samples. Furthermore, Euclidean distances behave strangely in high dimensional spaces as discussed in another article. Therefore, we could use PCA to reduce the dimensionality of the feature space by calculating the eigenvectors of the covariance matrix of the set of 1024-dimensional feature vectors, and then projecting each feature vector onto the largest eigenvectors.

Since the eigenvector of 2D data is 2-dimensional, and an eigenvector of 3D data is 3-dimensional, the eigenvectors of 1024-dimensional data is 1024-dimensional. In other words, we could reshape each of the 1024-dimensional eigenvectors to a 32×32 image for visualization purposes. Figure 6.2.1 shows the first four eigenvectors obtained by eigendecomposition of the Cambridge face dataset Source: `https://nl.wikipedia.org/wiki/Eigenface`,

Each 1024-dimensional feature vector (and thus each face) can now be projected onto the N largest eigenvectors, and can be represented as a linear combination of these eigenfaces. The weights of these linear combinations determine

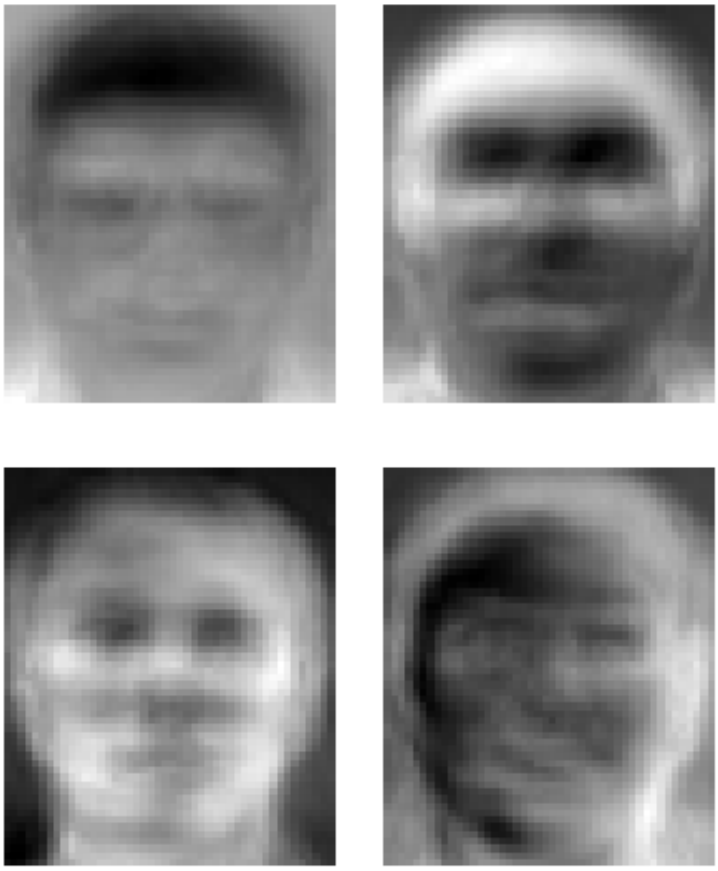


Figure 6.2.1. *The four largest eigenvectors, reshaped to images, resulting in so called EigenFaces. Credit to AT&T Laboratories Cambridge.*

the identity of the person. Since the largest eigenvectors represent the largest variance in the data, these eigenfaces describe the most informative image regions (eyes, noise, mouth, etc.). By only considering the first n (e.g. n=70) eigenvectors, the dimensionality of the feature space is greatly reduced.

The remaining question is now how many eigenfaces should be used, or in the general case; how many eigenvectors should be kept. Removing too many eigenvectors might remove important information from the feature space, whereas eliminating too few eigenvectors leaves us with the curse of dimensionality. Regrettably there is no straight answer to this problem. Although cross-validation techniques can be used to obtain an estimate of this hyperparameter, choosing the optimal number of dimensions remains a problem that is mostly solved in an empirical (an academic term that means not much more than "trial-and-error") manner. Note that it is often useful to check how much (as a percentage) of the variance of the original data is kept while eliminating eigenvectors. This is done by dividing the sum of the kept eigenvalues by the sum of all eigenvalues.

Based on the previous sections, we can now list the simple recipe used to apply PCA for feature extraction:

Center the data Before applying PCA rotate the data in order to obtain uncorrelated axes, any existing shift needs to be countered by subtracting the mean of the data from each data point. This simply corresponds to centering the data such that its average becomes zero.

Normalize the data The eigenvectors of the covariance matrix point in the direction of the largest variance of the data. However, variance is an absolute number, not a relative one. This means that the variance of data, measured in centimeters (or inches) will be much larger than the variance of the same data when measured in meters (or feet). Consider the example where one feature represents the length of an object in meters, while the second feature represents the width of the object in centimeters. The largest variance, and thus the largest eigenvector, will implicitly be defined by the first feature if the data is not normalized.

To avoid this scale-dependent nature of PCA, it is useful to normalize the data by dividing each feature by its standard deviation. This is especially important if different features correspond to different metrics.

Calculate the eigendecomposition One of the most widely used methods to efficiently calculate the eigendecomposition is Singular Value Decomposition (SVD); see above or [44, p. 535].

Project the data To reduce the dimensionality, the data is simply projected onto the largest eigenvectors. Let U be the $N \times n$ matrix whose columns contain the largest eigenvectors and let Y be the original data whose columns contain the different observations. Then the projected data $\hat{Y}$ is obtained as $\hat{Y} = U^\intercal Y U$. We can either choose the number of remaining dimensions, i.e. the columns of Y, directly, or we can define the amount of variance of the original data that needs to kept while eliminating eigenvectors. If only n eigenvectors are kept, and $\lambda_1 ... \lambda_n$ represent the corresponding eigenvalues, then the amount of variance that remains after projecting the original d-dimensional data can be calculated as:

$$s = \frac{\sum_{i=0}^{n} \lambda_i}{\sum_{j=0}^{N} \lambda_j}$$

In the previous discussion we saw how PCA decorrelates the data. In fact, we started the discussion by expressing our desire to recover the unknown, underlying independent components of the observed features. Indeed, PCA allows us to decorrelate the data, thereby recovering the *independent* components in case of Gaussianity. However, it is important to note that decorrelation only corresponds to statistical independency in the Gaussian case.

In general, PCA only uncorrelates the data but does not remove statistical dependencies. If the underlying components are known to be non-Gaussian, techniques such as *independent Component Analysis* (ICA) could be more appropriate.

6.3 ▪ Canonical Correlation Analysis

Bayesian regression with redundant data

Suppose you want to do linear regression of a high dimensional (zero-mean finite variance) random vector $\mathbf{y}$ of dimension m based on an input $\mathbf{x}$ which is possibly also high dimensional. A typical situation could be the prediction of some linear function of the future variables at certain present time t of a stochastic process given observations of an extended past sequence (before time t) which theoretically could even be infinitely long.
The regression data can be highly redundant on both sides and there is a natural data compression problem to be faced. The goal is to describe the interaction structure of the two vectors in the most succinct and efficient way.
Notation: In this lecture it will be convenient to use the simplified notation $\mathbb{E}^{\mathbf{X}}\boldsymbol{\xi}$ to denote the orthogonal projection of a random variable $\boldsymbol{\xi} \in \mathbf{H}$ onto a subspace $\mathbf{X} \subset \mathbf{H}$ (disregarding the hat on top of $\mathbb{E}$). Whenever a vector of generators say $\mathbf{x}$ is in evidence, so that $\mathbf{X} = \mathbf{H}(\mathbf{x})$, then the projection can be expressed as a linear function of $\mathbf{x}$.

Recall the Bayesian regression formula (conditional expectation in the Gaussian case) assuming all variables are zero mean:

$$\hat{\mathbf{y}} := \mathbb{E}[\mathbf{y} \mid \mathbf{x}] = \Sigma_{y,x}\Sigma_x^{-1}\mathbf{x} := A\mathbf{x} \tag{6.3.1}$$

where we have used the standard notation $\mathbb{E}\,\mathbf{y}\mathbf{x}^\top = \Sigma_{y,x} \equiv \Sigma$. This could be a very large matrix and you may think that there should be a way to extract the "essential interaction" information from the two long vectors which may possibly lead to a dimension reduction by extracting "principal components" from both vectors. Tis is in principle similar to what we did for PCA but now we have the different goal of describing the mutual interaction between the two random vectors which may possibly be concentrated in another random vector of much smaller dimension than the dimensions of $\mathbf{x}$ and $\mathbf{y}$. Motivated by the above discussion, consider then the following problem

Problem 6.3. *Given two zero mean random vectors $\mathbf{y}$ and $\mathbf{x}$ of dimensions m and p find the coefficients of linear combinations*

$$\boldsymbol{\eta} := a^\top\mathbf{y}, \qquad \textit{and} \qquad \boldsymbol{\xi} := b^\top\mathbf{x}$$

so that $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ have maximal cross correlation $\mathbb{E}\,\boldsymbol{\eta}\boldsymbol{\xi}$. Of course for the problem to make sense $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ should be normalized (otherwise the cross correlation could be made arbitrarily large). We ask that both norms (equivalently the variances) $\|\boldsymbol{\xi}\|$ and $\|\boldsymbol{\eta}\|$ should be equal to one.

In a sense the solution would yield a best linear model to predict a certain linear function of $a^\top\mathbf{y}$ from observations of $b^\top\mathbf{x}$. The error variance of estimating $a^\top\mathbf{y}$ would be $1 - (\mathbb{E}\,\boldsymbol{\eta}\boldsymbol{\xi})^2$ which would be minimal among all normalized linear functions $a^\top\mathbf{y}$, $b^\top\mathbf{x}$. Since the error variance is always nonnegative, the absolute value of the cross product must always be ≤ 1.

Since both $\mathbf{X} := \mathbf{H}(\mathbf{x})$ and $\mathbf{Y} := \mathbf{H}(\mathbf{y})$ are finite dimensional the maximum exists. It has in fact a nice geometrical interpretation which we shall explore later on.

Solution of the unconstrained problem

Let $\Sigma = U\Delta V^\top$ with $\Delta = \operatorname{diag}\{\sigma_1, \ldots, \sigma_n, 0 \ldots, 0\}$ be the singular value decomposition of Σ with the singular values ordered in decreasing magnitude. Since

$$\mathbb{E}\,\boldsymbol{\eta}\boldsymbol{\xi} = a^\top \Sigma\, b = a^\top U \Delta V^\top b$$

the (unconstrained) maximum of $\mathbb{E}\,\boldsymbol{\eta}\boldsymbol{\xi}$ is achieved for vectors a, b such that

$$a^\top U = \begin{bmatrix} 1 & 0, & \ldots & 0 \end{bmatrix}, \qquad V^\top b = \begin{bmatrix} 1 & 0, & \ldots & 0 \end{bmatrix}^\top$$

are both the first unit vectors (generically denoted e_1) of the canonical bases of dimensions m and p in $\mathbb{R}^m$ and $\mathbb{R}^p$. Since $\boldsymbol{\eta} = a^\top U\, U^\top \mathbf{y} = e_1^\top\, U^\top \mathbf{y}$ and similarly, $\boldsymbol{\xi} = b^\top V\, V^\top \mathbf{x} = e_1^\top\, V^\top \mathbf{x}$ the optimal linear combinations $\boldsymbol{\eta}$ and $\boldsymbol{\xi}$ are the first components of the rotated vectors

$$\hat{\mathbf{y}} := U^\top \mathbf{y}, \quad \text{and} \quad \hat{\mathbf{x}} := V^\top \mathbf{x}$$

which span the same spaces $\mathbf{H}(\mathbf{y})$ and $\mathbf{H}(\mathbf{x})$ but have **mutually orthogonal components** since $\mathbb{E}\,\hat{\mathbf{y}}\hat{\mathbf{x}}^\top = \Delta$ that is

$$\mathbb{E}\,\hat{\mathbf{y}}_j \hat{\mathbf{x}}_h = \sigma_h\, \delta_{j,h}, \qquad j, h = 1, \ldots, n$$

and zero otherwise, where $\delta_{j,h}$ is the Kronecker symbol. □

Solution of the constrained problem

The previous solution did not take into account the unit norm constraints on the variables $\boldsymbol{\eta}$ and $\boldsymbol{\xi}$. Note that in order to express the inner product of random elements in $\mathbf{X}$ and $\mathbf{Y}$ in terms of their coordinates, we must introduce appropriate weights in the corresponding Euclidean inner products. In fact, the inner product of two scalar elements $\boldsymbol{\eta}_i = a_i^\top \mathbf{y} \in \mathbf{Y}$, $i = 1, 2$, induces in $\mathbb{R}^m$ the inner product

$$\mathbb{E}\,\boldsymbol{\eta}_1\boldsymbol{\eta}_2 = a_1^\top \Sigma_{\mathbf{y}}\, a_2 = \langle a_1,\, a_2 \rangle_{\Sigma_{\mathbf{y}}},$$

where $\Sigma_{\mathbf{y}} := \mathbb{E}\,\{\mathbf{y}\mathbf{y}^\top\}$. Similarly, there is an inner product $\langle b_1, b_2 \rangle_{\Sigma_{\mathbf{x}}} := b_1^\top \Sigma_{\mathbf{x}} b_2$ corresponding to the basis $\boldsymbol{\xi}$ for $\mathbf{X}$. To obtain the standard Euclidean inner product the bases should need to be orthonormal (which in general is not the case) and it is only in this case that the matrix representation of the adjoint of the restricted orthogonal projection operator $\mathbb{E}^{\mathbf{X}}|_{\mathbf{Y}}$ is the transpose of the matrix representation of $\mathbb{E}^{\mathbf{Y}}|_{\mathbf{X}}$.

Let now $L_{\mathbf{x}}$ and $L_{\mathbf{y}}$ be the lower triangular Cholesky factors of the covariance matrices $\Sigma_{\mathbf{x}}$ and $\Sigma_{\mathbf{y}}$, respectively; i.e.,

$$L_{\mathbf{x}}\, L_{\mathbf{x}}^\top = \Sigma_{\mathbf{x}}, \qquad L_{\mathbf{y}}\, L_{\mathbf{y}}^\top = \Sigma_{\mathbf{y}}$$

and let the corresponding innovation vectors be denoted

$$\boldsymbol{\nu}_x := L_x^{-1}\mathbf{x}, \qquad \boldsymbol{\nu}_y := L_y^{-1}\mathbf{y} \tag{6.3.2}$$

which form orthonormal bases in $\mathbf{X}$ and $\mathbf{Y}$ respectively. Then in this new basis the cross covariance becomes the matrix

$$H := \mathbb{E}\,\{\boldsymbol{\nu}_{\mathbf{y}}\boldsymbol{\nu}_{\mathbf{x}}^\top\} = L_y^{-1}\mathbb{E}\,\{\mathbf{y}\mathbf{x}^\top\}(L_x^\top)^{-1}\,. \tag{6.3.3}$$

Consider again two scalar random variables

$$\boldsymbol{\eta} = a^\top \mathbf{y} = a^\top L_{\mathbf{y}} \boldsymbol{\nu}_y, \qquad \boldsymbol{\xi} = b^\top \mathbf{x} = b^\top L_{\mathbf{x}} \boldsymbol{\nu}_x$$

letting $\hat{a} := L_{\mathbf{y}}^\top a$ and $\hat{b} := L_{\mathbf{x}}^\top b$ we have $\|\boldsymbol{\eta}\|^2 = \|\hat{a}\|^2$ and $\|\boldsymbol{\xi}\|^2 = \|\hat{b}\|^2$ where the last norms are Euclidean, so that

$$\frac{\mathbb{E}\,\boldsymbol{\eta}\boldsymbol{\xi}}{\|\boldsymbol{\eta}\|\|\boldsymbol{\xi}\|} = \frac{a^\top L_y H L_x^\top b}{\|\hat{a}\|\|\hat{b}\|} = \frac{\hat{a}^\top H \hat{b}}{\|\hat{a}\|\|\hat{b}\|}, \quad \text{where} \quad \|\hat{a}\| = \|\hat{b}\| = 1$$

Hence, by the well-known Rayleigh-Ritz maximization theorem our constrained maximization problem 6.3 becomes the problem of finding the maximum singular value of H.

To complete the discussion, consider the (full rank) singular value decomposition of H

$$H = \hat{U} D \hat{V}^\top, \qquad \hat{U}\hat{U}^\top = I_m, \quad \hat{V}\hat{V}^\top = I_p$$

where $D = \operatorname{diag}\{\kappa_1, \ldots, \kappa_n\}$, n being the rank of H or of Σ. The elements of D are called the **canonical correlation coefficients** of the subspaces $\mathbf{X}$ and $\mathbf{Y}$ and the components of the n-dimensional random vectors

$$\mathbf{u} = \hat{V}^\top \boldsymbol{\nu}_x, \qquad \mathbf{v} = \hat{U}^\top \boldsymbol{\nu}_y. \tag{6.3.4}$$

are called the **canonical variables**. Obviously the components of $\mathbf{u}$ and $\mathbf{v}$ are still orthonormal and have diagonal covariance, that is

$$\mathbb{E}\,\mathbf{v}\mathbf{u}^\top = D = \begin{bmatrix} \kappa_1 & 0 & \cdots & 0 \\ 0 & \kappa_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \kappa_n \end{bmatrix}, \tag{6.3.5}$$

so that the optimal $\boldsymbol{\eta}$ and $\boldsymbol{\xi}$ are just the first components, $\mathbf{v}_1$ and $\mathbf{u}_1$ of the canonical vectors. Since the κ_k are all covariances and the canonical variables are normalized, we have

$$\mathbb{E}^{\,\mathbf{X}} \mathbf{v}_k = \kappa_k \mathbf{u}_k\,, \qquad \mathbb{E}^{\,\mathbf{Y}} \mathbf{u}_k = \kappa_k \mathbf{v}_k\,. \tag{6.3.6}$$

In particular, (6.3.5) follows from

$$\langle \mathbf{u}_k, \mathbf{v}_j \rangle = \langle \mathbb{E}^{\,\mathbf{Y}} \mathbf{u}_k, \mathbf{v}_j \rangle = \kappa_k \langle \mathbf{v}_k, \mathbf{v}_j \rangle = \kappa_k \delta_{kj}\,.$$

Note that the canonical correlation coefficients are positive and all less or equal to 1 while the singular values, σ_k, of Σ may be arbitrarily large. Uniqueness of the solution is guaranteed if and only if the singular values $\{\kappa_1, \kappa_2, \ldots, \kappa_n\}$ are distinct. □

What is the meaning of canonical variables in terms of our Bayesian regression problem? The answer is obvious if we want to predict a canonical vector based on our data,

Proposition 6.3. *It holds that*

$$\mathbb{E}\,\{\mathbf{v} \mid \mathbf{x}\} = \mathbb{E}\,\{\mathbf{v} \mid \mathbf{u}\} = D\,\mathbf{u} \tag{6.3.7a}$$

$$\mathbb{E}\,\{\mathbf{u} \mid \mathbf{y}\} = \mathbb{E}\,\{\mathbf{u} \mid \mathbf{v}\} = D\,\mathbf{v} \tag{6.3.7b}$$

and likewise, for the unnormalized canonical variables $\hat{\mathbf{x}} := D^{1/2}\mathbf{u}$ *and* $\hat{\mathbf{y}} := D^{1/2}\mathbf{v}$, *we have*

$$\mathbb{E}\{\hat{\mathbf{y}} \mid \mathbf{x}\} = \mathbb{E}\{\hat{\mathbf{y}} \mid \hat{\mathbf{x}} :\} = D\,\hat{\mathbf{x}} \tag{6.3.7c}$$

$$\mathbb{E}\{\hat{\mathbf{x}} \mid \mathbf{y}\} = \mathbb{E}\{\hat{\mathbf{x}} \mid \hat{\mathbf{y}}\} = D\,\hat{\mathbf{y}} \tag{6.3.7d}$$

Hence the (ordered) canonical variables have a decreasing sequence of regression coefficients equal exactly to their mutual canonical correlations. Depending on the magnitude of $\frac{\kappa_1}{\kappa_n}$ one may reasonably discard canonical regressors of sufficiently high order. We shall come back to this point as we first need to assess this phenomenon in terms of the original variables.

Note that from (6.3.3) we get the following **rank factorization** of the covariance matrix

$$\Sigma = L_y H L_x^\top = L_y \hat{U} D (L_x \hat{V})^\top := \Omega\, D\, \bar{\Omega}^\top \tag{6.3.8}$$

where $\Omega = L_y\hat{U} \in \mathbb{R}^{m\times n}$, $\bar{\Omega} = L_x\hat{V} \in \mathbb{R}^{p\times n}$ are full rank.This means that the components of $\mathbf{y}$ and of $\mathbf{x}$ along the unnormalized canonical variables are

$$\mathbb{E}\{\mathbf{y} \mid \mathbf{x}\} = \mathbb{E}\{\mathbf{y} \mid \mathbf{u}\} = \Omega\,\hat{\mathbf{x}} \tag{6.3.9}$$

$$\mathbb{E}\{\mathbf{x} \mid \mathbf{y}\} = \mathbb{E}\{\mathbf{x} \mid \mathbf{v}\} = \bar{\Omega}\,\hat{\mathbf{y}} \tag{6.3.10}$$

so that the solution of the regression problem of $\mathbf{y}$ in terms of $\mathbf{x}$ can be expressed as the orthogonal sum

$$\mathbf{y} = \Omega\,\hat{\mathbf{x}} + \mathbf{e} \tag{6.3.11}$$

where $\mathrm{Var}[\mathbf{e}] = \Omega[I - D]\,\Omega^\top$ as it follows from $\Sigma_{\mathbf{y}} = \Omega\Omega^\top$ and the standard formula for the error variance. Therefore,

Proposition 6.4. *The (prediction) error* $\mathbf{e}$ *in the model* (6.3.11) *has variance*

$$\mathrm{var}\{\mathbf{e}\} = \mathrm{Trace}\,([I - D]\,\Omega^\top\Omega) = \sum_{i=1}^{n}(1-\kappa_i)\,\mathrm{var}\;\mathbf{y}_i\,. \tag{6.3.12}$$

an analogous formula holding for the regression of $\mathbf{x}$ *in terms of* $\mathbf{y}$.

Proof. In computing the trace, one must compute the scalar product of the vector of the diagonal elements of $[I - D]$ times the vector of diagonal elements of $\Omega^\top\Omega$ which is the same as the vector of diagonal elements of $\Omega\Omega^\top = \Sigma_{\mathbf{y}}$, listing the variances of the scalar components of $\mathbf{y}$. □

For high dimensional problems one could then introduce low order approximations by neglecting the (ordered) components of the canonical interaction vector $\hat{\mathbf{x}}$ which have small variance.

Naturally to solve efficiently our regression problem we should find a sequential procedure (algorithm) to compute such a pair of orthonormal sequences of uncorrelated random variables, each spanning the input and output subspaces. To this end one may follow a "bilateral" Gram-Scmidt procedure. Introduce the notation $\mathbf{x}_1 \equiv \mathbf{x}$, $\mathbf{y}_1 \equiv \mathbf{y}$ and

$$\mathbf{X}_1 := \mathbf{H}(\mathbf{x}); \quad \text{and} \quad \mathbf{Y}_1 := \mathbf{H}(\mathbf{y})$$

which we shall assume of full dimensions p and m, respectively. Hence let $\mathbf{u}_1 \in \mathbf{X}$ and $\mathbf{v}_1 \in \mathbf{Y}$ be the (scalar random variables) solutions of our problem 6.3, that is

$$\begin{cases} \max & \langle \mathbf{v}_1, \mathbf{u}_1 \rangle \\ \mathbf{u}_1 \in \mathbf{X},\ \mathbf{v}_1 \in \mathbf{Y} \\ \|\mathbf{u}_1\| = 1,\ \|\mathbf{v}_1\| = 1 \end{cases} \tag{6.3.13}$$

and let κ_1 be the maximum.

Define next the orthogonal complements

$$\mathbf{X}_2 := \mathbf{X}_1 \ominus \mathbf{H}(\mathbf{u}_1)\,, \quad \text{and} \quad \mathbf{Y}_2 = \mathbf{Y}_1 \ominus \mathbf{H}(\mathbf{v}_1) \tag{6.3.14}$$

which are spanned by the random vectors

$$\mathbf{x}_2 := \mathbf{x}_1 - \mathbb{E}\,[\mathbf{x}_1 \mid \mathbf{u}_1]\,; \qquad \mathbf{y}_2 := \mathbf{y}_1 - \mathbb{E}\,[\mathbf{y}_1 \mid \mathbf{v}_1], \tag{6.3.15}$$

and look for normalized random variables $\mathbf{u}_2 \in \mathbf{X}_2$ and $\mathbf{v}_2 \in \mathbf{Y}_2$ which maximize $\mathbb{E}\,\mathbf{v}_2\mathbf{u}_2$. This is the same problem solved above but for the error subspaces $\mathbf{X}_2$ and $\mathbf{Y}_2$. Continuing in this way leads to a sequential maximization procedure which defines a sequence of orthonormal random variables $\{\mathbf{u}_k \in \mathbf{X}\}$ and $\{\mathbf{v}_k \in \mathbf{Y}\}$ solving:

$$\langle \mathbf{v}_{k+1}, \mathbf{u}_{k+1} \rangle = \max_{\mathbf{u} \in \mathbf{X}_k,\, \mathbf{v} \in \mathbf{Y}_k} \langle \mathbf{v}, \mathbf{u} \rangle \tag{6.3.16a}$$

subject to:

$$\begin{cases} \langle \mathbf{u}, \mathbf{u}_j \rangle = 0, \quad j = 1, \dots, k, \\ \langle \mathbf{v}, \mathbf{v}_j \rangle = 0, \quad j = 1, \dots, k, \\ \|\mathbf{u}_j\| = \|\mathbf{v}_j\| = 1, \quad \forall\, j \end{cases} \tag{6.3.16b}$$

It is not hard to show that this sequential maximization procedure leads exactly to the same canonical variables and canonical correlation coefficients defined above.

Theorem 6.2. *The sequential maximization procedure* (6.3.16) *defines two sequences of orthonormal random variables* $\{\mathbf{u}_1, \dots, \mathbf{u}_n\} \in \mathbf{X}$ *and* $\{\mathbf{v}_1, \dots, \mathbf{v}_n\}, \in \mathbf{Y}\}$ *such that*

$$\mathbb{E}\,\{\mathbf{u}_k\, \mathbf{v}_j\} = \kappa_k\, \delta_{kj}, \quad k = 1, \dots, n,\ j = 1, \dots, n.$$

where $n \le \min\{m, p\} = \operatorname{rank}\, \Sigma$. *The two random vectors* $\mathbf{u} := (\mathbf{u}_1, \dots, \mathbf{u}_n)^\top$ *and* $\mathbf{v} := (\mathbf{v}_1, \dots, \mathbf{v}_n)^\top$ *formed from the elements of the two bases, are the canonical vectors and the numbers* κ_k *are the canonical correlation coefficients of the two subspaces* $\mathbf{X}, \mathbf{Y}$. *Moreover*

$$\Sigma_{\mathbf{v},\mathbf{u}} = \mathbb{E}\,\{\mathbf{v}\mathbf{u}^\top\} = D\,.$$

where D is the diagonal matrix of ordered canonical correlation coefficients in (6.3.5).

Problem 6.4 (A matrix representation of the operator $\mathbb{E}^{\mathbf{Y}}|_{\mathbf{X}}$).

Choose an arbitrary pair of basis vectors $\mathbf{x}, \mathbf{y}$ in $\mathbf{X}$ and $\mathbf{Y}$, respectively and an arbitrary random variable $\xi = a^\top \mathbf{x} \in \mathbf{X}$; then the Bayesian estimation formula yields

$$\mathbb{E}^{\mathbf{Y}}\xi = a^\top \mathbb{E}\,\{\mathbf{x}\mathbf{y}^\top\}\Sigma_{\mathbf{y}}^{-1}\mathbf{y}, \qquad \Sigma_{\mathbf{y}} := \mathbb{E}\,\{\mathbf{y}\mathbf{y}^\top\},$$

Show that the representation of $\mathbb{E}^{\mathbf{Y}}|_{\mathbf{X}}$ in the chosen bases is matrix multiplication $a^\top \to a^\top \Sigma_{\mathbf{xy}} \Sigma_{\mathbf{y}}^{-1}$ acting from the right.

6.4 ▪ Bayesian regression on observed samples.

So far we have introduced canonical correlation analysis in a random variable setting, but in practice one should work with observed data.

Suppose we have sequences of i.i.d. observations of the output vector say $\{\mathbf{y}_1(\omega) = x_1, \mathbf{y}_2(\omega) = x_2, \dots \mathbf{y}_N(\omega) = x_N\}$ and corresponding i.i.d. observations of the input $\{\mathbf{x}_1(\omega) = x_1, \mathbf{x}_2(\omega) = x_2, \dots \mathbf{x}_N(\omega) = x_N\}$ where for simplicity we assume equal leght N. After the sample observed vectors $x_k\,; k = 1, \dots, N$ and $x_k\,; k = 1, \dots, N$ have been depurated from their sample mean, form the *data matrices*:

$$Y := \begin{bmatrix} y_1 & \dots & y_N \end{bmatrix} \in \mathbb{R}^{m\times N} \quad \text{and} \quad X := \begin{bmatrix} x_1 & \dots & x_N \end{bmatrix} \in \mathbb{R}^{p\times N}$$

from which one can get the (sample) covariance matrices

$$\hat{\Sigma} := \frac{1}{N} Y X^\top, \qquad \hat{\Sigma}_x := \frac{1}{N} X X^\top \tag{6.4.1}$$

which can be used as consistent estimates of the population parameters $\Sigma_{\mathbf{y},\mathbf{x}}$, $\Sigma_{\mathbf{x}}$ to obtain a sample estimate of the regression matrix A in (6.3.1). The sample canonical correlation coefficients could likewise be computed by first doing Choleski factorizations of the sample covariance matrices $\hat{\Sigma}_{\mathbf{x}}$ and $\hat{\Sigma}_{\mathbf{y}}$, respectively; i.e.,

$$\hat{L}_{\mathbf{x}}\, \hat{L}_{\mathbf{x}}^\top = \hat{\Sigma}_{\mathbf{x}}, \qquad \hat{L}_{\mathbf{y}}\, \hat{L}_{\mathbf{y}}^\top = \hat{\Sigma}_{\mathbf{y}}$$

and then doing SVD of the normalized sample cross covariance

$$\hat{H} := \hat{L}_y^{-1} \hat{\Sigma} (\hat{L}_x^\top)^{-1}\,. \tag{6.4.2}$$

The operation can actually be performed sequentially by implementing an evident sample version of the sequential maximization (6.3.16). It will sequentially produce the ordered components of the sample canonical vectors. The procedure lies at the background of a variety of algorithms called *Partial Least Squares* [**?**] .

The calculations involved, beside being costly for high dimensional data may suffer from bad conditioning. One should instead start directly from the data and apply a QR-type orthogonalization techniques similar to that which was suggested for least squares problems.

Let us consider the problem of estimating (by least squares) the matrix $A \in \mathbb{R}^{m\times p}$ in the regression model

$$Y = AX + E$$

from the observed samples X, Y. Here all data matrices, including the error E have N columns, the k-th row being interpreted as a list of i.i.d. observations from the random scalar components say $\mathbf{y}_k$ or $\mathbf{x}_k$. We assume that $\frac{1}{N} X E^\top \to 0$ as $N \to \infty$. The normalization factor $1/N$ in the covariance estimates like (8.3.3) will normally be neglected as it will cancel in forming the regression matrix.

Solution by LQ factorization

The LQ factorization, is the transpose of the better known QR factorization, a well-known procedure in numerical linear algebra. It states that any rectangu-

lar matrix M of full row rank can be factorized as

$$M = LQ^\top$$

where L is square lower triangular and $Q^\top$ has orthonormal rows, that is $Q^\top Q = I$. This factorization turns out to be very useful, especially in computations involving solutions of least squares problems and orthogonal projections on large-dimensional data spaces. For example, instead of using Cholesky factorizations we can compute the sample innovation matrices from the following LQ factorizations:

$$Y = L_Y N_Y^\top, \qquad X = L_X N_X^\top \tag{6.4.3}$$

where $N_X^\top \in \mathbb{R}^{p\times N}$ and $N_Y^\top \in \mathbb{R}^{m\times N}$ are the sample matrices generated by the random innovation vectors (6.3.2). For our regression problem we generalize as follows

Proposition 6.5. *There is an $m \times N$ matrix N_E such that*

$$\begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} L_X & 0 \\ L_{21} & L_E \end{bmatrix} \begin{bmatrix} N_X^\top \\ N_E^\top \end{bmatrix},$$

where and L_X, L_E are lower triangular, $N_X^\top N_X = I_p$, $N_X^\top N_E = 0$, $N_E^\top N_E = I_m$, The rows of $N_X^\top$ form an orthonormal basis for the rowspace $\mathcal{X} := \text{row-span } X$ and

$$\Pi\left[Y \mid \mathcal{X}\right] = Y N_X \left[N_X^\top N_X\right]^{-1} N_X^\top = L_{21} N_X^\top \tag{6.4.4a}$$

$$\Pi\left[Y \mid \mathcal{X}^\perp\right] = Y N_E \left[N_E^\top N_E\right]^{-1} N_E^\top = L_{22} N_E^\top \tag{6.4.4b}$$

where $\Pi\left[\cdot \mid \mathcal{X}\right]$ denotes row-wise orthogonal projection onto the rowspace $\mathcal{X}$.

From (6.5) we can get the (unnormalized) sample covariance as

$$N\,\hat{\Sigma}_{y,x} = YX^\top = L_{21} L_X^\top \in \mathbb{R}^{m\times p}$$

Assume that $\text{rank } X = p$, then L_X is invertible, and from (6.5) one obtains the estimate of the regression model $Y = AX + E$ in the form,

$$Y = L_{21} L_X^{-1} X + L_E N_E^\top \tag{6.4.5}$$

so that the sample estimate of the regression matrix is $A = L_{21} L_X^{-1}$ with sample regression error $E = L_E N_E^\top$. By construction the rows of $N_X^\top$ and of of $N_E^\top$ are a sequential orthonormal basis obtained by a Gram-Schmidt orthonormalization procedure starting from the first row of $\begin{bmatrix} X \\ Y \end{bmatrix}$ and proceeding downwards. $N_E^\top$ is seen to be an orthonormal basis of the row space of the residual error space $\mathcal{Y} \ominus \Pi\left[Y \mid \mathcal{X}\right]$.

Next the normalized matrix $\hat{H}$ is obtained by substituting the first of (6.4.3) into (6.4.5) to get the orthogonal decomposition

$$N_Y^\top = L_Y^{-1} L_{21} N_X^\top + L_Y^{-1} L_E N_E^\top \tag{6.4.6}$$

where evidently the first term must be the sample regression of the innovation matrix $N_Y^\top$ in terms of $N_X^\top$. In other words

$$\hat{H} = L_Y^{-1} L_{21} \tag{6.4.7}$$

Note that L_Y is lower triangular and the procedure to get $\hat{H}$ involves only operations on the data matrices without forming sample covariances and requires a much smaller amount of computation than the one alluded at in the previous paragraph. The "canonical" SVD decomposition could actually be done without normalizing by the inverse of L_Y (i.e. essentially directly on L_{21}) by a technique called the *Quotient, or Generalized, Singular Value Decomposition* [38, p. 318].

Hypothesis Testing for the rank n

In practice $n = \operatorname{rank}\ \Sigma$ is unknown and it is important to have a statistical procedure to estimate it. Since the canonical correlations are naturally a decreasing sequence, estimating n reduces to deciding if the i-th sample canonical correlation $\hat{\kappa}_i$ with $i \leq \min\{m, p\}$ is zero or not. One could then implement a testing procedure starting with the minimum possible index $i = \min\{m, p\}$ and proceeding upwards. It has been proven, see e.g. **[?]**, that for $N \to \infty$ the statistic

$$\mathbf{c}_i := -\left(N - 1 - \frac{1}{2}(m + p + 1)\right) \log \left\{ \prod_{k=i}^{\min\{m.p\}} (1 - \hat{\boldsymbol{\kappa}}_k^2) \right\}$$

has a χ^2 distribution with $(m - i + 1)(p - i + 1)$ degrees of freedom. Testing if $\mathbf{c}_i = 0$ is clearly the same as testing if $(1 - \hat{\kappa}_k^2) = 1, \text{ for } k = i, \ldots, \min\{m, p\}$.

Problem 6.5. *Analyze the case in which the canonical correlations $\kappa_1 = \ldots = \kappa_r$ are all equal to 1 and $\kappa_{r+1} < 1$. Show that this can happen if and only if $\mathbf{Y} \cap \mathbf{X} \neq \{0\}$ and has in fact exactly dimension r.*

What happens then to the rank of the matrix $\begin{bmatrix} Y \\ X \end{bmatrix}$ for $N \to \infty$?

6.5 ▪ Contimuous parameter: the Karhunen-Loève expansion

The Karhunen-Loève expansion is the analog of PCA for signals depending on a continuous real parameter, typically continuous-time. Although from a conceptual point of view there are no novelties, more sophisticated mathematics is needed; in particular the spectral decomposition of the covariance matrix Σ must be replaced by an eigenfunction expansion of a certain integral operator. We shall just quote the main results without justification.

Let us consider a continuous time random signal $\mathbf{y} := \{\mathbf{y}(t);\ t \in \mathbb{T}\}$, the variabile t now ranging on the interval $[0,\, T]\,, T \leq +\infty$ of the real line which for short we shall denote $\mathbb{T}$. As before we shall assume (w.l.o.g.) that $\mathbf{y}$ has zero mean. The covariance function

$$R(t,s) := \mathbb{E}\,\{\mathbf{y}(t)\mathbf{y}(s)\} \tag{6.5.1}$$

is assumed to be continuous in both arguments. This condition is equivalent to mean square continuity of the process. An essential technical assumption is that

$$\int_{\mathbb{T}}\int_{\mathbb{T}} R(t,s)^2\, dtds \;<\; \infty \tag{6.5.2}$$

Clearly, when $\mathbb{T}$ is a finite interval this condition is automatically satisfied. Let us now consider the inner product space space $C^2[\mathbb{T}]$ of continuous (deterministic) signals endowed with the inner product

$$\langle f,\, g\rangle \;:= \int_{\mathbb{T}} f(t)\, g(t)\, dt$$

This space is not complete (i.e. Hilbert) in general. It is immediate to check that, in force of condition (6.5.2), the linear operator Σ_R defined by

$$[\Sigma_R\, f](t) \;:=\; \int_{\mathbb{T}} R(t,s)\, f(s)\, ds \tag{6.5.3}$$

maps $C^2[\mathbb{T}]$ into itself. Iin fact, in force of condition (6.5.2), Σ_R turns out to be a *compact operator*. The eigenvalues and the corresponding eigenfunctions of an integral operator of this kind are pairs λ, φ with $0 < \|\varphi\|_{L^2[\mathbb{T}]} < \infty$, which satisfy

$$[\Sigma_R\, \varphi](t) \;=\; \int_{\mathbb{T}} R(t,s)\, \varphi(s)\, ds \;=\; \lambda\, \varphi(t) \qquad t \in \mathbb{T} \tag{6.5.4}$$

Although for a general linear operator, eigenvalues may not exist at all, it is well-known that under the compactness condition (6.5.2), the operator Σ_R does admit eigenvalues. In fact it behaves virtually like a finite dimensional operator described by a symmetric positive definite matrix. The following is the central result of the theory; see e.g. [4, 26] for the complete story.

Theorem 6.3 (Mercèr). *Under the stated assumptions, the following holds:*

1. *The eigenvalue problem* (6.5.4) *admits solutions and all eigenvalues are real and non-negative.*

2. *The eigenvalues of the problem* (6.5.4) *form a monotone nonincreasing sequence of positive numbers (not necessarily distinct), whose only accumulation point can*

be 0. *The corresponding eigenfunctions are all continuous and belong to* $C^2[\mathbb{T}]$*; they can be made orthonormal, that is, such that*

$$\int_{\mathbb{T}} \varphi_k(t)\varphi_j(t)\,dt \;=\; \delta_{k,j}\,.$$

3. *The covariance function* (6.5.1) *admits the following expansion*

$$R(t,s) \;=\; \sum_{k=0}^{\infty} \lambda_k\,\varphi_k(t)\varphi_k(s), \qquad t,s \;\in\; \mathbb{T}\times\mathbb{T} \tag{6.5.5}$$

the series being pointwise uniformly convergent on $\mathbb{T}\times\mathbb{T}$.

The eigenvalues can be defined by an iterated *Rayleigh quotient* algorithm which is summarized below.

1. There is a maximal eigenvalue λ_0 which is given by the formula

$$\lambda_0 = \max_{\|\varphi\|_{L^2[\mathbb{T}]}=1} \langle \Sigma_R\varphi,\,\varphi\rangle \;=\; \max_{\|\varphi\|_{L^2[\mathbb{T}]}=1} \int_{\mathbb{T}}\int_{\mathbb{T}} R(t,s)\,\varphi(t)\varphi(s)\,dtds \tag{6.5.6}$$

   The corresponding eigenfunction, $\varphi_0(t)$, is a continuous function and belongs to $C^2[\mathbb{T}]$.

2. The function $R_1(t,s) \;:=\; R(t,s) - \lambda_0\varphi_0(t)\varphi_0(s)$ is still a covariance function (of positive type) satisfying the compactness condition (6.5.2). Hence the eigenvalue problem

$$[\Sigma_{R_1}\,\varphi](t) \;:=\; \int_{\mathbb{T}} R_1(t,s)\,\varphi(s)\,ds \;=\; \lambda\,\varphi(t) \tag{6.5.7}$$

   still has a maximal eigenvalue, λ_1, given by

$$\lambda_1 = \max_{\|\varphi\|_{L^2[\mathbb{T}]}=1} \langle \Sigma_{R_1}\varphi,\,\varphi\rangle \;=\; \max_{\|\varphi\|_{L^2[\mathbb{T}]}=1} \int_{\mathbb{T}}\int_{\mathbb{T}} R_1(t,s)\,\varphi(t)\varphi(s)\,dtds \tag{6.5.8}$$

   and $\lambda_1 \le \lambda_0$.

3. The procedure can be iterated *ad infinitum*.

By truncating the expansion (6.5.5) to the first $n+1$ terms, one can obtain an approximation of rank $n+1$ of the covariance function $R(t,s)$,

$$R(t,s) \;\simeq\; R_n(t,s) \;:=\; \sum_{k=0}^{n} \lambda_k\,\varphi_k(t)\varphi_k(s), \tag{6.5.9}$$

It is possible to show that this approximation is the best possible in a variety of ways. For example, the linear operator Σ_{R_n} defined by

$$[\Sigma_{R_n}\,f](t) \;:=\; \int_{\mathbb{T}} R_n(t,s)\,f(s)\,ds$$

is the best approximant of rank $n+1$ of Σ_R, in the sense that it solves the constrained optimum problem

$$\min_{\text{rank }(\,\Sigma\,)\,=\,n+1} \|\Sigma_R \,-\, \Sigma\| \tag{6.5.10}$$

the minimum being exactly λ_{n+1}, the first neglected eigenvalue. Here the norm is the *operator norm* (see (D.0.6) in the appendix or e.g. [3]).

The following is the continuous time analog of Proposition 6.1

Proposition 6.6. *If the covariance function of the random process* $\mathbf{y}$ *is a continuous function satisfying* (6.5.2)*, then* $\mathbf{y}$ *admits the biorthogonal expansion* [17]

$$\mathbf{y}(t) \,=\, \sum_{k=0}^{+\infty} \sqrt{\lambda_k}\, \varphi_k(t)\, \mathbf{x}_k \tag{6.5.11}$$

where $\lambda_k;\ k = 0, 1, \ldots$ *are the eigenvalues of the operator* Σ_R*, ordered in decreasing magnitude,* $\varphi_k;\ k = 0, 1, \ldots$ *the corresponding normalized eigenfunctions and the random variables* $\mathbf{x}_k;\ k = 0, 1, \ldots$ *are defined by*

$$\mathbf{x}_k = \frac{1}{\sqrt{\lambda_k}} \int_{\mathbb{T}} \varphi_k(t)\, \mathbf{y}(t)\, dt \tag{6.5.12}$$

These random variables form an orthonormal basis for the Hilbert space $\mathbf{H}(\mathbf{y})$*, linearly generated by the process* $\mathbf{y}$ *(so in particular* $\mathbb{E}\,\mathbf{x}_k^2 = 1$ *for all* k*). The expansion converges in quadratic mean, uniformly in* $t \in \mathbb{T}$*.*

The above is the celebrated *Karhunen-Loève expansion* of the process $\mathbf{y}$. In analogy to the discrete time case, the expansion (6.5.11) is normally truncated to a finite number of terms, leading to the approximate description

$$\mathbf{y}_n(t) \,=\, \sum_{k=0}^{n} \sqrt{\lambda_k}\, \varphi_k(t)\, \mathbf{x}_k \tag{6.5.13}$$

in terms of the first $n+1$ modes. The quality of the approximation can be measured in terms of statistical energy content. Since

$$\mathbb{E} \int_{\mathbb{T}} \mathbf{y}(t)^2\, dt \,=\, \int_{\mathbb{T}} E\, \mathbf{y}(t)^2\, dt \,=\, \sum_{k=0}^{+\infty} \lambda_k$$

the energy of $\mathbf{y}_n$ is just given by the sum above truncated to the first $n+1$ terms. One sees that the energy of the approximation error $\mathbf{y} - \mathbf{y}_n$ decreases with n at the same rate as the residual sum of eigenvalues of the operator Σ_R. In relative terms the energy of the error can be expresed as the ratio

$$\frac{\sum_{k=n+1}^{+\infty} \lambda_k}{\sum_{k=0}^{+\infty} \lambda_k} \,=\, 1 - \frac{\sum_{k=0}^{n} \lambda_k}{\sum_{k=0}^{+\infty} \lambda_k}.$$

[17] The factor $\sqrt{\lambda_k}$ could be absorbed in the random variable $\mathbf{x}_k$ making its norm equal to λ_k, as was done in the discrete-parameter setting.

In fact, one can show, in perfect analogy to Theorem 6.1, that the truncated expansion (6.5.13) provides the best approximant of rank $n+1$ of $\mathbf{y}$ in the sense that it minimizes the norm

$$\|\mathbf{y} - \hat{\mathbf{y}}\|^2 := \int_{\mathbb{T}} \mathbb{E}\, |\mathbf{y}(t) - \hat{\mathbf{y}}(t)|^2\, dt$$

where $\hat{\mathbf{y}}$ is a m.s. continuous process with $H(\hat{\mathbf{y}}) \subset H(\mathbf{y})$ a subspace of dimension $n+1$.

Note that in general the expansion in sinusoidal modes provided by a pathwise Fourier analysis of the signal has worse approximation properties than the expansion (6.5.13) except in a special case.

Example 6.1 (The K-L expansion of stationary processes).

When $\mathbf{y}$ is a stationary process one has $R(t,s) = R(t-s)$ and it is easy to check that the condition (6.5.2) can hold only when $\mathbb{T}$ is a *bounded interval*. If the interval $\mathbb{T}$ is unbounded, say $\mathbb{T} = [0, +\infty)$, the process has no K-L expansion.

On a finite interval, say $\mathbb{T} = [-a, a]$, the covariance function can be expanded in Fourier series

$$\Sigma(\tau) = \sum_{k=0}^{+\infty} \sigma_k \cos \frac{k\pi\tau}{a}$$

and substituting this expression in the integral equation (6.5.4) and taking into account the orthogonality of the cosine functions, one readily sees that the eigenvalues of the operator Σ_R are simply the Fourier coefficients of R; i.e. $\lambda_k = \sigma_k$, while the normalized eigenfunctions are

$$\varphi_k(t) = \frac{1}{\sqrt{a}} \cos \frac{k\pi\tau}{a}$$

Hence the random variables $\mathbf{x}_k$ are just the (random) Fourier coefficients of the signal $\mathbf{y}$. In this case the K-L expansion coincides with the Fourier representation.

6.6 ▪ Reproducing Kernel Hilbert Spaces

The Mercer expansion of positive kernel functions (6.5.5) has given rise to a whole branch of Functional Analysis, in particular to the discovery of a very important class of Hilbert spaces called **Reproducing Kernel Hilbert Spaces (RKHS)**. The consequences have been striking both in Statistics and in Signal Processing Engineering. Important applications of the RKHS theory in these fields were first pointed out by Emmanuel Parzen [66] and successively studied by various authors, in particular, among the forerunners we quote Grace Wahba's influential book [101].

Consider a $N \times N$ symmetric positive definite matrix K. Any such matrix can be considered the covariance of an N-dimensional random signal. For convenience we shall write the entries of K as $K(i,j)\,;\, i,j = 1,\dots,N$. Let us introduce in $\mathbb{R}^N$ the inner product

$$\langle f,\, g\rangle_{K^{-1}} := f^\top K^{-1} g \tag{6.6.1}$$

and note that choosing g as the j-th column vector of K, i.e.

$$g(\cdot) := K(\cdot, j)$$

in this inner product space one has

$$\langle f,\, K(\cdot,\, j)\rangle_{K^{-1}} = f^\top K^{-1} K(\cdot, j) = f^\top e_j = f(j) \tag{6.6.2}$$

which is called the **reproducing property**. In particular we have

$$\langle K(\cdot,\, i),\, K(\cdot,\, j)\rangle_{K^{-1}} = K(i, j) \tag{6.6.3}$$

This inner product space is then called a **reproducing kernel space**. This construction can be generalized starting from an arbitrary symmetric covariance function $K(t, s)$ of a continuous-time process which satisfies the compactness condition (6.5.2) on some set $\mathbb{T}\times\mathbb{T}$. Recall that an arbitrary real covariance function $K(t, s)$ should be symmetric and *positive semidefinite*. Hereafter we shall assume that it satisfies the *strict positivity* condition

$$\sum_{i,j} a_i K(t_i,\, t_j) a_j > 0 \tag{6.6.4}$$

for all times t_i, t_j and non-zero finite sequences $\{a_i\}$.

Let $K(\cdot, \cdot)$ be a symmetric positive definite continuous function on some square interval $\mathbb{T} \times \mathbb{T}$ of $\mathbb{R}^2$ satisfying the integrabiliy condition (6.5.2). Define the linear vector space of all finite linear combinations

$$\tilde{\mathbf{H}} := \{\sum_k a_k K(\cdot, s_k),\; s_k \in \mathbb{T}\,, a_k \in \mathbb{R}\}$$

with inner product of functions

$$f(\cdot) = \sum_k a_k K(\cdot, t_k)\,, \qquad g(\cdot) = \sum_j b_j K(\cdot, t_j)$$

defined as

$$\langle f,\, g\rangle_K := \sum_{k,j} a_k K(t_k,\, t_j) b_j \tag{6.6.5}$$

which extends the property (6.6.3) to continuous parameter functions. Let $\mathbf{H}$ be the closure of the inner product space $\tilde{\mathbf{H}}$ with respect to this inner product.

Theorem 6.4 (Aronszajn). *The closure* $\mathbf{H}$ *is a Hilbert space of continuous functions and the functions* $\{K(\cdot,\, s), s \in \mathbb{T}\}$ *are generators of* $\mathbf{H}$*. The extension of the inner product* (6.6.5) *to* $\mathbf{H}$ *is*

$$\langle f,\, g\rangle_{\mathbf{H}} = \langle f, \mathbf{K}^{-1} g\rangle_{L^2} = \int_{\mathbb{T}} f(t)\,(\int_{\mathbb{T}} K^{-1}(t, s) g(s) ds)\, dt \tag{6.6.6}$$

which can be expressed in terms of the orthonormal expansion (6.5.5) *as*

$$\langle f,\, g\rangle_{\mathbf{H}} = \sum_{i=1}^{\infty} \frac{\langle f, \varphi_i\rangle_{L^2}\, \langle g, \varphi_i\rangle_{L^2}}{\lambda_i}. \tag{6.6.7}$$

The Kernel K *has the* **reproducing property***; i.e. for all* $f \in \mathbf{H}$

$$f(s) = \langle K(\cdot,\, s), f(\cdot)\rangle_{\mathbf{H}}\,. \tag{6.6.8}$$

A Hilbert space of functions with a reproducing kernel is called a **Reproducing Kernel Hilbert space (RKHS)**.

Reproducing Kernel Regression

We shall now study regularized least squares regression problems in the context of RKHS's. Consider a scalar ridge regression problem

$$\min_{f} \sum_{k=1}^{N} q_k[\, y_k - f(x_k)\,]^2 + \lambda \, \|f\|_{\mathbf{H}}^2 \tag{6.6.9}$$

where the x_k and y_k are scalar observed input-output pairs, f is the unknown regression function which we shall assume belongs to some Hilbert space $\mathbf{H}$ of continuous functions of the variable x. Introducing the standard N-vector notations; $y \in \mathbb{R}^N$, $\mathbf{f} : \mathbb{R} \mapsto \mathbb{R}^N$, this problem can be restated as

$$\min_{f \in \mathbf{H}} \{ \, \| \, y - \mathbf{f}(x) \, \|_Q^2 + \lambda \, \|f\|_{\mathbf{H}}^2 \} \tag{6.6.10}$$

where Q is some positive definite $N \times N$ symmetric matrix. This formulation could in principle also accomodate the case of vector valued data, say $x_k \in \mathbb{R}^n$ and $y_k \in \mathbb{R}^m$, $\mathbf{f}$ being a m-vector function of n real variables but for the sake of simplicity we shall refer this generalization to the literature. Actually, Schölkopf, Herbrich and Smola [80] relax the squared norm cost allowing the regularizer to be any strictly monotonically increasing function $g(\cdot)$ of the Hilbert space norm.

The space $\mathbf{H}$ encodes the a priori constraints (or the a priori information) available on the nonlinear regression function f. We shall assume that it is endowed with a RKHS structure with reproducing kernel $K(x, z)$ where (x, z) both run on some feature space $\mathcal{X}$ which here we shall assume to be a real interval containing all data $\{x_k\}$.

Theorem 6.5 (Representer Theorem). *The optimal solution f^* of the problem* (6.6.10) *admits a representation of the form:*

$$f^*(\cdot) = \sum_{i=1}^{N} \theta_i \, K(\cdot, x_i) \tag{6.6.11}$$

Letting K_N be the $N \times N$ positive definite matrix with entries $K(x_i, x_j)$; $i, j = 1, \ldots, N$, the N-dimensional real parameters $\{\theta_i\}$ can be found by solving the normal equations of a regularized least squares problem with quadratic norms induced by the matrices Q and K_N.

Proof. Denote by $\mathbb{R}^{\mathcal{X}}$ the vector space of real valued functions $\mathcal{X} \to \mathbb{R}$ and, for $x \in \mathcal{X}$, introduce the function $\Phi(x) = K(\cdot, x)$ which is considered as a mapping

$$\Phi(x) \, : \, \mathcal{X} \to \mathbb{R}^{\mathcal{X}}$$

that is, $\Phi(x)$ is just as a "section" of the function K at the value x of its second argument. Since K is a reproducing kernel, (6.6.8) with $f(\cdot) = \Phi(x)$ yields

$$\Phi(x)(x') = K(x', x) = \langle \Phi(x'), \Phi(x) \rangle,$$

where $\langle \cdot, \cdot \rangle$ is the inner product in $\mathbf{H}$.

Given $x_1, ..., x_N$, use orthogonal projection to decompose any $f \in \mathbf{H}$ into a sum of two functions, one lying in the linear subspace $\text{span}\{\Phi(x_1), ..., \Phi(x_N)\}$, and the other lying in the orthogonal complement, to get:

$$f = \sum_{i=1}^{N} \theta_i \Phi(x_i) + v,$$

where $\langle v, \Phi(x_i) \rangle = 0$ for all i.

Using the reproducing property $f(x) = \langle f, \Phi(x) \rangle$ on the above orthogonal decomposition shows that applying f to any training point x_j, one gets

$$f(x_j) = \left\langle \sum_{i=1}^{N} \theta_i \Phi(x_i) + v, \Phi(x_j) \right\rangle$$

which is independent of v. Consequently, the value of the quadratic cost in (6.6.10) is likewise independent of v. For the second term (assuming a generalized regularization term $g(\|f\|)$), since v is orthogonal to $\sum_{i=1}^{N} \theta_i \Phi(x_i)$ and g is strictly monotonic, we have

$$g\left(\|f\|\right) = g\left(\|\sum_{i=1}^{N} \theta_i \Phi(x_i) + v\|\right) = g\left(\sqrt{\|\sum_{i=1}^{N} \theta_i \Phi(x_i)\|^2 + \|v\|^2}\right)$$
$$\geq g\left(\|\sum_{i=1}^{N} \theta_i \Phi(x_i)\|\right).$$

Therefore setting $v = 0$ does not affect the first term of (6.6.10), while it is strictly decreasing the second term. Consequently, any minimizer f^* in (6.6.10) must have $v = 0$, i.e., it must be of the form

$$f^*(\cdot) = \sum_{i=1}^{N} \theta_i \Phi(x_i) = \sum_{i=1}^{N} \theta_i K(\cdot, x_i),$$

which is the desired result. □

This theorem states that an apparently infinite-dimensional variational problem like (6.6.10) has a finite-dimensional *parametric solution* and can therefore be solved by a finite dimensional algorithm. Note in particular that, as noticed by [101], the Smoothing Splines problem (3.5.7) can be recast in the present frame since the L^2 norm of the second derivative can be interpreted as a RKHS norm in a suitable function space[18].

Exactly as it happens for Smoothing Splines, the problem (6.6.10) can be restated in parametric form as a generalized ridge regression and solved by parametric techniques like those described in Corollary 3.1 of Chap.3. Note that, for any function of the form (6.6.11) one has

$$\|f\|_{\mathbf{H}}^2 = \sum_{i,j=1}^{N} \theta_i K(x_i, x_j) \theta_j = \theta^\top K_N \theta \tag{6.6.12}$$

[18] In fact a Sobolev space [1].

where K_N is the $N \times N$ positive definite matrix with entries $K(x_i, x_j)$.

Problem 6.6. *Use* (6.6.11) *and the definition of the inner product* (6.6.6) *to prove equation* (6.6.12) *and state in detail the parametric ridge regression problem for the calculation of the optimal f^* solution of the problem* (6.6.9).

Chapter 7

SOME NON LINEAR INFERENCE PROBLEMS

7.1 ▪ Introduction

While linearity is a well-defined concept, non-linearity is vague. Somebody has compared the attribute non-linear to *non-elephant* in animal classification. It is quite obvious that one cannot expect a general theory of non linear inference. Yet, there are particular areas and problems where the linear theory is clearly inadequate and sensible results can only be obtained by ad hoc techniques and a special problem formulation.

In most nonlinear problems it is usually hard to devise and set up a simple probabilistic setting which allows to do statistical inference in a consistent probabilistic framework, so one is usually content with treating the inference problem just as a *deterministic model fitting* from observed data. This philosophy lies for example at the grounds of the much praised and popular *Neural Network* methodology. In general, by accepting this kind of naive paradigm one must consciously realize that is going to give up any statistical error or performance analysis. There are however some notable exceptions like the one we are going to describe in the next section.

7.2 ▪ Direction estimation on the unit sphere

In this section we shall discuss a class of Bayesian estimation problems which cannot be treated effectively by the linear m.v. theory. We shall discuss problems where the variable to be estimated is a *direction*. These problems have some important applications in practice, occurring when using bearing-only sensors, for example in antenna array systems or single-camera optical sensors, in particular in problems of scene and motion reconstruction in computer vision. In computer vision systems one has in practice only access to the projections of 3-D points on the image plane of a camera. Digital images are formed on an array of CCD sensors, ideally superimposed to the focal plane of the system. The detected feature points on the image plane do not correspond exactly to straight perspective projections of the real target points in $\mathbb{R}^3$. This occurs because of distortion of the optical systems and noise of various kinds entering the signal detection and signal processing phase on the electronic image acquired on the

CCD array.

In the applications we have in mind the observations are angular measurements which are corrupted by noise. From them one wants to reconstruct the scene or track some moving target object in 3-dimensional space. To simplify the problem setting we shall just assume the target can be described by a single point P, which in general may move randomly in $\mathbb{R}^3$ (although in this section we shall address only the reconstruction of a fixed target). The sensor at our disposal cannot measure the distance of P from the observation center O (conventionally coinciding with the origin of the coordinate system) but can only measure the direction of the vector OP joining the target to the optical center.

Mathematically, the *direction* of a vector OP can be identified with the normalized *unit vector* $OP/\|OP\|$. In our setting, the only reconstructible feature of a target point P will be its direction, a unit vector denoted by the symbol $\mathbf{x}$. The observations $\mathbf{y}_k\,;|, k = `1, 2, \ldots$ are noisy directional data which are also modeled as unit vectors belonging to the unit sphere with center in O. Hence, in this setting we want to estimate the direction $\mathbf{x}$ of a target point P in $\mathbb{R}^3$ from noisy measurement $\mathbf{y}_k,\ k = 1, \ldots, m$, which are unit vectors randomly distributed about $\mathbf{x}$. One could then say that the problem can be formulated as an *estimation problem on the unit sphere*. In general, the unit sphere in $\mathbb{R}^n$ is defined as the set of vectors of $\mathbb{R}^n$ having unit Euclidean norm. It is a surface of dimension $n-1$ and is denoted by the symbol $\mathbb{S}^{n-1}$. Hence our problem can be posed just as estimation on the 2-sphere $\mathbb{S}^2$.

The precise nature of the observation noise affecting the measurements $\mathbf{y}_k,\ k = 1, \ldots, m$ will be discussed later; it should however be clear that the way the noise affects the ideal direction $\mathbf{x}$ can no longer be additive as in the standard linear model and a realistic formulation of the problem should depart sharply from the standard linear-Gaussian setup.

More general examples of estimation problems on manifolds which occur in computer vision are discussed in the work of Soatto et al. [89].

Some other perspective estimation problems, for example recovering lines moving in $\mathbb{R}^3$ by observing their projections on the image plane, give rise to estimation on high dimensional manifolds such as the *Grassmannian manifold*.

7.3 ▪ The Langevin Distribution

Here we shall discuss probability distributions on the unit sphere.

A family of probability distributions on the sphere which has many desirable properties is defined by the *Langevin density*

$$p(x) = \frac{\kappa}{4\pi \sinh \kappa} \exp \kappa \mu^\top x, \qquad x \in \mathbb{R}^3\,; \quad \|x\| = 1 \tag{7.3.1}$$

with respect to the spherical surface measure $d\sigma = \sin\theta d\theta d\phi$ on the unit sphere. Here $\theta\,, \phi$ are the polar angle and azimuth coordinates. The vector parameter $\mu \in \mathbb{S}^2$, conventionally normalized to unit length, is the *mode* of the distribution, while the nonnegative number $\kappa > 0$ is called the *concentration* of the distribution. For $\kappa \to 0$ the density becomes the uniform distribution while for $\kappa \to \infty$, $p(x)$ tends to a Dirac distribution concentrated at $x = \mu$. The density function (7.3.1), denoted $L(\mu, \kappa)$, was introduced by Paul Langevin in his statistical-mechanical model of magnetism [53]. Since then it has been rediscovered and used in statistics by a number of people, including von Mises whose name is

sometimes attached to the distribution, see [102]. Observe that the Langevin distributions form a one-parameter exponential family and that this family is *invariant with respect to rotations*, in the sense that, if $\mathbf{x}$ is Langevin distributed, then it is easy to check that for any $R \in SO(3)$ [19] the random vector $\mathbf{y} := R\mathbf{x}$ has still a Langevin distribution with the same concentration parameter as $\mathbf{x}$ and mode parameter $R\mu$. An important property of the Langevin distribution is the preservation of the functional form under multiplication

$$L(\mu_1, \kappa_1)\, L(\mu_2, \kappa_2) = L(\mu, \kappa) \tag{7.3.2}$$

where $L(\mu, \kappa)$ is Langevin with parameters

$$\mu = \frac{\kappa_1\, \mu_1 + \kappa_2\, \mu_2,}{\|\kappa_1\, \mu_1 + \kappa_2\, \mu_2\|} \quad \kappa = \|\kappa_1\, \mu_1 + \kappa_2\, \mu_2\|. \tag{7.3.3}$$

Introducing a coordinate system in $\mathbb{R}^3$ with unit vectors e_1, e_2, $e_3 = \mu$ and spherical polar coordinates (θ, ϕ), relative to this frame, we have

$$x = \sin\theta\cos\phi e_1 + \sin\theta\sin\phi e_2 + \cos\theta e_3$$

and the inner product $\mu^\top x$ is just equal to $\cos\theta$ whereby (7.3.1) takes the simple form

$$p(\theta, \phi) = \frac{\kappa}{4\pi \sinh\kappa} \exp(\kappa\cos\theta) \quad 0 \le \theta \le \pi.$$

which makes it clear that $L(\mu, \kappa)$ is rotationally symmetric around its mode μ.

The expression given in (7.3.1) is for a distribution on the unit sphere in $\mathbb{R}^3$. For higher dimension, only the normalization constant has a slightly more complicated expression. The Langevin distribution on $\mathbb{S}^{n-1}$, $n \ge 3$, is

$$p(x) = \frac{\kappa^{(n/2-1)}}{(2\pi)^{n/2} I_{n/2-1}(\kappa)} \exp \kappa\mu^\top x, \qquad \|x\| = 1 \tag{7.3.4}$$

where $I_{n/2-1}(x)$ is a modified Bessel function of the first kind. More generally, an arbitrary probability density functions on $\mathbb{S}^{n-1}$ can be expressed as the exponential of a finite expansion in spherical harmonics. These are discussed, for example, in [102, p. 80-88]. In this sense the Langevin density is a sort of first order approximation as only the first spherical harmonic, $\cos\theta$, is retained in the expansion and the others are assumed to be negligible. A more general approach than the one followed here could be to consider densities which are exponential of a finite sum of spherical harmonics. These are also of exponential type, have a set of finite dimensional sufficient statistics and could be treated by generalizing what is done in this section.

Rotation-invariant distributions like the Langevin distribution are natural for describing **random rotations**.

Let $\mathbf{x}$ be a fixed direction, represented as a point in $\mathbb{S}^2$, which is for example observed by a camera. The observation mechanism perturbs $\mathbf{x}$ in a random way, say because of lens distortion, pixel granularity etc. Since the output of the sensor, $\mathbf{y}$, is also a direction represented by a vector of unit length, the perturbation may always be seen as a random rotation described by a random rotation

[19] $SO(3)$ is the *special orthogonal group* of matrices R such that $RR^\top = I$ with $\det R = +1$.

matrix [20] $R = R(\mathbf{p}) \in SO(3)$, where $\mathbf{p}$ is the polar vector of the rotation, i.e. $R(\mathbf{p}) := \exp\{\mathbf{p}\wedge\}$ so that

$$\mathbf{y} := R(\mathbf{p})\,\mathbf{x} \tag{7.3.5}$$

In other words we can always model the noise affecting $\mathbf{x}$ as multiplication by a rotation matrix. The action of the "rotational observation noise" on directions $\mathbf{x} \in \mathbb{S}^2$ can in turn be described probabilistically by the *conditional density* function $p(y \,|\, \mathbf{x} = x)$ of finding the observation directed about a point y on the sphere, given that the "true" observed direction was $\mathbf{x} = x$. A very reasonable unimodal conditional distribution, rotationally symmetric about the starting direction $\mathbf{x}$ (no angular bias introduced by the observing device) is the Langevin-type density,

$$p(y \,|\, \mathbf{x}) = \frac{\kappa}{4\pi \sinh \kappa} \exp \kappa \mathbf{x}^\top y \tag{7.3.6}$$

In this framework we may therefore think of the ordinary distribution $L(\mu, \kappa)$ as a *conditional density* evaluated at the known conditioning direction $\mathbf{x} = \mu$.

Note that, since $\mu^\top y$ is just the cosine of the angle between the unit vectors μ and y on the sphere, the values of the conditional probability distribution $p(y \,|\, \mathbf{x})$ are invariant with respect to the action of the rotation group $SO(3)$ on $\mathbb{S}^2$, i.e. with respect to coordinate change on the sphere.

The Angular Gaussian Distribution

As we have seen, the functional form of the Langevin distribution is preserved under rotations which is a property resembling that of the Gaussian, which is preserved under linear maps in Euclidean spaces. In a sense the Langevin distribution is the natural analog on the unit sphere of the Gaussian distributions on a Euclidean space. There are various attempts in the literature to derive the Langevin distribution as the distribution function of some natural transformation of a Gaussian vector. Perhaps the easiest result in this vein is the observation, first made by R. A. Fisher [34], that the distribution of a normal random vector $\mathbf{x}$ having an isotropic Gaussian distribution $\mathcal{N}(\mu\,,\,\sigma^2 I)$, *conditional on the event* $\{\|\mathbf{x}\| = 1\}$ is Langevin with mode $\mu/\|\mu\|$ and concentration parameter $\|\mu\|/\sigma^2$.

A more useful result, discussed in [102, Appendix C] is the remarkable similarity of the so-called *Angular Gaussian* distribution to the Langevin. The angular Gaussian, denoted by the symbol Ag, is the probability density of the unit vector $\mathbf{x} := \boldsymbol{\xi}/\|\boldsymbol{\xi}\|$ when the random vector $\boldsymbol{\xi}$ has an isotropic Gaussian distribution, i.e. $\boldsymbol{\xi} \sim \mathcal{N}(\mu\,,\,\sigma^2 I)$. The distribution is obtained by computing the marginal of $\mathcal{N}(m\,,\,\sigma^2 I)$ on the unit sphere $\|x\| = 1$. It is shown in [102, Appendix C] that the angular Gaussian is a convex combination of Langevin densities with a varying concentration parameter κ,

$$Ag(x) = N \int_0^{+\infty} \kappa^{n-1} e^{-\frac{1}{2}\frac{\kappa^2}{\theta^2}} e^{\kappa \mu^\top x}\, d\kappa,$$

where

$$\mu = \frac{m}{\|m\|} \quad \alpha = \frac{\|m\|}{\sigma} \tag{7.3.7}$$

[20] The wedge $\wedge$ denotes cross product.

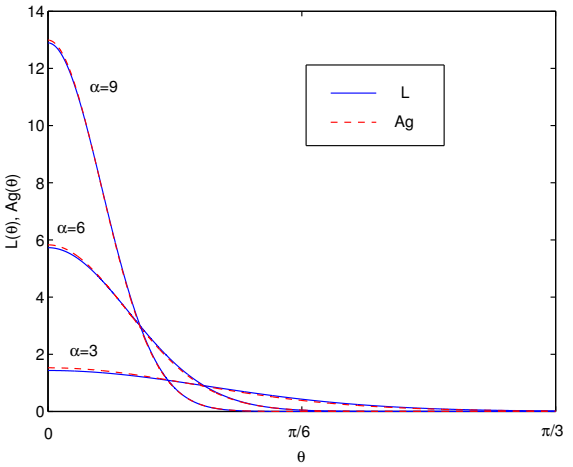


Figure 7.3.1. *Angular Gaussian vs Langevin*

and it is seen from this formula that Ag depends on μ , σ^2 only through the two parameters μ and α. We shall denote it by $Ag(\lambda, \alpha^2)$. The notation is convenient, since for either moderate or large values[21] of the parameter α, $Ag(\mu, \alpha^2)$ is, to all practical purposes, the *same thing as* $L(\mu, \alpha^2)$, where the parameters are related by (7.3.7). See Figure 7.3.1.

In fact the angular Gaussian approximates a Langevin distribution also for α small, when both of them are close to uniform (which is however a rather uninteresting case), but the relation between α and κ is different. In the following we shall just assume that $Ag(\mu, \alpha^2) = L(\mu, \alpha^2)$.
Note that all distributions $\mathcal{N}(\rho\mu\,,\, \rho^2\sigma^2 I)$, $\rho > 0$, give origin to the same angular Gaussian as $\mathcal{N}(\mu\,,\, \sigma^2 I)$. This is in fact precisely the family of isotropic Gaussians generating the same angular distribution.

The role of the angular Gaussian in modeling directional observations can be illustrated by the following example. Let $\boldsymbol{\xi}$, $\boldsymbol{\zeta}$ be independent Gaussian isotropic random vectors with $\boldsymbol{\xi} \sim \mathcal{N}(\mu\,,\, \sigma^2 I)$, $\boldsymbol{\zeta} \sim \mathcal{N}(0\,,\, \sigma_z^2 I)$ and assume we observe the direction of the vector

$$\boldsymbol{\eta} = C\boldsymbol{\xi} + \boldsymbol{\zeta} \ \sim \mathcal{N}(\mu\,,\, \sigma^2 CC^\top + \sigma_z^2 I). \tag{7.3.8}$$

If C is an orthogonal matrix, $CC^\top = I$ and the distribution of $\boldsymbol{\eta}$ is isotropic Gaussian, the direction $\mathbf{y} := \boldsymbol{\eta}/\|\boldsymbol{\eta}\|$, has an angular Gaussian distribution, namely $\mathbf{y} \approx L(\mu/\|\mu\|,\, \frac{\|\mu\|^2}{\sigma^2+\sigma_z^2})$.

Actually, no matter how $\boldsymbol{\xi}$, $\boldsymbol{\zeta}$ are correlated, it is easy to see that the conditional density $p(y \mid \boldsymbol{\xi} = \xi)$ is angular Gaussian. In fact, this follows since the conditional distribution of $\boldsymbol{\eta}$ given $\boldsymbol{\xi} = \xi$ is Gaussian with mean $C\xi$ and variance σ_z^2. Hence

$$p(y \mid \boldsymbol{\xi} = \xi) = Ag(C\xi/\|\xi\|,\, \|C\xi\|^2/\sigma_z^2) = Ag(Cx,\, \|\xi\|^2/\sigma_z^2) \tag{7.3.9}$$

[21] "Moderate or large" here means that $\kappa := \alpha^2$ should be greater than, say, 100 in order to have a fit within a few precent of the values of the two functions.

where $\mathbf{x}$ is the direction vector of $\boldsymbol{\xi}$.

In practice we are interested in the conditional density $p(y \mid \mathbf{x})$. We shall state the condition for this density to be angular Gaussian as follows.

Proposition 7.1. *If the additive perturbation $\boldsymbol{\zeta}$ is isotropic Gaussian with variance σ_z^2 proportional to $\|\xi\|^2$, i.e. $\sigma_z^2 = \sigma_0^2\|\xi\|^2$, then the conditional density $p(y \mid \mathbf{x})$ of the unit vector $\mathbf{y}$ of the observation* (7.3.8) *is angular Gaussian.*

Proof. Denote $r := \|\xi\|$ and note that $p(y \mid \boldsymbol{\xi} = \xi) = p(y \mid x, r)$. Since $p(y, r \mid x) = p(y \mid x, r)p(r \mid x)$, by a well-known formula in probability theory we have

$$p(y \mid x) = \int_0^\infty r^2 p(y \mid x, r)p(r \mid x)dr\,.$$

Then the claim follows from

$$p(y \mid x) = \int_0^\infty r^2 p(y \mid x, r)p(r \mid x)dr$$

since $p(y, r \mid x) = p(y \mid x, r)p(r \mid x) = p(y \mid \xi)p(r \mid x)$ and by the stated assumption, $p(y \mid x, r)$ does not depend on r and can be brought out of the integral sign. □

The fact that the variance of the additive noise in the model (7.3.8) must depend on ξ, which implies that this variance is in fact the *conditional* variance of $\boldsymbol{\eta}$ given $\boldsymbol{\xi} = \xi$, can be described as *condition of angular noise*. Note that this condition precludes the independence of $\boldsymbol{\xi}$ and $\boldsymbol{\zeta}$. This fact, which may look surprising at a a first glance, is in fact quite in line with the intuitive idea of angular noise. Every infinitesimal perturbation $d\boldsymbol{\xi}$ of a random vector $\boldsymbol{\xi}$ having a fixed length and direction $\mathbf{x} \in \mathbb{S}^2$, which should maintain the vector $\boldsymbol{\xi} + d\boldsymbol{\xi}$ of the same length (up to higher order infinitesimals), can be represented as the effect of a rotation about a certain infinitesimal polar vector $d\mathbf{p}$, that is

$$d\boldsymbol{\xi} = d\mathbf{p} \wedge \boldsymbol{\xi} = (d\mathbf{p} \wedge \mathbf{x})\,\|\boldsymbol{\xi}\| \tag{7.3.10}$$

where the symbol $\wedge$ indicates vector (or external) product [22]. In this way the differential $d\boldsymbol{\xi}$ lays on the tangent plane to to the sphere of ray $r = \|\boldsymbol{\xi}\|$, at the point $\boldsymbol{\xi}$ and has amplitude which is proportional to the norm $\|\boldsymbol{\xi}\|$ so that its variance is proportional to the square $\|\boldsymbol{\xi}\|^2$. Hence the perturbation $d\boldsymbol{\xi}$ must be related to both the direction and the norm of the vector $\boldsymbol{\xi}$ by the geometry of the space so it cannot be independent of the variable $\boldsymbol{\xi}$.

One may say that Proposition 7.1 asserts that the angular Gaussian distribution (and hence the Langevin distribution) describes statistically the effect of small angular perturbations on a fixed direction. Said in other words, for small angular perturbations, the conditional density $p(y \,|\, \mathbf{x} = x)$ of observing the direction y on the unit sphere, when the ideal observed direction is $\mathbf{x} = x$, is the

[22] If $p := [p_1 p_2 p_3]^\top$ then $p\wedge$ is the linear operator in $\mathbb{R}^3$ defined by the skew-symmetric matrix

$$p\wedge = \begin{bmatrix} 0 & -p_3 & p_2 \\ p_3 & 0 & -p_1 \\ -p_2 & p_1 & 0 \end{bmatrix}$$

.

Langevin distribution

$$p(y \,|\, \mathbf{x}) = \frac{\kappa}{4\pi \sinh \kappa} \exp \kappa \mathbf{x}^\top y \tag{7.3.11}$$

Therefore one can always interpret the distribution $L(\mu, \kappa)$ as being a *conditional distribution* based on the condition $\mathbf{x} = \mu$. That this distribution is unimodal and symmetric about the conditioning direction $\mathbf{x} = \mu$ may be interpreted as a description of a measurement process *without bias*.

When the angular perturbation is not infinitesimal and there may be large angular errors in the observations one should integrate the relation (7.3.10). One would find that a finite angular perturbation of a given direction $\mathbf{x} \in \mathbb{S}^2$ can be expressed by the formula

$$\mathbf{y} := R(\mathbf{p})\, \mathbf{x} \tag{7.3.12}$$

where $R(\mathbf{p}) := \exp\{\mathbf{p}\wedge\}$ is a rotation in $SO(3)$ (an orthogonal matrix of determinant $+1$) about the polar vector $\mathbf{p}$. It follows that the output of an unbiased noisy directional sensor can in general be represented mathematically as multiplication by a rotation matrix about a random polar vector: $R = R(\mathbf{p}) \in SO(3)$. In this general case the conditional distribution can of course be a quite arbitrary density function on the unit sphere.

A max Entropy characterization

The parameters (μ, κ) of a Langevin distribution can both be expressed as a function of the mean value of the distribution. On $\mathbb{S}^2$ it is not hard to check that one has

$$\mu = \frac{m}{\|m\|}\,, \qquad \frac{\cosh \kappa}{\sinh \kappa} - \frac{1}{\kappa} = \|m\| \tag{7.3.13}$$

and these formulas define a one-to-one correspondence between m and the pair (μ, κ). Hence in analogy to what happens with the Gaussian there is a vector parameter m which determines the distribution $L(\mu,\, \kappa)$ completely. The following proposition provides a characterization of the Langevin diostribution which is the analog on $\mathbb{S}^{n-1}$ of a well-know characterization of the Gaussian density as the one which, among all probability densities on $\mathbb{R}^n$ with fixed mean and variance, has maximum entropy.

Proposition 7.2. *Among all probability distributions on the unit sphere having a fixed mean vector, m, the Langevin distribution has maximum entropy.*

Proof. The entropy of a density f (or of an absolutely continuous distribution $dF(x) := f(x)\, d\sigma_x$) on the sphere is

$$H_f := -\int_{\mathbb{S}^{n-1}} \log f(x)\, f(x)\, d\sigma_x$$

Let's denote for brevity the Langevin distribution with mean m by the symbol $l(x)$. Using the expression (**??**), one gets

$$\begin{aligned} H_l &= -\textstyle\int_{\mathbb{S}^{n-1}} \log l(x)\, l(x)\, d\sigma_x \\ &= -\log \frac{\kappa^{(n/2-1)}}{(2\pi)^{n/2} I_{n/2-1}(\kappa)} - \kappa \mu^\top m \\ &= -\textstyle\int_{\mathbb{S}^{n-1}} \log l(x)\, f(x)\, d\sigma_x \end{aligned}$$

for an arbitrary distribution f of mean m. It follows that the difference

$$\begin{array}{l} H_l - H_f \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad = \\ = -\int_{\mathbb{S}^{n-1}} \log l(x)\, f(x)\, d\sigma_x + \int_{\mathbb{S}^{n-1}} \log f(x)\, f(x)\, d\sigma_x \\ = \int_{\mathbb{S}^{n-1}} \log \frac{f(x)}{l(x)}\, f(x)\, d\sigma_x \end{array}$$

is just the famous *Kullback-Leibler pseudo-distance* of f from l, see [52, 51]. This is also called *divergence* or *relative entropy* in Communication Theory, see [21] or the article by Van Schuppen and Stoorvogel in [10, p. 314]. In the first reference it is proved that this pseudo distance is positive unless $f = l$, in which case it is equal to zero. Hence $H_l \geq H_f$ for every f having mean m. □

Best Approximation by a Langevin distribution

The Kullback-Leibler (pseudo-) distance introduced in the previous section is a natural measure of distance among probability distributions. We shall use this distance to provide another interesting characterization of the Langevin distribution.

Let P be an arbitrary probability measure on the unit sphere, absolutely continuous with respect to the surface measure $d\sigma = \sin\theta\, d\theta\, d\varphi$; we want to approximate the density $f(x) = dP/d\sigma$ by means of a density of the Langevin type, i.e. by a density in the parametric class

$$\mathcal{L} = \{\ell(x) = \frac{\kappa}{4\pi \mathrm{sinh}(\kappa)}\, \exp\{\kappa \mu^\top x\} \quad , \quad \kappa \geq 0, \|\mu\| = 1\}; \tag{7.3.14}$$

, using as a criterion of fit the Kullback-Leibler's pseudo-distance,

$$K(f, \ell_{(\mu,\kappa)}) = \mathbb{E}_f \ln \frac{f(x)}{\ell_{(\mu,\kappa)}(x)} = \int_{\mathbb{S}^2} f(x) \ln \frac{f(x)}{\ell_{(\mu,\kappa)}(x)}\, d\sigma_x \tag{7.3.15}$$

The problem is to find the minimum:

$$\min_{(\mu,\kappa)\ :\ \kappa \geq 0, \|\mu\|=1} K(f, \ell(\mu, \kappa)) \tag{7.3.16}$$

This optimization problem can be solved by introducing Lagrange multipliers

$$\Lambda_f(\theta) = K(f, \ell_\theta) + \frac{\lambda}{2} \Phi(\theta) \tag{7.3.17}$$

where

$$\Phi(\theta) = \sum_{i=1}^{3} \mu_i^2 \,.$$

Taking derivatives with respect to μ and κ it can be shown that the minimum is attained for:

$$\begin{cases} \dfrac{\cosh \kappa}{\sinh \kappa} - \dfrac{1}{\kappa} - \mu^\top m_x = 0 \\ \kappa m_x - \lambda \mu = 0 \end{cases} \tag{7.3.18}$$

where m_x is the mean vector of P

$$m_x = \int_{\mathbb{S}^2} x f(x)\, d\sigma_x. \tag{7.3.19}$$

Explicitly, the optimal μ and κ are given by:

$$\begin{cases} \dfrac{\cosh\kappa}{\sinh\kappa} - \dfrac{1}{\kappa} = \|m_x\| \\ \mu = \dfrac{m_x}{\|m_x\|} \end{cases} \tag{7.3.20}$$

Note that for a Langevin density, the parameters (μ, κ) are completely determined by the mean vector m so that

$$\begin{cases} \dfrac{\cosh\kappa}{\sinh\kappa} - \dfrac{1}{\kappa} & = \|m\| \\ \mu & = \dfrac{m}{\|m\|} \end{cases} \tag{7.3.21}$$

Therefore:

Proposition 7.3. *The Langevin distribution which best approximates an arbitrary distribution P on the unit sphere according to the Kullback-Leibler distance, is the one having the same mean m of P.*

Hence our approximation problem is solved simply by equating the mean vectors of the two distributions. In other words the only thing we need to know to find the best Langevin approximant of P is its mean vector.

This result leads to a kind of wide-sense estimation theory on the unit sphere with the mean parameter playing the same role of the second order statistics in the Gaussian case. Recall that here both the mode (i.e. the "most probable direction") and the concentration parameter (telling us how data are scattered about the mode) are uniquely deductible from the mean. Obviously one expects reasonable results from this wide-sense theory only when the density to be approximated is unimodal and approximately symmetric about the mode.

7.4 ▪ MAP Estimation of directions

Assume we are measuring a direction $\mathbf{x}$ by a noisy sensor which is affected by angular noise and we know the conditional density, $p(y \mid x)$, which belongs to the Langevin class. Assume also that the a priori model for the unknown direction vector $\mathbf{x}$ is of the Langevin type say,

$$\mathbf{x} \sim L(x_0, \kappa_0)$$

For physical reasons it is reasonable to assume that $\mathbf{x}$ and the random rotation $d\mathbf{y}p\wedge$ which corrupts the observation of $\mathbf{x}$, are independent random variables. We can then compute the a posteriori density $p(x \mid y)$ using Bayes rule. The joint density is given by the expression

$$p(x\,,\, y) = p(y \mid x)p(x) = A(\kappa, \kappa_0) \exp \hat{\kappa}\, \hat{\mu}^\top x$$

where

$$\begin{aligned} A(\kappa, \kappa_0) \;&=\; \frac{\kappa}{4\pi \sinh\kappa} \frac{\kappa_0}{4\pi \sinh\kappa_0} \\ \hat{\kappa}\, \hat{\mu}^\top x &:= \kappa y^\top x + \kappa_0 x_0^\top x. \end{aligned}$$

In this formula, $\hat{\kappa} = \hat{\kappa}(y, x_0) > 0$ and $\hat{\mu} = \hat{\mu}(y, x_0)$ are functions of y and of the prior mode x_0 which are computed by the formulas

$$\hat{\mu} := \frac{\kappa y + \kappa_0 x_0}{\hat{\kappa}} \qquad \hat{\kappa} := \|\kappa y + \kappa_0 x_0\|. \tag{7.4.1}$$

Note that $\|\hat{\mu}\| = 1$.

Dividing by the marginal one obtains the a posteriori density:

$$p(x \,|\, \mathbf{y}) = \frac{\hat{\kappa}}{4\pi \sinh \hat{\kappa}} \exp \hat{\kappa}(\mathbf{y})\, \hat{\mu}^\top(\mathbf{y}) x$$

which is still of Langevin class. Te conditional mode $\hat{\mu}(\mathbf{y})$, that is the *Maximum a Posteriori Bayesian estimate* of $\mathbf{x}$, given the observation $\mathbf{y}$ and the conditional concentration $\hat{\kappa}(\mathbf{y})$ are given in (7.4.1). They are quite easy to figure out in this simple case. These formulas can, in certain cases, be generalized to a dynamic situation and lead to a nonlinear versions of the Kalman filter. We shall consider the simplest situation below.

Sequential Observations of a Fixed Target

Assume we have a sequence of observations

$$\mathbf{y}(t) := R(\mathbf{p}(t))\, \mathbf{x} = \exp\{\mathbf{p}(t)\wedge\}\, \mathbf{x} \quad t = 1, 2, \dots \tag{7.4.2}$$

where the $\mathbf{p}$'s are identically distributed independent random rotations which are also independent of the random vector $\mathbf{x}$. The $\mathbf{y}(t)$'s are conditionally independent given $\mathbf{x}$, and $p(y(t) \,|\, \mathbf{x}) = L(\mathbf{x}, \kappa)$, where κ is the concentration parameter of the angular noise. Hence, denoting

$$\mathbf{y}^t := [\mathbf{y}(1), \dots, \mathbf{y}(t)]^\top$$

we may write

$$p(y^t \,|\, \mathbf{x}) = \frac{\kappa^t}{(4\pi \sinh \kappa)^t} \exp \kappa \langle \mathbf{x}, \sum_{s=1}^{t} \mathbf{y}(s) \rangle \tag{7.4.3}$$

where $\langle\,.\,,\,.\,\rangle$ denotes inner product in $\mathbb{R}^3$. Assuming an a priori density of the same class, $\mathbf{x} \sim L(x_0, \kappa_0)$, one readily obtains the a posteriori measure

$$p(x \,|\, \mathbf{y}^t) = \frac{\hat{\kappa}(t)}{(4\pi \sinh \hat{\kappa}(t))} \exp \hat{\kappa}(t) \langle \hat{\mu}(t)\,,\, x \rangle \tag{7.4.4}$$

which is still of the Langevin class with parameters

$$\hat{\mu}(t) = \frac{1}{\hat{\kappa}(t)} (\kappa \sum_{s=1}^{t} y(s) + \kappa_0 x_0) \tag{7.4.5}$$

$$\hat{\kappa}(t) = \|\kappa \sum_{s=1}^{t} y(s) + \kappa_0 x_0\| \tag{7.4.6}$$

Note that in case of a uniform prior distribution ($\kappa_0 = 0$), the first formula reduces to

$$\hat{\mu}(t) = \frac{\sum_{s=1}^{t} y(s)}{\| \sum_{s=1}^{t} y(s) \|} \tag{7.4.7}$$

which should be compared with the Gaussian m.v. estimator which is just the arithmetic mean of the observations and of course does not preserve the unit norm of the summands.

These formulas can be easily rewritten as recursive relations which update the current estimate $\hat{\mu}(t)$, $\hat{\kappa}(t)$ for adjunction of the $t+1$-st measurement. At time $t+1$ one obtains,

$$\begin{aligned}\hat{\mu}(t{+}1) &= \frac{1}{\hat{\kappa}(t{+}1)}(\kappa \sum_{s=1}^{t} y(s) + \kappa_0 x_0 + \kappa y(t{+}1)) \\ \hat{\kappa}(t{+}1) &= \|\kappa \sum_{s=1}^{t} y(s) + \kappa_0 x_0 + \kappa y(t{+}1)\| \\ &= \|\hat{\kappa}(t)\hat{\mu}(t) + \kappa y(t{+}1)\|\end{aligned}$$

which can rewritten in a recursive form as in the proposition below.

Proposition 7.4. *The MAP estimate (conditional mode) $\hat{\mu}(t)$, of the fixed random direction $\mathbf{x}$, given observations corrupted by independent angular noise $\{\mathbf{p}(t)\}$ of concentration κ, propagates in time according to the recursions*

$$\hat{\mu}(t{+}1) = \frac{1}{\hat{\kappa}(t{+}1)}(\hat{\kappa}(t)\hat{\mu}(t) + \kappa y(t{+}1)) \tag{7.4.8}$$

$$\hat{\kappa}(t{+}1) = \|\hat{\kappa}(t)\hat{\mu}(t) + \kappa y(t{+}1)\| \tag{7.4.9}$$

with initial conditions $\hat{\mu}(0) = x_0$ and $\hat{\kappa}(0) = \kappa_0$.

These recursions look like a nonlinear version of the well-known Kalman-Filter updates for the sample mean which one would obtain in the Gaussian case. They can be generalized to the case of tracking directions which vary randomly. For a more general view of directional estimation see [69, 15, 6].

7.5 ▪ Introduction to Neural Networks

In these last few decades there has been a great amount of publications and an enormous promotion devoted to a class of nonlinear parametric models known as **Neural Networks**. We shall limit ourselves to a few citations, [78, 24, 71, 86, 91] but most of the books on Machine Learning nowdays devote thick chapters to the subject. The appeal and popularity of these models is probably due to the reference often made to neuro-biological models and to related new fashionable fields such a *Neuroscience, Brain science* etc. We would like to warn the reader that, contrary to what many writings of quite dubious scientific credentials tend to promote, this relation is mostly based on wording and is somewhat misleading.

Among the often claimed neuro-biological motivations, there is a primitive model of a neuron introduced in 1943 by McCulloch and Pitts [64] which has triggered the idea of *activaton* and *activation function*, a basic component of the Neural Networks philosophy. It should be said however that this model seems now really inadequate from a physiological point of view to explain the behaviour of a real neuron. Nevertheless this has led to the introduction in this field of a mystic/biological terminology which seems to have little to do with the actual scientific content of the subject and seems mainly serving to create audience.

7.6 ▪ Static Neural Networks

Let $\sigma : \mathbb{R} \to \mathbb{R}$ be a function, and let $m; n; p$ be positive integers. A *single hidden layer Neural Network* with a p-dimensional input u, m-dimensional output y, n hidden units, and activation function σ, is a function specified by a pair of matrices $B,\ C$ and a pair of vectors $b_0,\ c_0$, where B and C are real matrices of respective sizes $n \times p$ and $m \times n$, and b_0 and c_0 called the *shift* and the *offset*, are real vectors of size n and m.

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

$$\vec{\sigma}(x_1, \ldots, x_n) = \begin{bmatrix}\sigma_1(x) & \ldots & \sigma_n(x)\end{bmatrix}^\top := \begin{bmatrix}\sigma(x_1) & \ldots & \sigma(x_n)\end{bmatrix}^\top ,$$

so that the i-th component of the vector $\vec{\sigma}(x)$ depends only on the i-th coordinate of x. Then the 5-tuple

$$\Sigma \equiv \{B,\ C,\ b_0,\ c_0,\ \sigma\}$$

is meant to realize the **(one layer) Neural Network** map

$$f_\Sigma : \mathbb{R}^p \to \mathbb{R}^m : \quad u \mapsto C\,\vec{\sigma}(Bu + b_0) + c_0\,. \tag{7.6.1}$$

Sometimes this function is called the *behavior* of the net, leaving the name **Neural Network** to designate the graph representing the interconnections of the various units which compose the function.

In compact notation, a one-layer Neural Network is a composite map of the type

$$h \circ \vec{\sigma} \circ g, \tag{7.6.2}$$

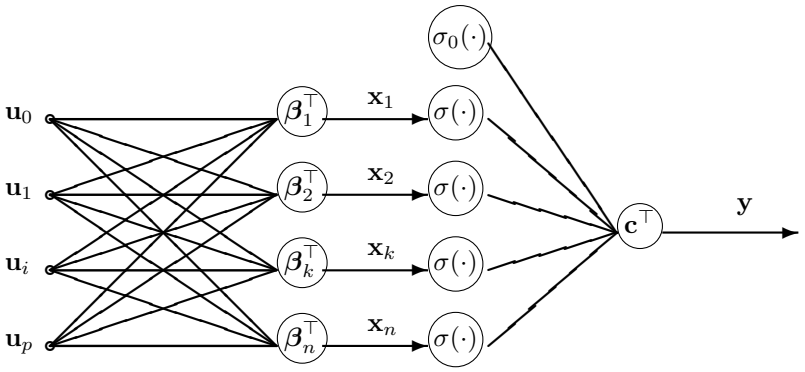


Figure 7.6.1. *One-layer Neural Network with a scalar output*

where $h(x) = Cx + c_0$ and $g(u) = Bu + b_0$ are affine maps. When $c_0 = 0$ the network is said to have no offset.

Remarks 7.1. One can always rewrite a behaviour function (7.6.1) as one without shift parameter b_0 by introducing an extra input $u_0 = 1$ and augmenting the matrix B by a zeroth-index column b_0. This notation has already been implemented in Figure 7.6.1 where the matrix B is specified row by row as

$$B = \begin{bmatrix} \boldsymbol{\beta}_1^\top \\ \dots \\ \boldsymbol{\beta}_n^\top \end{bmatrix}, \qquad \boldsymbol{\beta}_k \in \mathbb{R}^{p+1}.$$

Also we shall introduce an extra zeroth-index component σ_0 of the function $\vec{\sigma}$ identically equal to 1 so that the vector c_0 can also be absorbed in an enlarged matrix C of dimension $m \times (n+1)$.

Multiple Hidden-layer Networks are constructed by cascading in series many one-layer Neural Networks making each output to act as the input to the next Network. One can then obtain structures of arbitrary complication. A network with N hidden layers has as behavior function the successive composition of N one-layer networks:

$$f_\Sigma(u) = \mathbf{B}_N \vec{\sigma}(\mathbf{B}_{N-1}\vec{\sigma}(\dots\vec{\sigma}(\mathbf{B}_1 u)))$$

where the $\mathbf{B}_k = \begin{bmatrix} b_{0,k} & B_k \end{bmatrix} \in \mathbb{R}^{n_k \times (n_{k-1}+1)}$ are matrices defining the k-th layer with n_k activation units. The last $\mathbf{B}_N = \begin{bmatrix} c_0 & C \end{bmatrix}$ contains the output map parameters. Of course one should mantain the convention that the zero-th component of each $\vec{\sigma}$ is a constant equal to 1 so that $n_0 = p = \dim u$ and $n_N = \dim y$.

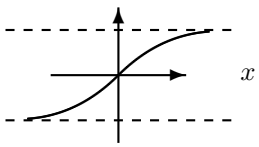
Since the first attempts people discovered experimentally, that with a structure of this kind one could approximate arbitrarily well an arbitrary continuous nonlinear function. The reason of this fact has however always been quite mysterious.

Universal approximating functions

The hidden layer always involves replicas of the same nonlinear *activation function* σ and one may wonder what special characteristics this function should have in order to obtain reasonable approximation properties. It turns out that the function could be chosen quite arbitrarily. In the literature there are different choices of σ which seem to work roughly the same. Popular classes of functions are the hyperbolic tangent, the so-called **sigmoid**

$$\sigma(x) = \frac{1}{1+e^{-ax}} = \tanh(\frac{a}{2}x) + 1\,, \qquad a > 0$$

which by choosing a large approximates the step function,



$$\sigma(x) = \tanh\, x$$

Figure 7.6.2. *Hyperbolic Tangent*

and the **radial functions**

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

where ϕ is an exponential, but one could use essentially any kind of function. This insensitivity to the shape of the activation function has led to extreme simplifications. In "deep networks" for classification it is now customary to use a *rectifier* characteristic like:

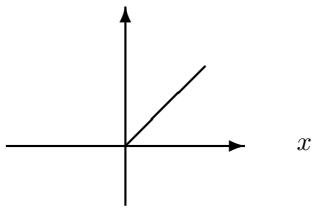


Figure 7.6.3. *Rectifier*

Every function seems to work except for the notable *exception of polynomials*. Why is this so?
It seems to be unappreciated by many that the success of Neural Networks as function approximation devices stems from the approximation capability of

linear combinations of **shifted versions** of the **same** activation function; say linear combinations the type,

$$\sum_k c_k \sigma(x + \tau_k)\,.$$

These can approximate arbitrary continuous nonlinear functions. In this respect an old and little known result of Norbert Wiener [104] states this fact in rigorous terms as follows.

Theorem 7.1 (Wiener 1932). *In order for the linear span 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 $\hat{f}(i\omega)$ 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 $\hat{f}(i\omega)$ is nonzero almost everywhere. This result has been generalized in several ways to continuous functions; see e.g. [24, Theorem 1].

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

Identifiability

As we have already hinted at, for any parameter estimation problem to be well-posed one needs at least local identifiability. Unfortunately the multi-layer Neural Networks which are often used in applications involve hundreds of parameters which are related to the model input-output map in a complicated way, making identifiability very hard to check. It seems in fact that in practice the condition is very seldom satisfied. Nevertheless due to the availability of easy and ready to implement computer routines, Neural Networks are an extremely popular model building device. Recently, *Deep Learning* has become a new paradigm of Machine Learning, advocating the use of extremely complicated networks with a very large number of hidden layers and an enormous number of parameters. We do not understand the rationale of this new paradigm; to us it seems to be just based on the gigantic "machine cranking" power afforded by nowadays colossal supercomputers than on true insight.

IdentifiabilIty of *one-layer* Neural Networks has been investigated in [29, 5]. One says that two (single layer) networks Σ and $\hat{\Sigma}$ having the same activation function, are *input/output equivalent* if

$$f_\Sigma = f_{\hat{\Sigma}}$$

that is when they realize the same function. The identifiability question for one-layer Neural Neworks is: when does $f_\Sigma = f_{\hat{\Sigma}}$ imply equality of the parameters

$\{B,\ C,\ b_0,\ c_0\}$?

The answer is that for sigmoid functions and offset $c_0 = 0$ this is true modulo signs [29]. So one at least has local identifiability.

In general however identifiability does not hold. In fact, there are *ad hoc* procedures, which we shall not describe here, to eliminate parameters which seem to be not influencing the input-output behaviour of the network. Of course based on semi-empirical judgements. This is called *pruning*, which should be done by *"optimal brain damage"* [55] algorithms.

7.7 ▪ Gradient descent and back-propagation

Assuming a fixed activation function σ, the behaviour function (7.6.1) of a single-layer network depends on $(n+1)m+(n+1)p$ parameters, which are the components $c_{k,j}$ of the output matrix C plus the offset vector $\mathbf{c}_0$ and the components $b_{j,h}$ of the augmented input matrix B including the the n-dimensional shift parameter $\mathbf{b}_0$. These scalar parameters, denoted θ, are generically called *weights* of the net and the output of (7.6.1) corresponding to an input sequence u is written $y = f_\Sigma(\theta,\ u)$ or simply as $y = f_\theta(u)$.

The "training" of a Neural Network is normally done by Least Squares. Given a training set

$$\{u(t),\ y(t)\,;\, t = 1, 2, \ldots, N\,\}$$

one wants to "learn" the network by minimizing the least squares distance

$$J_N(\theta) := \frac{1}{2}\sum_{t=1}^{N} \| \, y(t) - f_\theta(\mathbf{u}(t)) \, \|^2 \tag{7.7.1}$$

with respect to the parameter θ. Here we assume that the number of "neurons" n in the hidden layer is given to us but this is very seldom the case and one should really optimize also with respect to n in a certain range decided by taking into account the appropriate warnings for overfitting. Since $J(\theta)$ is a nonlinear function of the unknown parameter the minimization can only be done numerically. Note that the function is not convex and there may be several minima. The simplest optimization algorithm one could use is the *gradient descent* which we shall describe in the following.

For simplicity, we will assume from now on that $m = 1$; (single output network) and use the augmented variables notation explained in Remark 7.1 so that $C \equiv \mathbf{c}^\top$ will now be the row vector $\begin{bmatrix} c_0 & c^\top \end{bmatrix}$ of dimension $n+1$ and in the expression (7.6.2) we shall understand that there are $p+1$ inputs so that $g(u) = b_0 + Bu$, is a n-vector valued function and $\vec{\sigma}(x)$ is a $n+1$-vector function whose 0-th component is identically equal to 1. We shall first compute the derivatives of a simplified cost function $J(\theta) := \frac{1}{2}(\, y - \hat{y}_\theta \,)^2$ where

$$\hat{y}_\theta = \mathbf{c}^\top \vec{\sigma}(x)\,; \qquad x = g(u) = \begin{bmatrix} \boldsymbol{\beta}_1^\top \\ \boldsymbol{\beta}_2^\top \\ \ldots \\ \boldsymbol{\beta}_n^\top \end{bmatrix} \mathbf{u}$$

(where x has $n+1$ components) to obtain

$$\frac{\partial J(\theta)}{\partial c_k} = -\,(\,y-\hat{y}_\theta\,)\,\sigma(x_k)\;, \qquad k=1,\ldots,n$$
$$\frac{\partial J(\theta)}{\partial \boldsymbol{\beta}_i} = -\,(\,y-\hat{y}_\theta\,)\,\frac{\partial}{\partial \boldsymbol{\beta}_i}\sum_{k=0}^{n}\mathbf{c}_k\,\sigma(\boldsymbol{\beta}_k\mathbf{u}) = $$
$$= -\,(\,y-\hat{y}_\theta\,)\,\mathbf{c}_i\,\sigma'(x_i)\,\mathbf{u}\;, \qquad i=1,2,\ldots,n$$

where $\sigma(x_0)\equiv 1$ and $\sigma'(x)=\dfrac{d}{dx}\sigma(x)$. We can then apply these expressions to compute the gradient of the actual quadratic cost $J_N(\theta)$, getting

$$\frac{\partial J_N(\theta)}{\partial c_k} = -\sum_{t=1}^{N}[\,y(t)-f_\theta(\mathbf{u}(t))\,]\,\sigma(x_k(t))\;, \qquad k=0,1,\ldots,n$$
$$\frac{\partial J_N(\theta)}{\partial \boldsymbol{\beta}_i} = -\sum_{t=1}^{N}[\,y(t)-f_\theta(\mathbf{u}(t))\,]\,\mathbf{c}_i\,\sigma'(x_i(t))\;\mathbf{u}(t)\,.$$

The optimization is usually implemented by a gradient descent algorithm of which there are various versions. The simplest is probably the batch version which uses all the data at each step and has the form:

$$\tilde{\mathbf{c}}(k+1) = \tilde{\mathbf{c}}(k) - \gamma_c(k)\frac{\partial J_N[\theta(k)]}{\partial \tilde{\mathbf{c}}} \tag{7.7.2}$$
$$\boldsymbol{\beta}_i(k+1) = \boldsymbol{\beta}_i(k) - \gamma_{\mathbf{b}}(k)\frac{\partial J_N[\theta(k)]}{\partial \boldsymbol{\beta}_i}\;, \qquad i=0,1,2,\ldots,p \tag{7.7.3}$$

where for convergence, each sequence $\{\gamma(k)\,;\,k=1,2,\ldots\}$ of positive steplengths must go to zero slowly enough so as to make $\sum_k\gamma(k)=+\infty$ but $\sum_k\gamma^2(k)<\infty$. For example taking $\gamma(k)$ proportional to $\dfrac{1}{k}$ would do. There are several variants of gradient-like algorithms which are for example described in detail in the *Matlab Neural Networks toolbox guide* [**?**, Sect. 5.2]. Newton or Quasi-Newton methods are essentially inapplicable due to the large dimension of the Hessians.

When the neuron activation functions are sigmoids, or exponential radial functions, some of the calculations needed to compute the gradient with respect to the output parameters in (7.7.2), can be used in the gradient with respect to the input parameters in in (7.7.3). This is called *Back-propagation*.

More specifically, let $z_k(t)=\sigma(x_k(t))$ be the output of the k-th neuron and $\epsilon_\theta(t) := [\,y(t)-f_\theta(\mathbf{u}(t))\,]\,z_k(t)$. This may be interpreted as an approximation error of $y(t)$ based on the hidden layer output $x_k(t)$, with parameter value θ. Now since $\sigma'(x_k)=a\sigma(x_k)[\sigma(x_k)-1]$ we can substitute the computed $\epsilon_\theta(t)$, $\mathbf{c}_k$ and $x_k(t)$ into the gradient with respect to the input parameters

$$[\,y(t)-f_\theta(\mathbf{u}(t))\,]\,\mathbf{c}_i\,\sigma'(x_k(t))\;\mathbf{u}(t)$$

which is the *back substitution*. The generalization to multiple-layer and multiple-output cases is quite straightforward even if the notations tend to be complicated. In particular Back-Propagation applies exactly the same to the output-input parameters $\mathbf{c}$ and $\mathbf{b}$ of a cascaded multi-layer network. Of course considered in the reverse order, going backwards from outputs to inputs.

7.8 ▪ Bayesian Neural Networks

In this section we shall take a statistical approach and describe the training set as involving a sequence of **random** output measurement

$$\{u(t),\ \mathbf{y}(t)\,;\, t = 1, 2, \ldots, N\,\}$$

where the $\mathbf{y}(t)$'s have a structured parametric mean function $f_\theta(u(t))$ depending on a parameter θ and on the deterministic input $u(t)$. Each $\mathbf{y}(t)$ is corrupted by an additive random perturbation $\mathbf{w}(t)$ which we shall assume Gaussian and i.i.d.

The Gaussian log-likelihood function for N i.i.d. observations will then be the least squares distance

$$J_N(\theta) := \frac{1}{2\sigma^2} \sum_{t=1}^{N} \| y(t) - f_\theta(u(t)) \|^2, \tag{7.8.1}$$

involving the noise variance σ^2 and the parameter θ. Clearly imposing to $f_\theta(u(t))$ the Neural Network structure of the previous section, the problem of ML estimation of θ ends up with solving the same non linear least squares problem.

We shall now introduce the simplest prior distribution on the parameters, describing them as *zero mean independent Gaussian random variables* all having the same variance γ^2

$$p(\theta) \equiv \exp\{-\frac{1}{2\gamma^2} \|\theta\|^2\,\}\,.$$

Then, exactly as in the linear case, the MAP estimator of θ is found by solving the regularized minimization problem

$$\min_\theta \left\{ \sum_{t=1}^{N} \| y(t) - f_\theta(u(t)) \|^2 + \lambda^2 \|\theta\|^2 \right\} \tag{7.8.2}$$

which is a nonlinear ridge regression problem. Here $\lambda^2 = \dfrac{\sigma^2}{\gamma^2}$ has still the meaning of the inverse of a signal-to-noise power ratio. Hence very noisy measurements (that is $\sigma^2 \gg \gamma^2$) will lead to higher trust in the prior.

The regularization term induces the so-called **weight decay** (in the literature the parameters of a Neural Network are often called weigths). This can actually be seen from the gradient descent equation as follows.
Every time we update a weight parameter θ with the negative gradient ∇J with respect to θ, we must subtract from it a term $\lambda^2\theta$ to get a descent algorithm of the form

$$\theta(k+1) = \theta(k) - \gamma(k)\nabla J[\theta(k)] - \lambda^2\theta(k)\,,$$

This gives the weights (parameter estimates) a tendency to decay towards zero, hence the name of weight decay.

Intuitively, as in linear ridge regression the main reason this is done is to *prevent overfitting* which is an ever present problem in Neural Network modeling. When looking at regularization from this angle, one might even suggest a Lasso-type regularization but unfortunately the algorithmics seems to become extremely complicated.

Actually, as explained earlier, in the regularization cost one should not include the offset parameters in c_0 which correspond to the estimate of the mean value.

Bayesian prediction with Neural Networks

Our main interest in using neural networks is to predict the values of the output variable for new values of the input variables. From the independence of the sequence $\{\mathbf{y}(t)\}$ and the previous discussion we see that

$$p[y(t+1) \mid u^{t+1}, \mathbf{y}^t] = p[\, y(t+1) \mid u^{t+1}]$$

where the sequence $u^{t+1} := \{u(1), u(2), \dots, u(t+1)\}$ is of deterministic variables and the last conditioning sign has just the meaning of dependence. The probability on the right side actually depends on the unknown parameter θ. Therefore the predictions should be made by integrating over the a posteriori parameter density, so that

$$p[y(t+1) \mid u^{t+1}, y^t] = \int p[y(t+1) \mid u^{t+1}, \theta]\, p(\theta \mid u^{t+1}, y^t\,)\, d\theta$$

Here $p[y(t+1) \mid u^{t+1}, \theta]$ is just the Gaussian likelihood function with exponent $\frac{1}{2\sigma^2} \parallel y(t+1) - f_\theta(u(t+1)) \parallel^2$.

Since the posterior density is hard to compute, one should use some approximation. The usual trick is to assume that the posterior is peak-shaped at the MAP value so that it approximately behaves like a delta function centered at the value $\hat{\theta}_{MAP}$ in the integral, yielding

$$p[y(t+1) \mid u^{t+1}, y^t] \simeq p[y(t+1) \mid u^{t+1}, \hat{\theta}_{MAP}(u^t, y^t)]$$

which is a ($\simeq$conditional) Gaussian density with mean $f_{\hat{\theta}_{MAP}}(u(t+1))$.

Note that $\hat{\theta}_{MAP}(u^{t+1}, y^t)$ is computed by minimizing the Bayes cost (7.8.2) with only t output sample values, that is with $N = t$ so that it does not depend on $u(t+1)$ and hence $\hat{\theta}_{MAP}(u^{t+1}, y^t) = \hat{\theta}_{MAP}(u^t, y^t)$. In conclusion we have:

Proposition 7.5. *Assume that t is large enough and the posterior density $p(\theta \mid u^t, y^t)$ is unimodal and sharply pick shaped at its maximum $\hat{\theta}_{MAP}$, then the conditional mean of $\mathbf{y}(t+1)$ given the past and present observations u^{t+1}, y^t, is*

$$\mathbb{E}\,[\mathbf{y}(t+1) \mid u^{t+1}, y^t\,] = f_{\hat{\theta}_{MAP}}(u(t+1))$$

We shall leave the discussion of **the Linear case** to the problem below.

Problem 7.1. *Assume that in the linear model*

$$\mathbf{y}(t) = \theta_0 + \theta_1 u(t) + \mathbf{w}(t)\,, \qquad t = 1, 2, \dots$$

the noise $\mathbf{w}$ is Gaussian i.i.d. of zero mean and variance σ^2 and that you have a prior on the parameters which is also Gaussian, zero-mean with covariance $\gamma^2 I_2$. The inputs $u(t)$ form a deterministic sequence.

Find the formula for the conditional expectation $\mathbb{E}\,[\mathbf{y}(t+1) \mid u^{t+1}, y^t\,]$.

Dynamic Neural Networks

In time series modeling, a **Nonlinear Autoregressive eXogenous** model (NARX) is a nonlinear model which explains the current output $y(t)$ at time t by describing it as a nonlinear function of a finite sequence past outputs and inputs.

Something like

$$\mathbf{y}(t) = F\left[\mathbf{y}(t-1), \ldots, \mathbf{y}(t-n); u(t-1), \ldots, u(t-p)\right] + \mathbf{w}(t)\,; \qquad t = 1, 2, \ldots \tag{7.8.3}$$

which obviously generalizes the linear ARX models to be introduced in Sect. **??**. Given a (long) training set of data

$$\{y(t),\ \mathbf{u}(t)\,;\, t = 1, 2, \ldots, N\}$$

which for simplicity we assume made of scalar signals, one can fit a Neural Network to these data by least squares, using for example the same gradient algorithm seen in the previous section. The idea is based on pretending that the current input is now a $n + p$-dimensional vector sequence

$$\mathbf{v}(t) := \begin{bmatrix}\mathbf{y}(t-1) & \ldots & \mathbf{y}(t-n) & u(t-1) & \ldots & u(t-p)\end{bmatrix}^{\top}\,; \tag{7.8.4}$$

where $t = t_0,\ t_0+1, , t_0+2, \ldots$, with t_0 being a suitable initial time which allows the data in $\mathbf{v}(t_0)$ to be defined.

If the autoregressive part is non trivial ($n > 0$) these models may however suffer of serious stability problems.

7.9 ▪ Non Linear Support Vector Machines

We shall now go back to the (deterministic) classification problem studied in Sect. 4.5. The mathematical tools which we have introduced in Section 6.6 will allow us to address the **non linearly separable case**.

A natural attempt to deal with training sets which are not linearly separable is to use nonlinear discriminant functions. A situation like the one depicted in Fig 7.9 below may easily be solved by transforming the feature coordinates

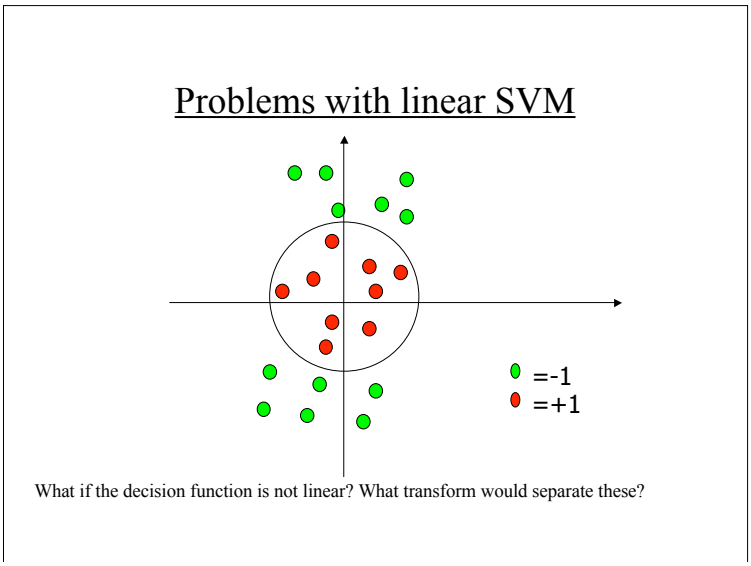


Figure 7.9.1. *Separability by polar coordinates*

into polar ones $\{\rho,\, \theta\}$ by which the red features are linearly separated from the green ones by a line $\{\rho = \rho_0\}$.

However, when there is no a priori geometric insight on the structure of the feature set one could just try to generalize the linear structure (4.4.5) to a generic nonlinear one, say a quadratic or, more generally, polynomial discriminant function, e.g.:

$$\varphi(y) = b + \sum_k \beta_k\, y_k + \sum_{k,j} \beta_{k,j}\, y_k y_j$$

which can produce quadratic (or polynomial) discriminant surfaces in $\mathbb{R}^p$. The quadratics are the same kind of separating surfaces which arise in the general multivariate Gaussian MLR (4.3.3). In this case the coefficients $\beta_k,\ \beta_{k,j}$ are components of the sample mean and sample variance matrices which are a priori estimated from the training set. When no probabilistic information is available, the unknown coefficients should be estimated from the training data. Since $\beta_k,\ \beta_{k,j}$ appear linearly, their estimation should be reducible to a linear problem. One may then be tempted to add cubic and higher order polynomial terms to get more flexible functions but should keep in mind that the number of coefficients grows exponentially with the degree, to a point where the required

computations become completely unrealistic. This *curse of dimensionality* can for example be seen already in the case one may want to obtain a decision surface corresponding to a polynomial of degree two. To stay in the framework of the linear discriminant theory one must create a larger feature space by renaming the feature coordinates as

$$\begin{aligned} z_1 \sim y_1, \ldots, z_p \sim y_p, \qquad & p \text{ coordinates}, \\ z_{p+1} \sim y_1^2, \ldots, z_p \sim y_p^2, \qquad & p \text{ coordinates}, \\ z_{2p+1} \sim y_1 y_2, ..., z_{\mathbf{N}} \sim y_p y_{p-1}, \qquad & \frac{p(p-1)}{2} \text{ coordinates}, \end{aligned} \tag{7.9.1}$$

which has dimension $\mathbf{N} = \frac{p(p+3)}{2}$ and may be already too large for large p.

In 1992, Bernhard E. Boser, Isabelle M. Guyon and Vladimir N. Vapnik suggested a way to create nonlinear classifiers by applying the so-called **kernel trick** (originally proposed by Aizerman et al.[2]) to maximum-margin hyperplanes. The resulting algorithm is formally similar to the linear one, except that every inner product of feature vectors is replaced by a nonlinear kernel function. This allows the algorithm to fit the maximum-margin hyperplane in a transformed feature space which has the structure of a *Reproducing Kernel Hilbert Space*. Although the classifier is still a hyperplane in the transformed feature space, it is generally nonlinear in the original input space and may allow to separate non linearly separable features. The mathematical background is discussed in Sect. 6.6. First we need to introduce transformations of the feature space.

Feature maps

A *feature map* is a continuous map $\varphi : \mathcal{X} \to \mathbf{F}$, where $\mathbf{F}$ is a separable Hilbert space which we will call the *extended feature space*. In practical applications $\mathbf{F}$ is, or is approximable by, a large dimensional Euclidean space. Given an orthonormal basis $\{e_k\,; k = 1, 2, \ldots, \mathbb{N}\}$ in $\mathbf{F}$, the image $z = \varphi(x)$ can be represented as a vector with $\mathbb{N}$ coordinates

$$z_k = \varphi_k(x) := \langle e_k, \varphi(x) \rangle_{\mathbf{F}}\,; \qquad k = 1, 2, \ldots, \mathbb{N}\,.$$

The dimension $\mathbb{N}$ can be finite or infinite but generally is much larger than the dimension p of the original features. Since the map φ is far from being one-to-one, it provides in general, a redundant new coordinatization of the feature space $\mathcal{X}$. A simple standard example is the quadratic map (7.9.1) where $\mathbb{N} = p(p+3)/2$ and $\mathbf{F}$ is simply the Euclidean space of dimension $\mathbb{N}$. In this section we provide a representation of RKHSs in terms of feature maps.

The key point is that every feature map defines a kernel via

$$K(x, y) := \langle \varphi(x), \varphi(y) \rangle_{\mathbf{F}} \quad x, y \in \mathcal{X}. \tag{7.9.2}$$

It follows from the properties of the inner product in $\mathbf{F}$ that K is symmetric and positive (semi-)definite. In fact, for any $a \in \mathbb{R}^{\mathbb{N}}$ with a finite number of non zero components (compact support), one has

$$\sum_{i,j=1}^{\mathbb{N}} a_i K(x_i, x_j) a_j = \langle \sum_{i=1}^{\mathbb{N}} a_i \varphi(x_i), \sum_{j=1}^{\mathbb{N}} a_j \varphi(x_j) \rangle_{\mathbf{F}} = \| \sum_{i=1}^{\mathbb{N}} a_i \varphi(x_i) \|_{\mathbf{F}}^2 \geq 0\,.$$

For example letting $\mathbf{F}$ to be a space of second order random variables on some probability space, say $\mathbf{F} = L^2(\Omega, \mathcal{A}, P)$, the inner product becomes correlation. Let further $\mathcal{X}$ be a time interval (finite or infinite), then $\{\varphi(x);\, x \in \mathcal{X}\}$ is a random signal (process) and $K(x, y)$ is just the correlation (or covariance in case of zero mean) function of a second-order process. Positivity of the kernel follows by the same argument showing positive definiteness of the covariance function of a stochastic process or, of the covariance matrix of a random vector whenever $\mathcal{X}$ is a discrete set. This example is discussed in depth in the literature, e.g. [66], [101].
Conversely, for every positive definite function the corresponding RKHS has infinitely many associated feature maps such that (7.9.2) holds. For example, one can take a family of generators $\{\varphi_x = K(\cdot, x)\,;\, x \in \mathcal{X}\}$ mapping $\mathcal{X}$ into the possibly infinite-dimensional RKHS defined by K. If $\mathcal{X}$ is a discrete set, they constitute a (not necessarily orthonormal) basis of the RKHS. Then positivity in (7.9.2) is satisfied by the reproducing property. Another classical and useful example is to introduce a feature map by exploiting the representation (6.5.5) of Mercer's theorem in the previous section[23] by taking $\mathbf{F} = \ell^2$ and a vector function $\varphi(x) = [\,\sqrt{\lambda_i}\varphi_i(x)\,]_{i=1,2,\ldots}$ which takes values in ℓ^2. In this way we get a canonical feature map of dimension $\mathbb{N} = \infty$ which can be approximated to finite dimension by truncation.

Application of RKHS to Non linear SVM

As we have just seen, every continuous feature map can naturally define a RKHS of a positive definite kernel function. The connection between kernels and feature maps provides us with a new way to understand linear separability in an extended feature space $\mathbf{F}$.

In fact, feature maps allow us to construct linear function spaces that generalize to the enlarged feature space $\mathbf{F}$, the idea of a linear subspace or affine hyperplane. One can generalize by thinking of the running variable in $\mathbf{F}$ as being the image of a feature map, function of $x \in \mathcal{X}$. Consider the possibly infinite-dimensional linear space

$$\mathbf{H}_\varphi = \{f : \mathcal{X} \to \mathbb{R} \mid \exists\, \boldsymbol{\beta} \in \mathbf{F},\ f(x) = \langle \boldsymbol{\beta}, \varphi(x) \rangle_{\mathbf{F}}, \forall x \in \mathcal{X}\}.$$

a subspace of linear functionals each defined by a vector parameter $\boldsymbol{\beta} \in \mathbf{F}$. It can be interpreted as an affine subspace (hyperplane) by assuming that $\varphi(x)$ has a constant component. We can define a norm on $\mathbf{H}_\varphi$ by identifying a functional with its parameters $\boldsymbol{\beta}$. Since the $\boldsymbol{\beta}$'s are only defined up to a multiplicative constant one takes the norm to be the minimal in all possible linear representations as

$$\|f\|_\varphi = \inf\{\|\boldsymbol{\beta}\|_{\mathbf{F}} : \boldsymbol{\beta} \in \mathbf{F},\ f(x) = \langle \boldsymbol{\beta}, \varphi(x) \rangle_{\mathbf{F}}, \forall x \in \mathcal{X}\}.$$

It can be shown that $\mathbf{H}_\varphi$ is itself a RKHS with kernel defined by

$$K(x, y) = \langle \varphi(x), \varphi(y) \rangle_{\mathbf{F}}$$

This representation implies that the elements of the RKHS $\mathbf{H}_\varphi$ are inner products of elements in the extended feature space and can accordingly be seen as

[23] There is a generalization of the thorem actually holding for argument t running on a real space of arbitrary finite dimension.

subspaces or hyperplanes. This view of the RKHS is related to the so-called *kernel trick* in machine learning [45], [74] which can be explained as follows.

Assume that the two classes, once reparameterized in the extended feature space, are linearly separable, as sketched in Fig. 7.9.2

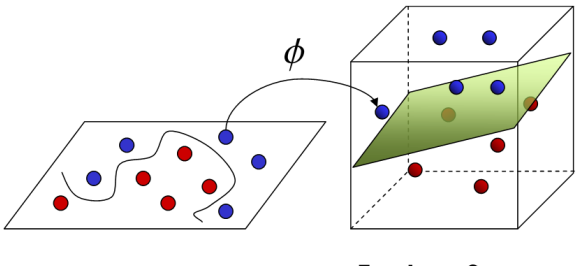


Figure 7.9.2. *Separability by extension*

Note that the value of the classification variable y associated to a feature x in the training set remains the same also after passing to a new coordinatization by a feature map since it is intrinsically related to the location of a *phisical feature* no matter how it is described by a coordinate system. Hence if say, x_k is classified with the label $y_k = 1$, that is as belonging to a pattern $\mathcal{P}_1$ instead of the alternative pattern $\mathcal{P}_2$, then $\varphi(x_k)$ will have to be associated to the same classification label $y_k = 1$.

We want now to consider the maximum margin problem in the extended feature space $\mathbf{F}$ where we assume we have obtained separability. By mimicking the linear geometry of a separable feature space in $\mathbb{R}^p$ which we have exploited in Section 4.5, the problem still reduces to minimizing the square norm of $\boldsymbol{\beta} \in \mathbf{F}$ subject to the empty slab constraint in the extended feature space $\mathbf{F}$, that is:

$$\min_{\boldsymbol{\beta},b} \left\{ \frac{1}{2} \|\boldsymbol{\beta}\|^2 \,;\, \text{subject to: } y_k(\boldsymbol{\beta}^\top \varphi(x_k) + b) \geq 1, \quad k = 1, \dots, N \right\} \tag{7.9.3}$$

which is again a convex quadratic programming problem with linear inequality constraints in the extended setting. The maximization can be performed resorting to the dual Lagrangian quadratic cost

$$L_D = \sum_{k=1}^{N} \lambda_k \; - \frac{1}{2} \sum_{i=1}^{N} \sum_{k=1}^{N} \lambda_i \lambda_k \, y_i y_k \langle z_i, z_k \rangle_{\mathbf{F}} \tag{7.9.4}$$

where $z_i,\ z_k \in \mathbf{F}$ are transformations of original feature points, so that

$$\langle z_i, z_k \rangle_{\mathbf{F}} = \langle \varphi(x_i),\, \varphi(x_k) \rangle_{\mathbf{F}} = K(x_i, x_k)\,.$$

If $\mathbb{N} < \infty$, the kernel can also be written as $K(x_i, x_k) = \varphi(x_i)^\top \varphi(x_k)$. The dual cost (7.9.4) is to be maximized with respect to the N multipliers λ_k. The "kernel

trick" consists just of the observation that the cost only depends on the kernel K, e.g.

$$L_D = \sum_{k=1}^{N} \lambda_k - \frac{1}{2} \sum_{i=1}^{N} \sum_{k=1}^{N} \lambda_i \lambda_k \, y_i y_k \, K(x_i, x_k) \tag{7.9.5}$$

so that in practice *one can choose the kernel from the outset and forget about the feature map* φ which is implicitly determined by the kernel itself, say via Mercer's eigenfunction expansion. The remarkable fact is that one does not need to compute the $\mathbb{N}$ eigenfunctions and the $\mathbb{N}$-dimensional inner products in the extended space. So in practice **you don't even need to see the feature map**!

The Non-linear decision boundary

In the extended feature space a discriminant function is described by an equation of the form

$$g(x) = \langle \beta, \, \varphi(x) \rangle_{\mathbf{F}} + b = \sum_{k=1}^{\mathbb{N}} \beta_k \varphi_k(x) + b$$

where φ is a feature map. The maximum margin problem in the extended space is now formulated as the optimization problem (7.9.5) in terms of a kernel function K. Hence, the support vectors in the original feature space are now just the vectors x_k for which $\lambda_k > 0$ in the (unique) solution of Problem (7.9.5). Denoting again by SV the index set of support vectors, the optimal parameter vector of the hyperplane can be written

$$\beta^* = \sum_{k \in SV} \lambda_k^* \, y_k \, \varphi(x_k)$$

which is clearly also an element of $\mathbf{F}$. Therefore we have

Theorem 7.2. *The optimal (maximum margin) hyperplane in* $\mathbf{F}$ *is described by the non-linear function of* $x \in \mathcal{X}$*:*

$$g^*(x) = \sum_{k \in SV} \lambda_k^* y_k \, K(x_k, x) + b^* \tag{7.9.6}$$

where the optimal Lagrange multipliers are the unique solution of the optimization problem (7.9.5)*.*

Hence the decision rule for the classification of a new feature x in the original space is

$$y = \operatorname{sign} \left[g^*(x) \right].$$

and the discriminant boundary between the two classes is the nonlinear curve defined by the equation $g^*(x) = 0$.

In principle the just exposed SVM procedure seems quite simple and well suited to the application to practical automatic classification problems. There are however some basic questions which need to be addressed:

1. How to choose the kernel ?

2. How to check linear separability in the extended feature space ?
3. How to deal with noise (overfitting) ?
4. Optimality: Could other methods (say Neural Networks) perform better?

There is a variety of kernels proposed in the literature such as

$$K(x,y) = (\langle x,y\rangle + 1)^d \tag{7.9.7a}$$

$$K(x,y) = \exp\{\frac{1}{\sigma^2}\|x-y\|^2\} \tag{7.9.7b}$$

$$K(x,y) = \tanh\{\kappa\langle x,y\rangle - \delta\} \tag{7.9.7c}$$

which have different names and seem to be used in the literature based on *ad hoc* and quite arbitrary principles. In particular (7.9.7b), called Gaussian Kernel, seems to be a quite popular choice while the algebraic Kernel (7.9.7a) leads to polynomial feature maps of degree d. The third is the famous *sigmoid* of Neural Networks (which is however positive only for some values of κ, δ).

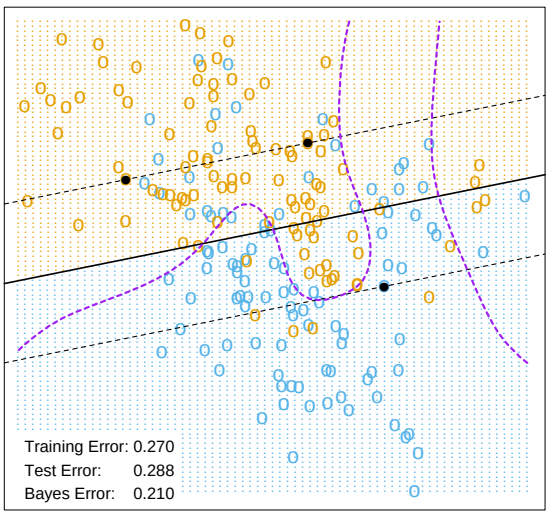


Figure 7.9.3. *Degree 4 polynomial kernel vs linear separation. From [44]*

Linear separability seems to hold for general kernels involving infinitely many eigenfunctions (feature maps), that is for $\mathbb{N}$ tending to ∞. This is an abstract topological question which seems to be still open. In practice it may hold only approximately. Note in any case that the Kernel expansion (7.9.6) always involves a finite number of (support vector) terms.

SVM regression

The maximum margin optimization problem can be transformed into a regularized least squares. Just recall that we are minimizing the distance from a

separating hyperplane described by the equation $\beta^\top x + b = 1$. The Lagrangian of the maximum margin problem:

$$\min_{\beta,b} \left\{ \frac{1}{2} \|\beta\|^2 \,;\, \text{subject to: } y_k(\beta^\top x_k + b) \geq 1, \quad k = 1, \ldots, N \right\}$$

can be rearranged into a regularized optimization problem:

$$\min_{\beta,b} \left\{ \sum_{k=1}^{N} V(1 - y_k(\beta^\top x_k + b)) + \frac{\gamma}{2}\|\beta\|^2 \right\}$$

where $V(\cdot)$ can be an arbitrary convex symmetric cost function, ex. $V(\cdot) = [\cdot]^2$. Both problems have a solution of the same form

$$g^*(x) = \sum_{k=1}^{N} \alpha_k \langle x_k, x \rangle + b^* \,.$$

where some of the α_k's may be zero. We can transfer this idea to the nonlinear extended feature setting.
Let $\varphi(x)$ be a feature map. The maximum margin problem on extended feature space:

$$\min_{\beta,b} \left\{ \frac{1}{2} \|\beta\|^2 \,;\, \text{subject to: } y_k[\beta^\top \varphi(x_k) + b] \geq 1, \quad k = 1, \ldots, N \right\}$$

can be reformulated as the optimization problem:

$$\min_{\beta,b} \left\{ \sum_{k=1}^{N} V(1 - y_k(\beta^\top \varphi(x_k) + b)) + \frac{\gamma}{2}\|\beta\|^2 \right\}$$

where again $V(\cdot)$ can be taken as $[\cdot]^2$. This problems also has a solution of the form

$$g^*(x) = \sum_{k=1}^{N} \alpha_k K(x_k, x) + b^*, \qquad K(x, z) = \langle \varphi(x), \varphi(z) \rangle_{\mathbf{F}}$$

Hence the solution to the SVM problem can also be obtained by a *regularized Kernel regression*. See [44] p. 426 for more details.

For more information on Support Vector Machines see [20], [44, p. 167-174 and 423-440], [45] and the excellent Vikipedia survey
`https://en.wikipedia.org/wiki/Reproducing_kernel_Hilbert_space`.

Chapter 8

ARX MODELING OF TIME SERIES

8.1 ▪ Introduction: Discrete-time signals

In this chapter we shall start addressing the study of *dynamic phenomena* in particular prepare for the study of statistical problems involving time, where the data of our inference problems will be sequences of observations (possibly infinite) indexed by time. These objects we shall later model as *trajectories of a stochastic process*. For now however it will be convenient to prepare the ground by studying operations on time trajectories in a completely deterministic framework.

We shall need the following definition: A *discrete-time signal* y is just a sequence of real or complex numbers indexed by a variable t which we shall call *(discrete) time*, running on the set of integers, $\mathbb{Z}$. Notation $y \equiv \{y(t)\,;\, t \in \mathbb{Z}\}$. Occasionally we shall need to deal with signals whose values $y(t)$ may be multiple numbers which for convenience are consider simultaneously say either $y(t) \in \mathbb{R}^m$ or $y(t) \in \mathbb{C}^m$, usually written as column vectors. Most of what we shall say will be for scalar signals but also applies to vector-valued ones.

Except perhaps for Economic or Econometric data, in Engineering and applied sciences discrete-time signals usually appear as periodically sampled versions of continuous time signals. The value of the signal is acquired by an acquisition device with a time interval T between successive time samples which is dictated by physical or technological constraints. We shall ignore this mechanism and denote the sampled version at time tT, say $\tilde{y}(tT)$, of a continuous time signal $\tilde{y}$, simply by the symbol $y(t)$ without mentioning T at all.

Discrete-time signals convey *information* about the temporal evolution of some physical phenomena, say a phone conversation coded and transmitted by broadcasting, the evolution of the flow rate of a river monitored in real time to predict possible overflow, the composition of the final product flowing out of a distillation column or of a chemical reactor etc. These signals may be related to diverse kinds of physical settings and the process of extracting this information by suitable algorithms is the scope of *Digital Signal Processing*. These algorithms are implemented through a series of mathematical operations which we shall briefly examine in this chapter. The reader should however be warned that most signals of interest in econometrics and technology are actually *ran-*

dom; since a deterministic signal, which is therefore a priori known, does not convey any information. In the next chapter we shall extend these operations to random signals.

Real or complex signals form a real (or complex) vector space on which the usual operations of sum and multiplication by a scalar take place in the obvious way. The **energy** of a signal is conventionally defined as the quantity

$$\sum_{t=-\infty}^{+\infty} |y(t)|^2$$

assuming that the sum converges. Finite energy signals form a Hilbert space which is denoted ℓ^2 or ℓ^2_m for m-dimensional signals. The (square) norm on these spaces is precisely the energy. These Hilbert spaces constitutes a suitable general framework to study discrete time signals. We refer the reader to the Appendix D for all concepts and notations regarding Hilbert spaces which shall be used in the following. In particular we should just mention here that there are other interesting spaces of discrete-time signals such as ℓ^1 and ℓ^∞ which are not Hilbert spaces but will nevertheless play a role in our study. We shall be mainly interested in *linear* operations on signals which will have the mathematical description of linear operators on ℓ^2-spaces.

A first important example of linear map on discrete time signals is the **translation operator $\boldsymbol{\sigma}$** also called the *shift*, which moves backwards by one time step the graph of the signal:

$$[\,\boldsymbol{\sigma} y\,](t) := y(t+1)\,, \qquad t \in \mathbb{Z} \tag{8.1.1}$$

By repeatedly applying the translation operator one can define powers $\boldsymbol{\sigma}^k$ which move backwards the graph by k steps. These powers have an inverse and form clearly a group; in fact within this group structure one can define the (backward or) *right translation operator* $\boldsymbol{\sigma}^{-1}$ which is the inverse of $\boldsymbol{\sigma}$ and moves the graph forward by one time step.
An important class of linear operators are those whose image does not depend on the particular date of application to the signal. They can formally be defined as follows;

Definition 8.1. *A* **time invariant operator** *A on ℓ^2 is a map which commutes with translation, that is*

$$A[\,\boldsymbol{\sigma}^k y\,] = \boldsymbol{\sigma}^k A y\,, \qquad y \in \ell^2 \tag{8.1.2}$$

for all $k \in \mathbb{Z}$.

Linear time invariant operations on discrete time signals will be the main object of study in this chapter. An important example is *convolution*.

Convolution is a bi-linear operation involving two functions, occurring in an enormous range of physical problems. In discrete time the convolution of two signals (or functions) h and u mapping $\mathbb{Z}$ to $\mathbb{R}$ is

$$(h * u)(t) := \sum_{-\infty}^{+\infty} h(t-s)u(s) = \sum_{-\infty}^{+\infty} h(s)u(t-s) \tag{8.1.3}$$

It will be shown below that this operation occurs when describing the response of a linear dynamical system to an *input function* u which is applied at an infinitely remote initial time ($t = -\infty$). In this case $y := h * u$ is called the *output*

of the system and h the *impulse response*. This name comes from the fact that when $u(t) = \delta(t)$ (the Kronecker delta signal, equal to 1 for $t = 0$ and zero otherwise) then the output of the system is just $y(t) = h(t)$.
Conditions for the existence of the convolution may be tricky, since a blow-up in u at infinity can be easily offset by sufficiently rapid decay in h. The question of existence thus may involve different conditions on h and u. Conditions on h are referred to as *stability* conditions on the system. A typical condition of interest in most contexts is the so called *Bounded-Input-Bounded-Output (BIBO)* stability, which corresponds to conditions on the mapping $u \to y$ to be a map of ℓ^∞ into himself.

Proposition 8.1. *The convolution relation* (8.1.3) *defines a bounded linear operator from ℓ^∞ into himself if and only if $h \in \ell^1$.*

If $h \in \ell^1$ then (8.1.3) *also maps ℓ^2 into itself and*

$$\|y\|_{\ell^2} \leq \|h\|_{\ell^1} \|u\|_{\ell^2} \tag{8.1.4}$$

This property is called ℓ^2-stability.

Bounded convolution operators exhaust, in a certain sense, the class of all bounded time-invariant operators on ℓ^2.

8.2 ▪ Stationary Time Series

Let t denote the discrete time variable taking values on the integer line. A (scalar) *time series* is just an ordered sequence of real numbers $y(t)\,;\, t = 0, 2, \ldots, N$ representing successive measurements of some variable. In science, econometrics and engineering, in order to construct reasonable predictors of this sequence one must confront the issue of modeling *serial correlation* as in virtually all situations of interest in science, econometrics and engineering the data cannot be modeled as independent (or uncorrelated) measurements. In particular, since the ordering of samples is of utmost importance, it is not appropriate to describe the data as an i.i.d. sequence as in classical Statistics. One needs instead to imagine it as a sample chunk of trajectory drawn from a **stochastic process** $\{\mathbf{y}(t)\}$. There is no space here to deal with stochastic processes in any depth; we shall just refer the reader to the notion of stochastic process reported in Sect. A.3 of the appendix. Mostly we shall refer to stationary processes because stationarity, most often to be understood as *wide sense stationarity*, implies that one can describe the data by *constant coefficient models* which makes them adapt to statistical estimation. In this section we want to discuss a very simple class of linear dynamical models with constant coefficients which are very often used in applied fields.

When there is serial correlation, the current variable of a stochastic process $\mathbf{y}(t)$, is in particular correlated with its past $\mathbf{y}(t-1)$, $\mathbf{y}(t-2)$, $\mathbf{y}(t-3), \ldots$ and a (dynamical) model should describe the influence of this past history on the current observed variable. In engineering or econometric applications there often are external forcing *exogenous variable*, that is *inputs*, denoted by the symbol $\mathbf{u}$, which influence the temporal behaviour of $\mathbf{y}(t)$ and one wants to describe how $\mathbf{y}(t)$ changes in time both as a consequence of the correlation with its own past but also as a consequences of time-varying exogenous variables. We won't

care much about modeling $\mathbf{u}$ itself since external forces are often assumed to be observed exactly.

The simplest generalization of linear regression models to describe a serially correlated time series is a linear relation of the following form,

$$\mathbf{y}(t) \;=\; \sum_{k=1}^{n} a_k \mathbf{y}(t-k) + \sum_{k=1}^{m} b_k \mathbf{u}(t-k) + \mathbf{w}(t)\,, \tag{8.2.1}$$

where $\mathbf{w} := \{\mathbf{w}(t),\, t \in \mathbb{Z}\}$ is a process of random errors which will here be assumed i.i.d or, more generally, just uncorrelated. This is called an **Auto-Regressive model with eXogenous input** and is denoted by the acronym ARX. If there is no $\mathbf{u}$ then the model is called (purely) **Auto-Regressive** and is referred to by the acronym AR. There are also more general models, called ARMA, ARMAX, which involve a *moving average* input noise component made of a linear combination of delayed noise samples such as

$$\mathbf{y}(t) \;=\; \sum_{k=1}^{n} a_k \mathbf{y}(t-k) + \sum_{k=1}^{m} b_k \mathbf{u}(t-k) + \sum_{k=0}^{r} c_k \mathbf{w}(t-k)\,, \tag{8.2.2}$$

whose study however requires more sophisticated tools than what we subsume in this course and we shall have to refer to the literature, see e.g. **[?]**.

An ARX model depends on $p := n + m$ unknown parameters which will be written as a column vector:

$$\theta := \begin{bmatrix} a_1 & \dots & a_n & b_1 & \dots & b_m \end{bmatrix}^\top$$

and on the unknown noise variance σ^2. It can be formally written in *psedo-linear regression* form as

$$\mathbf{y}(t) = \boldsymbol{\varphi}(t)^\top \theta + \mathbf{w}(t)\,, \qquad t \in \mathbb{Z} \tag{8.2.3}$$

where

$$\boldsymbol{\varphi}(t)^\top = \begin{bmatrix} \mathbf{y}(t-1) & \dots & \mathbf{y}(t-n) & \mathbf{u}(t-1) & \dots & \mathbf{u}(t-m) \end{bmatrix}$$

(so that $\boldsymbol{\varphi}(t)$ is a column vector depending on past data). Note that there is a very important difference with the classical linear model (2.2.7) namely now **the coefficient vectors $\boldsymbol{\varphi}(t)$ of the model depend on the (past) input-output variables**. For this reason we have chosen a different symbol than $s(t)$.

Assume we have a sequence of training data $\{y(t)\,,\, u(t)\,\}$ denoted

$$y^N := \{y(t)\,;\; t = t_0, t_0+1, \dots, N\}\,, \qquad u^N := \{u(t)\,;\; t = t_0, t_0+1, \dots, \dots, N\}\,,$$

which we want to describe by an ARX model. The data will always be assumed to have been suitably pre-processed e.g. by subtracting the sample mean so as to be compatible with zero-mean and stationarity. Imposing the ARX structure to these observed data we obtain a system of linear relations which, rewritten in vector form look like

$$\mathbf{y} = \Phi_N\, \theta + \mathbf{w} \tag{8.2.4}$$

where the random vectors $\mathbf{y}$ and $\mathbf{w}$ to have components $\mathbf{y}(t)$ and $\mathbf{w}(t)$ indexed by $t = 1, 2, \dots, N$ and Φ_N is an $N \times p$ matrix of past data of the form:

$$\Phi_N := \begin{bmatrix} \varphi(1)^\top \\ \vdots \\ \varphi(N)^\top \end{bmatrix}\,,$$

Assuming the initial time t_0 is far enough, we can fill in Φ_N with the available data so as to describe the output from time $t = 1$ to $t = N$. Often we do not know the probability distribution of the error process. We may assume it is Gaussian but, because of the dependence on the data of Φ_N we cannot implement a simple Maximum Likelihood procedure to estimate the parameter θ. The Gaussian assumption is therefore not so useful. We shall try to do by just assuming that $\mathbf{w}$ is an i.i.d. process.

The function of the past data

$$\hat{\mathbf{y}}_\theta(t \mid t-1) = \boldsymbol{\varphi}(t)^\top \theta \tag{8.2.5}$$

is called the (one step ahead) **predictor function** associated to the model. Note that the predictor function is a linear function of θ but now is also a function of the previous $n + m$ past samples of the joint data process.

PEM Identification of Time Series

To estimate the parameter θ of the model (8.2.4) from the observed data (y^N, u^N) we shall use the empirical **Prediction Error Minimization (PEM)** approach. This is a general estimation method, essentially the same as the *Empirical Training Error* or average sample error, minimization in Machine Learning, see [44, p. 221].

The procedure is quite general and applies to any dynamical model. It is based on the following steps:

1. For a generic value of θ, construct a *predictor* of the next output, say $y(t)$, based on the training data up to time $t-1$. For each fixed θ the predictor is a deterministic function of the past data, denoted $\hat{y}_\theta(t \mid t-1)$.
   For analysis purpose we may consider $\hat{y}_\theta(t \mid t-1)$ as a function (of θ) and of the past **random** observed data and denote it $\hat{\mathbf{y}}_\theta(t \mid t-1)$.

2. Form the *empirical prediction errors* incurred by using θ as a current parameter value:

   $$\varepsilon_\theta(t) := y(t) - \hat{y}_\theta(t \mid t-1)\,; \qquad t = 1, 2, \ldots, N\,.$$

   These errors are *numbers* but may also be interpreted as sample values of a random variable, written $\boldsymbol{\varepsilon}_\theta(t)$.

3. Minimize with respect to θ the sample **average (squared) prediction error**

   $$V_N(\theta) := \frac{1}{N}\sum_{t=1}^{N} \varepsilon_\theta(t)^2 \tag{8.2.6}$$

   or, more generally, one may choose any convex function of $\varepsilon_\theta(t)\,; t = 1, 2, \ldots, N$.

We may introduce a *discount factor* for past errors that is a positive sequence $q(N, t)$, and form

$$V_N(\theta) := \frac{1}{N}\sum_{t=1}^{N} q(N,t)\varepsilon_\theta(t)^2 \qquad q(t, N) > 0$$

For small N, the function $q(N,t)$ should give small weight to errors incurred at the beginning. One designs the weighting function so that for $N \to \infty$, $q(N,t) \to 1$.

The Minimal Prediction Error (PEM) parameter estimate

$$\hat{\theta}_N := \text{Arg} \min_{\theta} \; V_N(\theta)$$

becomes a function of the data $(y^N\,,\,u^N)$. As we shall see, for the ARX model it can be computed explicitly.
Next define as an estimate of $\sigma^2 = \text{var}\,\{\mathbf{w}(t)\}$, the *residual quadratic error*,

$$\hat{\sigma}_N^2 := \; V_N(\hat{\theta}_N) \tag{8.2.7}$$

where V_N is defined above.

ARX Identification and Least Squares

In the following it will be convenient to use vector notations. For an ARX model defined by a generic parameter vector θ, the N-dimensional vector of predictors is a linear function of the parameter; hence the predictor and prediction error vectors have the form

$$\hat{\mathbf{y}}_\theta = \Phi_N\,\theta\,, \qquad \boldsymbol{\varepsilon}_\theta = \mathbf{y} - \Phi_N\,\theta\;.$$

Hence, when no weighting is present, $V_N(\theta)$ is just the squared Euclidean norm of $\boldsymbol{\varepsilon}_\theta$,

$$V_N(\theta) = \frac{1}{N}\,\|\mathbf{y} - \Phi_N\,\theta\|^2\,.$$

The PEM estimation principle leads again to the solution of a Least Squares Problem. In our case either $Q = I_N$ ($N \times N$ identity matrix) or is a diagonal matrix with entries $q(N,k)\,;\; k = 1, \ldots N$. In the first case the PEM estimator of θ is just

$$\hat{\theta}_N = \left[\Phi_N^\top \Phi_N\right]^{-1}\,\Phi_N^\top\,\mathbf{y}$$

which can also be written in the form

$$\hat{\theta}_N = \left[\sum_{t=1}^{N} \varphi(t)\varphi(t)^\top\right]^{-1}\,\sum_{t=1}^{N} \varphi(t)\mathbf{y}(t)\,. \tag{8.2.8}$$

where we assume that the inverse exists for suitably large N.
• Note that $\hat{\theta}_N$ is a **non linear function of the observed data**. One may ask what are the statistical properties of this estimator.
Actually we don't even know when it may be unbiased. Even if $\mathbf{y}$ and $\mathbf{u}$ were Gaussian, the pdf of $\hat{\theta}_N$ for finite sample size is impossible to compute. One can only try to see what happens for $N \to \infty$. This we shall attempt to do next.

8.3 ▪ Strong consistency of the least squares AR estimator

Naturally it is now very difficult, if not impossible, to say anything about the statistical properties of the estimate (8.2.8) for a finite sample size N. Under certain circumstances however one can carry on an asymptotic analysis for $N \to \infty$

and prove statements regarding the consistency and asymptotic normality of the method. Here we shall only give a short preview for the case of no exogenous input ($\mathbf{u} \equiv 0$).

Theorem 8.1. *Assume there is a true AR model describing the data having the same order n as the candidate AR model and true parameter $\theta_0 := \begin{bmatrix} a_{0,1} & \dots & a_{0,n} \end{bmatrix}^\top$. Assume also that the true model is* **causal** *that is*

$$\mathbb{E}_{\theta_0} \mathbf{y}(s)\, \mathbf{w}(t) = 0\,; \qquad \forall t > s \in \mathbb{Z}\,; \tag{8.3.1}$$

and that $\mathbb{E}_{\theta_0} \boldsymbol{\varphi}(t) \boldsymbol{\varphi}(t)^\top > 0$; then

$$\lim_{N \to \infty} \hat{\boldsymbol{\theta}}_N = \theta_0$$

with probability one.

Proof. Rewrite $\hat{\boldsymbol{\theta}}_N$ as

$$\hat{\boldsymbol{\theta}}_N = \left[\frac{1}{N} \sum_{t=1}^{N} \varphi(t) \varphi(t)^\top \right]^{-1} \frac{1}{N} \sum_{t=1}^{N} \varphi(t) \mathbf{y}(t)\,; \tag{8.3.2}$$

and substitute $\mathbf{y}(t) = \varphi(t)^\top \theta_0 + \mathbf{w}(t)$ (true model). Then define the sample covariance matrix of $\boldsymbol{\varphi}(t)$

$$\hat{\boldsymbol{\Sigma}}_N := \frac{1}{N} \sum_{t=1}^{N} \boldsymbol{\varphi}(t) \boldsymbol{\varphi}(t)^\top \ \in \mathbb{R}^{n \times n}\,. \tag{8.3.3}$$

Next we shall need the following fact.

Lemma 8.1. *Assume* (8.3.1) *and $\mathbf{u} \equiv 0$; then if in the true model, the i.i.d. process $\{\mathbf{w}\}$ is not zero, $\{\mathbf{y}\}$ is ergodic and $\hat{\boldsymbol{\Sigma}}_N$ converges almost surely for $N \to \infty$ to the positive semidefinite covariance matrix*

$$\Sigma_n := \mathbb{E}_{\theta_0} \{ \begin{bmatrix} \mathbf{y}(t-1) \\ \dots \\ \mathbf{y}(t-n) \end{bmatrix} \begin{bmatrix} \mathbf{y}(t-1) & \dots & \mathbf{y}(t-n) \end{bmatrix} \}\,. \tag{8.3.4}$$

Under the stated assumptions, this matrix is in fact positive definite.

Proof. The ergodicity follows from the consequence (A.3.3) of Corollary A.2 but to use this result we shall need to prove that $\mathbf{y}(t)$ admits such a convolution representation. This follows from the fact that (8.3.1) implies that $\mathbf{w}$ is the stationary innovation of $\mathbf{y}$. We shall prove all of this later on. The positivity of the Toeplitz matrix Σ_n can be seen to follow from the following argument.
By Proposition 8.3, if Σ_n is singular then so must be Σ_{n+1} and there must be some nonzero $c \in \mathbb{R}^{n+1}$ such that $c^\top \Sigma_{n+1} c = 0$. This is the same as $\sum_{k=0}^{n} c_k \mathbf{y}(t-k) = 0$ (the zero random variable) which can hold true only in case $\mathbf{y}(t)$ satisfies the deterministic recursion $\sum_{k=0}^{n} c_k \mathbf{y}(t-k) = 0$, where we can without loss of generality assume $c_0 \neq 0$ so the process should satisfy an AR recursion where the i.i.d. input is absent, or, $\mathbf{y}$ should be a purely deterministic process. ∎

Now just go to the limit in formula (8.3.2), that is compute

$$\lim_{N\to\infty}\left[\frac{1}{N}\sum_{t=1}^{N}\boldsymbol{\varphi}(t)\boldsymbol{\varphi}(t)^{\top}\right]^{-1}\frac{1}{N}\sum_{t=1}^{N}\boldsymbol{\varphi}(t)\,\left(\boldsymbol{\varphi}(t)^{\top}\,\theta_0+\mathbf{w}(t)\right)$$

to get, by gity and in virtue if the two main assumptions

$$\lim_{N\to\infty}\hat{\boldsymbol{\theta}}_N=\Sigma_0^{-1}\Sigma_0\;\theta_0=\theta_0$$

since $\Sigma_0:=\mathbb{E}_{\theta_0}\boldsymbol{\varphi}(t)\boldsymbol{\varphi}(t)^{\top}>0$. This ends a formal proof of consistency. □

Of course the reason why the two main assumptions should hold needs to be investigated. The first (causality) is related to the notion of **innovation process** of $\mathbf{y}$ while, as we shall see, the non singularity of $\mathbb{E}_{\theta_0}\boldsymbol{\varphi}(t)\boldsymbol{\varphi}(t)^{\top}>0$ will hold automatically for AR models but for the ARX model (8.2.1) will require a discussion of the allowable input class.

8.4 ▪ The Innovation of a stationary random processes

One of the main motivations for time-series model identification is *prediction*. Prediction turns out to be an easy task for models with constant coefficients which are necessarily models describing stationary processes defined on the whole time axis. See the next section 8.5 for an analysis of stationary processes defined on a subinterval of $\mathbb{Z}$. There it is seen that in this situation modeling and prediction of a stationary process requires time-varying parameters.

Let $\mathbf{H}(\mathbf{y}^t)$ be the Hilbert space of random variables obtained as the closure with respect to the variance norm of the vector space $\tilde{\mathbf{H}}(\mathbf{y}^t)$ of linear statistics of the past history of the process at time t

$$\tilde{\mathbf{H}}(\mathbf{y}^t):=\{\sum_k a_k\mathbf{y}(t_k)\,;\,t_k\le t\,,a_k\in\mathbb{R}\,\}$$

where the sums are all finite. We are interested in the linear minimum variance predictor of a future random variable of the process, say $\mathbf{y}(t+k),\ k>0$, given the whole past history. This linear predictor is by definition the orthogonal projection onto the whole past history of the process:

$$\hat{\mathbf{y}}(t+k\mid t)=\hat{\mathbb{E}}\,[\mathbf{y}(t+k)\mid\mathbf{H}(\mathbf{y}^t)\,] \tag{8.4.1}$$

which is itself a random process, say $\{\hat{\mathbf{y}}_k(t)\}$.

Proposition 8.2. *If the process* $\mathbf{y}$ *is stationary, then the predictor process* (8.4.1) *is also stationary.*

Intuitively, this means that the predictor process can also be described by a model with constant coefficients. This will be true only if the predictor is based on the *infinite* past history. In Section 8.5 we show that the innovation (i.e the one step prediction error) of a stationary process, based on the past time interval $[\,0,\,t\,]$ cannot be stationary. Therefore this must also be true for the predictor.

In general the covariance function $\sigma(\tau)$ of a stationary process is supported on the whole time line so that the correlation of any two variables $\mathbf{y}(t)$ and $\mathbf{y}(s)$

will be non zero no matter of how far apart the time instants t and s. In particular, it is a fact that if $n > 0$ the output at time t of an AR model is correlated with the *infinite* past history of the process $\mathbf{y}$ and also with the *infinite* past history of the input process. We shall see this very clearly when we shall be studying difference equations in the next section. Therefore whenever talking about prediction we shall mean "prediction given the infinite past history".

Assume there is a true AR model of orders n and true parameter θ_0 describing the data. Then the causality condition (8.3.1) implies that $\mathbf{w}(t)$ must be the one step ahead prediction error of $\mathbf{y}(t)$ given the **finite** chunk of joint past history of the processes $\mathbf{y}$ from time $t-n$ to $t-1$. This fact follows easily from the orthogonal projection lemma, given that in the decomposition

$$\mathbf{y}(t) = \mathbf{w}(t) + \left[\sum_{k=1}^{n} a_k \mathbf{y}(t-k)\right]$$

$\mathbf{w}(t)$ is orthogonal to the term between square brackets. In this section we shall argue that if (8.3.1) holds, the orthogonality actually holds also with respect to the infinite joint past of the process $\mathbf{y}$, so that the term between brackets is actually the minimum variance predictor of $\mathbf{y}(t)$ *given the infinite joint past* of $\mathbf{y}$ from time $-\infty$ to $t-1$. In these circumstances $\mathbf{w}(t)$ is called the *stationary innovation process* of $\mathbf{y}$.

For easy reference we shall recall that the Gram-Schmidt procedure already discussed in (5.11.15) can be applied sequentially to an *infinite sequence* of random observations, that is, to a *stochastic process* $\{\mathbf{y}(t)\}$. The essential condition however to make the procedura applicable, is that there should be an *initial time* (conventionally assumed to be $t = 0$) to start the recursion. Theorem 5.7 and the relative corollary continue to hold also in this situation, provided of course that the $(t+1) \times (t+1)$ variance matrix $\Sigma_t := \mathbb{E}\,[\mathbf{y}_0^t (\mathbf{y}_0^t)^\top]$ stays positive definite for all t. Under this condition also the Cholesky factorization algorithm continues to hold irrespective of the fact that now $m \to \infty$. In this setting however the resulting innovation process *cannot be stationary*. This is a rather unpleasant fact since it does imply that the predictor processes cannot be stationary either and hence cannot be described by a constant coefficient algorithm (or model).

Recall that, based on the causal relation (5.11.17), one can give a *dynamic* interpretation of the Cholesky factorization procedure. Isolate the equation corresponding to the $t+1$-st row of the lower triangular matrix L and write it as a representation of $\mathbf{y}(t)$ in function of its normalized innovation process $\{\boldsymbol{\varepsilon}(t)\}$:

$$\mathbf{y}(t) = \ell_t\, \boldsymbol{\varepsilon}_0^t = \sum_{0}^{t} \ell(t,s)\, \boldsymbol{\varepsilon}(s)\,, \qquad t = 0, 1, 2, \dots \tag{8.4.2}$$

where $\ell_t = [\,\ell(t,0), \ell(t,1), \dots, \ell(t,t)]$ is the $t+1$-st row of L. This relation represents $\mathbf{y}(t)$ as a *time-dependent* linear function of the past innovation samples. In engineering language one could say that $\mathbf{y}(t)$ is the output of a *time-varying convolution filter having impulse response* $\ell(t,\cdot)$ and as input signal the normalized *white noise* $\{\boldsymbol{\varepsilon}(t)\}$. This is the prototype of the *innovation representation of a random process* which we shall have to generalize later on for stationary processes defined on the whole time axis. Representations of this kind are of fundamental importance for solving filtering and dynamic estimation problems. We stress that $\mathbf{y}(t)$ depends only on the past and present history, $\boldsymbol{\varepsilon}_0^t$, of the innovation process but not on its future values.

Conversely, we can express the innovation at time t as a causal linear function of the present and past variables of the observation vector. To clarify this point, just isolate the $t+1$-th row of the inverse L^{-1}. Since this inverse is also lower triangular we obtain a relation of similar structure

$$\varepsilon(t) = \boldsymbol{\gamma}_t \, \mathbf{y}_0^t = \sum_0^t \, \gamma(t,s) \, \mathbf{y}(s) \,, \qquad t = 0, 1, 2, \dots \tag{8.4.3}$$

where $\boldsymbol{\gamma}_t = \begin{bmatrix} \gamma(t,0) & \gamma(t,1) & \dots & \gamma(t,t) \end{bmatrix}$ is the non-zero subvector of the $t+1$-st row of L^{-1}. In a sense this is the the inverse of the convolution filter (8.4.2) which now transforms, by a causal operation, the process $\{\mathbf{y}(t)\}$ into the white noise $\{\varepsilon(t)\}$. In this sense one says that the process and its innovation are **causally equivalent**. The two signals carry the same information.

8.5 ▪ Innovations of stationary processes on $\mathbb{Z}_+$

As we have seen, in spite of stationarity, if a process is indexed on the half line $\mathbb{Z}_+$, its innovation representation will be *time varying*, that is, the coefficients of the linear representation map will not be constant in time. The unnormalized innovation itself will have a time varying variance and hence be non-stationary. Below we shall elaborate on this fact.

Consider a scalar zero-mean (weakly) stationary process $\{\mathbf{y}(t)\}$ defined on the time set $t = 0, 1, 2, \dots$ with a covariance function $\sigma(\tau) = \mathbb{E}\, \mathbf{y}(t+\tau)\, \mathbf{y}(t)$ which is a positive definite symmetric function of $\tau \in \mathbb{Z}$. We shall assume that all principal $n \times n$ submatrices, denoted Σ_n, of the infinite Toepltz matrix

$$\Sigma = \begin{bmatrix} \sigma(0) & \sigma_(1) & \dots & \sigma(n) & \dots \\ \sigma(1) & \sigma(0) & \sigma_(1) & \dots & \dots \\ & \sigma(1) & \sigma(0) & \sigma_(1) & \\ \sigma(n) & \dots & \dots & \ddots & \end{bmatrix} \quad , \tag{8.5.1}$$

are non singular.

Note that Σ_{n+1} can be obtained by bordering Σ_n according to the following scheme

$$\Sigma_{n+1} = \begin{bmatrix} \sigma(0) & \sigma_n \\ \sigma_n^\top & \Sigma_n \end{bmatrix} \quad , \qquad n = 1, 2, \dots \tag{8.5.2}$$

where $\sigma_n = [\sigma(1), \dots, \sigma(n)]$. Consider the *memory* n one-step ahead predictor of $\mathbf{y}(t)$ given the previous n variables of the process $\mathbf{y}(t-1), \dots, \mathbf{y}(t-n)$, which we shall denote

$$\hat{\mathbf{y}}_n(t) := \hat{\mathbb{E}}\, [\mathbf{y}(t) \mid \mathbf{y}(t-1), \dots, \mathbf{y}(t-n)\,]$$

and denote by λ_n^2 the variance of the prediction error $\mathbf{e}_n(t) := \mathbf{y}(t) - \hat{\mathbf{y}}_n(t)$. The process $\{\mathbf{e}_n(t)\}$ is called the *memory n innovation* of $\{\mathbf{y}(t)\}$.
Since by the orthogonality principle we must have $\mathbf{e}_n(t) \perp \{\mathbf{y}(t-1), \dots, \mathbf{y}(t-n)\}$, the variance matrix of the random vector $\begin{bmatrix} \mathbf{e}_n(t) & \mathbf{y}(t-1) & \dots & \mathbf{y}(t-n) \end{bmatrix}^\top$ must clearly be block-diagonal namely,

$$\begin{bmatrix} \lambda_n^2 & 0 \\ 0 & \Sigma_n \end{bmatrix} \tag{8.5.3}$$

However this matrix must be congruent to Σ_{n+1} since, repeating the argument used in the proof of Theorem 5.8, letting T_n be the matrix transforming the vector $\begin{bmatrix}\mathbf{y}(t) & \mathbf{y}(t-1) & \dots & \mathbf{y}(t-n)\end{bmatrix}^\top$ into $\begin{bmatrix}\mathbf{e}_n(t) & \mathbf{y}(t-1) & \dots & \mathbf{y}(t-n)\end{bmatrix}^\top$, namely

$$T_n = \begin{bmatrix} 1 & -\sigma_n\, \Sigma_n^{-1} \\ 0 & I \end{bmatrix} \quad . \tag{8.5.4}$$

it must hold that

$$\begin{bmatrix} \lambda_n^2 & 0 \\ 0 & \Sigma_n \end{bmatrix} = T_n \Sigma_{n+1} T_n^\top \tag{8.5.5}$$

Note however that $\det\, T_n = 1$ and hence the two matrices must have the same determinant, so the block-diagonalization (8.5.5) directly leads to a remarkable formula of prediction theory for stationary processes namely

$$\operatorname{var} \mathbf{e}_n(t) = \lambda_n^2 = \frac{\det\, \Sigma_{n+1}}{\det\, \Sigma_n} \quad . \tag{8.5.6}$$

From this formula one can draw several consequences among which a criterion for non-singularity of the sequence of covariance matrices $\Sigma_n\,;\, n = 1, 2, \dots$ whose proof follows immediately from (8.5.5).

Proposition 8.3. *If for some natural $n > 0$, $\lambda_n^2 = 0$ then also $\lambda_{n+k}^2 = 0$ for all $k \geq 0$. Assume $\sigma_0 > 0$, then Σ_n is non-singular if and only if all $\lambda_k^2\,;\, k = 1, 2 \dots, n$ are non zero.*

Constant coefficient models will appear only for stationary processes defined on the whole discrete time line $\mathbb{Z}$. Quite unfortunately in this case there is no finite initial time and the Gram-Schmidt idea is of no use. We shall have to use a different approach.

Example 8.1 (Innovation representation of a stationary MA process).
Consider a process $\{\mathbf{y}(t)\}$ described by the "moving average" (MA) model

$$\mathbf{y}(t) = \mathbf{w}(t) + a\, \mathbf{w}(t-1) \quad , \qquad t = 1, 2, \dots$$

where $\{\mathbf{w}(t)\}$ is zero-mean white noise of variance σ^2. We want to check if this noise process is the innovation of $\mathbf{y}$.

The covariance function of $\{\mathbf{y}(t)\}$ is readily computed as,

$$\sigma(\tau) = \mathbb{E}\; \mathbf{y}(t+\tau)\, \mathbf{y}(t) = \begin{cases} \sigma^2(1+a^2) & \tau = 0 \\ \sigma^2\, a & \tau = \pm 1 \\ 0 & |\tau| > 1 \end{cases}$$

whence the variance matrix $\Sigma_t := \mathbb{E}\; \mathbf{y}^t (\mathbf{y}^t)^\top$, $t = 1, 2, \dots$, has the tridiagonal structure

$$\Sigma_t = \sigma^2 \begin{bmatrix} 1+a^2 & a & 0 & \dots & 0 \\ a & 1+a^2 & \ddots & & \\ 0 & & \ddots & \ddots & \\ \vdots & & & \ddots & a \\ 0 & & a & & 1+a^2 \end{bmatrix}$$

It is not difficult to see that the normalized determinant $d_t := \frac{1}{\sigma^{2t}} \det \Sigma_t$ satisfies the linear difference equation

$$d_{t+1} = (1 + a^2)\, d_t - a^2\, d_{t-1} \quad , \qquad t \geq 1$$

which can be solved to obtain

$$d_t = \sum_0^t (a^2)^k \quad .$$

Clearly, d_t is non-zero for all t and we can hence apply to Σ_t the Cholesky algorithm to get the triangular factors $L_t = [\ell(i,j)]$ which define the innovation representation of $\{\mathbf{y}(t)\}$.

It is not hard to see that L_t should have a bi-diagonal structure like

$$L_t = \begin{bmatrix} \ell(1,1) & 0 & \dots & 0 \\ \ell(2,1) & \ell(2,2) & & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & \ell(t,t-1) & \ell(t,t) \end{bmatrix}$$

where the entries $\ell(t,t-1)$ and $\ell(t,t)$ are given by the formulas (5.11.22) and (5.11.23) imposing that $\ell_{t,s} := \ell(t,s) = 0$ for $s < t-1$. This leads to the following recursions

$$\ell(t,t) = \Big[\sigma(0) - \ell(t,t-1)^2\Big]^{\frac{1}{2}}$$

$$\ell(t,t-1) = \frac{1}{\ell(t-1,t-1)}\, \sigma(1)$$

which can be solved by iteration on the index $t = 2, 3, \ldots$, starting from the initial condition

$$\ell(1,1) = +\sqrt{\sigma^2(1+a^2)} \quad .$$

It is probably easier to recall the final argument in the proof of Theorem **??** where the diagonal terms $\ell^2(t,t)$ are identified with the one-step prediction error variance of $\mathbf{y}(t)$ based on the $t-1$ preceding observations, $\mathbf{y}^{t-1}$, of the process $\{\mathbf{y}(t)\}$. Using formula (8.5.6) derived in the previous section, one gets

$$\ell^2(t,t) = \frac{\det\, \Sigma_t}{\det\, \Sigma_{t-1}} = \sigma^2\, \frac{d_t}{d_{t-1}}$$

and

$$\ell(t,t-1) = \sigma a \sqrt{\frac{d_{t-2}}{d_{t-1}}} \quad .$$

Even if the process is stationary, these expressions turn out to be *time dependent*, implying that also the innovation representation

$$\mathbf{y}(t) = \ell(t,t)\, \boldsymbol{\varepsilon}(t) + \ell(t,t-1)\, \boldsymbol{\varepsilon}(t-1) \tag{8.5.7}$$

must be time-varying. We shall show later on however that the coefficients of the model (8.5.7) tend to become constants as $t \to \infty$. For this particular example, when $|a| < 1$, the ratio d_t/d_{t-1} converges to 1; when $|a| > 1$, one has

$$\frac{d_t}{d_{t-1}} = \frac{a^{2t}}{a^{2(t-1)}} \, \frac{\sum_0^t a^{-2k}}{\sum_0^{t-1} a^{-2k}} \longrightarrow a^2$$

while when $|a| = 1$, $d_t/d_{t-1} = (t+1)/t = 1 + 1/t$ which also converges to 1 although at a slower rate. In conclusion, for $t \to \infty$ the parameters of the model (8.5.7) converge, respectively, to

$$\ell(t,t) \to \ell_0 = \begin{cases} \sigma & \text{when } |a| \le 1 \\ |a|\,\sigma & \text{when } |a| > 1 \end{cases}$$

$$\ell(t,t-1) \to \ell_1 = \begin{cases} a\sigma & \text{when } |a| \le 1 \\ \dfrac{a}{|a|}\,\sigma & \text{when } |a| > 1 \end{cases}$$

and the innovation model becomes *asymptotically time-invariant*,

$$\mathbf{y}(t) = \ell_0 \, \boldsymbol{\varepsilon}(t) + \ell_1 \, \boldsymbol{\varepsilon}(t-1) \quad .$$

This kind of asymptotic behavior of the impulse response function $\ell(t,\cdot)$ is the rule for a wide class of stationary processes. ♢

8.6 ▪ A glimpse on Linear Difference Equations

Linear difference equations arise as deterministic mathematical models of many physical or economic systems. In particular, ARX models are just linear difference equations (with constant coefficients) with two kinds of inputs. In this section we shall review some basic facts about these models. For convenience a DE will be written as

$$y(t) + \sum_{k=1}^{n} a_k \, y(t-k) = f(t)\,, \qquad t \in \mathbb{Z} \tag{8.6.1}$$

or, equivalently as

$$y(t+n) + \sum_{k=1}^{n} a_k \, y(t+n-k) = g(t), \qquad t \in \mathbb{Z} \tag{8.6.2}$$

where $f(t)$ or $g(t) = f(t+n)$ are exogenous signals. One should keep in mind that a solution is a *sequence of real numbers indexed by* $t \in \mathbb{Z}$. To find a solution first look at the homogeneous case where $f(t) = 0$. One tries with a simple exponential function $y(t) = \lambda^t$ which leads to the algebraic equation

$$\lambda^{t+n} + \sum_{k=1}^{n} a_k \, \lambda^{t+n-k} = 0\,.$$

Assuming $\lambda \neq 0$ we can collect λ^t and end with an algebraic equation of degree n

$$\lambda^n + \sum_{k=1}^{n} a_k \, \lambda^{n-k} = 0$$

which is called the **characteristic equation** of the system. It has n complex solutions λ_k ; $k = 1, 2, \ldots, n$ not necessarily distinct. It is easy to check that the family of solution sequences of a linear homogeneous difference equation of order n is a linear vector space. By a linear algebra argument one can show that this vector space must have dimension n and hence the equation has exactly n linearly independent solution sequences. Therefore, assuming that all roots are distinct, any solution must be a linear combination of the n exponentials. In this case the general solution turns out to be

$$y(t) = \sum_{k=1}^{n} c_k \, \lambda_k^t$$

where the coefficients c_k can be determined by imposing n **initial conditions** for example the values $y(0) = y_0$, $y(1) = y_1, \ldots y(n-1) = y_{n-1}$, supposed given. In case of distinct roots this leads to a linear system

$$\begin{bmatrix} 1 & \ldots & 1 \\ \lambda_1 & \ldots & \lambda_n \\ \ldots & \ldots & \ldots \\ \lambda_1^{n-1} & \ldots & \lambda_n^{n-1} \end{bmatrix} \begin{bmatrix} c_1 \\ \ldots \\ c_n \end{bmatrix} = \begin{bmatrix} y_0 \\ \ldots \\ y_{n-1} \end{bmatrix}$$

where the sytem matrix is the Vandermonde matrix of the roots λ_k ; $k = 1, 2, \ldots, n$ which is non singular since they are all different.

Note that in general the λ_k may be complex numbers and in this case they must occur in complex conjugate pairs which may lead the solution to have oscillatory behaviour.

In case of multiple roots the general solution has a more complicated form involving linear combination of sequences of the form $t^{m_k} \lambda_k^t$ where m_k depends on the multiplicity of the root λ_k; see e.g. [**?**].

Non homogeneous equation and Convolution

Note that for zero initial conditions the solution sequence $y(\cdot)$ of a DE depends linearly on the input $f(\cdot)$. This means that if $y(\cdot)$ is a solution corresponding to input $f(\cdot)$ and $z(\cdot)$ a solution corresponding to input $g(\cdot)$ then $\alpha y(\cdot) + \beta z(\cdot)$ will be the solution corresponding to the input $\alpha f(\cdot) + \beta g(\cdot)$ and hence the input-output map defined by a DE is a **linear transformation** mapping the real vector space of input sequences $f(\cdot)$ into a real vector space of output sequences $y(\cdot)$. This is a simplest example of a **linear dynamical system**. Suppose now that you want to solve

$$y(t) + \sum_{k=1}^{n} a_k \, y(t-k) = \delta(t) \, ; \tag{8.6.3}$$

where the initial conditions for negative times are all zero and $\delta(t)$ is equal to 1 for t=0 and zero otherwise. The function δ is called the *elementary* or *unit impulse function*. One can work out an equivalent **homogeneous equation** to the

non-homogemeous (8.6.3) by introducing some "fake" initial conditions $\{y(k) = y_k\,;\, k = 0, 1, \ldots, n-1\}$ for positive times, which solve the system of equations obtained by writing (8.6.3) at times $t = 0, 1, \ldots, n-1$, namely

$$\begin{aligned} y_0 + \sum_{k=1}^{n} a_k\, y_{-k} &= \delta(0) = 1 \\ y_1 + \sum_{k=1}^{n} a_k\, y_{1-k} &= 0 \\ \ldots &= \ldots \\ y_{n-1} + \sum_{k=1}^{n} a_k\, y_{n-k-1} &= 0 \end{aligned}$$

This system can be solved by successive substitutions since the first equation yields $y_0 = 1$, the second $y_1 + a_1\, y_0 = 0$ etc. These induced n initial conditions determine uniquely the solution of (8.6.3) for $t \geq 0$. Let's denote this solution by $h(t)$; it is called the **impulse response of the system**. As we have just seen the impulse response of a difference equation of the form (8.6.3) is a particular solution of the homogeneous eqiuation. It must have a normalized first term; i.e. $h(0) = 1$.

Remarks 8.1. Note that imposing that the values $y(t)$ for n negative times should all be zero is equivalent to imposing that the impulse response sequence $h(t) = \sum_{k=1}^{n} \alpha_k\ \lambda_k^t$ should be identically zero *for all negative times* $t < 0$, while by construction $h(t)$ turns out to be non-zero for positive times. A sequence with this property is called **causal**. Note also that we could have chosen a dual set of initial conditions by imposing that the values $y(k)$ for the first n *positive* times should be all zero in which case we would have obtained an **anticausal** impulse response function which would then result to be identically zero for positive times.

We can now solve the equation for an arbitrary input $f(t)$. Since any input function f can be expressed as a (possibly infinite) linear combination of impulse functions located at all times $t = k$, that is

$$f(t) = \sum_{k=-\infty}^{+\infty} f(k)\, \delta(t-k)$$

by virtue of linearity, the response of the system can be written as a sum of infinitely many impulse responses to the impulses $\delta(t-k)$'s each located at times $t = k$ and weighted by amplitude $f(k)$. By time invariance the solution of (8.6.3) to a shifted impulse $\delta(t-k)$ is $h(t-k)$ and hence linearity leads to the (formal) **convolution representation**

$$y(t) = \sum_{k=-\infty}^{+\infty} f(k)\, h(t-k)\,; \qquad \text{or, equivalently} \qquad y(t) = \sum_{k=-\infty}^{+\infty} h(k)\ f(t-k)\,. \tag{8.6.4}$$

where of course one needs specific conditions on the system and input for the infinite sums to converge. The second formula is obtained by a change of variable.

Let's now go back to the ARX model. Assume for the moment that $\mathbf{u} = 0$ and we care only about the i.i.d. input process $\{\mathbf{w}(t)\}$. We can still write the solution as a convolution sum

$$\mathbf{y}(t) = \sum_{k=-\infty}^{+\infty} h(k)\, \mathbf{w}(t-k)$$

which is of the same form of the representation (A.3.3). We need to check under what circumstances the convergence condition $\sum_{k=-\infty}^{+\infty} h(k)^2 < \infty$ is satisfied.

Proposition 8.4. *If and only if all roots of the characteristic equation have modulus strictly less than 1; i.e.* $|\lambda_k| < 1\,;\ k = 1, 2, \ldots, n$, *then the (unique) causal impulse response of* (8.6.3) *i.e. such that* $h(t) = 0$ *for* $t < 0$ *is summable, that is*

$$\sum_{k=0}^{+\infty} h(k)^2 < \infty\,.$$

Proof. Clearly when $|\lambda_k| < 1$ then $\lim_{t\to+\infty}\ \lambda_k^t = 0$ exponentially fast and the impulse response is absolutely summable (and hence summable also in the ℓ^2 sense). This is clearly true also when there are multiple roots.

If some root has modulus greater than 1; i.e. $|\lambda_k| > 1$, then $\lim_{t\to+\infty}\ \lambda_k^t = \infty$ and the causal impulse response is *not summable*. Note that the anticausal impulse response may however turn out to be summable if *all* roots have modulus greater than one.

All complex roots, must always come as pairs of complex conjugate numbers say $\lambda_k, \bar{\lambda}_k = a_k \pm i b_k$ since the characteristic polynomial is real. Using the polar form

$$\lambda_k, \bar{\lambda}_k = \rho_k\, e^{\pm i\varphi_k}\,, \qquad \rho_k = |\lambda_k|\,; \quad \varphi_k = \arg(\lambda_k)$$

we can write $c_k\lambda_k^t + \bar{c}_k\bar{\lambda}_k^t$ in real form as

$$a_k\rho_k^t \cos(\varphi_k\, t) + b_k\rho_k^t \sin(\varphi_k\, t)\,.$$

for suitable real coefficients $a_k,\ b_k$. Purely imaginary roots, i.e such that $|\lambda_k| = 1$ like all complex roots, must also come as pairs of complex conjugate numbers. Therefore for an imaginary pair $\lambda, \bar{\lambda} = e^{\pm i\omega}$ we have

$$y(t) = c_k\, e^{+i\omega\, t} + \bar{c}_k\, e^{-i\omega\, t} = a_k \cos\omega\, t + b_k \sin\omega\, t\,,$$

and in this case $h(t)$ will have an oscillatory behaviour and the sum will not converge. It can be seen that in this case the process $\mathbf{y}$ is **not ergodic**. □

Example 8.2. Suppose we have the DE

$$y(t) - ay(t-1) = \delta(t)\,, \qquad t \in \mathbb{Z}$$

where $|a| < 1$ and with zero initial condition at $t = -1$. Show that

$$h(t) = \begin{cases} 0 \text{ for } t < 0 \\ a^t \text{ for } t \geq 0 \end{cases}\,.$$

What would you get for zero initial condition at $t = +1$?

Prediction and the innovation for purely AR processes

When $|\lambda_k| < 1\,;\, k = 1, 2, \ldots, n$ an AR process $\mathbf{y}$ with i.i.d. input $\mathbf{w}$ is ergodic but more is true. Because of causality $h(t) = 0$ for $t < 0$ and therefore

$$get\mathbf{y}(t) = \sum_{k=0}^{+\infty} h(k)\ \mathbf{w}(t-k)\ = \sum_{k=-\infty}^{t} h(t-k)\ \mathbf{w}(k) \tag{8.6.5}$$

so $\mathbf{y}(t)$ **depends only on the past history of** $\{\mathbf{w}(t)\}$. In general the past history will be *infinite*, since $h(t)$ is non zero for all $t \geq 0$. In abstract geometric terms (8.6.5) is equivalent to

$$\mathbf{H}(\mathbf{y}^t)\ \subset\ \mathbf{H}(\mathbf{w}^t)\,; \quad \text{for all} \quad t \in \mathbb{Z}\,. \tag{8.6.6}$$

On the other hand for an AR process we can express the input noise as a function of the past (and present) $n+1$ output samples as

$$\mathbf{w}(t) = \sum_{k=0}^{n} a_k \mathbf{y}(t-k)$$

which in fact implies that $\mathbf{w}(t) \in \mathbf{H}(\mathbf{y}^t)$ for all t and therefore

$$\mathbf{H}(\mathbf{w}^t)\ \subset\ \mathbf{H}(\mathbf{y}^t)\,; \quad \text{for all} \quad t \in \mathbb{Z}\,.$$

This relation, together with (8.6.6) implies *causal equivalence*:

$$\mathbf{H}(\mathbf{y}^t)\ =\ \mathbf{H}(\mathbf{w}^t)\,; \quad \text{for all} \quad t \in \mathbb{Z}\,. \tag{8.6.7}$$

so that,

Theorem 8.2. *For an AR model with all characteristic roots of modulus strictly less than one,* $\mathbf{w}$ **is the innovation process of** $\mathbf{y}$ *and the one step ahead linear predictor (conditional expectation in the Gaussian case) of* $\mathbf{y}(t)$ *given* $\mathbf{y}^{t-1}$ *is*

$$\hat{\mathbf{y}}(t \mid t-1) = \sum_{k=1}^{n} a_k \mathbf{y}(t-k) \tag{8.6.8}$$

while $\mathbf{w}(t)$ *is the innovatiopn; i.e. the one step ahead prediction error. The condition on the characteristic roots is also necessary.*

Problem 8.1. Consider the simple Auto-Regressive (AR) model

$$\mathbf{y}(t+1) = 0.8\,\mathbf{y}(t) + \mathbf{e}(t)$$

where $\mathbf{e}$ is an i.i.d. sequence of zero mean and variance equal to $\sigma^2_{\mathbf{e}}$. Suppose you start computing the solution process at time $t = t_0$ pretending you know the initial condition $\mathbf{y}(t_0) = y_0$. Write the expressions of $\mathbf{y}(t)$ for $t = t_0, t_0+1, \ldots,$ as a function of the error process $\{\mathbf{e}(t)\}$ leaving the powers of 0.8 indicated. Check that you always have $\mathbf{e}(t)$ uncorrelated with the previous random variables $\mathbf{y}(s)\,; s < t$. Consider the limit of the solution for $t_0 \to -\infty$. Does the limit exist?

Consider then the covariance function of the zero-mean stationary process $\mathbf{y}$ defined as

$$\sigma(k) := \mathbb{E}\,\mathbf{y}(t+k)\mathbf{y}(t)\,; \qquad k = 0, 1, \dots$$

so that $\sigma(0) = \mathbb{E}\,\mathbf{y}(t)^2$ is just the variance (which by stationarity is constant in time). Find the expression of the covariance of the process described by the AR model above.

Consider now the more realistic situation in which the true order is unknown so that the order n of the model is different form the true order n_0. We shall assume that n_0 is finite so that $\mathbf{y}(t)$ admits a representation

$$\mathbf{y}(t) = \mathbf{w}_0(t) + \hat{\mathbf{y}}_0(t \mid t-1)$$

where $\mathbf{w}_0(t)$ is the true i.i.d. innovation and

$$\hat{\mathbf{y}}_0(t \mid t-1) = \sum_{k=1}^{n_0} a_{0,k}\mathbf{y}(t-k) \tag{8.6.9}$$

is the true one-step-ahead linear predictor. Since the process is (stationary and) ergodic the average (squared) prediction error (8.2.6) converges as $N \to \infty$ and clearly we have,

$$\frac{1}{N}\sum_{t=1}^{N} \varepsilon_\theta(t)^2 = \mathbb{E}_{\,\theta_0}\boldsymbol{\varepsilon}_\theta^2(t) \tag{8.6.10}$$

where θ is the model parameter and

$$\boldsymbol{\varepsilon}_\theta = \mathbf{y}(t) - \hat{\mathbf{y}}_\theta(t \mid t-1) = \mathbf{w}_0(t) + [\hat{\mathbf{y}}_0(t \mid t-1) - \hat{\mathbf{y}}_\theta(t \mid t-1)\,]$$

so that the true variance of $\boldsymbol{\varepsilon}_\theta$ in (8.6.10) can be expressed as

$$\mathbb{E}_{\,\theta_0}\boldsymbol{\varepsilon}_\theta^2(t) = \sigma_0^2 + \mathbb{E}_{\,\theta_0}\,[\hat{\mathbf{y}}_0(t \mid t-1) - \hat{\mathbf{y}}_\theta(t \mid t-1)\,]^2 = \sigma_0^2 + \|\,\hat{\mathbf{y}}_0(t \mid t-1) - \hat{\mathbf{y}}_\theta(t \mid t-1)\,\|_{\mathbf{H}}^2 \tag{8.6.11}$$

Hence,

Theorem 8.3. *Under the above stated assumptions the limit as $N \to \infty$ of the PEM estimator $\hat{\theta}_N$ is the unique solution of the minimization problem*

$$\hat{\theta} = \operatorname{Arg}\min_\theta \mathbb{E}_{\,\theta_0}\,[\hat{\mathbf{y}}_0(t \mid t-1) - \hat{\mathbf{y}}_\theta(t \mid t-1)\,]^2 \tag{8.6.12}$$

which is a quadratic function of θ. The minimum exists and is unique.

Since the predictors are linear functions of the parameter the minimization can be performed explicitly. See Problems **8-4** and **8-5**. It is easy to show that if the true order n_0 is strictly greater than n we cannot have consistency, in the sense that $\hat{\theta}_N$ will not converge to the corresponding subvector of θ_0 having the same dimension n.

8.7 ▪ Prediction and innovation for ARX processes

If an exogenous input is also present the response of the ARX system has an additional term and the process can be represented as a sum of two convolutions of the form

$$\mathbf{y}(t) = \sum_{k=0}^{+\infty} h(k)\ \mathbf{w}(t-k) + \sum_{k=0}^{+\infty} g(k)\ \mathbf{u}(t-k) \tag{8.7.1}$$

where convergence of the first sum holds under the conditions stated in Proposition 8.4. This condition also guarantees that the second sum will be convergent for bounded deterministic inputs **[?]**. The calculations for finding the deterministic impulse response $g(t)$ originate from the substitution

$$\sum_{k=0}^{+\infty} h(k)\,(\sum_{j=1}^{m} b_j\,\mathbf{u}(t-k-j)) = \sum_{j=1}^{m} b_j \sum_{k=0}^{+\infty} h(k)\,\mathbf{u}(t-k-j) =$$

$$\sum_{k=-\infty}^{t}\,(\sum_{j=1}^{m} b_j h(t-j-k))\,\mathbf{u}(k)$$

by which $g(t) = \sum_{j=1}^{m} b_j h(t-j)$.
Assume the structure (8.2.1) where the "instantaneous coupling" coefficient b_0 between input and output is assumed to be zero. Initially we shall assume (although this is not strictly necessary and we may just do with a deterministic bounded sequence) that the input $\mathbf{u}$ is a zero-mean process jointly stationary with $\mathbf{y}$ having finite variance and **uncorrelated with the input noise** $\mathbf{w}$.

Also assume that both convolution sums in (8.7.1) are convergent in mean square. The one-step-ahead linear predictor now must be based on the joint past information $\mathbf{H}(\mathbf{y}^{t-1}, \mathbf{u}^{t-1})$.

Since $\mathbf{w}$ is an i.i.d. process uncorrelated with its own past and with the whole history of the input process,

$$\mathbf{w}(t) \perp \mathbf{H}(\mathbf{w}^{t-1}, \mathbf{u}^{t-1})\,. \tag{8.7.2}$$

This easily leads to a formula for the predictor.

Theorem 8.4. *Under the stated assumptions, the minimum variance linear predictor of* $\mathbf{y}(t)$ *based on the joint past history* $\mathbf{H}(\mathbf{y}^{t-1}, \mathbf{u}^{t-1})$ *is*

$$\mathbf{y}(t \mid t-1) = \sum_{k=1}^{n} a_k \mathbf{y}(t-k) + \sum_{k=1}^{m} b_k\,\mathbf{u}(t-k)\,. \tag{8.7.3}$$

and $\mathbf{w}(t)$ *is the one step ahead prediction error of* $\mathbf{y}(t)$ *bsed on the joint infinite past* $\mathbf{H}(\mathbf{y}^{t-1}, \mathbf{u}^{t-1})$.

Proof. We need to show that

$$\mathbf{H}(\mathbf{y}^{t-1}, \mathbf{u}^{t-1}) = \mathbf{H}(\mathbf{w}^{t-1}, \mathbf{u}^{t-1})\,, \tag{8.7.4}$$

which is the analog of causal equivalence in the present context. This will imply that (8.7.3) is exactly the orthogonal projection of $\mathbf{y}(t)$ onto $\mathbf{H}(\mathbf{y}^{t-1}, \mathbf{u}^{t-1})$. Now from the model equation it follow trivially that $\mathbf{w}(t)$ must be a (linear) function of the past and present output $\mathbf{y}^t$ and of the (strict) past $\mathbf{u}^{t-1}$. Therefore

$\mathbf{H}(\mathbf{w}^t, \mathbf{u}^t) \subset \mathbf{H}(\mathbf{y}^t, \mathbf{u}^t)$. On the other hand by assumption we have the representation (8.7.1) which implies that $\mathbf{y}(t) \in \mathbf{H}(\mathbf{w}^t, \mathbf{u}^t)$ for all t. Therefore we also have the inclusion $\mathbf{H}(\mathbf{y}^t, \mathbf{u}^t) \subset \mathbf{H}(\mathbf{w}^t, \mathbf{u}^t)$ which together with the previous one implies (8.7.4). Hence by (8.7.2) we have $\mathbf{w}(t) \perp \mathbf{H}(\mathbf{y}^{t-1}, \mathbf{u}^{t-1})$ and since the model relation (8.2.1) can be written as the orthogonal sum

$$\mathbf{y}(t) = \mathbf{w}(t) + \left[\sum_{k=1}^{n} a_k \mathbf{y}(t-k) + \sum_{k=1}^{m} b_k \, \mathbf{u}(t-k) \right] \tag{8.7.5}$$

by the usual reasoning based on the orthogonality principle we have the conclusion. □

As for AR processes, the orthogonal decomposition (8.7.5) may lead to think that we might have guessed the formula (8.7.3) from the beginning, without worrying about all work spent for characterizing the innovation. One should however realize that the orthogonality can only lead to the conclusion that (8.7.3) is the predictor given the **finite past** spanned only by the $n + m$ past variables $\mathbf{y}(t-k)$, $\mathbf{u}(t-k)$, while we have shown that it is actually the predictor based on the *whole infinite joint past history of* $\mathbf{y}$ *and* $\mathbf{u}$. The dependence on only the finite $n + m$ past variables $\mathbf{y}(t-k)$, $\mathbf{u}(t-k)$ is a lucky circumstance which does not happen with more complex models like e.g. ARMAX structures. In the language of Statistics one may say that the $n + m$ past variables $\mathbf{y}(t-k)$, $\mathbf{u}(t-k)$ constitute a **sufficient statistic** of the whole past history for the prediction of $\mathbf{y}(t)$.

8.8 ▪ Strong consistency of the Least Squares ARX estimator

In order to generalize the consistency theorem 8.1 to ARX models we need to discuss the nature of the input signal. It is clear that not all input signals can lead to a convergent estimate. In fact some signals, like for example a deterministic signal constant in time, are structurally uncapable of discriminating some parameter values. An easy way out is to assume that $\mathbf{u}$ is a stationary ergodic process independent of the i.i.d noise $\mathbf{w}$.

Theorem 8.5. *Assume there is a true ARX model describing the data having orders n, m equal to those of the candidate ARX model and true parameters θ_0, σ_0^2 and that $\mathbf{u}$ is a jointly stationary ergodic process independent of the i.i.d noise $\mathbf{w}_0$. Assume also that the true model is* **causal** *so that*

$$\mathbb{E}_{\theta_0} \mathbf{y}(s) \; \mathbf{w}_0(t) = 0 \,; \qquad \forall s < t \in \mathbb{Z} \,; \tag{8.8.1}$$

and that

$$\Sigma_0 := \mathbb{E}_{\theta_0} \boldsymbol{\varphi}(t) \boldsymbol{\varphi}(t)^\top > 0; \tag{8.8.2}$$

then both $\hat{\boldsymbol{\theta}}_N$ and the sample innovation variance (8.2.7) *converge to the respective true values*

$$\lim_{N \to \infty} \hat{\boldsymbol{\theta}}_N = \theta_0 \,, \qquad \lim_{N \to \infty} \hat{\sigma}_N^2 = \sigma_0^2$$

with probability one.

Proof. The sample covariance matrix of $\boldsymbol{\varphi}(t)$, now defined as in (8.2.3),

$$\hat{\boldsymbol{\Sigma}}_N := \frac{1}{N}\sum_{t=1}^{N} \boldsymbol{\varphi}(t)\boldsymbol{\varphi}(t)^\top \in \mathbb{R}^{(n+m)\times(n+m)}\,.$$

can be partitioned in four blocks

$$\hat{\boldsymbol{\Sigma}}_N = \begin{bmatrix} \hat{\boldsymbol{\Sigma}}_N(y) & \hat{\boldsymbol{\Sigma}}_N(yu) \\ \hat{\boldsymbol{\Sigma}}_N(uy) & \hat{\boldsymbol{\Sigma}}_N(u) \end{bmatrix} \in \mathbb{R}^{(n+m)\times(n+m)}\,. \tag{8.8.3}$$

which, because of ergodicity converges as $N \to \infty$ to Σ_0. Then if (8.8.2) holds true we can use the same argument as in the proof of Theorem 8.1. □

A weaker assumption on the imput process is that it is a deterministic signal as in the following definition.

Definition 8.2. *A deterministic sequence $u = \{u(t)\,; t \in \mathbb{Z}\}$ is said to be* second order stationary *if the limit*

$$\lim_{N\to\infty} \frac{1}{N}\sum_{t=1}^{N} u(t+\tau)u(t) := r(\tau) \tag{8.8.4}$$

exists for all $\tau \geq 0$.

A second order stationary signal, u, is persistently exciting of order (at least) n *if the Toeplitz matrix*

$$\mathbf{R}_n := \begin{bmatrix} r(0) & r(1) & \dots & & r(n-1) \\ r(1) & r(0) & r(1) & \dots & r(n-2) \\ \vdots & & & \vdots & \\ r(n-1) & r(n-2) & \dots & & r(0) \end{bmatrix} \tag{8.8.5}$$

is positive definite

Hence if u is persistently exciting of order (at least) m then the lower diagonal block in the matrix (8.8.3) converges to a positive definite covariance. Then, by Cauchy-Schwartz inequality one can show that the sample cross covariances also converge.
This is evidently a necessary condition for the Least Squares algorithm to provide a consistent estimate of the b_k parameters in the model (8.2.1).

Note that $\mathbf{R}_n$ is always at least positive semidefinite since for an arbitrary polynomial in the delay operator z^{-1}, say $p(z^{-1}) := \sum_{k=0}^{n-1} p_k z^{-k}$ one has

$$\lim_{N\to\infty} \frac{1}{N}\sum_{t=1}^{N} [p(z^{-1})u(t)]^2 = \sum_{k,j=0}^{n} p_k\, r(k-j)\, p_j = p^\top \mathbf{R}_n\, p \geq 0$$

where $p := [p_0\, p_1\, \dots\, p_{n-1}]^\top$ is the vector of the coefficientas of $p(z^{-1})$.

Continuous time second order stationary signals have been studied by Norbert Wiener in 1930 [103, 105]). Since the function $\tau \to r(\tau)\,;\, \tau \in \mathbb{Z}$ is *positive*

definite it can be treated as the correlation function of a stationary process. It can be shown that there must exist a function, $F_u(e^{i\omega})$, monotone non decreasing on the intervalo $[-\pi,\, \pi]$, such that a Fourier-like representation holds

$$r(\tau) = \int_{-\pi}^{\pi} e^{i\omega\tau}\, dF_u(e^{i\omega})\,.$$

We shall quote without proof the following theorem.

Theorem 8.6. *A second order stationary signal u is persistently exciting of order exactly n if and only if the unique points of increase of the function $F_u(e^{i\omega})$ are n jumps, that is, there are exactly n spectral lines at different frequencies $\{\omega_1, \omega_2, \dots, \omega_n\}$ in the interval $(-\pi,\, \pi)$.*

It can be shown that second order stationary signals defined on $\mathbb{Z}$ form a Hilbert space with inner product

$$\langle u,\, v\rangle := \lim_{N\to\infty} \frac{1}{2N} \sum_{t=-N}^{N} u(t)v(t)$$

and that quasi periodic signals, defined in Section A.4, of arbitrary order, form a dense subset of this space.

Example 8.3. Use the Law of Large Numbers to show that any trajectory of an i.i.d. process is a second-order stationary signal of infinite order and describe $\mathbf{R}_n$ in this case.

Example 8.4 (quasi periodic signals). The linear combination of N sinusoidal signals of different frequencies

$$u(t) = \sum_{k=1}^{N} A_k \sin(\omega_k t + \phi_k) \qquad \omega_k \neq \omega_j \tag{8.8.6}$$

is second order stationary. Its correlation function is

$$r(\tau) = \sum_{k=1}^{N} \frac{A_k^2}{2} \cos\omega_k\tau \tag{8.8.7}$$

see Sect. A.4 and the spectral function $F_u(e^{i\omega})$ has exactly $2N$ steps (that is, the derivative of F_u has exactly $2N$ δ functions) supported in $\{\pm\omega_k\}$. The signal is therefore persistently exciting of order $2N$.

In particular a discrete time signal periodic of period N, being the sum of N sinusoidal components of frequency $\omega_1 = \dfrac{2\pi}{N}$, $\omega_2 = 2\dfrac{2\pi}{N}, \dots, \omega_N = 2\pi$ is P.E. of order $2N$.

Example 8.5 (Pseudo-Random Binary Sequences). A PRBS (*Pseudo Random Binary Sequence*) is a particular computer-generated periodic signal which approximates a sample trajectory of an i.i.d. sequence. It is necessarily periodic but the period is a very large natural number. See [90, p. 124-125].

Persistently exciting input sequences of high enough order play the same role as ergodic input processes.

Corollary 8.1. *Assume there is a true ARX model describing the data having true parameter θ_0, orders n, m equal to those of the candidate ARX model and that $\mathbf{u}$ is a second-order stationary sequence of order greater or equal to m. Assume also that the true model is* **causal** *so that*

$$\mathbb{E}_{\theta_0}\mathbf{y}(s)\ \mathbf{w}(t) = 0\,; \qquad \forall s < t \in \mathbb{Z}\,; \tag{8.8.8}$$

then

$$\lim_{N\to\infty}\ \hat{\boldsymbol{\theta}}_N = \theta_0$$

with probability one.

Under these assumptions the positivity condition that

$$\Sigma_0 := \lim_{N\to\infty}\ \hat{\boldsymbol{\Sigma}}_N > 0; \tag{8.8.9}$$

is in fact guaranteed.

Problem 8.2. *There is an analog of Theorem 8.3 for ARX models. State it and give a proof.*

The asymptotic Variance

Recall that the variance of a consistent estimator must tend to zero as $N \to \infty$. However since data are always finite, we would like to have approximate "asymptotic" expressions and an idea of how fast the variance tends to zero. The concept of *asymptotic variance* of a consistent estimator is meant to capture this idea. It must however be defined properly. Here is one possible definition.

Definition 8.3. *Let $\{\phi_N(\mathbf{y})\,;\ N = 1, 2, \ldots\}$ be a consistent sequence of estimators of the parameter θ and $d(N)$ a function of N which is increasing to $+\infty$ with N and strictly positive. One say that $\phi_N(\mathbf{y})$ has asymptotic variance Σ if*

$$\sqrt{d(N)}\,[\,\phi_N(\mathbf{y}) - \theta_0\,] \xrightarrow{L} D(0, \Sigma)$$

where the convergence is in Law, to a pdf $D(0, \Sigma)$ which has variance Σ, possibly depending on θ_0, which is finite and strictly positive definite.

Hence for N large the variance of $\phi_N(\mathbf{y})$ can be approximated by $\frac{1}{d(N)}\Sigma$. In many situations, the central limit theorem applies and the estimator is asymptotically normal. It can be shown that in these situations $d(N)$ can be taken equal to N.

The condition of strict positivity $\Sigma > 0$ is essential since it excludes the possibility of linear combinations of the components of $\phi_N(\mathbf{y})$ having variance which tends to zero. This just means that the order of infinitesimal of the variance of these combinations will be different from $O(\frac{1}{d(N)})$.

Let us look at the ARX case, assuming that the conditions of the consistency theorem 8.5 hold. We shall also assume that for N large enough the inverse of

the sample covariance (8.3.3) exists. Then it can be shown that the asymptotic distribution of

$$\sqrt{N}\,(\hat{\boldsymbol{\theta}}_N - \theta_0) = \hat{\Sigma}_N^{-1} \frac{1}{\sqrt{(N)}} \sum_{t=1}^{N} \boldsymbol{\varphi}(t)\, \mathbf{w}_0(t)$$

is Gaussian with variance $\sigma_0^2 \Sigma_0^{-1}$, where Σ_0 is the limit (8.8.9). This expression can therefore be identified with the asymptotic variance of $\hat{\boldsymbol{\theta}}_N$ according to Definition 8.3. In practice, the formula

$$\text{Var}\,\left(\hat{\boldsymbol{\theta}}_N\right) \simeq \frac{\hat{\sigma}_N^2}{N}\, \hat{\Sigma}_N^{-1} \tag{8.8.10}$$

can be used as a consistent estimate of the variance of $\hat{\boldsymbol{\theta}}_N$.

8.9 ▪ Bayesian recursive estimators for ARX models

We consider the ARX model (8.2.1) written as a "pseudo-linear regression"

$$\mathbf{y}(t) = \boldsymbol{\varphi}(t)^\top\, \theta + \mathbf{w}(t)\,, \qquad t \in \mathbb{Z}_+ \tag{8.9.1}$$

where now we shall assign a Gaussian prior distribution to the parameter which becomes a random p-dimensional vector

$$\boldsymbol{\theta} \sim \mathcal{N}(\theta_0,\, P_0)$$

with θ_0 some "nominal" value and P_0 a variance matrix which as we shall see, does not need to be known with great confidence. The noise $\mathbf{w}(t)$ is also assumed Gaussian i.i.d. with variance σ^2, independent of $\boldsymbol{\theta}$ for all t.

For simplicity we shall assume that $\boldsymbol{\varphi}(t)$ is only a function of $\mathbf{y}^{t-1}$ so that initially (8.9.1) will be a purely Auto Regressive model. At the end we shall consider some generalizations of this model, in particular allow $\boldsymbol{\varphi}(t)$ to depend also on the input $\mathbf{u}$. In this case we shall need to require that $\mathbf{u}(t)$ and $\mathbf{w}(s)$ are independent for all $t, s \in \mathbb{Z}_+$.

The model (8.9.1) is a difference equation which can be solved recursively starting from some initial values $\boldsymbol{\varphi}(0)^\top = \begin{bmatrix}\mathbf{y}(-1) & \dots & \mathbf{y}(-n)\end{bmatrix}^\top$, which we assume are zero mean random variables independent of future values of the noise $\{\mathbf{w}(t)\,;\, t \geq 0\}$, yielding a solution

$$\mathbf{y}(t) = h(\,\boldsymbol{\theta},\, \mathbf{w}^t)\,; \qquad t \geq 0$$

which is a function of the parameter, the initial conditions, and the past noise from time zero up to time t. From the independence of $\mathbf{w}(t+1)$ and $\mathbf{w}^t$ we see immediately that,

Lemma 8.2. *For all $t \geq 0$ the random variable $\mathbf{w}(t+1)$ is independent of the past observations $\mathbf{y}^t$; in fact it is also independent of $(\,\mathbf{y}^t,\, \boldsymbol{\theta}\,)$.*

We shall say that a random variable $\mathbf{x}$ is *conditionally Gaussian* given a family of random variables $\{\mathbf{z}_\alpha\,;\, \alpha \in A\}$ if $\mathbf{x}$ admits a conditional distribution given $\{\mathbf{z}_\alpha\,;\, \alpha \in A\}$ which is Gaussian. Naturally the mean and variance of this distribution will be the conditional mean and variance of $\mathbf{x}$ given $\{\mathbf{z}_\alpha\,;\, \alpha \in A\}$.

In what follows it will be convenient to use vector notations. By stacking the system equations (8.9.1) ordered for increasing time $t = 0, 1, \dots$ we obtain a relation among random vectors

$$\mathbf{y}^t = \begin{bmatrix} \boldsymbol{\varphi}(0)^\top \\ \boldsymbol{\varphi}(1)^\top \\ \dots \\ \boldsymbol{\varphi}(t)^\top \end{bmatrix} \boldsymbol{\theta} + \mathbf{w}^t \;:=\; \boldsymbol{\Phi}_t\, \boldsymbol{\theta} + \mathbf{w}^t \tag{8.9.2}$$

where the matrix $\boldsymbol{\Phi}_t$ is a function of the initial conditions and past outputs up to time $t-1$.

Theorem 8.7. *Assume t is large enough and that $\boldsymbol{\Phi}_t$ has almost surely a left inverse $\boldsymbol{\Phi}_t^{-L}$. Then, the random variables $(\,\mathbf{y}(t+1),\, \boldsymbol{\theta}\,)$ are jointly conditionally Gaussian given $\mathbf{y}^t$.*

Proof. We shall first show that $p(\theta \mid \mathbf{y}^t)$ is a conditionally Gaussian distribution. Left-multiply (8.9.2) by $\boldsymbol{\Phi}_t^{-L}$ to get

$$\boldsymbol{\theta} = \boldsymbol{\Phi}_t^{-L}\mathbf{y}^t - \boldsymbol{\Phi}_t^{-L}\mathbf{w}^t$$

which shows that the conditional distribution of $\boldsymbol{\theta}$ [24] given $\mathbf{y}^t$, is Gaussian with (conditional) mean vector $\boldsymbol{\Phi}_t^{-L}\mathbf{y}^t$ and conditional variance equal to $\sigma^2\boldsymbol{\Phi}_t^{-L}\,[\,\boldsymbol{\Phi}_t^{-L}\,]^\top$. Then the statement follows from Bayes rule

$$p(y(t+1),\, \theta \mid \mathbf{y}^t) = p(y(t+1) \mid\, \boldsymbol{\theta},\, \mathbf{y}^t)\, p(\theta \mid \mathbf{y}^t)$$

since the first factor on the right is clearly a conditional Gaussian distribution with mean $\boldsymbol{\varphi}(t+1)^\top\theta$ and variance equal to $\mathrm{var}\,\{\mathbf{w}(t+1)\}$. If we condition with respect to some fixed observation $\mathbf{y}^t(\omega) = y^t$ these are just usual regular Gaussian densities. □

In spite of its appearance the model (8.9.2) is non-linear and it is not immediately clear what could be a reasonable estimation strategy. One option could be empirical prediction error minimization (PEM) as described in Section 2.5 with a ridge penalty term related to the a priori variance of $\boldsymbol{\theta}$. We shall instead propose a *Bayesian recursive solution* which is much in the same spirit of the algorithm developed in Sect. 2.6. Consider then the following

Problem 8.3. *Find a recursive updating algorithm to compute the conditional mean and conditional variance of the random parameter vector $\boldsymbol{\theta}$:*

$$\hat{\boldsymbol{\theta}}(t) := \mathbb{E}\,[\boldsymbol{\theta} \mid \mathbf{y}^t]\,, \qquad \Sigma(t) := \mathrm{Var}\,[\,\boldsymbol{\theta} \mid \mathbf{y}^t] \tag{8.9.3}$$

We shall rely on Theorem 8.7 and on the full rank assumption of $\boldsymbol{\Phi}_t$. Consider the conditional expectation of an arbitrary random variable $\mathbf{x}$ given two other random variables, $\mathbf{y}_1, \mathbf{y}_2$ the first of which is kept fixed while the second may vary. We shall need to consider the regression function $\mathbb{E}\,[\,\mathbf{x} \mid \mathbf{y}_1, \mathbf{y}_2]$, which

[24] Since this θ depends on the choice of the left inverse, more correctly one should say: any random vector θ satisfying (8.9.2).

is by definition a measurable function of both variables, when $\mathbf{y}_1$ is kept fixed. This we shall consider as a function of $\mathbf{y}_2$ only and denote it by the symbol $\mathbb{E}_{\mathbf{y}_1}[\,\mathbf{x} \mid \mathbf{y}_2]$. Obviously with this convention $\mathbb{E}_{\mathbf{y}_1}[\,\mathbf{x}\,] = \mathbb{E}[\,\mathbf{x} \mid \mathbf{y}_1]$. Suppose $\mathbf{x}$ is conditionally Gaussian given $(\mathbf{y}_1, \mathbf{y}_2)$, then the Gaussian conditional expectation formula (5.3.3) yields

$$\begin{aligned}\mathbb{E}[\,\mathbf{x} \mid \mathbf{y}_1, \mathbf{y}_2] = \mathbb{E}_{\mathbf{y}_1}[\,\mathbf{x}\,] + \operatorname{Cov}\{\mathbf{x}, \mathbf{y}_2 \mid \mathbf{y}_1\} \operatorname{Var}\{\mathbf{y}_2 \mid \mathbf{y}_1\}^{-1} \\ \times \operatorname{Cov}\{\mathbf{y}_2, \mathbf{x} \mid \mathbf{y}_1\} \{\mathbf{y}_2 - \mathbb{E}[\,\mathbf{y}_2 \mid \mathbf{y}_1]\,\}\end{aligned} \tag{8.9.4}$$

which can be justified just by thinking that the conditional density with respect to both variables $p(x \mid y_1, y_2)$ is the Gaussian density $p_{y_1}(x) := p(x \mid y_1)$ conditioned with respect to $\mathbf{y}_2$. Consider now the estimate at time $t+1$

$$\hat{\boldsymbol{\theta}}(t{+}1) := \mathbb{E}[\boldsymbol{\theta} \mid \mathbf{y}(t{+}1),\, \mathbf{y}^t] = \mathbb{E}_{\mathbf{y}^t}[\,\boldsymbol{\theta} \mid \mathbf{y}(t{+}1)\,]$$

where the operator $\mathbb{E}_{\mathbf{y}^t}$ is as defined above. Then applying formula (8.9.4) we obtain

$$\begin{aligned}\mathbb{E}[\boldsymbol{\theta} \mid \mathbf{y}(t{+}1),\, \mathbf{y}^t] = \mathbb{E}_{\mathbf{y}^t}[\,\boldsymbol{\theta}\,] + \operatorname{Cov}\{\boldsymbol{\theta}, \mathbf{y}(t{+}1) \mid \mathbf{y}^t\} \operatorname{Var}\{\mathbf{y}(t{+}1) \mid \mathbf{y}^t\}^{-1} \\ \times \operatorname{Cov}\{\mathbf{y}(t{+}1), \boldsymbol{\theta} \mid \mathbf{y}^t\} \{\mathbf{y}(t{+}1) - \mathbb{E}[\,\mathbf{y}(t{+}1) \mid \mathbf{y}^t]\,\}\end{aligned} \tag{8.9.5}$$

where

$$\mathbb{E}_{\mathbf{y}^t}[\,\boldsymbol{\theta}\,] = \hat{\boldsymbol{\theta}}(t)\,, \qquad \mathbb{E}[\,\mathbf{y}(t{+}1) \mid \mathbf{y}^t] = \boldsymbol{\varphi}(t{+}1)^\top \hat{\boldsymbol{\theta}}(t)$$

the last equality following since $\mathbf{w}(t{+}1)$ and $\mathbf{y}^t$ are independent (Lemma 8.2). The last equation describes the (one step ahead) predictor of $\mathbf{y}(t{+}1)$ given $\mathbf{y}^t$ which for short we denote $\hat{\mathbf{y}}(t+1 \mid t)$. The estimator (8.9.5) is a linear function of the (one step ahead) **prediction error**

$$\mathbf{e}(t) := \mathbf{y}(t) - \hat{\mathbf{y}}(t \mid t-1) = \boldsymbol{\varphi}(t)^\top [\,\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}(t-1)\,] + \mathbf{w}(t) \tag{8.9.6}$$

which is a process with uncorrelated (and hence independent) variables by the orthogonality principle. Its conditional variance is just

$$\begin{aligned}\operatorname{var}[\,\mathbf{e}(t) \mid \mathbf{y}^{t-1}\,] = \operatorname{var}[\,\mathbf{y}(t) \mid \mathbf{y}^{t-1}] = \mathbb{E}\{\,[\boldsymbol{\varphi}(t)^\top(\,\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}(t-1)\,) + \mathbf{w}(t)\,]^2 \mid \mathbf{y}^{t-1}\} \\ = \boldsymbol{\varphi}(t)^\top \Sigma(t-1) \boldsymbol{\varphi}(t) + \sigma^2\end{aligned}$$

The variable $\mathbf{e}(t{+}1)$ is just the part of $\mathbf{y}(t{+}1)$ which is unpredictable based on the past $\mathbf{y}^t$; the sequence $\{\mathbf{e}(t)\}$ is actually the **innovation process** of $\{\mathbf{y}(t)\}$. For the covariance matrices in (8.9.4) we obtain

$$\begin{aligned}&\operatorname{Cov}\{\boldsymbol{\theta}, \mathbf{y}(t{+}1) \mid \mathbf{y}^t\} = \mathbb{E}\{[\,\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}(t)\,][\mathbf{y}(t{+}1) - \hat{\mathbf{y}}(t+1 \mid t)\,] \mid \mathbf{y}^t\} = \\ &= \mathbb{E}\{[\,\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}(t)\,][\,\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}(t)\,]^\top \boldsymbol{\varphi}(t{+}1) \mid \mathbf{y}^t\} + \mathbb{E}\{[\,\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}(t)\,]\mathbf{w}(t{+}1) \mid \mathbf{y}^t\} = \\ &= \mathbb{E}\{[\,\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}(t)\,][\,\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}(t)\,]^\top \mid \mathbf{y}^t\} \boldsymbol{\varphi}(t{+}1) = \Sigma(t) \boldsymbol{\varphi}(t{+}1)\,.\end{aligned}$$

The last equality follows from the independence of $\boldsymbol{\theta}$ and $\mathbf{w}(t)$ and Lemma 8.2. The conditional variance $\operatorname{var}[\,\mathbf{y}(t{+}1) \mid \mathbf{y}^t]$ has been computed above for time t instead of $t+1$. The updating formula (8.9.5) can therefore be written as

$$\hat{\boldsymbol{\theta}}(t{+}1) = \hat{\boldsymbol{\theta}}(t) + k(t{+}1)[\,\mathbf{y}(t{+}1) - \boldsymbol{\varphi}(t{+}1)^\top \hat{\boldsymbol{\theta}}(t)\,] \tag{8.9.7}$$

where the *gain vector* $k(t+1)$ is given by

$$k(t+1) = \Sigma(t)\boldsymbol{\varphi}(t+1)\big[\boldsymbol{\varphi}(t+1)^\top\Sigma(t)\boldsymbol{\varphi}(t+1) + \sigma^2\big]^{-1} \tag{8.9.8}$$

The recursion is driven by the innovation $\mathbf{e}(t+1)$.The initial condition can be taken as $\hat{\boldsymbol{\theta}}(0) = \theta_0$.
We still need an updating equation for $\Sigma(t)$. Using again the iterated conditioning formula (8.9.4) and (5.3.4) we get

$$\begin{aligned}
&\Sigma(t+1) = \mathrm{Var}\,[\,\boldsymbol{\theta} \mid \mathbf{y}^{t+1}] = \\
&= \mathrm{Var}\,[\,\boldsymbol{\theta} \mid \mathbf{y}^{t}\,] - \mathrm{Cov}\,\{\boldsymbol{\theta}, \mathbf{y}(t+1) \mid \mathbf{y}^t\}\,\mathrm{var}\,[\,\mathbf{y}(t+1) \mid \mathbf{y}^t]^{-1}\,\mathrm{Cov}\,\{\mathbf{y}(t+1), \boldsymbol{\theta} \mid \mathbf{y}^t\} = \\
&= \Sigma(t) - \Sigma(t)\boldsymbol{\varphi}(t+1)\big[\boldsymbol{\varphi}(t+1)^\top\Sigma(t)\boldsymbol{\varphi}(t+1) + \sigma^2\big]^{-1}\boldsymbol{\varphi}(t+1)^\top\Sigma(t)
\end{aligned}$$

with initial condition the a priori covariance $\Sigma(0) = \mathrm{Var}\,[\,\boldsymbol{\theta}\,] = P_0$.

Remarks 8.2. The reader should compare this Bayesian algorithm with the one derived in Sect. 2.6. Note that now the gain and the variance matrix are functions of the past data $\mathbf{y}^t$ making the algorithm a truly non-linear recursion. Evidently the prior information does modify the algorithm but the formulas here would be hard to derive from the one-shot regularized solution.

There are various extensions of the algorithm to more complicated models. One easy step is to consider ARX models where the input process $\mathbf{u}(t)$ enters as in (8.2.1) and the parameter is now $n + m$-dimensional. Since in general for physical reasons there cannot be instantaneous effect of the input $\mathbf{u}(t)$ on the variable $\mathbf{y}(t)$ the input parameter b_0 in the model is normally set to zero. Hence the information available at time t is now constituted by the joint input-output sequences $\mathbf{z}^t := (\mathbf{u}^{t-1}, \mathbf{y}^t)$. The reader should work out the derivation considering all conditional expectations with respect to this joint information flow assuming that the input and the noise processes are independent.

Theorem 8.8. *Consider the ARX model* (8.2.1) *with a Gaussian noise* $\mathbf{w}$ *independent of the input process* $\mathbf{u}$*. Then the estimator* $\hat{\boldsymbol{\theta}}(t)$ *which minimizes the conditional error variance* $\Sigma(t) := \mathrm{Var}\,[\,\boldsymbol{\theta} \mid \mathbf{z}^t\,]$ *evolves in time according to the following recursion*

$$\hat{\boldsymbol{\theta}}(t+1) = \hat{\boldsymbol{\theta}}(t) + k(t+1)[\,\mathbf{y}(t+1) - \boldsymbol{\varphi}(t+1)^\top\hat{\boldsymbol{\theta}}(t)\,] \tag{8.9.9}$$

where the gain vector $k(t+1)$ *is given by*

$$k(t+1) = \Sigma(t)\boldsymbol{\varphi}(t+1)\big[\boldsymbol{\varphi}(t+1)^\top\Sigma(t)\boldsymbol{\varphi}(t+1) + \sigma^2\big]^{-1} \tag{8.9.10}$$

The process $\mathbf{e}(t+1) := \mathbf{y}(t+1) - \boldsymbol{\varphi}(t+1)^\top\hat{\boldsymbol{\theta}}(t)$ *driving the recursion is the one step ahead prediction error of* $\mathbf{y}(t+1)$ *given the past* $\mathbf{z}^t$ *(that is the innovation). The initial condition can be taken as* $\hat{\boldsymbol{\theta}}(0) = \theta_0$.
The conditional error variance $\Sigma(t)$ *can be updated by the following matrix recursion*

$$\Sigma(t+1) = \Sigma(t) - \Sigma(t)\boldsymbol{\varphi}(t+1)\big[\boldsymbol{\varphi}(t+1)^\top\Sigma(t)\boldsymbol{\varphi}(t+1) + \sigma^2\big]^{-1}\boldsymbol{\varphi}(t+1)^\top\Sigma(t) \tag{8.9.11}$$

with initial condition the a priori variance $\Sigma(0) = P_0$.

8.10 ▪ Problems

8-0 *Find the impulse response of the system described by the DE*

$$y(t) + \frac{1}{4}y(t-2) = u(t)\,.$$

8-1 *Let* $\{\mathbf{e}(t)\}$ *be a zero-mean uncorrelated process; i.e.* $\mathbb{E}\,\mathbf{e}(t)\mathbf{e}(s) = 0$ *for* $t \neq s$*, show that the transformation*

$$\mathbf{w}(t) := e^{i\omega t}\mathbf{e}(t)\,, \qquad t \in \mathbb{Z}$$

produces another (complex) uncorrelated process. Is this also true for $\mathbf{z}(t) := \sin(\omega t)\mathbf{e}(t)$*?*

8-2 *Show that the deterministic signal* $s(t) = A\sin(\omega t + \varphi)$ *satisfies a second order difference equation of the form*

$$s(t) + a_1\, s(t-1) + a_2\, s(t-2) = 0$$

and find the coefficients a_1*,* a_2*. Show that* $a_2 = 1$ *and only* a_1 *depends on the angular frequency* ω*.*

8-3 *Consider a process* $\mathbf{y}$ *described by the "true" AR model*

$$\mathbf{y}(t) + a_0\mathbf{y}(t-1) = \mathbf{w}_0(t) \qquad |a_0| < 1$$

where $\mathbf{w}_0$ *is i.i.d. with* $\operatorname{var}[\mathbf{w}_0(t)] = \sigma_0^2$*. For identification you are using a class of AR models of the same type, namely*

$$\mathbf{y}(t) + \theta\mathbf{y}(t-1) = \mathbf{w}(t)\,,$$

and estimate the parameter θ *by implementing a PEM (least squares) algorithm. Show that* $\hat{\sigma}_N^2$ *in* (8.2.7) *converges to* σ_0^2 *when* $N \to \infty$*.*

8-4 *Consider the same process* $\mathbf{y}$ *described in the previous problem. Now you are using a class of more complicated AR models, namely*

$$\mathbf{y}(t) + \theta_1\mathbf{y}(t-1) + \theta_2\mathbf{y}(t-2) = \mathbf{w}(t)\,,$$

and estimate the vector parameter θ *by implementing a PEM (least squares) algorithm. Where will the estimate* $\hat{\boldsymbol{\theta}}_N$ *converges when* $N \to \infty$ *?*

8-5 *Same question if the true process* $\mathbf{y}$ *is described by the AR model*

$$\mathbf{y}(t) + a_{0,1}\mathbf{y}(t-1) + a_{0,2}\mathbf{y}(t-2) = \mathbf{w}_0(t)$$

where the roots of the characteristic equation have modulus less than one and the model has only one free parameter

$$\mathbf{y}(t) + a\mathbf{y}(t-1) = \mathbf{w}(t)\,.$$

Appendix A

SOME FACTS FROM PROBABILITY THEORY

A.1 ▪ A quick review of Proability Theory

[◦] A **Probability Space**: $\{\Omega, \mathcal{A}, \mathbf{P}\}$ is composed by the set of *elementary events* $\omega \in \Omega$ chosen by "nature", the sigma-algebra $\mathcal{A}$ which contains all subsets of Ω (Events) of which you can compute the probability and a countably additive set function

$$\mathbf{P} : \mathcal{A} \to [\,0,\,1\,]\,.$$

[◦] **Random variables** are (mesurable) functions $\mathbf{x} : \Omega \to \mathbb{R}$.
The *Probability distribution function* of $\mathbf{x}$ is

$$F(x) := \mathbf{P}\{\omega \mid \mathbf{x}(\omega) \leq x\}\,; \qquad x \in \mathbb{R}$$

a right-continuous non-decreasing monotonic function.
[◦] The **Expectation** of a random variable is the integral

$$\mathbb{E}\,\mathbf{x} := \int_\Omega \mathbf{x}(\omega)\,\mathbf{P}(d\omega) = \int_{\mathbb{R}} x\,dF(x)\,.$$

Notations

◦ Random variables are denote by **lower case boldface symbols** such as $\mathbf{x}$, $\mathbf{y}$, ... etc. The notation using Upper case symbols like X, Y .. is bad. Upper case symbols are standard for MATRICES such as covariances or loading matrices in linear models. We need to work with *multivariate statistics* and this notation would produce confusion.
◦ The sample size is usually denoted by N: lower case n or m etc. is often used for dimension of vectors (either random or non-random) or degrees of freedom. So in general n is fixed while N may tend to ∞.
◦ Acronyms: PDF means a probability distribution function; $\mathbf{x} \sim F$ means that the random variable $\mathbf{x}$ has probability distribution F. In discrete probability spaces $F(x)$ is a staircase function. Continuous variables admit a **probability density function** (pdf) $p(x) := \dfrac{d\,F(x)}{d\,x}$.

If $\mathbf{P}$ is any probability measure defined on $\mathcal{A}$, the *probability* $P_{\mathbf{y}}$, *induced* by an m-dimensional random vector $\mathbf{y}$, on its sample space $\{\mathbb{R}^m, \mathcal{B}^m\}$ is defined

by the position

$$P_{\mathbf{y}}(E) := \mathbf{P}\{\omega \mid \mathbf{y}(\omega) \in E\}\,; \qquad E \in \mathcal{B}^m \tag{A.1.1}$$

which can be written more economically as,

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

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 $\{y \in \mathbb{R}^m \mid y_1 \leq \eta_1, \ldots, y_m \leq \eta_m\}$ which we shall write symbolically as $\{y \leq \eta\}$. That this is so follows from the fact that all sets in the Borel σ-algebra $\mathcal{B}^m$ are limits of sequences of sets obtained by Boolean operations on such intervals [59].
Hence $P_{\mathbf{y}}$ is uniquely determined by its probability distribution function (PDF) $F : \mathbb{R}^m \to [0,1]$

$$F(x) = F(x_1, \ldots, x_m) = P\{\omega \mid \mathbf{y}_1(\omega) \leq x_1, \ldots, \mathbf{y}_m(\omega) \leq x_m\} \tag{A.1.2}$$

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

$$\tilde{\mathbf{y}} : \mathbb{R}^m \to \mathbb{R}^m \quad , \qquad \tilde{\mathbf{y}}_i\,(x_1, \ldots, x_m) := x_i \quad , \qquad i = 1, \ldots, m \quad , \tag{A.1.3}$$

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

$$P_{\mathbf{y}}\{x \mid \tilde{\mathbf{y}}_1(x) \leq x_1, \ldots, \tilde{\mathbf{y}}_m(x) \leq x_m\}$$

is exactly the same as the original PDF, $F(x)$, of $\mathbf{y}$ defined in (A.1.2). It follows that $\mathbf{y}$ and $\tilde{\mathbf{y}}$ are indistinguishable as the probability of any event $E \in \mathcal{B}^m$ is computed by integrating the PDF's of $\mathbf{y}$ and $\tilde{\mathbf{y}}$ over E and hence must coincide. It follows that $\tilde{\mathbf{y}}$ and $\mathbf{y}$ can be regarded as *the same random variable*.

Convergence of sequences of random variables

There are three standard kinds of convergence:

1. **Almost sure convergence**: is ordinary convergence of functions $\mathbf{x}_N(\omega) \to \mathbf{x}(\omega)$ for all $\omega \in \Omega$, except perhaps a subset of ω's of probability zero. Written $\mathbf{x}_N \overset{a.s.}{\to} \mathbf{x}$
2. **Quadratic mean convergence**: means that $\mathbb{E}\,|\mathbf{x}_N - \mathbf{x}|^2 \to 0$ This is written $\mathbf{x}_N \overset{q.m.}{\to} \mathbf{x}$
3. **Convergence in probability**: $\mathbf{P}\{\omega \mid |\mathbf{x}_N(\omega) - \mathbf{x}(\omega)| > \epsilon\} \to 0$ for all $\epsilon > 0$. This is written $\mathbf{x}_N \overset{\mathbf{P}}{\to} \mathbf{x}$ or $\mathbf{P} - \lim \mathbf{x}_N = \mathbf{x}$.

Implications:

$$\mathbf{1.} \Rightarrow \mathbf{3.} \qquad \mathbf{2.} \Rightarrow \mathbf{3.}$$

The last implication is proven by **Chebyshev inequality :** Suppose $\mathbf{x}$ and $\mathbf{y}$ have finite second moment, then for all $\varepsilon > 0$

$$\mathbf{P}\left\{\,|\mathbf{x} - \mathbf{y}\,|^2 \geq \varepsilon\right\} \leq \frac{1}{\varepsilon^2}\,\mathbb{E}\left[(\mathbf{x} - \mathbf{y})^2\right]$$

Proposition A.1. *Let* $\mathbf{x}_N \overset{q.m.}{\to} \mathbf{x}$ *then* $\mathbf{x}_N \overset{\mathbf{P}}{\to} \mathbf{x}$

Proof: just let $\mathbf{x} \equiv \mathbf{x}_N$ and $\mathbf{y} \equiv \mathbf{x}$.

Theorem A.1 (Weak law of large numbers). *If the random variables* $\{\mathbf{x}_k\,;\, k = 1, 2, \ldots\}$ *are* **independent identically distributed (i.i.d.)** *then*

$$\bar{\mathbf{x}}_N \overset{P}{\to} \mu = \mathbb{E}\,\mathbf{x}_k$$

that is, the sample mean is a (weakly) consistent estimator *of the mean.*

Proof. By Chebyshev inequality

$$\mathbf{P}\left\{\,|\bar{\mathbf{x}}_N - \mu\,|^2 \geq \varepsilon\right\} \leq \frac{1}{\varepsilon^2}\,\mathbb{E}\left[(\bar{\mathbf{x}}_N - \mu)^2\right]\,.$$

The quantity $\mathbb{E}\left[(\bar{\mathbf{x}}_N - \mu)^2\right]$ is the variance of the sample mean:

$$\text{var}\left\{\frac{1}{N}\sum_{k=1}^{N}\mathbf{x}_k\right\} = \frac{1}{N^2}\;\text{var}\left\{\sum_{k=1}^{N}\mathbf{x}_k\right\} = \frac{1}{N^2}\;N\,\sigma^2$$

where $\sigma^2 = \text{var}\,(\mathbf{x}_k)$. Then $\mathbb{E}\left[(\bar{\mathbf{x}}_N - \mu)^2\right] \to 0$ (q.m convergence). Also

$$\mathbf{P}\left\{\,|\bar{\mathbf{x}}_N - \mu\,|^2 \geq \varepsilon\right\} \leq \frac{1}{N}\;\sigma^2 \to 0$$

for all $\varepsilon > 0$ as $N \to \infty$. Hence $\bar{\mathbf{x}}_N \overset{P}{\to} \mu = \mathbb{E}\,\mathbf{x}_k$. □

Generalizations to sequences of random vectors hold by exactly the same arguments. Just substitute the absolute values with norms.

Definition A.1. *A sequence of PDF's* $\{F_N\}$ *(possibly multivariable),* converges in law *to a PDF* F*; notation:* $F_N \overset{L}{\to} F$*, if the functions* $\{F_N(x)\}$*, converge to a PDF* $F(x)$ *at all points* x *where* F *is continuous.*

One also talks abut **convergence in distribution** (or also *in law*) of *random variables*: a sequence $\{\mathbf{x}_N\}$ (maybe vector valued), converges in distribution: $\mathbf{x}_N \overset{L}{\to} \mathbf{x}$ if the PDF's of $\{\mathbf{x}_N\}$ converge in law to the PDF of $\mathbf{x}$.
This is a weaker notion than convergence of random variables as defined above.

WARNING: To talk about convergence of *random variables* $\{\mathbf{x}_N\}$ and $\mathbf{x}$ must be **defined in the same probability space** (the same random experiment). Otherwise $F_N \to F$ does not necessarily mean that $\{\mathbf{x}_N\}$ with $\mathbf{x}_N \sim F_N$ converges to a limit random variable in any reasonable sense.

Theorem A.2. *Convergence in probability implies convergence in distribution.*

Convergence in distribution is weaker than (implied by) convergence in probability *except when the limit is a constant (nonrandom) variable.*
A *degenerate* PDF is

$$F(x) := \mathbf{1}(x - c) = \begin{cases} 1 & \text{if } x \geq c \\ 0 & \text{if } x < c \end{cases}$$

this is the PDF of a constant (nonrandom) variable $\mathbf{x}(\omega) = c$ for all $\omega \in \Omega$.

Theorem A.3. *Convergence in law to a degenerate PDF (that is convergence in distribution to a constant) implies and is hence* equivalent to convergence in probability *to the same constant :*

$$\mathbf{x}_N \xrightarrow{L} c \;\Rightarrow\; \mathbf{x}_N \xrightarrow{P} c$$

whenever c is a (nonrandom) constant.

Theorem A.4 (Weak Convergence). *The sequence of random variables $\{\mathbf{x}_N\}$ converges in distribution to $\mathbf{x}$ if and only if*

$$\mathbb{E}\ f(\mathbf{x}_N) \to \mathbb{E}\ f(\mathbf{x})\,; \quad \textit{that is} \quad \int f(x)\, dF_N(x) \to \int f(x)\, dF(x)$$

for all bounded continuous **real valued** *functions f. In fact for all real valued functions f which are bounded and continuous in a set of probability one for the PDF of $\mathbf{x}$.*

Characteristic Functions

Definition:

$$\phi_{\mathbf{x}}(i\,t) := \int e^{i\,tx}\, dF(x) = \mathbb{E}\ e^{i\,t\,\mathbf{x}}$$

NOTE: The imaginary argument of the exponential here is essential to guarantee boundedness, as $|\,e^{i\,tx}| = 1$.
Therefore *convergence in distribution implies pointwise convergence of the characteristic functions*

$$\phi_{\mathbf{x}_N}(i\,t) := \mathbb{E}\ e^{i\,t\,\mathbf{x}_N} \to \phi_{\mathbf{x}}(i\,t) := \mathbb{E}\ e^{i\,t\,\mathbf{x}}, \quad \text{for all } \ t \in \mathbb{R}\,.$$

Actually this result can be inverted

Theorem A.5 (Levy-Helly Bray). *The convergence of characteristic functions is necessary and sufficient (and hence* equivalent *to) convergence in distribution.*

This is a very useful fact. Used for example in the proof of the CLT.

The moments of a PDF are derivatives of the characteristic function computed at $t = 0$.

$$\phi^{(k)}(i\,t) := i^{\,k} \int x^k\, e^{i\,tx}\, dF(x) \;\Rightarrow\; \phi^{(k)}(0) := i^{\,k} \int x^k\, dF(x) = i^{\,k}\mu_k$$

Convergence $\phi_n(i\,t) \to \phi(i\,t)$ does not necessarily imply convergence of the derivatives at $t = 0$. In general *convergence in law does not imply convergence of the moments* . Means, variances etc.., etc., of a sequence $\{\mathbf{x}_N\} \xrightarrow{L} \mathbf{x}$, do not necessarily converge to means, variances etc.., of the limit.

Theorem A.6 (Billingsley p.32). *Let $\mathbf{x}_N \xrightarrow{L} \mathbf{x}$, and*

$$\sup_N \mathbb{E}\ \mathbf{x}_N^2 < \infty \tag{A.1.4}$$

then all existing moments of $\mathbf{x}_N$ converge to the respective moments of the limit distribution.

A.2 ▪ The χ^2 and related distributions

One says that a scalar random variable $\mathbf{y}$ has a $\chi^2(n)$ distribution if its pdf is supported on the nonnegative real line and has the following strucure

$$P(x \leq \mathbf{y} < x + dx) = \frac{1}{2^{n/2}\,\Gamma\left(\frac{n}{2}\right)}\, x^{\left(\frac{n}{2}\right)-1}\, e^{-x/2}\, dx\,, \qquad x \geq 0\,. \tag{A.2.1}$$

In this expression n is a natural number called *the number of degrees of freedom* of the distribution. One sees that the χ^2 is a special case of the Gamma distribution. Its characteristic function can be easily obtained from ordinary tables of Fourier transforms by just recalling that the sign in the exponential factor of the characteristic function integral is the opposite; yielding

$$\phi(it) := \mathbb{E}\; e^{it\,\mathbf{y}} = (1 - 2it)^{-n/2}\,. \tag{A.2.2}$$

From this expression one can derive formulas for the moments of the distribution. The first few *central* moments are

$$\begin{aligned} \mu_1 &= n \\ \mu_2 &= 2n \\ \mu_3 &= 8n \\ \mu_4 &= 48n + 12n^2 \qquad \text{ecc...} \end{aligned} \tag{A.2.3}$$

Lemma A.1. *For large n a $\chi^2(n)$ random variable tends in distribution to a Gussian variable with pdf $\mathcal{N}(n, 2n)$.*

Proof. Let $\mathbf{y} \sim \chi^2(n)$; introduce a standardized random variable

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

which for all n has mean zero and unit variance. Of course $\mathbf{z}_n$ is no longer a χ^2 (as this could happen only for *Gaussian* random variables!). We shall show that the limit in distribution, $L - \lim_{n\to\infty} \mathbf{z}_n$, is a standard $\mathcal{N}(0,1)$ density. By the Levy-Helly-Bray Theorem the statement follows since the characteristic function, $\phi_n(t)$, of $\mathbf{z}_n$ can be written as,

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

where $\lim_{n\to\infty}\, \psi(n) = 0$. By a well known formula in Analysis the limit $\lim_{n\to\infty}\, \phi_n(t)$ is equal to

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

which is the characteristic function of a standard Gaussian distribution. □

The χ^2 distribution plays a role in many questions of statistical inference, especially entering in the pdf of estimators.

Proposition A.2. *The sum of N independent random variables $\mathbf{y}_i \sim \chi^2(n_i)$ is distributed as $\chi^2(n)$ where*

$$n = \sum_{i=1}^{N} n_i \quad , \tag{A.2.4}$$

that is, when summing i.i.d. χ^2's, the degrees of freedom add up.

Proof. Recall that the pdf of the sum $\sum_1^N \mathbf{y}_i$ of i.i.d. random variables is just the N-fold convolution of the respective p.d.f's, so that the characteristic functions $\phi_i(t)$, of the $\mathbf{y}_i$'s get multiplied together. It is then clear that multiplying functions like (A.2.2) the exponents at the denominators must add up. □

The following is a partial converse of this statement.

Proposition A.3. *Let $\mathbf{y} = \mathbf{y}_1 + \mathbf{y}_2$ be the sum of two independent random variables. Assume that $\mathbf{y} \sim \chi^2(n)$ and $\mathbf{y}_2 \sim \chi^2(n_2)$ where $n > n_2$. Then $\mathbf{y}_1 \sim \chi^2(n - n_2)$.*

Proof. By independence the characteristic function of $\mathbf{y}$ is $\phi = \phi_1\phi_2$ so that

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

and by substituting the relative expressions (A.2.2) one sees that the statement must be true. □

Proposition A.4. *The pdf of the random variable*

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

where $\mathbf{y}_i \sim \mathcal{N}(\mu, \sigma^2)$ i.i.d. is $\chi^2(n)$.

Proof. Just note that, with $\mathbf{y} \sim \mathcal{N}(\mu, \sigma)$, the pdf of $\mathbf{z} := (\mathbf{y} - \mu)^2/\sigma^2$ is $\chi^2(1)$ and then use Proposition A.2. Note also that $\mathbf{z} = \mathbf{x}^2$ with $\mathbf{x} \sim \mathcal{N}(0, 1)$. Using the well-known rules for the pdf of a function of random variable, say $z = f(x)$ with $f(x) = x^2$, one obtains

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

which is indeed $\chi^2(1)$. □

Proposition A.5. *Let $\mathbf{y}_i \sim \mathcal{N}(\mu, \sigma^2)$, $i = 1, \dots, n$, i.i.d. Then the pdf of the normalized sample variance:*

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

is $\chi^2(n-1)$.

Proof. This is a consequence of the following remarkable result:

Lemma A.2. *Under the above hypotheses, the statistics $\bar{\mathbf{y}}_n$ and $\hat{\boldsymbol{\sigma}}_n^2$ are independent.*

Proof. We just need to show that $\bar{\mathbf{y}}_n$ and $\mathbf{y}_i - \bar{\mathbf{y}}_n$ are uncorrelated for all i's. By Gaussianity, this will imply independence.

Define $\tilde{\mathbf{y}}_i = \mathbf{y}_i - \mu$ and $\tilde{\mathbf{y}} = \bar{\mathbf{y}}_n - \mu$, so that $\mathbf{y}_i - \bar{\mathbf{y}}_n = \tilde{\mathbf{y}}_i - \tilde{\mathbf{y}}$ ed $\mathbb{E}\ \bar{\mathbf{y}}_n(\mathbf{y}_i - \bar{\mathbf{y}}_n) = \mathbb{E}\ \tilde{\mathbf{y}}(\tilde{\mathbf{y}}_i - \tilde{\mathbf{y}}) = \mathbb{E}\,(\tilde{\mathbf{y}}\tilde{\mathbf{y}}_i) - \mathbb{E}\,(\tilde{\mathbf{y}})^2$. Independence of the variables $\mathbf{y}_i$ implies

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

so that, comparing with $\mathbb{E}\,(\tilde{\mathbf{y}})^2 = \sigma^2/n$, one gets the conclusion. □

By the usual identity

$$\sum_1^n (\mathbf{y}_i - \mu)^2 = \sum_1^n (\mathbf{y}_i - \bar{\mathbf{y}}_n)^2 + n(\bar{\mathbf{y}}_n - \mu)^2 \tag{A.2.5}$$

one has

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

where the two random variables in the right member are *independent*. We know from Proposition A.4 that $n\bar{S}^2/\sigma^2 \sim \chi^2(n)$ and that $(\bar{\mathbf{y}}_n - \mu)^2/(\sigma^2/n) \sim \chi^2(1)$ (which also follows from Proposition A.4 with $n = 1$). By Proposition A.3 the pdf of first summand in the second member must be $\chi^2(n-1)$. □

So far we have been discussing the case of scalar variables. Suppose $\mathbf{y}$ is an m-dimensional random vector. We are interested to find out when the pdf of quadratic forms like $\mathbf{y}^\top Q\mathbf{y}$ with $Q = Q^\top$, is χ^2. The most obvious situation in which this happens is the following.

Proposition A.6. *Let $\mathbf{y} \sim \mathcal{N}(\mu, \Sigma)$ with $\mu \in \mathbb{R}^m$ and $\Sigma \in \mathbb{R}^{m\times m}$ positive definie; then*

$$(\mathbf{y} - \mu)^\top\ \Sigma^{-1}(\mathbf{y} - \mu) \sim \chi^2(m)\,. \tag{A.2.6}$$

Proof. One just needs to standardize $\mathbf{y}$, by setting $\mathbf{z} := \Sigma^{-1/2}(\mathbf{y} - \mu)$; so that $\mathbf{z} = [z_1, \dots, \mathbf{z}_m]^\top$ is $\mathcal{N}(0, I)$, in particular $\mathbf{z}_1, \dots, \mathbf{z}_m$ are i.i.d. and $\mathcal{N}(0, 1)$. It follows that

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

and the last member is $\chi^2(m)$ by Proposition A.2. □

A less obvious characterization which is used frequently is the following.

Proposition A.7. *Let $\mathbf{z} \sim \mathcal{N}(0, I_m)$ and $Q \in \mathbb{R}^{m\times m}$. Then the quadratic form $\mathbf{z}^\top Q\mathbf{z}$ is χ^2 distributed if and only if Q is idempotent; i.e. $Q = Q^2$. In this case the number of degrees of freedom is equal to $r = \operatorname{rank} Q$.*

Proof. The proof is based on diagonalization of Q. Indeed since Q is symmetric (and can always be assumed to be such) and idempotent, it is really an orthogonal projection in $\mathbb{R}^m$. Its non-zero eigenvalues are all equal to 1 and there are exactly $r = \operatorname{rank} Q$ of them. The spectral decomposition of Q can therefore be written

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

that is

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

where U_1 is an $m \times r$ matrix formed by the first r (orthonormal) columns of U. Hence

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

where the r-dimensional random vector $\mathbf{z}_1 := U_1^\top \mathbf{z}$ is distributed as $\mathcal{N}(0, I_r)$. Proposition A.2 then yields the conclusion. □

The Maxwell-Boltzmann Distribution

One may wonder about the origin of the term "degrees of freedom" of the χ^2 distribution. It seems to be related physics, in particular to the famous *Maxwell–Boltzmann distribution* which is a cornerstone of the kinetic theory of gases, which provides an explanation of many fundamental gaseous properties, including pressure and diffusion. The distribution was first derived by Maxwell in 1860 on heuristic grounds. Ludwig Boltzmann later, in the 1870s, carried out precise investigations into the physical origins of this distribution. The derivation using the properties of the χ^2 distribution is almost immediate.We shall roughly follow [60].

The Maxwell–Boltzmann distribution applies to the *magnitude v of the velocity* of the particles. The actual speed of a particle selected at random in an ensemble of rarified ideal gas particles is the effect of a very large number of impacts from neighbouring particles and can be described with excellent approximation as a random variable having a Gaussian distribution.

One makes the assumption that the three Cartesian components v_1, v_2, v_3 of the (vector) velocity of a particle are independent random variables each having the same Gaussian distribution

$$p(v_i) = (2\pi\, kT/m)^{-1/2} \exp -\frac{mv_i^2}{2kT}$$

where $\dfrac{kT}{m} \equiv \sigma^2$ is the squared most probable velocity. Hence $\mathbf{v}_i/\sigma \sim \mathcal{N}(0, 1)$

and

$$\frac{\mathbf{v}^2}{\sigma^2} := \sum_{i=1}^{3} (\mathbf{v}_i/\sigma)^2 \sim \chi^2(3)\,.$$

By the change of variable $x = \frac{v^2}{\sigma^2}$ in the expression of $\chi^2(3)$ we obtain

$$p(v) = 4\pi(m/2\pi kT)^{3/2} v^2 \exp -\frac{mv^2}{2kT}$$

which is the Maxwell-Boltzmann distribution. Note that $\mathbf{u} := \frac{1}{2}m\mathbf{v}^2 = \frac{1}{2}kT\frac{\mathbf{v}^2}{\sigma^2}$ is the kynetic energy of a particle and from the mean of the $\chi^2(3)$ distribution we immediately get the relation

$$\mathbb{E}\,\frac{1}{2}m\mathbf{v}^2 = \frac{3}{2}kT$$

which is a well-known relation in the kinetic theory of gases.

The Student distribution

Let $\mathbf{y} \sim \mathcal{N}(0\,,\,1)$ and $\mathbf{x} \sim \chi^2(n)$ be independent. Then the ratio

$$\mathbf{t} := \frac{\mathbf{y}}{\sqrt{\mathbf{x}/n}} \tag{A.2.7}$$

has the pdf

$$p_n(t) = \frac{1}{\sqrt{n}\,B(1/2\,,\,n/2)} \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}} \qquad t \in \mathbb{R} \tag{A.2.8}$$

called a *Student distribution with n degreesof freedom*, which we shall denote by the symbol $\mathcal{S}(n)$. In (A.2.8) B is the Euler Beta function:

$$B(p\,,\,q) := \int_0^1 x^{p-1}(1-x)^{q-1}\,dx \;= \frac{\Gamma(p)\,\Gamma(q)}{\Gamma(p+q)}$$

where the function Γ is the well-known generalization of the factorial. When n is an integer greater than 1, $\Gamma(n) = (n-1)!$.

The Student pdf has a curious history which is reported in all textbooks of classical Statistics see e.g. **[?]**. For $n = 1$ it reduces to the Cauchy distribution

$$\mathcal{S}(1) \equiv \frac{1}{\pi(1+t^2)}\,.$$

It can be shown that $\mathcal{S}(n)$ has finite moments only up to order $n-1$; given by the formulas

$$\begin{aligned} \mu_r &= \quad 0 \quad \text{if } r \text{ is odd and } \; r < n \\ \mu_r &= \frac{\Gamma(\frac{1}{2}n - r)\Gamma(r + \frac{1}{2})}{\Gamma(\frac{1}{2}n)\Gamma(\frac{1}{2})} \quad \text{is } r \text{ is even and } \; 2r < n\,. \end{aligned}$$

It is also not hard to show that for $n \to \infty$ the distribution $\mathcal{S}(n)$ converges to $\mathcal{N}(0\,,\,1)$.

The F distribution

Let $\mathbf{x}_1 \sim \chi^2(n_1)$ and $\mathbf{x}_2 \sim \chi^2(n_2)$ be independent. Then the ratio

$$\mathbf{z} := \frac{\mathbf{x}_1/n_1}{\mathbf{x}_2/n_2} \tag{A.2.9}$$

has the pdf

$$p_{n_1,\,n_2}(z) = \left[\frac{\Gamma(\frac{n_1+n_2}{2})}{\Gamma(\frac{n_1}{2})+\Gamma(\frac{n_2}{2})}\right](\frac{n_1}{n_2})^{\frac{n_1}{2}}\,\frac{z^{\frac{n_1}{2}-1}}{\left(1+\frac{n_1}{n_2}z\right)^{\frac{n_1+n_2}{2}}} \qquad z\in\mathbb{R}_+ \tag{A.2.10}$$

called F *distribution* with n_1 and n_2 degrees of freedom. It is denoted by the symbol $\mathcal{F}(n_1,\,n_2)$.
The derivation of the expression (A.2.9) can be found in many statistical textbooks; see e.g. [46]. Together with the Gaussian, the F distribution is perhaps one of the most important pdf's in classical Statistics. Tables of the F distribution can be found in many websites; e.g.

For $n_1 = 1$ the random varible $\mathbf{z}$ in (A.2.9) is the square, $\mathbf{t}^2$, of a Student random variable with n_2 degrees of freedom. The mean μ and the mode, m, of $\mathcal{F}(n_1,\,n_2)$ exist only if n_1 and n_2 are strictly greater than 1 and are given by the formulas:

$$\mu = \frac{n_2}{n_2-2}\,, \qquad m = \frac{n_2(n_1-2)}{n_1(n_2+2)}\,.$$

It can be shown that

$$L-\lim_{n_2\to\infty} n_1\mathbf{z} = \chi^2(n_1)\,, \tag{A.2.11}$$

(limit in distribution). Moreover if $\mathbf{z}\sim\mathcal{F}(n_1,\,n_2)$ and $a := a(n_1,\,n_2)$ is defined as

$$\mathbb{P}(\mathbf{z}\geq a) = \alpha$$

then the abscissa b for which

$$\mathbb{P}(\mathbf{z}\leq b) = \alpha$$

is the reciprocal of a, computed by exchanging the degrees of freedom in $\mathcal{F}(n_1,\,n_2)$i; i.e.

$$b(n_1\,,\,n_2) = a(n_2\,,\,n_1)\,.$$

See for example `http://econtools.com/jevons/java/Graphics2D/FDist.html`.

A.3 ▪ Stationarity and Ergodicity

Let us pretend that we have an infinite sequence of observations indexed by (discrete) time, extending from $t = -\infty$ (the infinite past) to the infinite future $t = +\infty$. This is called a (discrete-time) **stochastic process** denoted

$$\mathbf{y} = \{\mathbf{y}(t)\}, \qquad t \in \mathbb{Z}$$

the symbol $\mathbb{Z}$ (Zhalen in German) stands for the set of integer numbers.

Definition A.2. *A stochastic process $\{\mathbf{y}(t)\}$ is* **stationary** *(in the strict sense) if all PDF's relative to $\mathbf{y}(t_1)$, $\mathbf{y}(t_2), \dots \mathbf{y}(t_n)$ say $F_n(x_1, \dots, x_n, t_1, \dots, t_n)$ are invariant for temporal translation, that is for every n it must hold that,*

$$F_n(x_1, \dots, x_n, t_1 + \Delta, \dots, t_n + \Delta) = F_n(x_1, \dots, x_n, t_1, \dots, t_n) \quad ,$$

(same function of $x_1, \dots, x_n$, $t_1, \dots, t_n$), whatever the time shift $\Delta \in \mathbb{Z}$.

Consequences:

- The PDF $F(x, t)$ of any variable $\mathbf{y}(t)$ cannot depend on t; that is the random variables $\mathbf{y}(t)$, $t \in \mathbb{Z}$, are *identically distributed*;
- The *second order* joint Pdf $F_2(x_1, x_2, t_1, t_2)$ of the variables$\mathbf{y}(t_1)$, $\mathbf{y}(t_2)$, only depends on the difference $\tau = t_1 - t_2$ and not on the date.
  In particular, $\mu(t) := \mathbb{E}\ \mathbf{y}(t)$, is a constant equal to $\mu \in \mathbb{R}^m$ and the Covariance function:

  $$\Sigma(t_1, t_2) := \mathbb{E}\ [\mathbf{y}(t_1) - \mu(t_1)]\ [\mathbf{y}(t_2) - \mu(t_2)]^\top$$

  depends only on the difference $\tau = t_1 - t_2$.

Wide sense stationarity just requires that the covariance function should depend on the difference of the arguments; i.e. on $\tau = t_1 - t_2$. This is clearly a less demanding condition which is often assumed in applications.

The Ergodic Theorem

Let $f(\mathbf{y})$ denote a statistic, function of any number of random variables of the process, *which does not depend on time*. Denote by $f_k(\mathbf{y})$ the same function in which all time indices of these variables are shifted by k units.

Theorem A.7 (Birkhoff Ergodic Theorem). *Let $\{\mathbf{y}(t)\}$ be a strictly stationary process. The limit*

$$\bar{\mathbf{z}} := \lim_{T \to \infty} \frac{1}{T} \sum_{k=1}^{T} f_k(\mathbf{y}) \tag{A.3.1}$$

exists with probability one for all functions f such that $\mathbb{E}\,|f(\mathbf{y})| < \infty$

The limit can either be random or constant. If it is random it must be a "very special" random variable. These are called *invariant random variables*. We shall not investigate them.

If the limit is a constant then the process is called **Ergodic**.

Note now that for all $T > 0$,

$$\mathbb{E}\,\{\frac{1}{T}\sum_{k=1}^{T} f_k(\mathbf{y})\} = \frac{1}{T}\sum_{k=1}^{T}\mathbb{E}\, f_k(\mathbf{y}) = \mathbb{E}\, f(\mathbf{y})$$

since $\mathbf{z}(k) = f_k(\mathbf{y})$ is itself a strictly stationary process. Hence the expectation of the time average in the second members of (A.3.1) is constant and hence converges as $T \to \infty$ and one finds

$$\mathbb{E}\,\bar{\mathbf{z}} = \mathbb{E}\, f(\mathbf{y})\,.$$

Corollary A.1. *If $\{\mathbf{y}(t)\}$ is ergodic*

$$\lim_{T\to\infty}\ \frac{1}{T}\ \sum_{k=1}^{T}\ f_k(\mathbf{y}) = \mathbb{E}\ f(\mathbf{y}) \tag{A.3.2}$$

with probability one whatever may be $f(\mathbf{y})$ having finite expectation.

Proof: In fact $\bar{\mathbf{z}}$ must be a constant and hence coincides with its own expectation $\bar{\mathbf{z}} = \mathbb{E}\ \bar{\mathbf{z}} = \mathbb{E}\ f(\mathbf{y})$. □

Let $\mathbf{y}$ be an ergodic process and $\mathbf{z}(t) := f_t(\mathbf{y})$ a sequence of translates having finite expectation. Then it is not difficult to check that the process $\{\mathbf{z}(t)\}$ is itself stationary and ergodic.

Proposition A.8. *An ergodic process cannot admit limit for $t \to \pm\infty$ unless it reduces to a deterministic sequence (with probability 1).*

In fact such a limit should be a constant random variable.

The strong law of large numbers

This is a special case of ergodicity.

Theorem A.8 (Kolmogorov). *Every i.i.d. process having finite expectation is ergodic.*

The following is an important consequence.

Corollary A.2. *Let $\mathbf{e}$ be a i.i.d. process, $\mathbf{z}(0) := f(\mathbf{e})$ a function of the process, possibly of infinitely many variables, having finite expectation and $\mathbf{z}(t) := f_t(\mathbf{e})$ be the same function of translates by t units of time; i.e. $\mathbf{e}(k) \to \mathbf{e}(t+k)$. Then the process $\{\mathbf{z}(t)\}$ is stationary and ergodic. In other words, the time translates of every time-invariant function of an i.i.d. process form an ergodic process.*

For example if $\mathbf{e}$ is i.i.d. of finite variance and $\sum_{-\infty}^{+\infty} |c_k|^2 < \infty$, the time-translated random variables of $\mathbf{z}(0) := \sum c_k \mathbf{e}(k)$, namely

$$\mathbf{z}(t) := \sum_{-\infty}^{+\infty} c_k\, \mathbf{e}(t+k) = \sum_{-\infty}^{+\infty} c_{-k}\, \mathbf{e}(t-k)\,; \qquad t \in \mathbb{Z} \tag{A.3.3}$$

form an ergodic process.
Convergence of the sum follows from Cauchy-Schwartz inequality

$$\mathbb{E}\,|\sum_{-N}^{+N} c_k\,\mathbf{e}(t+k)| \le \sum_{-N}^{+N} |c_k|^2\;\mathbb{E}\,|\mathbf{e}(t+k)|^2 = \sum_{-N}^{+N} |c_k|^2\,\sigma_{\mathbf{e}}^2\,.$$

A.4 ▪ Stationary random oscillations

It will initially be convenient to work with complex-valued random variables and processes. The covariance of two zero-mean complex variables $\mathbf{x}$ and $\mathbf{z}$ is $\mathbb{E}\,\mathbf{x}\bar{\mathbf{z}}$ where the bar denotes complex conjugate. With this definition the variance norm will be positive. A *quasi periodic complex process* is the sum of ν elementary complex random harmonic oscillations,

$$\mathbf{z}(t) = \sum_{k=1}^{\nu} \mathbf{z}_k e^{j\omega_k t}\,, \qquad t \in \mathbb{Z}$$

where $\omega_k \in [-\pi,\,\pi]$ are real angular frequencies which can be assumed distinct and the $\mathbf{z}_k,\ k = 1,\dots,\nu$ are (complex) random variables having finite variance. We want $\{\mathbf{z}(t)\}$ to be a *stationary process*.This can be true if and only if the correlations of the harmonic components

$$\mathbb{E}\,[\mathbf{z}_k\bar{\mathbf{z}}_h]\,e^{j\omega_k t - j\omega_h s} \qquad k,h = 1,2,\dots,\nu$$

depend on $t-s$, which can happen only if $k = h$ while, for $k \neq h$ one must necessarily have $\mathbb{E}\,[\mathbf{z}_k\bar{\mathbf{z}}_h] = 0$ hence, for stationarity the variables $\{\mathbf{z}_k\}$ must be uncorrelated. Then

$$r(t,s) = \mathbb{E}\,\mathbf{z}(t)\bar{\mathbf{z}}(s) = \sum_{k=1}^{\nu} \sigma_k^2 e^{j\omega_k(t-s)}, \qquad \sigma_k^2 = \mathbb{E}\,|\mathbf{z}_k|^2$$

that is $r(t,s) = r(t-s)$ and $\{\mathbf{z}(t)\}$ is (wide sense) stationary.

Assume now that the process is *real* ; then the harmonic components must appear in complex conjugate pairs

$$\mathbf{z}(t) = \frac{\mathbf{z}(t)+\bar{\mathbf{z}}(t)}{2} = \sum_{k=-\nu}^{\nu} \frac{1}{2}\mathbf{z}_k e^{j\omega_k t}\,, \qquad \omega_{-k} = -\omega_k \qquad \mathbf{z}_{-k} = \bar{\mathbf{z}}_k$$

where the $\{\mathbf{z}_k\}$ must be uncorrelated. The term with $k = 0$ is a zero frequency ($\omega_0 = 0$) term which is real.
Write $\mathbf{z}_k = \mathbf{x}_k + i\mathbf{y}_k$, and use negative indexing ($-k$) to denote the conjugate $\mathbf{z}_{-k} = \mathbf{x}_k - i\mathbf{y}_k$. The orthogonality of random coefficients with different indices implies

$$\mathbb{E}\,\{(\mathbf{x}_k + i\mathbf{y}_k)\overline{(\mathbf{x}_k - i\mathbf{y}_k)}\} = \mathbb{E}\,\{(\mathbf{x}_k^2 - \mathbf{y}_k^2) + 2i\mathbf{x}_k\mathbf{y}_k\} = 0$$

That is

$$\mathbb{E}\,\mathbf{x}_k^2 = \mathbb{E}\,\mathbf{y}_k^2\,, \qquad \mathbb{E}\,\mathbf{x}_k\mathbf{y}_k = 0\,.$$

Now write each harmonic component in real form

$$\mathbf{z}_k(t) := \frac{1}{2}\{\mathbf{z}_k e^{j\omega_k t} + \bar{\mathbf{z}}_k e^{-j\omega_k t}\} = \mathbf{x}_k \cos\omega_k t - \mathbf{y}_k \sin\omega_k t \qquad k = 1, \dots, \nu.$$

so that $\sigma_k^2 := \mathbb{E}\ \mathbf{z}_k(t)^2 = \mathbb{E}\ \mathbf{z}_k(0)^2 = \mathbb{E}\ \mathbf{x}_k^2 = \mathbb{E}\ \mathbf{y}_k^2$. The signal can be represented by a vector difference equation of the form

$$\begin{bmatrix}\mathbf{x}_k(t+1)\\ \mathbf{y}_k(t+1)\end{bmatrix} = \begin{bmatrix}\cos\omega_k & -\sin\omega_k\\ \sin\omega_k & \cos\omega_k\end{bmatrix}\begin{bmatrix}\mathbf{x}_k(t)\\ \mathbf{y}_k(t)\end{bmatrix}$$
$$\mathbf{z}_k(t) = \begin{bmatrix}1 & 0\end{bmatrix}\begin{bmatrix}\mathbf{x}_k(t)\\ \mathbf{y}_k(t)\end{bmatrix}$$

with uncorrelated initial conditions $\mathbf{x}_k(0) = \mathbf{x}_k,\ \mathbf{y}_k(0) = \mathbf{y}_k$ of equal variance~σ_k^2. By stacking the models of the $\nu+1$ elementary real components described above $\mathbf{z}(t)$ can be represented as

$$\mathbf{s}(t+1) = A\,\mathbf{s}(t) \tag{A.4.1}$$
$$\mathbf{z}(t) = c^\top \mathbf{s}(t) \tag{A.4.2}$$

wher $\mathbf{s}(t)$ is a state vector of dimension $2\nu + 1$ obtained by listing in a single column vector the elementary components $\begin{bmatrix}\mathbf{x}_k(t+1) & \mathbf{y}_k(t+1)\end{bmatrix}^\top$ for $k = 0, 1, 2, \dots, \nu$ and $c^\top = \begin{bmatrix}1 & 0 & \dots & 1 & 0\end{bmatrix}^\top$.

The zero-frequency component $\mathbf{z}_0(t) \equiv \mathbf{z}_0(0) \equiv \mathbf{z}_0$ is just a constant random variable which can normally be eliminated. In this case the matrix A is block-diagonal $A = \operatorname{diag}\{A_1, \dots A_\nu\}$ with all A_k of dimension 2×2, being orthogonal matrices. Because of uncorrelation of the various 2-dimensional components the variance matrix of $\mathbf{s}(t)$ is diagonal:

$$\Sigma = \mathbb{E}\,\mathbf{s}(0)\mathbf{s}(0)^\top = \operatorname{diag}\{\sigma_1^2 I_2, \dots, \sigma_\nu^2 I_2\}$$

and the covariance function can then be expressed as

$$\sigma(\tau) = c^\top A^\tau \Sigma c = \sum_{k=1}^{\nu} \sigma_k^2 \cos\omega_k\tau\,. \tag{A.4.3}$$

Remarks A.1. The 2-dimensional representations for possible components which have frequency $\omega_k = \pm\pi$ are clearly redundant. These components can obviously be represented by 1-dimensional models.

Non-Ergodicity of quasi-periodic signals

Assume we are observing one single trajectory of an elementary harmonic signal, which can obviously be written also as

$$z_k(t) = A_k \sin(\omega_k t + \varphi_k) \quad A_k^2 = x_k^2 + y_k^2, \quad \tan\varphi_k = \frac{y_k}{x_k}\,.$$

Because of stationarity the limit of the sample covariance exists (in fact irrespective of the probability distributions of $\mathbf{x}_k, \mathbf{y}_k$). We have

Theorem A.9.

$$\lim_{N\to\infty} \frac{1}{N}\sum_{t=1}^{N} z_k(t+\tau)z_k(t) = A_k^2 \cos\omega_k\tau \tag{A.4.4}$$

and, if all frequencies are different and non zero nor $\pm\pi$*,*

$$\lim_{N\to\infty} \frac{1}{N}\sum_{t=1}^{N} z(t+\tau)z(t) = \sum_{k=1}^{\nu} A_k^2 \cos\omega_k\tau\,.$$

Proof. Using the following trigonometric identity [25]

$$2\sin(\omega t{+}\omega\tau{+}\varphi)\sin(\omega t{+}\varphi) = \cos\omega\tau{-}\cos(2\omega t{+}\omega\tau{+}2\varphi) := \cos\omega\tau{-}\cos(2\omega t{+}\psi)\,.$$

where ψ does not depend on t. Now $\cos(2\omega t+\psi) = \Re e\left[e^{i2\omega t}e^{i\psi}\right]$ so that

$$\frac{1}{N}\sum_{t=1}^{N}\cos(2\omega t+\psi) = \frac{1}{N}\Re e\left[e^{i\psi}e^{i\omega}\frac{1-e^{i2\omega N}}{1-e^{i2\omega}}\right]$$

but the absolute value of the term between square brackets can be bounded by a constant independent of N so that the limit for $N\to\infty$ is zero.
The second relation follows from (A.4.4) and can be proven by imposing the uncorrelation of the components; see [90, p. 108]. □

This proves that a random quasi periodic process is **not ergodic** as the limit of the sample correlation of each component depends on the amplitude of the sample trajectory since it is $\frac{A_k^2}{2}\cos\omega_k\tau$ while the true correlation function is $r(\tau) = \sigma_k^2\cos\omega_k\tau$.

[25] From e.g. `https://en.wikipedia.org/wiki/List_of_trigonometric_identities`

Appendix B

SOME FACTS FROM MATRIX ALGEBRA

B.1 ▪ Inner products and Adjoints in finite-dimensional Vector Spaces

In the following we shall mostly work with complex Vector Spaces. The bar will denote complex conjugation.

Definition B.1. *A square matrix $A \in \mathbb{C}^{n\times n}$ is* **Hermitian** *if*

$$\bar{A}^\top = A$$

and **positive semidefinite** *if $\bar{x}^\top Ax \geq 0$ for all $x \in \mathbb{C}^n$. The matrix is called* **positive definite** *if $\bar{x}^\top Ax$ can be zero only when $x = 0$.*

There are well-known tests of positive definiteness, such as the *Sylvester's criterion* based on checking the signs of the determinants of the principal minors. They should all be positive for positive definiteness.

Given an Hermitian positive definite matrix Q we define the *weighted* inner product $\langle \cdot, \cdot \rangle_Q$ in the coordinate space $\mathbb{C}^n$ by setting

$$\langle x,\, y\rangle_Q := \bar{x}^\top Q y$$

This clearly satisfies the axioms of inner product.

Problem B.1. *Show that any inner product in $\mathbb{C}^n$ must have this structure for a suitable Q. Is Q uniquely defined ?*

Consider a linear map $A : \mathscr{X} \to \mathscr{Y}$, both finite-dimensional vector spaces endowed with inner products $\langle \cdot, \cdot \rangle_{\mathscr{X}}$ and $\langle \cdot, \cdot \rangle_{\mathscr{Y}}$ respectively.

Definition B.2. *The* adjoint, *of $A : \mathscr{X} \to \mathscr{Y}$ is a linear map $A^* : \mathscr{Y} \to \mathscr{X}$, defined by the relation*

$$\langle y, Ax\rangle_{\mathscr{Y}} = \langle A^* y, x\rangle_{\mathscr{X}}, \qquad x \in \mathscr{X},\ y \in \mathscr{Y} \tag{B.1.1}$$

Problem B.2. *Prove that A^* is well-defined by the condition* (B.1.1).

Hint: here you may use the fact that $\mathscr{X}$ and $\mathscr{Y}$ are finite-dimensional.
Example: Let $A : \mathbb{C}^n \to \mathbb{C}^m$ where the spaces are equipped with weighted inner products, say

$$\langle x_1, x_2\rangle_{|\mathbb{C}^n} = \bar{x}_1^\top Q_1 x_2, \qquad \langle y_1, y_2\rangle_{|\mathbb{C}^m} = \bar{y}_1^\top Q_2 y_2$$

where Q_1, Q_2 are Hermitian positive definite matrices. Then we have

Proposition B.1. *The adjoint of the linear map $A : \mathscr{X} \to \mathscr{Y}$ defined by a matrix $A : \mathbb{C}^n \to \mathbb{C}^m$ with weighted inner products as above, is*

$$A^* = Q_1^{-1}\bar{A}^\top Q_2 \tag{B.1.2}$$

where $\bar{A}^\top$ is the conjugate transpose of A.

Problem B.3. *Prove proposition B.1.*

Let $A : \mathbb{C}^n \to \mathbb{C}^m$ and assume that Q_1 and $Q_2 = I$ are both identity matrices. Both inner products in this case are Euclidean inner products. Then

$$A^* = \bar{A}^\top$$

i.e. the adjoint is the Hermitian conjugate. In particular, *for a real matrix the adjoint is just the transpose*. For any square Hermitian matrix the adjoint coincides with the original matrix. The linear map defined by the matrix is then called a *self-adjoint operator*. In the real case self-adjoint operators are represented by symmetric matrices. Note that all this is true only if the inner products are Euclidean.

Completing the square in n dimensions

This is one of those things that are re-derive 100 times in the literature, and so we are posting it here for ease of reference. It applies in particular to the solution formula of quadratic algebraic equations.

Take a quadratic polynomial for a vector $\mathbf{x}$,

$$a + \mathbf{b}^\top \mathbf{x} + \frac{1}{2}\mathbf{x}^\top C\mathbf{x}.$$

You want to convert this into the form

$$\frac{1}{2}(\mathbf{x} - \mathbf{m})^\top M(\mathbf{x} - \mathbf{m}) + v.$$

What are $M, \mathbf{m}$, and v?

Assume C is symmetric and non singular. Then we have, in decreasing order of obviousness,

$$M = C\,, \quad \mathbf{m} = -C^{-1}\mathbf{b}\,, \quad v = a - \frac{1}{2}\mathbf{b}^T C^{-1}\mathbf{b}\,.$$

B.2 ▪ The Singular value decomposition (SVD)

We shall first do the SVD for real matrices. In Section B.2 we shall generalize to general linear maps in inner product spaces.

Problem B.4. *Let $A \in \mathbb{R}^{m\times n}$ and $r := \min\{n, m\}$. Show that $AA^\top$ and $A^\top A$ share the first r eigenvalues. How are the eigenvectors related?*

Theorem B.1. *Let $A \in \mathbb{R}^{m\times n}$ of rank $r \le \min(m, n)$. One can find two orthogonal matrices $U \in \mathbb{R}^{m\times m}$ and $V \in \mathbb{R}^{n\times n}$ and positive numbers $\{\sigma_1 \ge, \dots, \ge \sigma_r\}$, the* **singular values** *of A, such that*

$$A = U\Delta V^\top \qquad \Delta = \begin{bmatrix}\Sigma & 0\\ 0 & 0\end{bmatrix}, \qquad \Sigma = \operatorname{diag}\{\sigma_1, \dots, \sigma_r\} \tag{B.2.1}$$

Let $U = \begin{bmatrix}U_r & \tilde{U}_r\end{bmatrix}$, $V = \begin{bmatrix}V_r & \tilde{V}_r\end{bmatrix}$ where the submatrices U_r, V_r keep only the first r columns of U, V. We get a *Full-rank factorization* of A

$$A = U_r\,\Sigma\,V_r \;=\; [u_1, \dots, u_r]\,\Sigma\,[v_1, \dots, v_r]^\top$$

where

$$U_r^\top U_r = I_r = V_r^\top V_r, \quad \text{but} \quad U_r U_r^\top \neq I_m, \quad V_r V_r^\top \neq I_n\,.$$

The proof is based on eigenvalue-eigenvector decomposition of the symmetric matrices $AA^\top$ and $A^\top A$. See the next section for the full proof. Here we just do a verification. Assume that (B.2.1) holds. Then

$$AA^\top = U\Delta^2 U^\top; \qquad A^\top A = V\Delta^2 V^\top$$

hence $U = [u_1, \dots, u_m]$ = normalized eigenvectors of $AA^\top$;
and $V := [v_1, \dots, v_n]$ = normalized eigenvectors of $A^\top A$
while $\{\sigma_1^2 \ge, \dots, \ge \sigma_r^2\}$ are the (non zero) eigenvalues of $AA^\top$ (or of $A^\top A$). Since

$$Ax = U\begin{bmatrix}\Sigma & 0\\ 0 & 0\end{bmatrix}\begin{bmatrix}V_r^\top\\ \tilde{V}_r^\top\end{bmatrix}x = U_r\,\Sigma\,V_r^\top x$$

where $\Sigma > 0$, we have $Ax = 0 \Leftrightarrow V_r^\top x = 0 \Leftrightarrow x \in \operatorname{span}\{\tilde{V}_r\}$. Hence we obtain the *dyad formulas*

$$Ax = \sum_{k=1}^{r} u_k\,\sigma_k\langle v_k\,,\,x\rangle\,, \qquad A^\top y = \sum_{k=1}^{r} v_k\,\sigma_k\langle u_k\,,\,y\rangle$$

In particular A acts on the singular vectors like multiplication by a rank one matrix

$$A\,v_j = \sum_{k=1}^{r} u_k\,\sigma_k\langle v_k\,,\,v_j\rangle = \sigma_j\,u_j\,, \quad A^\top u_j = \sum_{k=1}^{r} v_k\,\sigma_k\langle u_k\,,\,u_j\rangle = \sigma_j\,v_j \tag{B.2.2}$$

Hence the SVD can be seen as a far reaching generalization of the spectral decomposition of symmetric matrices.

Useful Features of the SVD

Range and Nullspace of A:

$$\begin{aligned}\operatorname{Im}(A) = \operatorname{Im}(U_r) &= \operatorname{span}([u_1, \dots, u_r]), & [\operatorname{Im}(A)]^\perp &= \operatorname{Im}(\tilde{U}_r)\\ \ker(A) = [\operatorname{Im}(V_r)]^\perp &= \operatorname{span}([v_{r+1}, \dots, v_n]) & [\ker(A)]^\perp &= \operatorname{Im}(V_r)\end{aligned}$$

Approximation properties: **the best approximant of rank** k of A is the matrix

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

with approximation errors

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

$$\min_{\text{rank }(B)=k} \|A - B\|_F^2 = \|A - A_k\|_F^2 = \sigma_{k+1}^2 + \ldots + \sigma_r^2$$

Matrix Norms

Let $A \in \mathbb{R}^{n\times m}$. For now $\mathbb{R}^m$ and $\mathbb{R}^n$ are equipped with the inner product $\langle u, v\rangle := u^\top v$ inducing the Euclidean norm $\|u\| := \sqrt{u^\top u}$.
The Euclidean norms on $\mathbb{R}^m$ and $\mathbb{R}^n$ induce a norm on the set of linear maps from $\mathbb{R}^m$ to $\mathbb{R}^n$ which is defined as follows:

$$\|\cdot\|_2 : \mathbb{R}^{n\times m} \longrightarrow \mathbb{R}_+ , \qquad \|A\|_2 := \sup_{v\neq 0} \frac{\|Av\|}{\|v\|}$$

The definition is quite general and applies to linear maps between arbitrary inner product spaces. If $A \in \mathbb{R}^{n\times m}$ it descends from Schwarz inequality $u^\top v \leq \|u\|\|v\|$ that there is a constant k such that $\|Av\| \leq k\|v\|$. The 2-norm of A is in fact the smallest such k.

Problem B.5. *Let $A \in \mathbb{R}^{n\times m}$. Show that:*
1.The sup *in the definition of the induced norm is indeed a* max*, i.e.*

$$\|A\|_2 = \max_{v\neq 0} \frac{\|Av\|}{\|v\|} \quad \textit{and} \quad \|A\|_2 = \max_{\|v\|=1} \|Av\|$$

2. $\|A\|_2$ is equal to σ_1, the first (i.e. the largest) singular value of A. For this reason this norm is also called spectral norm*.*

The second question relates to the very instructive maximization of the so-called *Rayleigh quotients*.
The solution of the following problem follows instead quite trivially from $\text{Tr}\,(A) = \sum \lambda_k(A)$.

Problem B.6. *The square of the* **Frobenius norm** $\|A\|_F^2 = \text{Trace}\,(A^\top A) = \sum_{i,j} a_{i,j}^2$ *can be also computed by the formula*

$$\|A\|_F^2 = \sigma_1^2 + \ldots + \sigma_r^2 \, .$$

Denote by $\Sigma(A)$ the set of singular values of A.

Problem B.7. *Let A be square. Show that:*
1. If $A = A^\top \geq 0$ then $\Sigma(A) = \sigma(A)$.

2. $0 \in \sigma(A)$ *if and only if* $0 \in \Sigma(A)$.
3. If $A = A^\top$ *then* $\Sigma(A) = \{|s| : s \in \sigma(A)\}$.
4. $\sigma(A)$ *and* $\Sigma(A)$ *can be quite different (discuss the intersection* $\sigma(A) \cap \Sigma(A)$*).*
5. What are the singular values of a skew-symmetric matrix?

If a matrix A is far from being symmetric (in fact far from normal), for example if A is lower triangular, then the singular values can be very different from the eigenvalues. Give some examples.

Generalization of the SVD

Let $\mathscr{X}$, $\mathscr{Y}$ be finite-dimensional inner product spaces of dimensions n and m.

Lemma B.1. *Let* $Q : \mathbb{R}^n \to \mathscr{V}$ *be a unitary map. Then* $\dim \mathscr{V} = n$ *and there is an orthonormal basis* $u_1, \dots, u_n$ *in* $\mathscr{V}$ *such that* $Qx = \sum u_k \xi_k$ *where* $\xi_k =$ *coordinates of* x.

Proof: Let $\{e_k\}$ be a canonical basis in $\mathbb{R}^n$ and $Qe_k := u_k$, $k = 1, \dots, n$. Then the u_k form an orthonormal basis. By linearity $Q \sum \xi_k e_k = \sum \xi_k Q e_k$. □

Theorem B.2. *Let* $A : \mathscr{X} \to \mathscr{Y}$ *of rank* $r \le \min(m, n)$*. There are two unitary maps* $U : \mathbb{R}^m \to \mathscr{Y}$, $V : \mathbb{R}^n \to \mathscr{X}$ *and a sequence of positive real numbers ordered in decreasing magnitude,* $\{\sigma_1 \ge \dots \ge \sigma_r\}$*, called the* singular values *of* A*, such that*

$$A = U \Delta V^*, \qquad \Delta = \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix}, \qquad \Sigma = \operatorname{diag}\{\sigma_1, \dots, \sigma_r\} \tag{B.2.3}$$

The matrix $U = [u_1, \dots, u_m]$, $u_k \in \mathscr{Y}$ *is made by the the normalized eigenvectors of* AA^**; dually, the columns of* $V := [v_1, \dots, v_n]$, $v_k \in \hat{}$ *are the normalized eigenvectors of* A^*A*. The squared singular values* $\{\sigma_1^2 \ge \dots \ge \sigma_r^2\}$ *are the non-zero eigenvalues of* AA^* *(or* A^*A*).*

Proof. Let $[v_1, \dots, v_n]$, be normalized eigenvectors of A^*A so that

$$A^*Av_k = \sigma_k^2 v_k \qquad k = 1, \dots n$$

with $A^*Av_k = 0$ for $k > r$. Note that these last eigenvectors are essentially arbitrary in the nullspace of A. Multiplying from the left by A one gets

$$AA^*(Av_k) = \sigma_k^2 (Av_k) \qquad k = 1, \dots n$$

so the vectors

$$u_k := \frac{1}{\sigma_k} Av_k \qquad k = 1, \dots r\,,$$

are normalized eigenvectors of AA^*. In fact,

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

Completing the family $\{u_1, \dots, u_r\}$ with $m - r$ (eigen)vectors in the nullspace of AA^* we obtain an orthonormal basis in $\mathscr{Y}$. Then

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

for $k, j \leq r$ and $\langle u_k,\, Av_j \rangle = 0$ otherwise. These relations are equivalent to $U^*AV = \Delta$. □

The following *full-rank SVD factorization* of A is obtained by eliminating all the zero blocks in (B.2.3)

$$A = [u_1, \ldots, u_r]\, \bar{\Sigma}\, [v_1, \ldots, v_r]^* := U_1 \bar{\Sigma} V_1^* \tag{B.2.4}$$

where $\bar{\Sigma} = \operatorname{diag}\{\sigma_1, \ldots, \sigma_r\}$ and U_1, V_1 are the submatrices obtained by keeping only the first r columns of U and V. Note that U_1 and V_1 still have orthonormal columns

$$U_1^* U_1 = I_r = V_1^* V_1\,.$$

Corollary B.1. *The image space and the nullspace of A are :*

$$\operatorname{Im}(A) = \operatorname{Im}(U_1) = \operatorname{span}\{u_1, \ldots, u_r\}, \qquad \ker(A) = \ker(V^*) = \operatorname{span}\{v_{r+1}, \ldots, v_n\}$$

Moreover, the 2-norm and the Frobenius norms of A are

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

The map

$$A_k := \sum_{i=1}^{k} \sigma_i u_i \langle v_i, \cdot \rangle \qquad k \leq r$$

is the best rank $k(\leq r)$ approximation of A in a variety of norms; in fact,

$$\min_{B\,;\,\operatorname{rank}(B)=k} \|A - B\|_2 = \|A - A_k\|_2 = \sigma_{k+1} \tag{B.2.5}$$

and

$$\min_{B\,;\,\operatorname{rank}(B)=k} \|A - B\|_F^2 = \|A - A_k\|_F^2 = \sigma_{k+1}^2 + \ldots + \sigma_r^2 \tag{B.2.6}$$

Note that $(A - A_k)x = \sum_{i=k+1}^{r} \sigma_i u_i \langle v_i, x \rangle$ and hence $\|A - A_k\|_2 = \sigma_{k+1}$. A similar argument holds for the Frobenius norm.

The proof that A_k is the actual minimizer is tricky. See [Golub-Van Loan] p. 19-20.

Problem B.8. *Is the SVD of A unique? discuss the case where there are multiple eigenvalues of AA^* (or of A^*A). Assume that the σ_i's are all distinct. Is the SVD unique in this case?*

Let $A = U \operatorname{diag}\{\sigma_1, \ldots, \sigma_n\} V^*$ where U and V are arbitrary orthonormal matrices and $\sigma_1 \geq \ldots \geq \sigma_n \geq 0$. Is this necessarily the SVD of A ? In any case, are $\sigma_1, \ldots \sigma_n$ the singular values of A ?

There is an equivalent statement where the singular values are $p = \min(n, m)$ but some of them $(\sigma_{r+1}, \ldots, \sigma_p)$ are allowed to be zero.

SVD and the Pseudoinverse

The theorem below provides a general rule to compute a special pseudoinverse called the *Moore-Penrose pseudoinverse*.

Theorem B.3. *let A admit the SVD*

$$A = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_1^* \\ V_2^* \end{bmatrix}, \qquad \Sigma > 0$$

Then the (moore-Penrose) pseudoinverse of A is

$$A^\dagger = \begin{bmatrix} V_1 & V_2 \end{bmatrix} \begin{bmatrix} \Sigma^{-1} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} U_1^* \\ U_2^* \end{bmatrix} = V_1 \Sigma^{-1} U_1^*$$

Lemma B.2. *If* $\Delta = \mathrm{diag}\,\{\Sigma,\, 0\}$ *then* $\Delta^\dagger = \Delta^+ = \mathrm{diag}\,\{\Sigma^{-1},\, 0\}$.

Proof: Identify the subspaces in Fig **??** and note that Δ is symmetric and $\ker(\Delta)$ is a reducing subspace for Δ. On the orthogonal complement $\Delta \equiv \Sigma$ is invertible. □

Appendix C

A QUICK REVIEW OF CONSTRAINED OPTIMIZATION

This is just a quick review. The whole story is beautifully exposed in the book [11]. The notation $x \succeq 0$ means that all components of the vector x are nonnegative. We shall give for granted the notions of convex set and convex function. In particular just recall that a differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if and only if

$$f(y) - f(x) \geq \nabla f(x)^\top (y - x)$$

for all x, y in its domain. The following is a basic fact from Convex Optimization Theory.

Theorem C.1.

Assume f is a smooth convex function and the feasible set $\mathcal{X}$ is a convex subset of $\mathbb{R}^n$ with piecewise smooth boundary. Consider the optimization problem $\min_x f(x)$ subject to the constraint $x \in \mathcal{X}$. Then at the optimum the gradient $\nabla f(x^*)$ must be the normal of a supporting (possibly tangent) hyperplane to the feasible set. In other words $x^* \in \mathcal{X}$ is optimal if and only if

$$-\nabla f(x^*)^\top (y - x^*) \leq 0 \quad \textit{for all} \quad y \in \mathcal{X}$$

that is, the opposite gradient makes an obtuse angle with all vectors $y - x^*; y \in \mathcal{X}$, see Fig. C below.

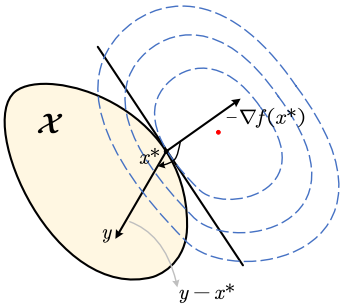


Figure C.0.1. *Convex Optimization*

Let now $f_0, \ldots, f_m,\ h_1, \ldots, h_p$ be real functions; consider a general optimization problem

$$\begin{aligned} \underset{x\in\mathbb{R}^n}{\text{minimize}} \quad & f_0(x) \\ \text{subject to} \quad & f_i(x) \le 0,\ i = 1, \ldots, m, \\ & h_i(x) = 0,\ i = 1, \ldots, p. \end{aligned} \tag{P}$$

Let p^* denote the optimal value $f_0(x^*)$.

The **Lagrangian** of problem P is defined as

$$L(x, \lambda, \nu) := f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x)$$

and the *Lagrange dual function* as

$$g(\lambda, \nu) := \inf_{x\in\mathbb{R}^n} L(x, \lambda, \nu) = \inf_{x\in\mathbb{R}^n} \left(f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x) \right),$$

which may possibly take the value $-\infty$. The function g is concave even when problem (P) is not convex.

Theorem C.2. *It holds that $g(\lambda, \nu) \le p^*$ for any $\lambda \succeq 0$ and any ν.*
Let the Lagrange dual problem

$$\begin{aligned} & \textit{maximize} \quad g(\lambda, \nu) \\ & \textit{subject to} \quad \lambda \succeq 0. \end{aligned} \tag{D}$$

have optimal value d^*. Then $d^* \le p^*$ (weak duality).

Theorem C.3 (Karush-Kuhn-Tucker (KKT)). *If x^* is a local minimizer and the functions $f_0, \ldots, f_m, h_1, \ldots, h_p$ are continuously differentiable at x^*, then there exists vector multipliers (λ^*, ν^*) such that*

$$\begin{aligned} \nabla f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla f_i(x^*) + \sum_{i=1}^{p} \nu_i^* \nabla h_i(x^*) = 0. \\ f_i(x^*) \le 0,\ i = 1, \ldots, m \\ h_i(x^*) = 0,\ i = 1, \ldots, p \\ \lambda_i^* \ge 0,\ i = 1, \ldots, m \\ \lambda_i^* f_i(x^*) = 0,\ i = 1, \ldots, m \end{aligned} \tag{C.0.1a}$$

In particular, x^* minimizes $L(x, \lambda^*, \nu^*)$.
These are first order necessary conditions for optimality. If the problem is convex they are also sufficient.

Consider now a particular convex problem:

$$\begin{aligned} \underset{x\in\mathbb{R}^n}{\text{minimize}} \quad & f_0(x) \\ \text{subject to} \quad & f_i(x) \le 0,\ i = 1, \ldots, m, \\ & Ax = b, \end{aligned}$$

with $f_0, \ldots, f_m$ convex.
Assume *Slater's constraint qualification* condition: strict feasibility

$$\exists\, x \text{ such that } f_i(x) < 0,\ i = 1, \ldots, m, \quad Ax = b.$$

Implications:

1. strong duality, $d^* = p^*$;
2. existence of optimal Lagrange multiplier (λ^*, ν^*) with $g(\lambda^*, \nu^*) = d^*$.

Existence and uniqueness of the Lasso solution

$$\begin{array}{ll} \underset{\theta}{\text{minimize}} & \dfrac{1}{2N}\|\mathbf{y} - S\theta\|_2^2 \\ \text{subject to} & \|\theta\|_1 \leq t, \end{array} \tag{PL}$$

This is a convex problem. The existence of the primal optimum is implied by the Weierstrass extreme value theorem on a closed set. Uniqueness is a consequence of convexity.
Existence of the dual optimum is guaranteed by Slater's condition, which would validate the claim. However, $\|\theta\|_1$ is not differentiable.

KKT in the subdifferentiable case

Need to introduce the *subgradient*...

Inequality constrained convex problem:

$$\begin{array}{ll} \underset{x \in \mathbb{R}^n}{\text{minimize}} & f_0(x) \\ \text{subject to} & f_i(x) \leq 0,\ i = 1, \ldots, m, \end{array}$$

If strong duality holds, x^*, λ^* are primal, dual optimal iff

1. x^* is primal feasible
2. $\lambda \succeq 0$
3. $\lambda_i^* f_i(x^*) = 0$ for $i = 1, \ldots, m$
4. x^* is a minimizer of $L(x, \lambda^*) = f_0(x) + \sum_{i=1}^m \lambda_i^* f_i(x)$ that is:

$$0 \in \partial f_0(x^*) + \sum_{i=1}^m \lambda_i^* \partial f_i(x^*)$$

Appendix D

SOME FACTS FROM HILBERT SPACE THEORY

A *Hilbert space* is an inner product space $(\mathbf{H}, \langle \cdot, \cdot \rangle)$ which is complete with respect to the metric induced by the inner product. In other words every Cauchy sequence has a limit in $\mathbf{H}$. To establish notation we shall give examples of Hilbert spaces which are frequently used in this book:

1. *The space of square summable m-dimensional sequences ℓ^2_m.* The elements of this space are sequences $x = \{x(t)\}_{t\in\mathbb{Z}}$ of real or complex (m-dimensional vectors $x(t)$, which we shall write as *column vectors*, indexed by the integer-valued parameter t, satisfying

$$\|x\|^2 := \sum_{t=-\infty}^{+\infty} x(t)^*\, x(t) < \infty,$$

where * denotes complex conjugate transpose. In signal processing this norm is sometimes called the "energy" of the signal x. It is induced by the inner product

$$\langle x, y\rangle := \sum_{t=-\infty}^{+\infty} x(t)^*\, \mathbf{y}(t).$$

A simple proof that ℓ^2_m is complete can be found in standard text books, e.g., [111].

2. *The Lebesgue space L^2_m.* Let $[a, b]$ be an interval (not necessarily bounded) of the real line. We shall denote by $L^2_m([a, b])$ the space of functions taking values in $\mathbb{C}^n$ (or $\mathbb{R}^n$) which are square integrable on $[a, b]$ with respect to the Lebesgue measure. The values, $f(t)$, of the functions will also be written as column vectors. It is well-known that this space is a Hilbert space under the inner product

$$\langle f, g\rangle := \int_a^b f(t)^*\, g(t)dt.$$

3. The space $L^2_{m\times n}([a, b])$ of matrix-valued functions with values in $\mathbb{C}^{m\times n}$

and with inner product

$$\langle F,\, G\rangle := \int_a^b \mathrm{Tr}\,\{F(t)^* G(t)\}dt. \tag{D.0.1}$$

This is the natural inner product which makes $L^2_{m\times n}([a,b])$ into a Hilbert space. The functions of this space are square integrable on $[a,b]$ with respect to the Lebesgue measure, in the sense that

$$\|F\|^2 := \int_a^b \mathrm{Tr}\,\{F(t)^* F(t)\}dt \,<\, \infty\,.$$

4. *The space $L^2(\Omega,\mathcal{A},P)$ of second-order random variables.* This is a Hilbert space often used in this book. It has the inner product

$$\langle \xi,\eta\rangle = \mathbb{E}\,\{\bar{\xi}\eta\},$$

where $\mathbb{E}\,\{\xi\} = \int_\Omega \xi dP$ denotes mathematical expectation. If we restrict to *zero-mean* random variables we obtain a subspace which does not contain deterministic constants (except zero).

Notations: No subscript will be used to denote the scalar ℓ^2 and L^2 spaces. For vector-valued functions say $f(t)\in\mathbb{R}^p$ the Euclidean norm will be denoted with the same symbol as absolute value; i.e. $|f(t)|^2 := \sum_{k=1}^p f_k(t)^2$; this will allow to use $\|\cdot\|$ for the Hilbert space norm. However when needed we shall use subscripts. The are instances in which using *row-vector* notation is more natural. The reason being that the elements of the functional Hilbert spaces ℓ^2_m and L^2_m naturally appear as multipliers in the combination of vector *random vector* variables. Other important examples of Hilbert spaces (e.g. the Hardy spaces H^2_m) will be introduced later on.

In this book the term *subspace* of a Hilbert space $\mathbf{H}$, will mean *closed* subspace. Given two subspaces $\mathbf{X},\mathbf{Y}\in\mathbf{H}$, the vector sum $\mathbf{X}\vee\mathbf{Y}$ is the smallest subspace containing both $\mathbf{X}$ and $\mathbf{Y}$; i.e., it is the *closure* of

$$\mathbf{X}+\mathbf{Y} := \{x+y \mid x\in\mathbf{X}, y\in\mathbf{Y}\}.$$

In fact, if both $\mathbf{X}$ and $\mathbf{Y}$ are infinite-dimensional, $\mathbf{X}+\mathbf{Y}$ may fail to be closed. A classical example illustrating this can be found in [43, p. 28]. The symbol $\oplus$ will be used for *direct sum*, i.e., $\mathbf{X}\oplus\mathbf{Y} = \mathbf{X}+\mathbf{Y}$ with the extra condition that $\mathbf{X}\cap\mathbf{Y}=0$. In particular, when $\mathbf{X}\perp\mathbf{Y}$, we have an *orthogonal direct sum*, which we write $\mathbf{X}\oplus\mathbf{Y}$. An orthogonal sum of subspaces is always closed. The *linear vector space generated by a family of elements* $\{x_\alpha\}_{\alpha\in\mathbb{A}}\subset\mathbf{H}$, denoted $\mathrm{span}\,\{x_\alpha \mid \alpha\in\mathbb{A}\}$ 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 $\overline{\mathrm{span}}\{x_\alpha \mid \alpha\in\mathbb{A}\}$.

Important examples of subspaces of ℓ^2_m are (in the language of signal processing) the subspaces of **causal** signals, ℓ^{2+}_m, which are zero for negative values of t ($f(t)=0,\ t<0$) and the **anticausal** signals, ℓ^{2-}_m, which are instead zero for positive values of t, ($f(t)=0,\ t>0$). These two subspaces have a non-empty intersection which is (isomorphic to) $\mathbb{R}^m$ (or $\mathbb{C}^m$). The orthogonal complement,

$\ell_m^{2+\perp}$, of ℓ_m^{2+} in ℓ_m^2 is the subspace of *strictly anticausal* functions which are zero also for $t = 0$. Evidently we have the orthogonal decomposition

$$\ell_m^2 = \ell_m^{2+} \oplus \ell_m^{2+\perp}. \tag{D.0.2}$$

We shall often have the occasion to deal with series of orthogonal random variables. A simple but basic result on convergence of these series is the following.

Lemma D.1. *A series of orthogonal elements in a Hilbert space,*

$$\sum_{k=0}^{\infty} x_k, \quad x_k \perp x_j, \quad k \neq j,$$

converges if and only if

$$\sum_{k=0}^{\infty} \|x_k\|^2 < \infty \tag{D.0.3}$$

i.e., the series of the square norms of the elements converges.

Proof. In fact the series converges if and only if

$$\left\| \sum_{k=0}^{m} x_k - \sum_{k=0}^{n-1} x_k \right\| \to 0$$

as $n, m \to \infty$ which is the same as $\|\sum_{k=n}^{m} x_k\|^2 \to 0$ which in turn is equivalent to $\sum_{k=n}^{m} \|x_k\|^2 \to 0$ as $n, m \to \infty$. □

Let $\{e_k\}$ be an an orthonormal sequence in a Hilbert space $\mathbf{H}$. Since, for an arbitrary $x \in \mathbf{H}$, the "approximation error"

$$\left\| x - \sum_{k=0}^{N} \langle x,\, e_k \rangle e_k \right\|^2 \leq \|x\|^2 - \sum_{k=0}^{N} |\, \langle x,\, e_k \rangle\, |^2$$

is non-negative, we have

$$\sum_{k=0}^{N} |\, \langle x,\, e_k \rangle\, |^2 \leq \|x\|^2 \qquad \text{for all } N$$

and hence the series $\sum_{k=0}^{\infty} \langle x,\, e_k \rangle\, e_k$ converges. An immediate consequence of this fact is that the sequence of the Fourier coefficients $f(k) := \langle x,\, e_k \rangle,\ k = 1, \ldots$ belongs to ℓ^2.

Corollary D.1. *Let $\{e_k\,;\, k \in \mathbb{Z}\}$ be an orthonormal sequence in a Hilbert space $\mathbf{H}$ and let $c := \{c_k\,;\, k \in \mathbb{Z}\}$ be a sequence of complex numbers. The series $\sum_{-\infty}^{+\infty} c_k e_k$ is convergent if and only if*

$$\sum_{-\infty}^{+\infty} |\, c_k\, |^2 < \infty$$

that is if and only if $c \in \ell^2$.

Let us recall that a set of orthonormal elements $\{e_\alpha\,;\,\alpha\in\mathbb{A}\}$ in a Hilbert space $\mathbf{H}$ is called *complete* if

$$\langle x, e_\alpha\rangle = 0, \quad \forall\alpha\in\mathbb{A}\Rightarrow x=0$$

that is, the only element of $\mathbf{H}$ which can be orthogonal to all $\{e_\alpha\,;\,\alpha\in\mathbb{A}\}$ is the zero element. A countable complete system, $\{e_k\,;\,k\in\mathbb{Z}\}$, is called an *orthonormal basis* of $\mathbf{H}$. A Hilbert space which admits an orthonormal basis is called *separable*. The result below is sometimes called the *Riesz-Fisher Theorem*.

Theorem D.1. *In a separable Hilbert space* $\mathbf{H}$*, every element* x *admits a representation*

$$x=\sum_{-\infty}^{+\infty} c_k e_k, \qquad c_k=\langle x, e_k\rangle \tag{D.0.4}$$

with respect to an orthonormal basis $\{e_k\}$*. The sequence of coefficients* $c:=\{c_k\}$ *is unique and belongs to* ℓ^2 *(is a finite energy signal). Conversely, for every finite energy signal* c*, and orthonormal basis* $\{e_k\}$*, the series* (D.0.4) *converges to an element of* $\mathbf{H}$*.*

For a fixed orthonormal basis $\{e_k\}$ *the correspondence* $c\to x$ *is a unitary map from* ℓ^2 *onto* $\mathbf{H}$*; in particular* $\|x\|_{\mathbf{H}}=\|c\|_{\ell^2}$*.*

Inclusion theorems for L^p spaces

if $\hat{x}$ is a finite measure space then for $1\le p<q\le\infty$

$$L^\infty(\hat{x},\mathscr{A},\mu)\subset L^q(\hat{x},\mathscr{A},\mu)\subset L^p(\hat{x},\mathscr{A},\mu) \tag{D.0.5}$$

proof by Hölder inequality. Example: L^2 *is properly contained in* L^1.

For ℓ^p spaces: If $p<q\le\infty$

$$\ell^p\subset\ell^q$$

Proof: $f\in\ell^p$ implies that for $t\to\infty$

$$|f(t)|^p\to 0 \quad \text{and hence} \quad |f(t)|\to 0$$

so for t large enough $|f(t)|<1$ and when $p<q$

$$|f(t)|^q\le|f(t)|^p$$

so that

$$\sum|f(t)|^q\le\sum|f(t)|^p\,.$$

Hence ℓ^1 *is properly contained in* ℓ^2 !! Example the harmonic sequence $\frac{1}{t}$ for $t\neq 0$.

Operators and their adjoints

A *linear operator* T from a Hilbert space $\mathbf{H}_1$ to another Hilbert space $\mathbf{H}_2$, is a linear map between the two spaces. In general T may not be defined on all of $\mathbf{H}_1$; think, for example, of the differentiation operator in L^2_m. When T is defined for all elements of $\mathbf{H}_1$, one says that T is defined *on* $\mathbf{H}_1$. The simplest linear operators are the *continuous*, also called *bounded* operators, which are defined on the whole space and satisfy an inequality of the type

$$\|Tx\|_2\le k\|x\|_1, \qquad x\in\mathbf{H}_1$$

for some constant k, the subscripts referring to the different norms in the two Hilbert spaces. As one can see, a continuous linear operator is actually uniformly continuous. The infimum of all k for which the inequality holds is called the *norm* of the operator T and is denote by $\|T\|$.

Proposition D.1. *Let $T : \mathbf{H}_1 \to \mathbf{H}_2$ be a bounded operator, then the suprema*

$$\|T\| = \sup \left\{ \frac{\|Tf\|}{\|f\|}, \quad f \in \mathbf{H}_1 \right\} \tag{D.0.6}$$

or, equivalently

$$\|T\| = \sup \left\{ \frac{|\langle Tf,\, g\rangle|}{\|f\|\|g\|}, \quad f \in \mathbf{H}_1,\, g \in \mathbf{H}_2 \right\}.$$

are both finite and equal to the norm of the operator T. In fact the supremum is a maximum

$$\|T\| = \max_{\{f \in \mathbf{H}_1;\, \|f\|=1\}} \|Tf\|.$$

A linear operator mapping $\mathbf{H}$ into the real or complex numbers is called a *linear functional*. The following is the fundamental representation theorem for bounded linear functionals.

Theorem D.2 (F. Riesz). *Let $T; \mathbf{H} \to \mathbb{C}$ be a bounded linear functional on the Hilbert spoace $\mathbf{H}$. Then there is a unique element $h \in \mathbf{H}$ such that*

$$T(x) = \langle h,\, x\rangle$$

for all $x \in \mathbf{H}$. The norm of T is $\|h\|$.

If T is bounded, it is quite easy, using the Riesz representation, to see that there is a unique bounded linear operator $T^* : \mathbf{H}_2 \to \mathbf{H}_1$, which satisfies

$$\langle Tx, z\rangle_2 = \langle x, T^* z\rangle_1 \qquad \forall x \in \mathbf{H}_1,\ z \in \mathbf{H}_2.$$

The operator T^* is called the *adjoint* of T. It holds that $\|T^*\| = \|T\|$; i.e. a bounded operator and its adjoint have the same norm. Unbounded operators may also have adjoints under suitable conditions (in general involving an extension to a larger space of the original operator). A linear operator from $\mathbf{H}$ into itself, for which $T^* = T$ is called *selfadjoint*. On a finite dimensional space the concept of adjoint corresponds to taking the transpose (or the Hermitian conjugate) of the matrix representing the operator with respect to an orthonormal basis. (Warning: this is *no longer true* if the basis is not orthonormal!).

Important examples of linear operators on L^2 spaces are *multiplication operators*. For this we need to recall the definition of L^∞ spaces.

Definition D.1. *A scalar measurable function f defined on the interval $[a, b]$ is* essentially bounded *(with respect to Lebesgue measure μ) if there is some constant $\alpha < \infty$ such that $|f(t)| \leq \alpha$ almost everywhere; i.e., except possibly for points t which form a subset of measure zero. The smallest such constant, denoted*

$$\operatorname*{ess\,sup}_{t \in [a,b]} f := \inf \alpha$$

is called the essential supremum *of f on $[a, b]$.The essential supremum is actually a norm which makes the vector space of essentially bounded functions on $[a, b]$ into a Banach space, denoted $L^\infty([a,\, b])$.*

Similarly, the vector space of $\mathbb{C}^{m\times n}$-valued matrix functions F, such that

$$\|F\|_\infty := \operatorname{ess\,sup}_{t\in[a,\,b]} \|F(t)\| < \infty$$

is a Banach space under the norm $\|\cdot\|_\infty$ defined above. This Banach space is denoted $L^\infty_{m\times n}([a,\, b])$.

Note that in the definition we have chosen the operator norm of the matrices $F(t)$. Since any two norms on a finite dimensional vector space are equivalent, the choice of matrix norm is immaterial for the definition of the space $L^\infty_{m\times m}([a,\, b])$. This choice turns out to be convenient when we regard the action of F on functions of $L^2_m([a,\, b])$, as a linear *multiplication operator*

$$M_F : L^2_m([a,\, b]) \to L^2_n([a,\, b])\,, \qquad M_F : f \mapsto fF$$

Proposition D.2. *A multiplication operator M_F on $L^2_m([a,\, b])$ by a $\mathbb{C}^{m\times n}$-valued matrix function F, is a bounded linear operator into $L^2_n([a,\, b])$ if and only if $F \in L^\infty_{m\times n}([a,\, b])$. The norm of the operator M_F is then,*

$$\|M_F\| = \|F\|_\infty\,. \tag{D.0.7}$$

The bound is a consequence of the multiplicative inequality (**??**), since

$$\begin{aligned}\langle f(t)F(t),\, f(t)F(t)\rangle &= \operatorname{Tr}\left(f(t)F(t)F(t)^*f(t)^*\right) = \operatorname{Tr}\left(F(t)F(t)^*f(t)^*f(t)\right)\\ &\le \|F(t)F(t)^*\|\|f(t)f(t)^*\|_F = \|F(t)\|^2\|f(t)\|^2.\end{aligned}$$

The *image* or *range* of an operator $T : \mathbf{H}_1 \to \mathbf{H}_2$, is the linear manifold $\operatorname{Im} T := \{Tx \mid x \in \mathbf{H}_1\}$. This manifold does not need to be closed, i.e. a subspace of $\mathbf{H}_2$, but if this is the case T is said to have *closed range*. The *kernel* or *nullspace* of an operator T, $\ker T := \{x \mid Tx = 0\}$, is always closed. Operators for which $\overline{\operatorname{Im} T} = \mathbf{H}_2$ will be called *densely onto*. The following simple but important result, is a generalization of an analogous one valid for finite dimensional inner product spaces.

Theorem D.3. *Let $T : \mathbf{H}_1 \to \mathbf{H}_2$ be a bounded operator from the Hilbert space $\mathbf{H}_1$ to the Hilbert space $\mathbf{H}_2$. Then*

$$\mathbf{H}_1 = \ker T \oplus \overline{\operatorname{Im} T^*} \tag{D.0.8a}$$

$$\mathbf{H}_2 = \ker T^* \oplus \overline{\operatorname{Im} T} \tag{D.0.8b}$$

A bounded operator T is *left-invertible* if there exists a bounded operator S such that $ST = I_1$ and *right-invertible* if there exists a bounded operator R such that $TR = I_2$. Clearly, right-invertibility implies that T is surjective (i.e., maps onto $\mathbf{H}_2$) while left-invertibility implies that T is injective (i.e., one-to-one). In fact it can be shown that *a bounded operator T is right-invertible if and only if it is onto*. However the dual statement for left-invertibility is in general false.

Theorem D.4. *A bounded linear operator from one Hilbert space to another is left-invertible if and only if it is injective and has closed range.*

If T is both left- and right- invertible it is called *invertible* tout-court. Note that left- or right- inverses are in general non-unique. However a two-sided inverse is unique.

A linear map U between two Hilbert spaces preserving the inner products, i.e. a map for which

$$\langle Ux, Uy\rangle_2 = \langle x, y\rangle_1, \qquad x,\, y \in \mathbf{H}_1$$

is called an *isometry*. An isometry is always an injective map. The following basic result is used repeatedly in this book.

Theorem D.5. *Every isometry defined on a subset of elements $\tilde{\mathbf{X}} := \{x_\alpha \mid \alpha \in \mathbb{A}\}$ of a Hilbert space $\mathbf{H}$ can be extended by linearity and continuity to the whole Hilbert space $\mathbf{X} := \overline{\text{span}}\{x_\alpha \mid \alpha \in \mathbb{A}\}$ linearly generated by the family $\{x_\alpha\}$, preserving the property of isometry. The isometric extension is unique.*

Proof. We shall follow [77, p.14-15].
We first show that an isometry is necessarily linear on $\tilde{\mathbf{X}}$. Suppose that $x_1, \ldots, x_m$ are elements of $\tilde{\mathbf{X}}$ and that $x = \sum_k \alpha_k x_k$ also belongs to $\tilde{\mathbf{X}}$. Pick an arbitrary $x^\top \in \tilde{\mathbf{X}}$, then by linearity of the scalar product and isometry of U

$$\langle x,\, x^\top\rangle = \sum_k \alpha_k \langle Ux_k,\, Ux^\top\rangle = \langle \sum_k \alpha_k Ux_k,\, Ux^\top\rangle$$

but the left hand side is also equal to $\langle Ux,\, Ux^\top\rangle$ and hence

$$\langle Ux - \sum_k \alpha_k Ux_k,\, Ux^\top\rangle = 0$$

for all $x^\top \in \tilde{\mathbf{X}}$, in particular for $x^\top = x, x_1, \ldots, x_m$. Therefore

$$\langle Ux - \sum_k \alpha_k Ux_k,\, Ux - \sum_k \alpha_k Ux_k\rangle = \|x - \sum_k \alpha_k Ux_k\|^2 = 0$$

that is $Ux = \sum_k \alpha_k Ux_k$; i.e. U is a linear operator on $\tilde{\mathbf{X}}$. Now every element $h \in \mathbf{X}$ is the limit of a sequence of linear combinations $\{h_n\}$ of elements of $\tilde{\mathbf{X}}$, that is $\lim_{n\to\infty} \|h - h_n\| = 0$ which is equivalent to

$$\lim_{n,m\to\infty} \|h_m - h_n\| = \lim_{n,m\to\infty} \|Uh_m - Uh_n\| = 0$$

and hence $\lim_{n\to\infty} Uh_n$ exists and belongs to $\mathbf{X}$. We define $Uh := \lim_{n\to\infty} Uh_n$ so that U is defined for all $h \in \mathbf{X}$. Isometry of the extension can be proved by a limit argument. Uniqueness is obvious. □

Note that isometric operators satisfy the relation $\langle x, U^*Ux\rangle_1 = \langle x, x\rangle_1$, from which $U^*U = I_1$ (the identity operator in $\mathbf{H}_1$). If U is surjective ($U\mathbf{H}_1 = \mathbf{H}_2$) one sees that

$$U^* = U^{-1}.$$

A surjective isometry is called a *unitary operator*. Two linear operators $A : \mathbf{H}_1 \to \mathbf{H}_1$ and $B : \mathbf{H}_2 \to \mathbf{H}_2$ which are related by

$$A = U^{-1}BU$$

where U is unitary, are *unitarily equivalent*. Unitary equivalence is a relation which preserves the fundamental characteristics of a linear operator, among them the spectrum [3, 4]. The *Fourier transform* defined in $L^2(\mathbb{R})$, is an example of a unitary operator $L^2(\mathbb{R}) \to L^2(\mathbb{I})$.

A subspace $\mathbf{X} \subset \mathbf{H}$ is *invariant* for the operator T if $T\mathbf{X} \subset \mathbf{X}$. If a subspace $\mathbf{X}$ is invariant for T we denote by $T_{|\mathbf{X}}$ the *restriction* of T to the subspace $\mathbf{X}$. A subspace $\mathbf{X}$ is said to be *reducing* for a linear operator T if it is invariant for T and there is a complementary subspace $\mathbf{Y}$, satisfying the direct sum decomposition

$$\mathbf{H} = \mathbf{X} \oplus \mathbf{Y},$$

which is also invariant. In this case T has a matrix representation

$$T = \begin{bmatrix} T_{|\mathbf{X}} & 0 \\ 0 & T_{|\mathbf{Y}} \end{bmatrix}$$

with respect to the decomposition $\mathbf{H} = \mathbf{X} \oplus \mathbf{Y}$.

Lemma D.2. *Let T be a linear operator on a Hilbert space $\mathbf{H}$. Then*

$$T\mathbf{X} \subset \mathbf{X} \Leftrightarrow T^*\mathbf{X}^\perp \subset \mathbf{X}^\perp$$

If T is self-adjoint, both $\mathbf{X}$ and $\mathbf{X}^\perp$ are reducing for T.

Proof. First note that $\mathbf{X}$ is T-invariant if and only if $\langle Tx, y\rangle = 0$ for all $x \in \mathbf{X}$ and $\mathbf{y} \in \mathbf{X}^\perp$. Then just apply the definition of adjoint. □

Bibliography

[1] R. Adams. *Sobolev Spaces*. Academic Press, New York and London, 1970. (Cited on p. 234)

[2] Mark A. Aizerman, Emmanuel M. Braverman, and Lev I. Rozonoer. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, 25:821–837, 1964. (Cited on p. 258)

[3] N. I. Akhiezer and I. M. Glazman. *Theory of linear operators in Hilbert space. Vol. I.* Translated from the Russian by Merlynd Nestell. Frederick Ungar Publishing Co., New York, 1961. (Cited on pp. 230, 328)

[4] N. I. Akhiezer and I. M. Glazman. *Theory of linear operators in Hilbert space. Vol. II.* Translated from the Russian by Merlynd Nestell. Frederick Ungar Publishing Co., New York, 1963. (Cited on pp. 228, 328)

[5] F. Albertini and E.D. Sontag. For neural networks, function determines form. *Neural Networks*, 6:975–990, 1993. (Cited on p. 251)

[6] Giorgio Picci Alessandro Chiuso and Stefano Soatto. Wide-sense estimation on the special orthogonal group. *Commun. on Information Systems*, 8, 3:185–200, 2008. (Cited on p. 247)

[7] Bekker P. A. and J. De Leeuw . The rank of reduced dispersion matrices. *Psychometrika*, **52**, 125–135, 1987. (Not cited)

[8] Paul A.BekkerJos M.F.ten Berge. Generic global identification in factor analysis. *Linear Algebra and its Applications*, 264:255–263, October 1997. (Cited on p. 182)

[9] Mario Bertero and Patrizia Boccacci. *Introduction to inverse problems in imaging*. CRC press, 1998. (Cited on pp. 95, 97)

[10] S. Bittanti and G. Picci. *Identification, Adaptation, Learning (the science of learning models from data)*, volume 153 of *Springer Verlag NATO-ASI series*. Springer Verlag, 1996. (Cited on p. 244)

[11] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. (Cited on p. 317)

[12] J. R. Bunch and C.P. Nielsen. Updating the singular value decomposition. *Numer. Math.*, 31:111–129, 1978. (Cited on p. 153)

[13] J. R. Bunch, C.P. Nielsen, and D. Sorensen. Rank one modification of the symmetric eigenvalue problem. *Numer. Math.*, 31:31–48, 1978. (Cited on p. 153)

[14] Rodney Burmeister and Edwin Dobell. *Mathematical Theories of Economic Growth*. Mac Millan, 1971. (Cited on p. 75)

[15] A. Chiuso and G. Picci. Tracking of point features in computer vision as estimation on the unit sphere. In W. Haeger D. Kriegman and S. Morse, editors, *The Confluence of Vision and Control*, 1998. (Cited on p. 247)

[16] V. Ciccone, A. Ferrante, and M. Zorzi. An alternating minimization algorithm for factor analysis. *arXiv:1806.04433*, 2019. (Cited on p. 182)

[17] J.M.C. Clark. The consistent selection of parametrization in system identification. In *Proc of the Joint Automatic Control Conference*, Purdue University, Indiana, 1976. (Cited on p. 17)

[18] W. G. Cochran. *Sampling Techniques (Third Ed.)*. Wiley, 1977. (Cited on p. 5)

[19] D. Commenges. The deconvolution problem: Fast algorithms including the preconditioned conjugate-gradient to compute a map estimator. *IEEE Trans. Automat. Contr.*, AC-29:229–243, 1984. (Cited on p. 95)

[20] Corinna Cortes and Vladimir Vapnik. Support vector networks. *Machine Learning*, 20:273–297, 1995. (Cited on p. 263)

[21] T.M. Cover and J.A. Thomas. *Elements of Information Theory*. Wiley, New York, 1991. (Cited on p. 244)

[22] Harald Cramèr. *Mathematical Methods of Statistics*. Princeton University Press, 1946. (Cited on p. 24)

[23] F. Cucker and S. Smale. On the mathematical foundations of learning. *Bulletin of the American Math. Society*, 39:1–49, 2001. (Cited on p. xiii)

[24] G. Cybenko. Approximation by sigmoidal functions. *Mathematics of Control Signals and Systems*, pages 303–314, 1989. (Cited on pp. 248, 251)

[25] Carl de Boor. *A Practical Guide to Spline*, volume Volume 27. 01 1978. (Not cited)

[26] J. Dieudonnè. *Foundations of Modern Analysis*. Academic Press, New York, 1960. (Cited on p. 228)

[27] J. L. Doob. *Stochastic Processes*. Wiley, 1953. (Cited on p. 173)

[28] Bradley Efron. Bayesians, frequentists, and scientists. *Journal of the American Statistical Association*, 100:469:1–5, 2005. (Cited on p. 157)

[29] E.D. Sontag F. Albertini and V. Maillot. Uniqueness of weights for neural networks. In R. Mammone, editor, *Artificial Neural Networks for Speech and Vision*, pages 115–125. Chapman and Hall, 1993. (Cited on pp. 251, 252)

[30] T. Ferguson. *A Course in Large Sample Theory*. Chapman and Hall, 1996. (Cited on pp. 26, 42, 189)

[31] Bruno De Finetti. *Theory of probability*. John Wiley & Sons, New. York, 1975. (Cited on p. 7)

[32] R. A. Fisher. The correlation between relatives on the supposition of mendelian inheritance. *Philosophical Transactions of the Royal Society of Edinburgh*, 52:399–433, 1918. (Cited on p. 124)

[33] R. A. Fisher. On the "probable error" of a coefficient of correlation deduced from a small sample. *Metron*, 1:3–32, 1921. (Cited on p. 124)

[34] R. A. Fisher. Dispersion on a sphere. *Proc. Royal Soc. London ser. A*, 217:295–305, 1953. (Cited on p. 240)

[35] Sir Fisher, Ronald Aylmer. Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society*, 22:700–725, 1925. (Cited on p. 7)

[36] N. Gauraha. Introduction to the lasso. *Resonance*, 23:439–464, 2018. (Cited on pp. 101, 107)

[37] K.F. Gauss. *Theoria Motus Corporum Coelestium*. Julius Springer, Berlin, 1901. (Cited on pp. 18, 55)

[38] G. Golub and C van Loan. *Matrix Computations*. John Hopkins U.P., N.Y., 2000. (Cited on p. 227)

[39] G.H. Golub. Numerical methods for solving linear least-squares problems. *Numerische Mathematik*, 7:206–216, 1965. (Cited on pp. 87, 92)

[40] G.H. Golub and C.R. Van Loan. *Matrix Computations 3rd Ed.* The Johns Hopkins University Press, 1996. (Cited on p. 217)

[41] 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. (Cited on p. 153)

[42] Ulf Grenander. Eight lectures on statistical inference in stochastic processes. Technical Report 2, Division of Applied Mathematics, Brown University, Providence R.I., 1967. (Cited on p. 39)

[43] P.l R. Halmos. *Introduction to Hilbert space and the theory of spectral multiplicity*. AMS Chelsea Publishing, Providence, RI, 1998. Reprint of the second (1957) edition. (Cited on pp. 171, 322)

[44] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning, 2nd Ed*. Springer Series in Statistics. Springer New York Inc., 2009. (Cited on pp. xiii, 38, 97, 99, 101, 135, 139, 206, 211, 214, 217, 219, 262, 263, 269)

[45] Thomas Hofmann, Bernhard Schölkopf, and Alexander J. Smola. Kernel methods in machine learning. *Ann. Statist.*, 36(3):1171–1220, 06 2008. (Cited on pp. 260, 263)

[46] R. V. Hogg and A. T. Craig. *Introduction to Mathematical Statistics 3d edition*. Macmillan Co. New York, 1969. (Cited on p. 302)

[47] A. H. Jazswinsky. *Stochastic Processes and Filtering Theory*. Academic Press, New York, 1970. (Cited on p. 208)

[48] T. Kailath. A general likelihood ratio formua for random signals in gaussian noise. *IEEE Trans Inform. Theory*, 15:350–361, 1969. (Cited on p. 41)

[49] R. E. Kalman. Algebraic geometric aspects of the class of linear systems of constant dimension. In *Proc 8th annual Conf. on Information Sciences and Systems*, 1974. (Cited on p. 17)

[50] A. N. Kolmogoroff. Stationary sequences in Hilbert space. *Bolletin Moskovskogo Gosudarstvenogo Universiteta. Matematika*, 2:40pp, 1941. (Cited on p. 169)

[51] S. Kullback. *Information Theory and Statistics*. Wiley, Republished by Dover Publications, 1959. (Cited on p. 244)

[52] S. Kullback and R.A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86, 1951. (Cited on p. 244)

[53] P. Langevin. Magnetisme et theorie des electrons. *Ann. de Chim et de Phys.*, 5:70–127, 1905. (Cited on p. 238)

[54] C.L. Lawson and R.J. Hanson. *Solving Least Squares Problems*. Prentice Hall, Englewood Cliffs, 1974. (Cited on pp. 90, 92)

[55] Yann LeCun, John S. Denker, and Sara A. Solla. Optimal brain damage. In D. S. Touretzky, editor, *Advances in Neural Information Processing Systems 2*, pages 598–605. Morgan-Kaufmann, 1990. (Cited on p. 252)

[56] E. Lehmann. *Testing Statistical Hypotheses (second Ed.)*. Wiley, 1986. reprinted by Springer Verlag. (Cited on pp. 33, 42)

[57] Bernard C. Levy. *Principles of Signal Detecton and Parameter Estimation*. Springer, 2008. (Cited on pp. 33, 41)

[58] Anders Lindquist and Giorgio Picci. *Linear stochastic systems; A geometric approach to modeling estimation and identification*. Springer New York Inc., 2016. (Cited on pp. 80, 173, 179, 199)

[59] M. Loeve. *Probability Theory*. Van Nostrand Reinhold, 1963. Reprinted by Springer Verlag in 1986. (Cited on pp. 208, 294)

[60] Ian MacDonald. A relation between the maxwell-boltzmann and chi-squared distributions. *J. Chem. Educ.*, 63, 7:575, 1986. (Cited on p. 300)

[61] David J. C. MacKay. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, 2004. (Cited on p. xiii)

[62] J. Mandel. Use of the singular value decomposition in regression analysis. *The Americam Statistician*, 36:15–24, 1982. (Cited on pp. 92, 152, 153)

[63] MATLAB. *Using MATLAB Version 6*. The MathWorks Inc., 2002. (Cited on p. 45)

[64] W. McCulloch and W Pitts. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5:115–133, 1943. (Cited on p. 248)

[65] K. Murphy. *Machine Learning, A Probabilistic Perspective*. MIT Press, Cambridge, USA, 2012. (Cited on pp. xiii, 139)

[66] E. Parzen. An approach to time series analysis. *The Annals of Mathematical Statistics*, 32:951–989, 1961. (Cited on pp. 231, 259)

[67] K. Pearson. Contributions to the mathematical theory of evolution ii: Skew variation. *Philosophical Transactions of the Royal Society of London*, 1895. (Cited on p. 26)

[68] David L. Phillips. A technique for the numerical solution of certain integral equations of the first kind. *J. ACM*, 9(1):84–97, January 1962. (Cited on pp. 95, 96)

[69] G. Picci. Dynamic vision and estimation on spheres. In *Proceedings of the 36th Conf. on Decision and Control*, pages 1140–1145. IEEE Press, 1997. (Cited on p. 247)

[70] Giorgio Picci. *Filtraggio Statistico (Wiener, Levinson, Kalman) e Applicazioni*. Ed. Libreria Progetto, Padova, 2007. (Cited on p. 13)

[71] T. Poggio and F. Girosi. Networks for approximation and learning. *Proc. IEEE*, 78(9):1481–1497, 1990. (Cited on p. 248)

[72] V. M. Popov. Invariant description of linear time-invariant controllable systems. *SIAM Journal on Control*, 10:254–264, 1972. (Cited on p. 17)

[73] C.R. Rao. *Linear Statistical Inference and its Applications*. Wiley, New York, 1973. (Cited on pp. 15, 75)

[74] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006. (Cited on p. 260)

[75] David G. Stork Richard O. Duda, Peter E. Hart. *Pattern Classification, 2nd Edition*. Wiley, 2001. (Cited on pp. xiii, 132, 135, 137, 138)

[76] J. Rissanen. Basis of invariants and canonical forms for linear dynamical systems. *Automatica*, 10:175–182, 1974. (Cited on p. 17)

[77] Y. A. Rozanov. *Stationary Random Processes*. Holden-Day, San Francisco, 1967. (Cited on p. 327)

[78] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representations by error propagation. In David E. Rumelhart, James L. McClelland, and PDP Research Group, editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1*, pages 318–362. MIT Press, 1986. (Cited on p. 248)

[79] H. Scheffe'. *The Analysis of Variance*. Wiley, New York, 1959. (Cited on p. 123)

[80] Smola A.J. Schölkopf B., Herbrich R. A generalized representer theorem. In Williamson B. Helmbold D., editor, *Computational Learning Theory. COLT 2001*, volume 2111, pages 757–775. Springer Lecture Notes in Computer Science, Berlin, Heidelberg, 2001. (Cited on p. 233)

[81] F.C. Schweppe. Evaluation of likelihood functions for gaussian signals. *IEEE Trans Inform. Theory*, 11:61–70, 1965. (Cited on p. 41)

[82] C. E Shannon. A mathematical theory of communication. *Bell System Tech. Bell Syst. Technical Journal*, 27:379–423, 623–656, 1948. (Cited on p. 5)

[83] C.E. Shannon and W. Weaver. *The Mathematical Theory of Communication*. Univ. of Illinois Press, 1949. (Cited on p. 5)

[84] S. Sherman. Non-mean square error criteria. *IRE Trans. on Information Theory Inform. Theory*, 4:125–126, 1958. (Cited on p. 161)

[85] A. Stuart Sir Maurice Kendall and J.K. Ord. *The advanced theory of statistics : Voll. 1,2,3*. Griffin, High Wycombe U.K., 1983. (Cited on pp. 126, 133)

[86] J. Sjoberg, 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. (Cited on p. 248)

[87] Alexander Shapiro, Rank-reducibility of a Symmetric Matrix and sampling Theory of minimum Trace Factor Analysis *Psychometrika*, 47, 2, 187-199, June 1982 (Cited on p. 182)

[88] Alexander Shapiro, Identifiability of factor analysis: some results and open problems, *Linear Algebra Appl.*, 70 (1985), 1-7. Erratum, 125 (1989), 149. (Cited on p. 182)

[89] Stefano Soatto, Ruggero Frezza, and Pietro Perona. Motion estimation on the essential manifold. In *Computer Vision — ECCV 94*. 1994. (Cited on p. 238)

[90] T. Söderström and P. Stoica. *System Identification*. Prentice-Hall, 1989. (Cited on pp. 151, 286, 307)

[91] E. Sontag. Neural Networks for Control, chapter in *Essays on Control*, Birkhauser, Basel, Switzerland, 1993. (Cited on p. 248)

[92] G.W. Stewart. Collinearity and least squares regression. *Statistical Science*, 2:68–100, 1987. (Cited on pp. 152, 153)

[93] G.W. Stewart and J.G. Sun. *Matrix Perturbation Theory*. Academic Press, Boston, 1990. (Cited on p. 93)

[94] R. Tibshirani T. Hastie and M. Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman and Hall/CRC, 2015. (Cited on pp. 101, 104, 105)

[95] S. Theodoridis. *Machine Learning: A Bayesian and Optimization approach*. Academic Press, 2015. (Cited on pp. xiii, 38, 204)

[96] A. N. Tikhonov and V. Y. Arsenin. *Solutions of ill-posed problems*. V. H. Winston & Sons, Washington, D.C., 1977. (Cited on p. 96)

[97] A.N. Tikhonov. Solution of incorrectly formulated problems and the regularization method. *Soviet Mathematics, Translated from Russian*, 4, 1963. (Cited on p. 96)

[98] P. van Overchee and B. De Moor. Subpace algorithms for the stochastic identification problem. *Automatica*, 29:649–660, 1993. (Cited on p. 80)

[99] H. T. van Trees. *Detection Estimation and Modulation Theory Vol. I*. Wiley, 1976. (Cited on p. 37)

[100] V. Vapnik. *The nature of Statistical Learning*. Springer Texts in Statistics, 995. (Cited on pp. xiii, 38, 139)

[101] G. Wahba. *Spline methods for observational data*. SIAM CBMS-NSF series, Philadelphia, 1990. (Cited on pp. 97, 113, 115, 155, 184, 231, 234, 259)

[102] G. S. Watson. *Statistics on Spheres*. Wiley, New York, 1983. (Cited on pp. 239, 240)

[103] N. Wiener. Generalized harmonic analysis. *Acta Mathematica*, 55:117–258, 1930. (Cited on p. 285)

[104] N. Wiener. Tauberian theorems. *Annals of Mathematics*, 33:1–100, 1932. (Cited on p. 251)

[105] N. Wiener. *The Fourier integral and certain of its appplications*. Cambridge U.P., 1933. (Cited on p. 285)

[106] N. Wiener and P. Masani. The prediction theory of multivariate stochastic processes. I. The regularity condition. *Acta Math.*, 98:111–150, 1957. (Cited on p. 196)

[107] J.C. Willems. Relfections on armax systems. In *Conference on Econometrics, Time Series Analysis and Systems Theory*, Vienna, 2009. (Cited on p. 54)

[108] Eugene Wong and Bruce Hajek. *Stochastic processes in engineering systems*. Springer Texts in Electrical Engineering. Springer-Verlag, New York, 1985. (Cited on p. 208)

[109] R. F. Wynn and K. Holden. *An Introduction to Applied Econometric Analysis*. Halsted Press, New York, 1974. (Cited on p. 75)

[110] B. Yang. Projection approximation subspace tracking. *IEEE Transactions on IEEE Transactions on Signal Procesing*, 43:95–107, 1995. (Cited on p. 216)

[111] N. Young. *An Introduction to Hilbert Space*. Cambridge University Press, Cambridge, 1988. (Cited on p. 321)

[112] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, 68:49–67, 2006. (Cited on p. 99)

[113] P.W. Zehna. Invariance of maximum likelihood estimators. *Annals of Math. Statist.*, 37:744, 1966. (Cited on p. 23)