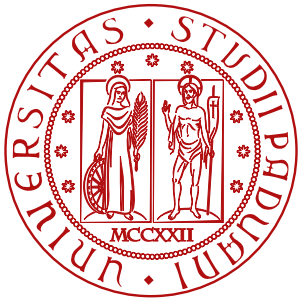

UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI INGEGNERIA CIVILE, EDILE E AMBIENTALE
Department Of Civil, Environmental and Architectural Engineering

Corso di Laurea Magistrale in Mathematical Engineering

TESI DI LAUREA

An Empirical Bayes approach to ARX estimation

Laureando:

Timofei Leahu

Matricola 2039113

Relatore:

Prof. Giorgio Picci

Anno accademico 2023/2024

Abstract

Empirical Bayes inference is based on estimates of the a priori parameter distribution obtained from the observed data. There are claims in the literature that in certain cases this technique may even be superior to Maximum Likelihood. In our opinion however this claim does not seem to be substantiated in sufficient generality as it may depend on various factors such as model class, data length, randomly varying parameters etc. In this work we compare Empirical Bayes estimators with a standard estimation technique for linear Autoregressive models with inputs (called ARX models) for time series. Such a comparison, can only make sense for a (realistic) finite data length; In this setting, it turns out that Empirical Bayes tends indeed to behave slightly better and so also in the case of slowly varying random parameters.

Contents

1

Introduction

Time series modeling is usually based on fitting certain stationary linear models called AR, ARX or ARMA to observed data. The standard estimation technique is based on least squares, but this technique requires stationary data to provide reliable results and may be optimal only asymptotically in the data length N. Even with a reasonable data size one may sometimes obtain unsatisfactory models. This is due to a variety of reasons such as poor order selection, nonlinearity and, especially, non-stationary errors which make the parameters of the model randomly variable in time. One may hope that describing the parameters as random variables and using a Bayesian approach could be a reasonable way to tackle the last problem. However the choice of a prior on the parameter space turns out to be a difficulty which may vanify this hope. In most cases of practical interest no a priori knowledge of a reasonable probabilistic description of the parameters is available. This research tries to attack the problem by using an *Empirical Bayes* philosophy. Empirical Bayes tries to estimate a prior distribution from the observed data. It looks like a naive idea and in fact it was proposed without much success, a long time ago. It is only recently that the idea has been revitalized due to availability of massive computer power and suitable algorithms see e.g. [3],[5],[19], [16], [14], [1], [15].

In this work we shall attempt to analyze the behavior of Empirical Bayes estimates of ARX models comparing their performance with a competing method based on standard Bayesian philosophy. Admittedly this is the simplest class of linear models of time series and its choice is essentially due to simplicity of the analysis. More general or more specific model classes require *ad hoc* methods

which are outside the scope of this thesis. See however [15] for an example.
In the chapter 2 an introduction background is reported in order to have some basics about probability, linear models, Bayesian statistics and time series, on which is constructed the next chapter.
Since asymptotically all estimation methods for ARX models tend to coincide, to work out a meaningful comparison we shall have to deal with *finite data* estimates. For this reason in the first part of the Chapter 3 we shall review and somewhat deepen some known results on estimation of static linear models. The performance of Empirical Bayes estimates and its comparison with another estimation method (essentially the standard Bayesian maximum a posteriori) will be based on this background material.
The chapter 4 contains a comparison example between the two methods mentioned above through the Mean Squared Error (MSE). The analysis is done using two different settings of parameters: fixed and randomly varying. Then some interesting results arise from the analysis and confirm the developed theory.

2 Background

2.1 Elements of Probability Theory & Notations

In this section we shall just list some basic concepts which are referred to in the next sections. The main reference is Shiryaev treatise [17].
A **Probability Space** $\{\Omega, \mathcal{A}, P\}$ 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 $P\colon \mathcal{A} \to [0,1]$.
Random variables are (mesurable) functions $\mathbf{x}\colon \Omega \to \mathbb{R}$. The *Probability distribution function* of $\mathbf{x}$ is

$$F(x) \coloneqq P\{\omega \mid \mathbf{x}(\omega) \leq x\}, \quad x \in \mathbb{R},$$

a right-continuous non-decreasing monotonic function.
The **Expectation** of a random variable is the integral

$$\mathbb{E}\,\mathbf{x} \coloneqq \int_{\Omega} \mathbf{x}(\omega) P(d\omega) = \int_{\mathbb{R}} x\, dF(x).$$

Random variables are denoted by **lower case boldface symbols** such as $\mathbf{x}$, $\mathbf{y}$, ... etc. Other notations like using Upper case symbols like X, Y, ... are not convenient since upper case symbols are standard for MATRICES such as covariances or loading matrices in linear models. As one need to work with multivariate statistics this notation would produce confusion. The sample size

is usually denoted by N: lower case n or m etc is normally 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 ∞. The acronym 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) $p(x) := \frac{dF(x)}{dx}$.

2.2 Classical Statistical Inference

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 assignment of a probability is *never certain*. There are several criteria which may lead to a decision that a certain P describes "well" the results of an experiment but these criteria may have different purposes and merits and may even not be comparable on an objective basis.
Typically, a statistical inference problem consists of:

- A space of experiments $\{\Omega, \mathcal{A}\}$;
- A family $\mathcal{P}$, or a number of disjoint families $\mathcal{P}_i$, $i = 1, \ldots, n$ (where n is *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.

2.2.1 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

$$\mathcal{P} = \{P_\theta \mid \theta \in \Theta\}, \tag{2.1}$$

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 n disjoint subsets (Θ_i, $i = 1, \ldots, n$) of Θ such that $\mathcal{P}_i = \{P_\theta \mid \theta \in \Theta_i\}$, $i = 1, \ldots, n$. The coin tossing problem 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.

2.2.2 Random Variables

In general a scientist shall need to consider groups of random variables describing simultaneous measurements of physical or economic variables from the same physical experiment, so there will be a need to consider *vector-valued*, say $\mathbb{R}^m$-valued, random variables. They will occasionally be called *random vectors* but this only when there is a need to distinguish sharply whether $m > 1$ or not.

When $m = 1$ we shall talk about *scalar* random variables. The abbreviation r.v. will sometimes also be used. Random variables (i.e. vectors) will be written as *column vectors* and always be denoted by **boldface** letters, such as $\mathbf{x}, \mathbf{y}$ etc. The components of $\mathbf{x}$ will generally be indexed by a subscript, say $\mathbf{x}_k$; $k = 1, 2, \ldots, m$. *Sequences* of random variables (possibly vector valued) will be denoted as functions of the discrete index t like $\{\mathbf{x}(t)\,;\, t = 1, 2, \ldots\}$. Of course a finite sequence can be organized as a vector and there is not a sharp distinction of the two objects in this case. The *variance* of a random variable $\mathbf{x}$ having (m-vector valued) mean μ is the $m \times m$ matrix

$$\mathrm{Var}\{\mathbf{x}\} := \mathbb{E}\,(\mathbf{x} - \mu)(\mathbf{x} - \mu)^\top \tag{2.2}$$

Which is always positive semidefinite and symmetric. The *Trace* of the variance matrix is called the *scalar variance* of $\mathbf{x}$ and is denoted by the lower case symbol:

$$\mathrm{var}\{\mathbf{x}\} := \mathrm{Tr}\,[\mathrm{Var}\{\mathbf{x}\}]\,.$$

Obviously $\mathrm{var}\{\mathbf{x}\}$ is the sum of the variances of its components.

The value $x \in \mathbb{R}^m$ taken by the random variable $\mathbf{x}$ corresponding to the exit ω of a random experiment, i.e. $\mathbf{x}(\omega) = x$ is called the *sample value* of $\mathbf{x}$.
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 $\mathbb{R}^m$. This space is always considered 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 $\{\mathbb{R}^m, \mathcal{B}^m\}$. There is a "canonical" representation of a random variable on its sample space as the identity function. This representation is very handy since it permits to identify **y just by assigning its** PDF. This is often tacitly understood, 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 $\mathbb{R}$

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

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 written as

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

where the "functional form" (i.e. the analytic dependence on the independent variable y) of each F_θ is a priori known, so that 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 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 *variance matrix* $\Sigma \in \mathbb{R}^{m\times m}$ are the parameters which the family depends on. Here we have assumed that the variance matrix Σ of $\mathbf{y}$ is invertible, which is a condition of *non-redundancy* of the components of $\mathbf{y}$ (i.e. non redundant observations in an inference problem).

2.2.3 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 $\{p(\cdot,\theta) \mid \theta \in \Theta\}$ and let x_0 be an observed value of $\mathbf{x}$.

Definition 2.2.1. *The* Likelihood function *of the observation x_0 is the function $L(x_0,\cdot) : \Theta \to \mathbb{R}_+$ (the nonnegative reals) defined by*

$$L(x_0,\theta) := p(x_0,\theta) \quad . \tag{2.3}$$

The "Maximum Likelihood principle", was introduced (of course not under this name) by Carl Friedrich Gauss in 1809 [**Gauss-809**] who was trying to estimate the orbit of Ceres, a Jupiter satellite, to show that it was a conic (in fact an ellipse). To deal with the unavoidable measurement errors, he first figured out the mathematical form of the distribution function of these errors, inventing what we now call a Gaussian density, and stated the principle that a best way to deal with these errors was to maximize the probability distribution (which he derived for this purpose) after substituting the actual measurements in place of the variables of the function. The justification of this optimization problem is described in much detail (in latin) in the famous memory *Theoria Motus corporun coelestium* [**Gauss-809**] [1] and led for the first time to a theoretical justification of Least Squares as a solution of the optimization problem. Actually Least Squares have been around earlier, sometimes attributed to Legendre, but before Gauss they were presented just as an empirical method of approximation without any rational justification.

The idea was successively broadened and popularized by R.A. Fisher, [**Fisher-25**] who formally stated the likelihood principle to assume as estimate of θ, corresponding to the observation x_0, the parameter value $\hat{\theta} \in \Theta$ which maximizes

[1] Reprinted in Vol 11 of the 12 volume series of his *Collected Works* by Julius Springer [**Gauss-01**].

$L(x_0, \cdot)$, that is

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

implicitly assuming that a maximum exists. The parameter value $\hat{\theta}$ renders *a posteriori* x_0 the most probable observation one could see, according to the family $\{p(\cdot, \theta) \mid \theta \in \Theta\}$.

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

$$L\Big(\mathbf{x}, \hat{\theta}(\mathbf{x})\Big) = \max_{\theta \in \Theta} p(\mathbf{x}, \theta) \quad . \tag{2.4}$$

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

$$\ell(\mathbf{x}, \cdot) = \log L(\mathbf{x}, \cdot) \tag{2.5}$$

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

$$\frac{\partial \ell}{\partial \theta_k}(\mathbf{x}, \theta) = 0\,, \qquad k = 1, \ldots, p \quad , \tag{2.6}$$

and then checking which solutions correspond to a maximum of $\ell(\mathbf{x}, \cdot)$. In general however one can only solve (2.6) 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 (2.6) provide the equations

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

the first of which is depending only on θ_1 which yields

$$\hat{\theta}_1 = \frac{1}{N} \sum_{1}^{N} y_i = \bar{y}_N \quad . \tag{2.7}$$

Substituting this expression in the second equations we easily find

$$\hat{\theta}_2^2 = \frac{1}{N} \sum_{1}^{N} (y_i - \bar{y}_N)^2 = \hat{\sigma}_N^2 \quad . \tag{2.8}$$

that is, the maximum likelihood estimator of θ_2^2 is the sample variance. It is immediate to check that the expressions (2.7) and (2.8) provide an absolute maximum of $\ell(x, \cdot)$. Summarizing:

Proposition 2.2.1 (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.*

This is a theoretical justification of **Least Squares** since to maximize the

likelihood, the estimator of the mean θ_1 must minimize

$$\frac{1}{\theta_2^2}\sum_{k=1}^{N}(y_k-\theta_1)^2\,.$$

As we shall see, the result holds practically unchanged in the multivariable case.

ML estimators for the multivariate Gaussian density

Theorem 2.2.1. *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

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

so that the negative log-likelihhod can be written

$$-\ell_N(\mathbf{y},\mu,\Sigma)=\frac{N}{2}\,\{\log\det\Sigma+\text{Trace}\ \hat{\Sigma}_N\Sigma^{-1}\}+\sum_{k=1}^{N}(\bar{\mathbf{y}}_N-\mu)^\top\Sigma^{-1}(\bar{\mathbf{y}}_N-\mu)\,. \quad (2.9)$$

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

$$\text{Trace}\ Y-\log\det Y\geq p \quad (2.10)$$

Proof. In fact, from the spectral decomposition $Y=U^*\,\text{diag}\{\lambda_1,\ldots,\lambda_p\}U$ with

$UU^* = I_p$ the inequality (2.10) 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 (2.10). □

[INTEGRATE WITH WHAT I NEED, AFTER REPORTED THEORY IN THE NEXT SECTIONS] [Start in the next section with Linear Statistical Models, empirical prediction-error minimization principle,]

2.3 Parameter Estimation for Linear Models

2.3.1 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

$$\mathbf{y} = h(\theta, \mathbf{w}) \quad , \tag{2.11}$$

where h is a known function and $\mathbf{w}$ is a random vector having 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 observed data. 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 [**Willems-09**]. 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 (2.11) 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, at least 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 (2.11). Probably the earliest example is the model used in experimental Physics called *theory of errors* which was originated by Gauss [**Gauss-809**] 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

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

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's 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

$$\mathbf{y} = s(\theta) + \mathbf{w} \quad , \tag{2.12}$$

where

$$s(\theta) = [s_1(\theta), \dots, s_N(\theta)]^\top \ , \qquad \mathbf{w} = \begin{bmatrix} \mathbf{w}_1 & \dots & \mathbf{w}_N \end{bmatrix}^\top \tag{2.13}$$

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 metrology and digital communication theory. 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 covariance matrix of the full vector $\mathbf{y}$ described by the model (2.13) is the same as the covariance of the noise vector

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

which may take into account 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

$$\mathbf{y} = s(\theta) + \sigma \mathbf{w} \quad , \tag{2.14}$$

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 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 (2.14) by scaling all members of (2.12) multiplying from the left both members by the inverse of a square root of R; i.e.

$$R^{-1/2}\mathbf{y} = R^{-1/2} s(\theta) + R^{-1/2}\, \mathbf{w} \quad , \tag{2.15}$$

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

$$s(\theta) = S\theta \quad , \qquad S \in \mathbb{R}^{N \times p} \quad , \tag{2.16}$$

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

$$\mathbf{y} = S\theta + \mathbf{w}\,, \qquad \mathbf{w} \sim \mathcal{N}(0, \sigma^2 R) \tag{2.17}$$

which naturally should be compared with the deterministic least-squares model.

2.4 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 (2.17); 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

$$\begin{aligned}\ell(y,\theta,\sigma^2) &= -\frac{N}{2}\,\log 2\pi - \frac{1}{2}\,\log\left[\det(\sigma^2 R)\right] - \frac{1}{2}\,(y-S\theta)^\top\,(\sigma^2 R)^{-1}\,(y-S\theta)\\ &= -\frac{N}{2}\,\log 2\pi - \frac{N}{2}\,\log\sigma^2 - \frac{1}{2}\log\det R - \frac{1}{2\sigma^2}\,(y-S\theta)^\top\,R^{-1}\,(y-S\theta)\,,\end{aligned} \tag{2.18}$$

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

$$I(\theta,\sigma) = \mathbb{E}_{\theta,\sigma}\begin{bmatrix}\mathbf{z}_\theta\mathbf{z}_\theta^\top & \mathbf{z}_\theta\mathbf{z}_\sigma\\ \mathbf{z}_\theta^\top\mathbf{z}_\sigma & \mathbf{z}_\sigma^2\end{bmatrix}. \tag{2.19}$$

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. $\frac{\|\tilde{\mathbf{y}}\|^2}{\sigma^2} \sim \chi^2(N)$. Hence, since the expected value of $\frac{\|\tilde{\mathbf{y}}\|^2}{\sigma^2}$ is exactly N it follows that

$$\mathbb{E}_{\theta,\sigma}\, \mathbf{z}_\sigma^2 = \mathbb{E}_{\theta,\sigma} \left\{ \frac{1}{2\sigma^2} \left[\frac{\|\tilde{\mathbf{y}}\|^2}{\sigma^2} - N \right] \right\}^2 = \frac{1}{4\sigma^4}\, \mathrm{Var} \left[\frac{\|\tilde{\mathbf{y}}\|^2}{\sigma^2} \right] = \frac{N}{2\sigma^4}\,. \tag{2.20}$$

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

$$I(\theta, \sigma^2) = \begin{bmatrix} \frac{1}{\sigma^2}\, S^\top R^{-1} S & 0 \\ 0 & \frac{N}{2\sigma^4} \end{bmatrix}. \tag{2.21}$$

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.

Proposition 2.4.1. *Let (2.17) be a model with $N \geq p$ scalar observations. Then θ is globally identifiable if and only if*

$$\operatorname{rank} S = p. \tag{2.22}$$

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 $\operatorname{rank} S = p$ 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:

$$\hat{\theta}(y) = [S^\top R^{-1} S]^{-1} S^\top R^{-1} y \,. \tag{2.23}$$

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 **??**. We shall use again the compact notation

$$\hat{\theta}(y) = Ay \quad , \qquad A := [S^\top R^{-1} S]^{-1} S^\top R^{-1} \,. \tag{2.24}$$

Theorem 2.4.1. *The ML estimator* (2.23) *of the parameter θ in the linear model* (2.17)

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*

$$\mathrm{Var}\{\hat{\theta}(\mathbf{y})\} = \sigma^2 [S^\top R^{-1} S]^{-1} \tag{2.25}$$

   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,*

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

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

$$AS = I \quad . \tag{2.26}$$

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. □

2.4.1 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

$$\hat{\mathbf{y}}_{N+1} = s_{N+1}\hat{\theta}(\mathbf{y})\,. \tag{2.27}$$

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 2.4.2. *The* Prediction Error *incurred by the ML (or by the Markov) predictor* (2.27)*, namely*

$$\mathbf{y}_{N+1} - \hat{\mathbf{y}}_{N+1} = s_{N+1}\left[\,\theta - \hat{\theta}(\mathbf{y})\right] + \sigma\mathbf{w}_{N+1} \tag{2.28}$$

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. □

2.5 Bayesian Statistics

2.5.1 Introduction to Bayesian estimation

In this section 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, in general infinitely many, possible values of the unknown parameter, described as determinations of a random variable (or of a finite-dimensional random vector[2]) $\mathbf{x}$ 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 $\mathbf{x}$ which is sufficient to justify the adoption of the Bayesian approach. For example, there is in general a known "nominal value" of $\mathbf{x}$ 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 θ (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 it. 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 $\mathbf{x}$ is by its very nature not directly measurable, it

[2]To adhere to standard conventions the dimension of this random vector will now be n instead of p.

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 [**Efron-05**]. It would be a mistake to disregard it.

We shall now proceed to describe formally what is meant by a Bayesian estimation problem. Recall that in this work 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 we shall, for the moment, 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

$$p_{\mathbf{x}}(x) = p_{\mathbf{x}}(x_1, \ldots, x_n) \tag{2.29}$$

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

$$f(y \mid x) = P(y \leq \mathbf{y} \leq y + dy \mid \mathbf{x} = x)/dy\,. \tag{2.30}$$

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 (2.29) and (2.30) 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 (2.29) and (2.30) one can get the

conditional density of $\mathbf{x}$ given the observation $\mathbf{y} = y$ as,

$$p(\mathbf{x} = x \mid \mathbf{y} = y) = \frac{f(y \mid x)\, p_{\mathbf{x}}(x)}{\int_{\mathbb{R}^n} f(y \mid x)\, p_{\mathbf{x}}(x)\, dx} = \frac{p_{\mathbf{y},\mathbf{x}}(y, x)}{p_{\mathbf{y}}(y)} \tag{2.31}$$

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 (2.31) is commonly called the *a posteriori* probability density of $\mathbf{x}$.

However Bayes formula (2.31) 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

$$\mathbf{y} = h(\mathbf{x}, \mathbf{w}) \tag{2.32}$$

where the parameter θ in (2.11) 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 $\mathbf{w} = 0$ (2.32) 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 (2.32) the conditional density $f(y \mid x)$ and the posterior (2.30) 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 (2.32), namely

$$y = h(x, w) \quad , \tag{2.33}$$

In this scheme, the problem of Bayesian point estimation appears as that of solving the equation (2.33) 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 (2.33). 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 2.5.1. *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

$$c(z - x) = \|z - x\|_Q^2 \quad , \tag{2.34}$$

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*,

$$R(z, y) := \mathbb{E}\left[c(z - \mathbf{x}) \mid \mathbf{y} = y\right] = \int_{\mathbb{R}^n} c(z - x)\, f(x \mid y)\, dx \quad , \tag{2.35}$$

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 $R(z, y)$,

$$\hat{x} = \operatorname{Arg} \min_z R(z, y) \tag{2.36}$$

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 2.5.1. *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 (2.36) is the conditional mean of* $\mathbf{x}$ *given* $\mathbf{y} = y$*,*

$$\mu(y) = \mathbb{E}(\mathbf{x} \mid \mathbf{y} = y). \tag{2.37}$$

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

$$R(z, y) = \mathbb{E}\left[c(z - \mathbf{x}) \mid \mathbf{y} = y\right] = \mathbb{E}\left[c(z + \mathbf{x}) \mid \mathbf{y} = y\right] = \mathbb{E}\left[c(\mathbf{x} - z) \mid \mathbf{y} = y\right]$$

the last equality following from the symmetry of c. Therefore, by strict convexity it also holds that

$$R(z, y) = \mathbb{E}\left[\frac{c(\mathbf{x} - z) + c(\mathbf{x} + z)}{2} \mid \mathbf{y} = y\right] > \mathbb{E}\left[c(\mathbf{x}) \mid \mathbf{y} = y\right] = R(0, y), \quad z \neq 0$$

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 $\mathbb{E}\left[\Delta\mathbf{x} \mid \mathbf{y} = y\right] = 0$ and $\Delta z - \Delta\mathbf{x} = z - \mathbf{x}$ which implies

$$R(z, y) = E\left[c(\Delta z - \Delta\mathbf{x}) \mid \mathbf{y} = y\right] > E\left[c(\Delta\mathbf{x}) \mid \mathbf{y} = y\right] = R(\mu(y), y), \quad z \neq \mu(y)$$

that is $\mu(y) = \text{Arg } \min_z R(z, y)$. □

The result may be different for non-symmetric probability distributions.

Proposition 2.5.1. *Let* $\mathbf{x}$ *be a scalar random variable and let* $c(z) = |z|$*, then*

$$\text{Arg} \min_z \mathbb{E}\left[|z - \mathbf{x}| \mid \mathbf{y} = y\right]$$

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

$$\mathbb{E}\,|\,\mathbf{x} - z\,| = \int_z^{+\infty} (x - z)\,p(x)\,dx + \int_{-\infty}^{z} (z - x)\,p(x)\,dx = \\ = \int_z^{+\infty} x\,p(x)\,dx - z(1 - F(z)) + zF(z) - \int_{-\infty}^{z} x\,p(x)\,dx$$

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 2.5.1 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 [**Sherman-58**]. 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

$$c(z, x) = \|z - x\|^2\,, \tag{2.38}$$

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 2.5.2 below.

Proposition 2.5.2. *For a quadratic cost function like* (2.38) *and, more generally, such as* (2.34)*, 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 (2.38) 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.

Remark 2.5.1. Since under our assumptions on c there always is uniqueness of the minimum (2.36), 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. 2.1.

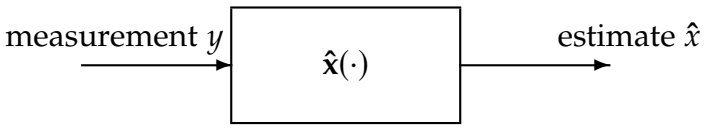

Figure 2.1: Estimator.

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 2.5.1, 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 2.5.2. *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,*

$$\mathbb{E}(\mathbf{x} \mid \mathbf{y}) = \mathrm{Arg} \min_{g(\cdot)} \mathbb{E}\,\|\mathbf{x} - g(\mathbf{y})\|_Q^2 \tag{2.39}$$

for an arbitrary symmetric semi positive 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

$$\mathbb{E}\,\|\mathbf{x}-g(\mathbf{y})\|_Q^2 = \int_{\mathbb{R}^m} \mathbb{E}\left[\,\|\mathbf{x} - g(\mathbf{y})\|_Q^2 \mid \mathbf{y} = y\right] p_{\mathbf{y}}(y)\,dy = \mathbb{E}\{\mathbb{E}\left[\|\mathbf{x} - g(\mathbf{y})\|_Q^2 \mid \mathbf{y}\right]\} \tag{2.40}$$

and then use the identity

$$\mathbb{E}\left\{\|\mathbf{x}-\mathbb{E}(\mathbf{x}\mid\mathbf{y})+\mathbb{E}(\mathbf{x}\mid\mathbf{y})-g(\mathbf{y})\|_Q^2\mid\mathbf{y}\right\}=\mathbb{E}\left\{\|\mathbf{x}-\mathbb{E}(\mathbf{x}\mid\mathbf{y})\|_Q^2\mid\mathbf{y}\right\}+$$
$$+2\mathbb{E}\left\{[\mathbf{x}-\mathbb{E}(\mathbf{x}\mid\mathbf{y})]^\top Q\left[\mathbb{E}(\mathbf{x}\mid\mathbf{y})-g(\mathbf{y})\right]\mid\mathbf{y}\right\}+\mathbb{E}\left\{\|\mathbb{E}(\mathbf{x}\mid\mathbf{y})-g(\mathbf{y})\|_Q^2\mid\mathbf{y}\right\}$$

where the second term on the right is zero since

$$\mathbb{E}\left\{[\mathbf{x}-\mathbb{E}(\mathbf{x}\mid\mathbf{y})]^\top Q\left[\mathbb{E}(\mathbf{x}\mid\mathbf{y})-g(\mathbf{y})\right]\mid\mathbf{y}\right\}=[\mathbb{E}(\mathbf{x}\mid\mathbf{y})-\mathbb{E}(\mathbf{x}\mid\mathbf{y})]^\top Q\left[\mathbb{E}(\mathbf{x}\mid\mathbf{y})-g(\mathbf{y})\right]$$

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*[3], 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, the equality above is an obvious and 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.

2.5.2 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

[3]Note that this is a different notion than unbiasedness in the Fisherian approach.

the type

$$G(z, y) := \mathbb{E}\left[\gamma(z - \mathbf{x}) \mid \mathbf{y} = y\right] = \int_{\mathbb{R}^n} \gamma(z - x)\, f(x \mid y)\, dx \quad , \tag{2.41}$$

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

$$\hat{x}_{MAP}(y) := \operatorname{Arg}\max_z f(z \mid y) \tag{2.42}$$

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. It is actually closely related to the maximum likelihood estimator in classical parametric Statistics see Example 2.5.1 below.
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 2.5.1 (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 summary, 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 details below.

2.5.3 Conditional expectation of Gaussian random vectors

Theorem 2.5.3. *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,*

$$\mu_{\mathbf{z}} = \begin{bmatrix} \mu_{\mathbf{x}} \\ \mu_{\mathbf{y}} \end{bmatrix} \tag{2.43}$$

and Covariance matrix

$$\Sigma_{\mathbf{z}} = \begin{bmatrix} \Sigma_{\mathbf{x}} & \Sigma_{\mathbf{xy}} \\ \Sigma_{\mathbf{yx}} & \Sigma_{\mathbf{y}} \end{bmatrix} \quad , \tag{2.44}$$

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:

$$\mathbb{E}(\mathbf{x} \mid \mathbf{y}) = \mu_{\mathbf{x}} + \Sigma_{\mathbf{xy}} \Sigma_{\mathbf{y}}^{-1} (\mathbf{y} - \mu_{\mathbf{y}}) \tag{2.45}$$

$$\mathrm{Var}(\mathbf{x} \mid \mathbf{y}) = \Sigma_{\mathbf{x}} - \Sigma_{\mathbf{xy}} \Sigma_{\mathbf{y}}^{-1} \Sigma_{\mathbf{yx}} \quad . \tag{2.46}$$

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 (2.44). Introduce a linear transformation of the variables $(\bar{\mathbf{x}}, \bar{\mathbf{y}})$ of the following form

$$\begin{cases} \tilde{\mathbf{x}} = \bar{\mathbf{x}} + A\bar{\mathbf{y}} \\ \tilde{\mathbf{y}} = \bar{\mathbf{y}} \end{cases} \tag{2.47}$$

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

$$A = -\Sigma_{\mathbf{xy}}\, \Sigma_{\mathbf{y}}^{-1} \quad . \tag{2.48}$$

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 $\mathbb{E}\,(\bar{\mathbf{x}} \mid \bar{\mathbf{y}})$. To this end use the first equality in (2.47) and note that by Gaussianness $\tilde{\mathbf{x}}$ and $\bar{\mathbf{y}}$ are independent so that $\mathbb{E}\,(\tilde{\mathbf{x}} \mid \bar{\mathbf{y}}) = \mathbb{E}\,(\tilde{\mathbf{x}}) = 0$ 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

$$\mathbb{E}\,(\bar{\mathbf{x}} \mid \bar{\mathbf{y}}) = -A\bar{\mathbf{y}} = \ \Sigma_{\mathbf{xy}}\, \Sigma_{\mathbf{y}}^{-1}\, \bar{\mathbf{y}} \tag{2.49}$$

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 (2.47) is an upper triangular matrix with an identity on the main diagonal it has a determinant equal to one. Then

$$p_{\bar{\mathbf{z}}}(z) = p_{\tilde{\mathbf{z}}}(\tilde{z})\Big|_{\substack{\tilde{x} = x + Ay \\ \tilde{y} = y}} = p_{\tilde{\mathbf{x}}}(x + Ay)\; p_{\bar{\mathbf{y}}}(y) \tag{2.50}$$

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 (2.49) and the expression

for the conditional covariance is

$$\begin{aligned} \mathrm{Var}\,(\bar{\mathbf{x}} \mid \bar{\mathbf{y}}) = \Sigma_{\tilde{\mathbf{x}}} &= \mathbb{E}\,(\bar{\mathbf{x}} + A\bar{\mathbf{y}})\,(\bar{\mathbf{x}} + A\bar{\mathbf{y}})^\top \\ &= \mathbb{E}\,\bar{\mathbf{x}}\bar{\mathbf{x}}^\top + \mathbb{E}\,A\,\bar{\mathbf{y}}\bar{\mathbf{x}}^\top \\ &= \Sigma_{\mathbf{x}} - \Sigma_{\mathbf{xy}}\,\Sigma_{\mathbf{y}}^{-1}\,\Sigma_{\mathbf{yx}} \quad . \end{aligned} \tag{2.51}$$

Finally (2.45) is obtained by reintroducing the mean values in (2.49). □

When $\Sigma_{\mathbf{y}}$ is singular one can give similar expressions where the inverse is replaced by the (Moore-Penrose) pseudoinverse of $\Sigma_{\mathbf{y}}$.

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 (2.47) is the *residual estimation error*

$$\tilde{\mathbf{x}} := \bar{\mathbf{x}} - \Sigma_{\mathbf{xy}}\,\Sigma_{\mathbf{y}}^{-1}\,\bar{\mathbf{y}} = \mathbf{x} - \mathbb{E}\,(\mathbf{x} \mid \mathbf{y}) \quad . \tag{2.52}$$

As shown in the proof of Theorem 2.5.3 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 (2.46) 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 (2.45) or (2.49), since

$$\mathrm{Var}\,\mathbb{E}\,(\mathbf{x} \mid \mathbf{y}) = \Sigma_{\mathbf{xy}}\,\Sigma_{\mathbf{y}}^{-1}\,\Sigma_{\mathbf{yx}} \quad . \tag{2.53}$$

Therefore the difference (2.46) 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 (2.53)) is to $\Sigma_{\mathbf{x}}$, the more efficient is the estimator.

Bayesian inference for Gaussian pdf's.

The calculation of the conditional density in (2.50) 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 2.5.3 that the a posteriori density $f(\theta \mid \mathbf{y} = y)$ is still Gaussian, with (conditional) mean and (conditional) variance given by formulas (2.45) and (2.46). 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

$$\hat{\mathbf{y}}(\mathbf{x}) = S\mathbf{x} \tag{2.54}$$

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

$$\mathbf{y} = S\mathbf{x} + \mathbf{w} \tag{2.55}$$

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 (2.55) 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 (2.55). One way to go is to exchange the role of $\mathbf{x}$ and $\mathbf{y}$ in the proof of Theorem 2.5.3, 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 (2.55). Finally, assuming a zero-mean prior we get

$$\hat{\mathbf{x}}(\mathbf{y}) = \Sigma_{\mathbf{x}}S^{\top}\left[\Sigma_{\mathbf{w}} + S\Sigma_{\mathbf{x}}S^{\top}\right]^{-1}\mathbf{y}\,, \tag{2.56}$$

$$\mathrm{Var}\left\{\mathbf{x} - \hat{\mathbf{x}}(\mathbf{y})\right\} = \Sigma_{\mathbf{x}} - \Sigma_{\mathbf{x}}S^{\top}\left[\Sigma_{\mathbf{w}} + S\Sigma_{\mathbf{x}}S^{\top}\right]^{-1}S\Sigma_{\mathbf{x}} \tag{2.57}$$

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 2.5.5, in the context of linear estimation. In particular see the remark 2.1.

Example 2.5.2. *A scalar random variable* $\mathbf{x}$ *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*

$$f(y \mid \mathbf{x} = \theta) \equiv \mathcal{N}\left(\mathbf{1}_N\theta, \sigma^2 I_N\right). \tag{2.58}$$

Assume the random variable 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* (2.58) *.*
Solution: The posterior is still Gaussian with conditional mean and conditional variance derived by formulas (2.56) and (2.57). We have $S = \mathbf{1}_N$ and $\Sigma_{\mathbf{w}} = \sigma^2 I_N$ so that

$$\hat{\mathbf{x}}(\mathbf{y}) = \mu + \tau^2\mathbf{1}_N^{\top}\left(\sigma^2 I_N + \tau^2\mathbf{1}_N\mathbf{1}_N^{\top}\right)^{-1}(\mathbf{y} - \mathbf{1}_N\,\mu) \tag{2.59}$$

$$\sigma^2_{\tilde{\mathbf{x}}} = \tau^2 - \tau^2\mathbf{1}_N^{\top}\left(\sigma^2 I_N + \tau^2\mathbf{1}_N\mathbf{1}_N^{\top}\right)^{-1}\mathbf{1}_N\,\tau^2 \tag{2.60}$$

At first sight the calculation of the inverse needed in these expressions looks quite demanding. They are facilitated by the use of the *Matrix Inversion Lemma*.

2.5.4 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 2.5.2. *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

$$\min_{A\ ,\ b} \left\{ \mathbb{E}\ \|A\mathbf{y} + b - \mathbf{x}\|^2 \ \mid A \in \mathbb{R}^{n\times m} \ , \ b \in \mathbb{R}^n \right\}. \tag{2.61}$$

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 $\hat{\mathbf{x}} = \widehat{\bar{\mathbf{x}}} + \mu_{\mathbf{x}}$ and substitute back $\bar{\mathbf{y}} = \mathbf{y} - \mu_{\mathbf{y}}$ to obtain the formula valid for non zero mean values (see Problem **??** below for a formal justification). We are henceforth to consider the minimization problem with zero-mean variables,

$$\min_{A} \ \mathbb{E} \ \|A\bar{\mathbf{y}} - \bar{\mathbf{x}}\|^2 \quad . \tag{2.62}$$

2.5.5 The linear model

Let us consider observation models (2.32) 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

$$\mathbf{y} = S\mathbf{x} + \mathbf{w} \tag{2.63}$$

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 (2.63) 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

$$\hat{\mathbf{x}}(\mathbf{y}) = \mu_{\mathbf{x}} + \Sigma_{\mathbf{xy}}\, \Sigma_{\mathbf{y}}^{-1}\, (\mathbf{y} - \mu_{\mathbf{y}}) \quad . \tag{2.64}$$

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 (2.63). 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 (2.63) by $\mathbf{x}$ and taking expectation, as

$$\Sigma_{\mathbf{yx}} = S\, \Sigma_{\mathbf{x}} = SP$$

whence

$$\Sigma_{\mathbf{xy}} = PS^{\top} \quad . \tag{2.65}$$

It then follows that

$$\Sigma_{\mathbf{y}} = SPS^{\top} + R \tag{2.66}$$

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

$$\hat{\mathbb{E}}(\mathbf{x} \mid \mathbf{y}) = PS^{\top}(SPS^{\top} + R)^{-1}\,\mathbf{y} \quad . \tag{2.67}$$

The residual error variance, denoted Λ, follows readily from the general formula

$$\mathrm{Var}\,[\mathbf{x} - \hat{\mathbf{x}}(\mathbf{y})] = \Sigma_{\mathbf{x}} - \Sigma_{\mathbf{xy}}\,\Sigma_{\mathbf{y}}^{-1}\,\Sigma_{\mathbf{yx}} \tag{2.68}$$

and is given by

$$\Lambda = P - PS^{\top}(SPS^{\top} + R)^{-1}\,SP \quad . \tag{2.69}$$

Remark 2.1. One may ask how general is the model (2.63) 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 (2.63).

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

$$\mathbf{y} = \hat{\mathbb{E}}(\mathbf{y} \mid \mathbf{x}) + \tilde{\mathbf{y}} \tag{2.70}$$

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 (2.70) is formally identical to (2.63). 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 2.5.3 for the derivation of the posterior density in case of jointly Gaussian variables. ◊

There is an alternative expression of the formulas (2.67) and (2.69) 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.

Theorem 2.5.4. *Assume that the a priori variance matrix P of the random vector $\mathbf{x}$ in the model* (2.63) *is invertible. Then the linear m.v. estimator of $\mathbf{x}$ can also be expressed in the form*

$$\hat{\mathbb{E}}(\mathbf{x} \mid \mathbf{y}) = \mu_{\mathbf{x}} + (P^{-1} + S^{\top}R^{-1}S)^{-1}\ S^{\top}R^{-1}(\mathbf{y} - \mu_{\mathbf{y}}) \tag{2.71}$$

and the relative error variance matrix as

$$\Lambda = (P^{-1} + S^{\top}R^{-1}S)^{-1} \quad . \tag{2.72}$$

Proof. Formula (2.72) is immediately obtained from (2.69) just by setting $A = P^{-1}$, $B = S^{\top}$ and $C = R^{-1}$ in the the Matrix Inversion Lemma formula (**??**).

The expression (2.71), can be obtained by the following sequence of steps

$$\begin{aligned}
\hat{E}(\mathbf{x} \mid \mathbf{y}) &= PS^{\top}\Big[R^{-1} - R^{-1}\,S(S^{\top}R^{-1}S + P^{-1})^{-1}\,S^{\top}R^{-1}\Big]\,\mathbf{y} \\
&= \Big[P - PS^{\top}R^{-1}\,S(S^{\top}R^{-1}S + P^{-1})^{-1}\Big]\ S^{\top}R^{-1}\mathbf{y} \\
&= \Big[P(S^{\top}R^{-1}\,S + P^{-1}) - PS^{\top}R^{-1}S\Big](S^{\top}R^{-1}S + P^{-1})^{-1}\,S^{\top}R^{-1}\,\mathbf{y}
\end{aligned}$$

and noting that the last term between square brackets is the identity. □

Formulas (2.71) and (2.72) 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 (2.71), assuming S is of full column rank, reduces to

$$\hat{\mathbb{E}}(\mathbf{x} \mid \mathbf{y}) = (S^{\top}R^{-1}S)^{-1}\ S^{\top}R^{-1}\mathbf{y} \tag{2.73}$$

which is the weighted least-squares estimate of classical parametric Statistics (2.23) with weighting matrix equal to R^{-1}, which has variance

$$\Lambda = (S^{\top}R^{-1}S)^{-1} \quad . \tag{2.74}$$

2.5.6 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 **y** and **x** is

$$p_{\mathbf{y},\mathbf{x}}(y,\theta) = p(y \mid \mathbf{x} = \theta)p_{\mathbf{x}}(\theta) \tag{2.75}$$

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 **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 **y** and **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* **x** and $\mathbf{w} \sim \mathcal{N}(\mu, \tau^2)$ is another random variable of mean μ and variance τ^2. The multiplicative relation (3.12) implies that **e** and **w** must be **independent** (show this). Therefore from

$$\mathbf{y} = \mu + \mathbf{w} + \mathbf{e}\,,$$

one can easily conclude that **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".

2.6 Time Series

2.6.1 Introduction: Discrete-time signals

In this section 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*.

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 mul-

tiple 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 scalars but also applies to vector-valued signals modulo notational complicatios.

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 section. The reader should however be warned that most signals of interest in econometrics and technology are actually *random*; since a deterministic signal, which is therefore by definition a priori known, does not convey any information. In the next section we shall extend these operations to random signals.

2.6.2 Stationary Time Series

In science, econometrics and engineering, discrete-time signals are usually noisy i.e. not describable as deterministic objects and of course not predictable as such unless some additional information is present. In order to construct reasonable predictors of these sequences one must confront the issue of modeling *serial correlation* of successive data. We shall therefore imagine the signal to be a *chunk of trajectory drawn from a stochastic process* $\{\mathbf{y}(t)\}$. In this context the signal is then called a **Time Series**. There is no space here to deal with stochastic processes in depth. Mostly we shall refer to *stationary processes* because stationarity permits simple modes. Most often it will be understood as *wide sense stationarity*.

This implies that one can describe the data by *constant coefficient models* which makes them adapt to statistical estimation.
In virtually all situations of interest the data cannot be modeled as the result of independent (or uncorrelated) measurements. In particular, since the ordering of samples here is of utmost importance, it will not be appropriate to describe the data as an i.i.d. sample sequences as in classical Statistics.

In this section we want to discuss a very simple class of linear dynamical models with constant coefficients which are very often used to describe time series 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,

$$\mathbf{y}(t) \;=\; \sum_{k=1}^{n} a_k \mathbf{y}(t-k) + \sum_{k=1}^{m} b_k \mathbf{u}(t-k) + \mathbf{w}(t)\,, \tag{2.76}$$

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, or ARMAX, which involve a *moving average* input noise component made of a linear combination of delayed noise samples such as

$$\mathbf{y}(t) \;=\; \sum_{k=1}^{n} a_k \mathbf{y}(t-k) + \sum_{k=1}^{m} b_k \mathbf{u}(t-k) + \sum_{k=0}^{r} c_k \mathbf{w}(t-k)\,, \tag{2.77}$$

whose estimation however becomes a *nonlinear problem* which requires different

tools than those subsume in this course. For them we shall have to refer to the literature, see e.g. [12],[18].

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

$$\mathbf{y}(t) = \boldsymbol{\varphi}(t)^\top \theta + \mathbf{w}(t)\,, \qquad t \in \mathbb{Z} \tag{2.78}$$

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 (2.17) 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

$$\mathbf{y} = \Phi_N\, \theta + \mathbf{w} \tag{2.79}$$

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

$$\hat{\mathbf{y}}_\theta(t \mid t-1) = \boldsymbol{\varphi}(t)^\top \theta \tag{2.80}$$

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 (3.5) 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 [**Hastie-T-F-09**].

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**
   $$V_N(\theta) := \frac{1}{N}\sum_{t=1}^{N} \varepsilon_\theta(t)^2 \tag{2.81}$$
   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*,

$$\hat{\sigma}_N^2 := \; V_N(\hat{\theta}_N) \tag{2.82}$$

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

$$\hat{\theta}_N = \left[\sum_{t=1}^{N} \varphi(t)\varphi(t)^\top\right]^{-1} \sum_{t=1}^{N} \varphi(t)\mathbf{y}(t)\,. \tag{2.83}$$

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.

Strong consistency of the least squares AR estimator

Naturally it is very difficult, if not impossible, to say anything about the statistical properties of the estimate (2.83) 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 2.6.1. *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*

$$\mathbb{E}_{\theta_0} \mathbf{y}(s)\, \mathbf{w}(t) = 0\,; \qquad \forall t > s \in \mathbb{Z}\,; \tag{2.84}$$

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

$$\hat{\boldsymbol{\theta}}_N = \left[\frac{1}{N} \sum_{t=1}^{N} \varphi(t)\varphi(t)^\top \right]^{-1} \frac{1}{N} \sum_{t=1}^{N} \varphi(t)\mathbf{y}(t)\,; \tag{2.85}$$

and substitute $\mathbf{y}(t) = \varphi(t)^\top \theta_0 + \mathbf{w}(t)$ (true model). Then define the sample covariance matrix of $\boldsymbol{\varphi}(t)$

$$\hat{\boldsymbol{\Sigma}}_N := \frac{1}{N} \sum_{t=1}^{N} \boldsymbol{\varphi}(t)\boldsymbol{\varphi}(t)^\top \ \in \mathbb{R}^{n\times n}\,. \tag{2.86}$$

Next we shall need the following fact.

Lemma 2.6.1. *Assume* (2.84) *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*

$$\Sigma_n := \mathbb{E}_{\theta_0}\{ \begin{bmatrix} \mathbf{y}(t-1) \\ \dots \\ \mathbf{y}(t-n) \end{bmatrix} \begin{bmatrix} \mathbf{y}(t-1) & \dots & \mathbf{y}(t-n) \end{bmatrix} \}\,. \tag{2.87}$$

Under the stated assumptions, this matrix is in fact positive definite.

Proof. The ergodicity follows from the consequence (B.3) of Corollary B.0.2 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 (2.84) 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.
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 (2.85), 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. □

3

Comparison of Empirical Bayes & Bayesian Estimators

3.1 ARX models and pseudo-Linear Regression

Assume we have a sequence of training data $\{y(t),\, u(t)\}$ where the (discrete) time index t runs from some initial time t_0 on, which we want to describe by a simple linear finite-dimensional stochastic 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 means describing them by a linear stochastic difference equation of the form

$$\mathbf{y}(t) \;=\; \sum_{k=1}^{n} a_k \mathbf{y}(t-k) + \sum_{k=1}^{m} b_k \mathbf{u}(t-k) + \mathbf{w}(t)\,, \tag{3.1}$$

where $\mathbf{w} := \{\mathbf{w}(t),\, t \in \mathbb{Z}\}$ is a process of random errors which is assumed Gaussian uncorrelated (white noise). An ARX model depends on $p := n + m$ unknown parameters which will be written as a column vector:

$$\theta := \begin{bmatrix} a_1 & \dots & a_n & b_1 & \dots & b_m \end{bmatrix}^\top \tag{3.2}$$

and on the unknown noise variance σ^2. It can be formally written in *pseudo-linear regression* form as

$$\mathbf{y}(t) = \boldsymbol{\varphi}(t)^\top \theta + \mathbf{w}(t)\,, \qquad t \in \mathbb{Z} \tag{3.3}$$

where $\boldsymbol{\varphi}(t)$ is a column vector depending on past data:

$$\boldsymbol{\varphi}(t)^\top = \begin{bmatrix} \mathbf{y}(t-1) & \dots & \mathbf{y}(t-n) & \mathbf{u}(t-1) & \dots & \mathbf{u}(t-m) \end{bmatrix} \tag{3.4}$$

Note that there is an important difference with the classical linear regression model, namely that now *the coefficient vectors $\boldsymbol{\varphi}(t)$ of the model are random, depending on the (past) input-output variables*.

Stacking sequentially the N equations (3.3) on top of each other we obtain a system of relations which, rewritten in vector form looks like

$$\mathbf{y} = \Phi_N\, \theta + \mathbf{w} \tag{3.5}$$

where the N-dimensional random vectors $\mathbf{y}$ and $\mathbf{w}$ 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

$$\Phi_N := \begin{bmatrix} \varphi(1)^\top \\ \vdots \\ \varphi(N)^\top \end{bmatrix}, \tag{3.6}$$

assuming the initial time t_0 is far enough, so that we can fill in Φ_N with the available data so as to describe the output from time $t = 1$ to $t = N$. Even with the assumption of Gaussian errors, because of the dependence of Φ_N on the observed data, the implementation of an *exact* Maximum Likelihood procedure to estimate the parameter θ turns out to be complicated, see e.g. [2, Chap. 8], [18, p.253] or [6] and no explicit variance expressions are available. One can however show that, asymptotically for $N \to \infty$ Maximum Likelihood for stable ARX models is equivalent to an empirical *prediction-error minimization* principle (to be recalled below) which is essentially Least Squares, see [18, p.207].
The function of the past data

$$\hat{\mathbf{y}}_\theta(t \mid t-1) = \boldsymbol{\varphi}(t)^\top\, \theta \tag{3.7}$$

is called the (one step ahead) *predictor function* associated to the model. Note that the predictor function is a linear function of θ but is also a function of the

previous $n + m$ past samples of the joint data process.

For an ARX model defined by a generic parameter vector θ, the N-dimensional vector of predictors formally looks like a linear function of the parameter and the prediction error vector has the form

$$\boldsymbol{\varepsilon}_\theta := \mathbf{y} - \Phi_N\, \theta\ .$$

The *average squared prediction error* $V_N(\theta)$ is then the squared Euclidean norm of $\boldsymbol{\varepsilon}_\theta$,

$$V_N(\theta) = \frac{1}{N}\, \|\mathbf{y} - \Phi_N\, \theta\|^2\, .$$

Some generalizations of this formula may involve a weighting of the errors perhaps variable in time. The Minimal Prediction Error (PEM) estimation principle [12] hence leads to the solution of a simple Least Squares problem and 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

$$\hat{\theta}_N = \left[\sum_{t=1}^{N} \varphi(t)\varphi(t)^\top\right]^{-1}\, \sum_{t=1}^{N} \varphi(t)\, \mathbf{y}(t)\, . \tag{3.8}$$

where we assume that the inverse exists for suitably large N.

Clearly $\hat{\theta}_N$ is a *non linear function of the past observed data* so we don't even know when it may be unbiased. Even if $\mathbf{y}$ and $\mathbf{u}$ were Gaussian, the variance (and indeed the pdf) of $\hat{\theta}_N$ for finite sample size is impossible to compute. In the literature one can mostly see analysis of what happens for $N \to \infty$. This is e.g. what is usually done in System Identification [12]. In this work we shall instead attempt to find *conditional* parameter estimates based on a fixed chunk of past observations. This approach will be the background of the next sections.

3.2 Conditional Linear Models and Marginal Gaussians

Let $\mathcal{P}_t$ denote the (strict) past observations $\{y(s), u(s); s < t\}$ at time t. For a Gaussian noise process we can interpret the estimator (3.8) as a *conditional* Maximum Likelihood estimator given $\mathcal{P}_t$. In fact, under conditioning with respect to

$\mathcal{P}_t$, which in this and in the next sections we shall maintain fixed throughout, we can assimilate the model (3.5) to a classical Gaussian linear model describing the random vector $\mathbf{y}$ as depending on a p-dimensional parameter vector θ, written

$$\mathbf{y} = \Phi\theta + \mathbf{w} \tag{3.9}$$

where the entries of Φ_N, written Φ for short, will be assumed known and treated as if they were **deterministic**. The error $\mathbf{w}$ is an unobserved N-dimensional zero-mean Gaussian vector with uncorrelated components of equal variance σ^2, which we write $\mathbf{w} \sim \mathcal{N}(0, \sigma^2 I_N)$, with $\mathbf{w}(t)$ independent of $\mathcal{P}_t$ at all times. The model (3.9) is sometimes called a "fixed effects" model since it depends on an unknown but deterministic parameter vector θ.

In a (conditional) Bayesian setting the parameter θ is modeled as a random vector which will be here denoted $\mathbf{x}$. The model (3.9) with $\mathbf{x}$ in place of θ

$$\mathbf{y} = \Phi\mathbf{x} + \mathbf{w} \tag{3.10}$$

is called a "random effects" model. The density of the random vector $\mathbf{y}$ having mean $\Phi\theta$ and variance $\sigma^2 I_N$ is now written as a *conditional density* [1] $p(y \mid \mathbf{x} = \theta)$ where the random vector $\mathbf{x}$ is described by some other density, $p_{\mathbf{x}}(x)$ which we shall also assume Gaussian, in general depending on unknown parameters say

$$p_{\mathbf{x}} \equiv \mathcal{N}(\theta, \sigma^2\Pi) \tag{3.11}$$

where $\theta \in \mathbb{R}^p$ and for convenience we have introduced a normalized *a priori* variance matrix $\Pi \in \mathbb{R}^{p\times p}$ which we shall assume positive definite. Both θ, and $\sigma^2\Pi$ are unknown. The *joint density* of $\mathbf{y}$ and $\mathbf{x}$ is

$$p_{\mathbf{y},\mathbf{x}}(y, x) = p(y \mid \mathbf{x} = x)p_{\mathbf{x}}(x) \tag{3.12}$$

which is again Gaussian and depends on the unknown a priori parameters. The *marginal distribution* of $\mathbf{y}$ could formally be obtained by integrating with respect to x (of course this distribution does not depend on $\mathbf{x}$ any more). The calculation in terms of density functions for Gaussian variables can be done in

[1] Naturally this density is conditioned also with respect to $\mathcal{P}_t$ but to simplify notations we shall not indicate this explicitely.

a straightforward way for linear models by expressing both $\mathbf{y}$ and $\mathbf{x}$ by means of two linear models as

$$\mathbf{y} = \Phi\mathbf{x} + \mathbf{w}\,, \qquad \mathbf{x} = \theta + \mathbf{e}$$

where $\mathbf{w} \sim \mathcal{N}(0, \sigma^2 I_N)$ and $\mathbf{e} \sim \mathcal{N}(0, \sigma^2\Pi)$ must be *independent* because of the multiplicative property (3.12). Therefore a trivial substitution to eliminate $\mathbf{x}$ leads to

$$\mathbf{y} = \Phi\theta + \mathbf{w} + \Phi\mathbf{e}\,. \tag{3.13}$$

This is a classical linear model from which one can easily get the **marginal distribution** of $\mathbf{y}$ which has mean $\Phi\theta$ and (conditional) variance $\sigma^2(I_N + \Phi\Pi\Phi^\top)$. In formulas

$$p_{\mathbf{y}} \equiv \mathcal{N}(\Phi\theta, \sigma^2 R) \tag{3.14}$$

where we have introduced the symbol $R := I_N + \Phi\Pi\Phi^\top$.

The marginal model (3.14) now depends on the unknown parameters θ, σ^2 and Π which are called *hyperparmeters* of the distribution. Hence, as suggested in the literature, e.g. [9, p. 262], inference on the Bayesian model (3.10) can be recast as inference on an equivalent fixed-effect model with an augmented set of parameters which are precisely the hyperparameters mentioned above. In theory their estimation could therefore be approached by applying the maximum likelihood principle to the *marginal likelihood* of (3.14).

In the present setting we can express the (conditional) maximum likelihood estimate of θ and its variance as an explicit function of the unknown hyperparameter Π. We just use the well-known Maximum Likelihood formulas for the linear model (3.13) to get the formal estimates:

$$\hat{\boldsymbol{\theta}} = \left(\Phi^\top R^{-1}\Phi\right)^{-1}\Phi^\top R^{-1}\mathbf{y} \tag{3.15}$$

$$\mathrm{Var}[\hat{\boldsymbol{\theta}}] = \sigma^2\left(\Phi^\top R^{-1}\Phi\right)^{-1}\,. \tag{3.16}$$

where the variance expression is derived from (3.15), taking into account that the variance of $\mathbf{y}$ as described by the model (3.13) is precisely $\sigma^2 R$ so that (3.16) comes directly from

$$\mathrm{Var}[\hat{\boldsymbol{\theta}}] = \left(\Phi^\top R^{-1}\Phi\right)^{-1}\Phi^\top R^{-1}\,\mathrm{Var}(\mathbf{y})R^{-1}\Phi\left(\Phi^\top R^{-1}\Phi\right)^{-1}\,.$$

The matrices in formulas (3.15) and (3.16) are (nonlinear) functions of the observations $\{y(s), u(s); s < N\}$ for some fixed sample size N. In particular $\mathbf{y}$ stands

for the random vector obtained by listing the components of the observed data available at time N, $\{y(s); s \leq N\}$, listed in increasing order on top of each other.

Naturally Π in these formulas is an unknown parameter and a question which should be addressed at this point is the *identifiability* of Π in the model (3.14). This requires to check the unique dependence (injectivity) of the covariance $\sigma^2(I_N + \Phi\Pi\Phi^\top)$ on the matrix variable Π which in turn is equivalent to checking whether the equation $\Phi\Delta\Phi^\top = 0$ has the unique solution $\Delta = 0$. This will clearly be true if the sample size is large enough to make Φ of full rank p almost surely, an assumption which guarantees identifiability of θ as well and will be assumed to hold all through. We should however warn the reader that this is a *finite sample* requirement. For sample size tending to infinity Π becomes non-identifiable. We shall comment on this later on in Sect. 3.5.

3.3 Relation to the Bayesian Estimate

To relate the marginal likelihood formulas (3.15), (3.16) to the Bayesian (a posteriori) estimate $\hat{\mathbf{x}} = \mathbb{E}[\mathbf{x} \mid \mathbf{y}]$ of the a priori variable $\mathbf{x}$ in the model (3.10), assuming for the moment that the a priori parameters σ^2, Π are known, one should rely on the standard formulas valid for the linear Bayesian model (3.10). After cancellation of the common factor σ^2, these standard formulas become

$$\hat{\mathbf{x}} = (\Phi^\top\Phi + \Pi^{-1})^{-1}\Phi^\top\mathbf{y} \tag{3.17}$$

$$\mathrm{Var}[\mathbf{x} \mid \mathbf{y}] = \mathrm{Var}(\mathbf{x} - \hat{\mathbf{x}}) = \sigma^2(\Phi^\top\Phi + \Pi^{-1})^{-1} \tag{3.18}$$

where Π is assumed invertible. Note, in contrast to (3.16), that (3.18) is actually the residual estimation **error** variance of the Bayes estimator $\hat{\mathbf{x}}$ since from (3.10) and the error variance formula

$$\mathrm{Var}[\mathbf{x} - \hat{\mathbf{x}}] = \mathrm{Var}(\mathbf{x}) - \mathrm{Cov}(\mathbf{x}, \mathbf{y})\,\mathrm{Var}(\mathbf{y})^{-1}\,\mathrm{Cov}(\mathbf{y}, \mathbf{x})$$

one gets precisely

$$\frac{1}{\sigma^2}\mathrm{Var}[\mathbf{x} - \hat{\mathbf{x}}] = \Pi - \Pi\Phi^\top\left(\Phi\Pi\Phi^\top + I\right)^{-1}\Phi\Pi = (\Pi^{-1} + \Phi^\top\Phi)^{-1} \tag{3.19}$$

where the last equality follows by the matrix inversion lemma.[2] Therefore (3.18) is not the variance of $\hat{\mathbf{x}}$ but the variance of the estimation error as in (3.19), that is the **conditional variance** of $\mathbf{x}$ given $\mathbf{y}$.

It turns out that the Bayesian estimator $\hat{\mathbf{x}}$ and the marginal ML estimator $\hat{\boldsymbol{\theta}}$ are not the same.

Proposition 3.3.1. *The conditional ML estimator of θ for the marginal model has variance*

$$\mathrm{Var}[\hat{\boldsymbol{\theta}}] = \sigma^2 \left(\Phi^\top R^{-1}\Phi\right)^{-1} = \sigma^2[(\Phi^\top\Phi)^{-1} + \Pi]\,. \tag{3.20}$$

Therefore $\hat{\mathbf{x}}$ and $\hat{\boldsymbol{\theta}}$ are in general different. In fact while $\hat{\boldsymbol{\theta}}$ is unbiased, in general $\mathbb{E}\,\hat{\mathbf{x}} \neq \theta$.

Proof. Follows from the identity

$$R^{-1} = [I + \Phi\Pi\Phi^\top]^{-1} = I - \Phi[\Pi^{-1} + \Phi^\top\Phi]^{-1}\Phi^\top$$

which, again by the matrix inversion lemma gives

$$\Phi^\top R^{-1}\Phi = \Phi^\top\Phi - \Phi^\top\Phi[\Phi^\top\Phi + \Pi^{-1}]^{-1}\Phi^\top\Phi = [(\Phi^\top\Phi)^{-1} + \Pi]^{-1}\,. \tag{3.21}$$

□

Hence, the variance of the estimator of θ based on the marginal model is the sum of the variance of the parameter in the classical fixed effect model plus the variance of the prior $\mathbf{x}$ (in fact of the noise $\mathbf{e}$). Clearly when $\Pi \to 0$ the variance of the conditional ML estimator tends to the classical fixed effect one while the variance (3.19) of the Bayes estimator tends to zero since the parameter becomes a *known* deterministic quantity.

3.4 The Empirical Bayes estimator

Assume now that we are able to compute some estimate of the a priori model noise parameters, say $\hat{\sigma}^2(\mathbf{y}), \hat{\Pi}(\mathbf{y})$, then following [9, p.262-263] the *Empirical Bayes estimator* of θ is obtained by substituting these estimates in place

[2] The lemma states that

$$[A + BCD]^{-1} = A^{-1} - A^{-1}B\,[C^{-1} + DA^{-1}B\,]^{-1}DA^{-1}$$

assuming that all indicated inverses exist.

of the unknown hyperparameters in the marginal density of $\mathbf{y}$ or, equivalently in the (conditional) Maximum Likelihood formulas (3.15), (3.16) as it turns out (and should be nearly obvious) that the Empirical Bayes principle leads to substitute the estimated quantities $\hat{\boldsymbol{\sigma}}^2(\mathbf{y}), \hat{\Pi}(\mathbf{y})$ in place of the unknown σ^2, Π in the expressions of the "marginal" ML estimates (3.15), (3.16) of the mean and variance of the "fixed effect" marginal model.
On the other hand, one could substitute the hyperparameter estimates also in the posterior density of $\mathbf{x}$ and use them in the expression of the Bayesian maximum a posteriori (MAP) estimate of $\mathbf{x}$. As we have seen, this in general may lead to different estimators. These two estimators will be called the *Empirical Bayes* and the *Bayes a posteriori* estimators.

A question that one would like to address in this context is the comparison of the MSE of Empirical Bayes estimators for ARX models with that of plain ML estimators. A folk conjecture in the literature is that in some circumstances the Empirical Bayes estimator may have a smaller mean square error (MSE) than the Maximum Likelihood estimator [8]. Unfortunately an explicit expression of the variance of the exact maximum likelihood estimator for ARX models for a finite sample size is unknown, of course unless one decides to refer to the *conditional* likelihood or introduces some extra assumption say, assuming that the noise variance σ^2 in (3.3) is known. It seems to us that the comparison would be unfair under such assumptions. For a fair comparison one should not involve unverifiable assumptions.

Since one can dispose of explicit formulas for the (mean and variance of the) two estimators only assuming that the past data have been observed, that is, conditioning everything on past data also in the Bayesian formulas (3.17), (3.18), it seems reasonable to restrict our analysis to a comparison of Empirical Bayes, with Bayes a posteriori, where the second estimator could generally be considered as an improvement on a conditional ML estimator. To address the comparison we first need expressions for the MSE of the two estimators. The next Lemma is instrumental for this calculation.

Lemma 3.4.1. *Let θ_0 be the true parameter in the fixed effect model* (3.9) *and assume that the normalized a priori variance Π of $\mathbf{x}$ is positive definite. Let $\Delta_N := \Phi_N^\top \Phi_N$, then the bias of the Bayes a posteriori estimator is*

$$\mathbb{E}\,(\hat{\mathbf{x}}) - \theta_0 = -[\Delta_N + \Pi^{-1}]^{-1}\Pi^{-1}\theta_0\,. \tag{3.22}$$

Proof. Follows from (3.17) by which $\mathbb{E}(\hat{\mathbf{x}}) - \theta_0 = [\Delta_N + \Pi^{-1}]^{-1}\Delta_N\theta_0 - \theta_0$ and the identity

$$\begin{aligned}[\Delta_N + \Pi^{-1}]^{-1}\Delta_N - \Delta_N^{-1}\Delta_N &= [\Delta_N + \Pi^{-1}]^{-1}[I_N - (\Delta_N + \Pi^{-1})\Delta_N^{-1}]\Delta_N \\ &= -[\Delta_N + \Pi^{-1}]^{-1}\Pi^{-1}.\end{aligned}$$

□

At this point we can display the formulas for the theoretical Mean Squared Errors of the two estimates.

Proposition 3.4.1. *The (scalar) MSE of the Bayes a posteriori estimator is*

$$\|\mathbb{E}(\hat{\mathbf{x}}) - \theta_0\|^2 + \operatorname{Tr}\operatorname{Var}(\hat{\mathbf{x}}) = \|[\Delta_N + \Pi^{-1}]^{-1}\Pi^{-1}\theta_0\|^2 + \sigma^2 \operatorname{Tr}[\Delta_N + \Pi^{-1}]^{-1} \quad (3.23)$$

while the marginal ML (which is to become Empirical Bayes) estimator is unbiased and its MSE coincides with the variance (3.20) *which can be written as* $\sigma^2 \operatorname{Tr}[\Delta_N^{-1} + \Pi]$.

The first formula follows from (3.19) (3.22). Note that both formulas assume that the hyperparameter Π is known which of course is never the case. Later we shall have to substitute the theoretical value with a suitable estimate. See Section 3.5.

To get some intuition on this formula let us first consider the case in which $\theta_0 = 0$ (which seems play the role of unbiasedness, but one should recall that the Bayes estimator is never unbiased) in which case we just compare variances. Then it is clear that for $\Pi \to 0$ the Empirical Bayes (i.e. the marginal ML) is always worse than the Bayesian estimator since

$$\Delta_N^{-1} + \Pi \geq (\Delta_N + \Pi^{-1})^{-1}$$

while for Π large the EB variance $\Delta_N^{-1} + \Pi$ diverges linearly while that of the Bayes estimator remains bounded above by Δ_N^{-1}.
However, since the Bayes estimator is never unbiased we need to analyze the effect of the squared bias term. Note that for small $\Pi \to 0$ the bias term reduces to $\|\theta_0\|^2$ which can be much larger than $\operatorname{Tr}\Delta_N^{-1}$ especially if N is large enough. Therefore if $\|\theta_0\|^2 > \operatorname{Tr}\Delta_N^{-1}$ there is room for the MSE of the Bayes estimate to be larger than that of the Empirical Bayes. This analysis is shown graphically in the Figure 3.1.

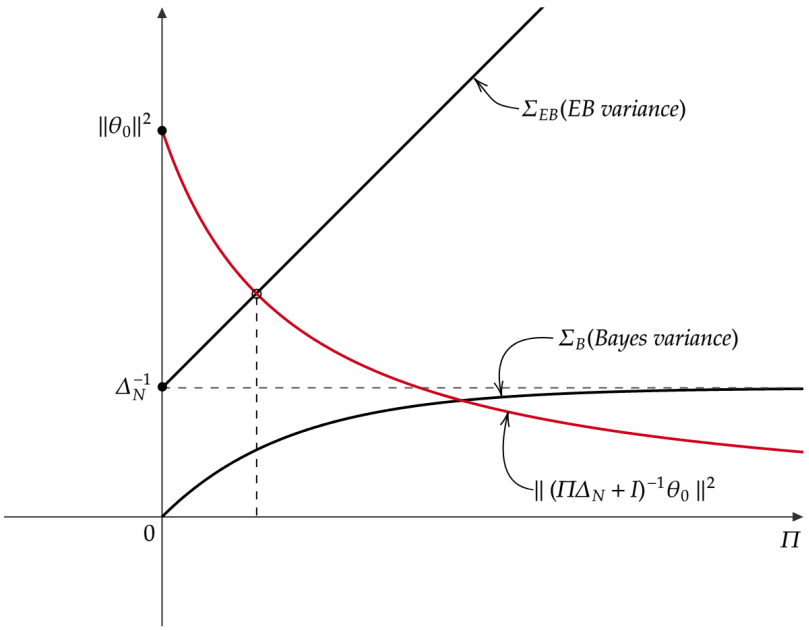

Figure 3.1: Behavior of the two variances and the bias norm squared (in red).

Moreover, the comparison between the two scalar MSEs is shown in the Figure 3.2. The important part to be appreciated in this figure is the not negligible interval in which the EB estimator gives better performance.

Example 3.4.1. Consider the case of a stationary AR model with a scalar parameter θ:

$$\mathbf{y}(t) = \theta \mathbf{y}(t-1) + \mathbf{w}(t), \qquad t = 1, \dots, N$$

in which $|\theta| < 1$, the normalized prior is simply denoted π and the scalar Δ_N is denoted $\delta_N^2 > 0$. The normalized mean square error of the Bayes estimate is

$$e_B^2(N) := \left(\frac{\pi^{-1}}{\delta_N^2 + \pi^{-1}} \right)^2 \theta^2 + \frac{1}{\delta_N^2 + \pi^{-1}} = \frac{\theta^2 + \pi + \delta_N^2 \pi^2}{1 + 2\delta_N^2 \pi + \delta_N^4 \pi^2},$$

which is to be compared with the normalized mean square error of the Empirical Bayes, $e_{EB}^2(N) = 1/\delta_N^2 + \pi$.

Let us first analyze the two errors in function of the prior π. For $\pi \to 0$, $e_B^2(N) \to \theta^2$ while $e_{EB}^2(N) \to 1/\delta_N^2$ while for $\pi \to \infty$, $e_B^2(N) \to 1/\delta_N^2$ while $e_{EB}^2(N)$ diverges to $+\infty$. As predicted earlier on, for large values of the prior

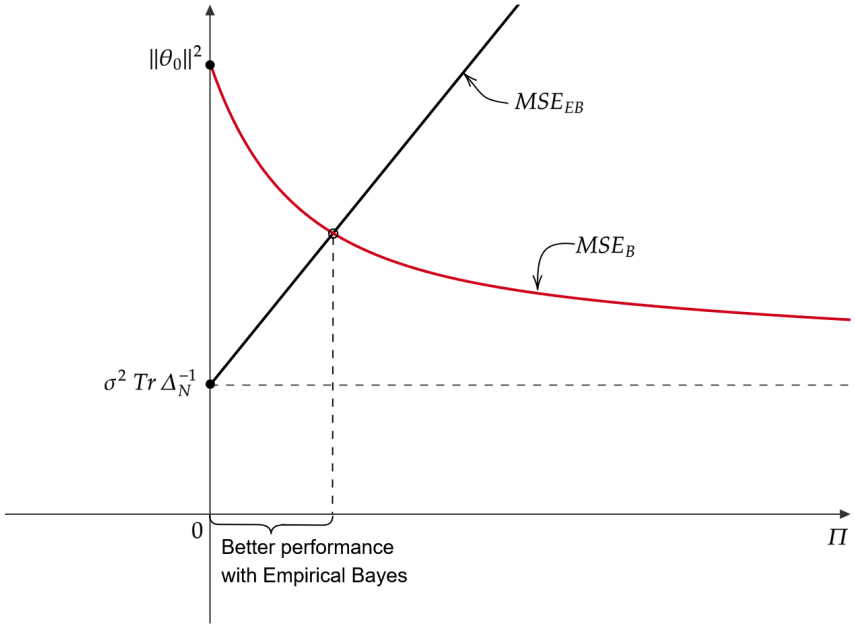

Figure 3.2: Comparison between Bayes MSE and Empirical Bayes MSE.

variance, $e^2_{EB}(N)$ is certainly larger than $e^2_B(N)$.
However for large values of N and small values of π the mean square error of the Bayes estimator can be much larger than $e^2_{EB}(N) \simeq 1/\delta^2_N$ since for $N \longrightarrow \infty$ the square norm δ^2_N must diverge (a necessary condition for ergodicity) and hence $1/\delta^2_N$ must converge to zero and be smaller than any preassigned parameter value θ^2. Incidentally this guarantees that the Bayes estimator is asymptotically unbiased (and its MSE tends to zero).

In conclusion we may say that for small values of the prior variance π and large enough sample size the Empirical Bayes can be a better estimator than the a posteriori Bayes (and hence of the Maximum Likelihood).

3.5 Hyperparameter estimation

We shall now consider the estimation of the hyperparameters of ARX models based on finite-data pseudo-linear regression models. It will be convenient to denote the a priori variance $\sigma^2\Pi$ by P_0.

The a posteriori Bayes estimates are given by

$$\hat{\mathbf{x}}(t) = \left(\Phi(t)^\top \Phi(t) + \Pi^{-1}\right)^{-1} \Phi(t)^\top \mathbf{y}^t, \tag{3.24}$$

$$\hat{P}(t) = \sigma^2 \left(\Phi(t)^\top \Phi(t) + \Pi^{-1}\right)^{-1}, \tag{3.25}$$

$$\hat{\sigma}^2(t) = \frac{1}{t} \|\mathbf{y}^t - \Phi(t)\hat{\mathbf{x}}(t)\|^2, \tag{3.26}$$

the first two of which require knowledge of both σ^2 and Π, the latter being the normalized a priori variance of $\mathbf{x}$.
Now the question is how should we estimate the hyperparameters, in particular the a priori variance Π or P_0. Can we recover them from a suitably long chunk of observations of the process $\mathbf{y}(t)$?

The estimates (3.24), (3.25), (3.26) can be computed recursively by a *conditional Kalman filter* algorithm which is recalled in the appendix. See (A.15), (A.11), (A.12). The unknown parameter $P_0 = \sigma^2\Pi$ plays the role of initial covariance data in the Riccati-type algorithm (A.12). Inserting this parameter estimate in the Bayesian formula (3.25) leads to a natural estimator of P_0^{-1} based on N data, given by

$$\hat{P}_0^{-1}(N) = \hat{P}(N)^{-1} - \frac{1}{\hat{\sigma}^2(N)} \Phi_N^\top \Phi_N \quad \text{or} \quad \hat{\Pi}^{-1}(N) = \hat{\sigma}^2(N)\hat{P}(N)^{-1} - \Phi_N^\top \Phi_N \tag{3.27}$$

which at first sight looks like a reasonable solution. However a first warning here is that for large N this difference could be quite small leading to error amplification.
On the other hand, N is usually assumed large enough to make $P(N)$ well defined and invertible. Note also that the sample of the joint process $\{\mathbf{y}(t), \mathbf{u}(t)\}$, once inserted in the p-dimensional data vector (3.4), should satisfy the following limit relation:

$$\lim_{t\to\infty} \frac{1}{t} \sum_{s=1}^{t} \boldsymbol{\varphi}(s)\boldsymbol{\varphi}(s)^\top = \Sigma_p \tag{3.28}$$

almost surely, where Σ_p is positive definite. This condition is essentially *second order ergodicity* of the joint process $\{\mathbf{y}(t), \mathbf{u}(t)\}$ which is always needed in time-series Statistics and can be shown to hold under stability of the AR part of the model [10, p. 510]. It implies that $\Phi(t)^\top \Phi(t)$ must also become positive definite for a suitable large t and, in fact, that this matrix must diverge as $t \to \infty$ at a linear rate with t which in turn, by a limit argument in the recursion (A.12) of

the appendix, implies that

$$\lim_{t\to\infty} P(t) = 0 \tag{3.29}$$

independently of the initial condition P_0. This even if $\hat{\sigma}_N^2(t)$ may converge to a constant (nonzero) positive value.

Hence, in the limit, of an *infinitely long* sequence of observations, the recover of P_0 would be ill-posed or impossible. For the limit (3.29) holds for **any positive semidefinite initial condition** P_0 and we may say that in the limit any positive semidefinite P_0 including $P_0 = 0$, could be a valid a priori variance. Therefore here we incur in *non-identifiability* of P_0. This also agrees with the fact that when $\Phi_N^\top \Phi_N$ becomes very large the influence of the prior in the expressions (3.25) and (3.24) becomes negligible and the estimates tend to coincide with maximum likelihood (i.e. plain least squares), where there is no prior involved.

To estimate the unknown parameters θ, σ^2 and Π we shall need a different idea. We shall propose a sequential procedure which is essentially a *Backward Kalman Filter* [10] adapted to the conditional ARX model. For simplicity we shall from here assume that the model (3.10) is from a purely autoregressive (AR) representation. The generalization to ARX models is straightforward and will be left to the reader.

The first step is to rewrite the pseudo-linear regression model (3.3) as a backward recursion. This is the content of the following lemma.

Lemma 3.5.1. *Let the process* $\{\mathbf{y}(t)\}$ *be represented by an n-th order AR model with vector coefficient* $\theta = \begin{bmatrix} a_1 & \dots & a_n \end{bmatrix}^\top$ *and let*

$$\bar{\varphi}(t) : \begin{bmatrix} \mathbf{y}(t+1) & \dots & \mathbf{y}(t+n) \end{bmatrix}^\top, \qquad t \in [0, N] \tag{3.30}$$

then $\mathbf{y}$ *can be also represented by a backward recursion of the form*

$$\mathbf{y}(t) = \bar{\varphi}(t)^\top \theta + \bar{\mathbf{w}}(t), \qquad t \in \mathbb{Z} \tag{3.31}$$

where $\bar{\mathbf{w}}(t)$ *is a stationary white noise with finite variance* $\bar{\sigma}^2$*, uncorrelated with all future variables* $\{\mathbf{y}(s);\, s > t\}$.

Proof. Let $S(z)$ be the spectral density of a stationary zero-mean process $\mathbf{y}$ represented by an AR model of order n; it is well known that such spectral density

must admit the spectral factorization

$$S(z) = \sigma^2 \frac{1}{a(z)} \frac{1}{a(z^{-1})} \tag{3.32}$$

where $a(z^{-1}) = 1 - a_1 z^{-1} - \ldots - a_n z^{-n}$ (the minus signs are for convenience) is a polynomial such that $z^n a(z)$ has all its zeros inside the unit circle while $a(z)$ has zeros of modulus greater than one, exactly in the reciprocal locations of those of $z^n a(z)$. The transfer function $\frac{1}{a(z^{-1})}$ leads to a familiar *causal* AR representation

$$\mathbf{y}(t) = \sum_{k=1}^{n} a_k \mathbf{y}(t-k) + \mathbf{w}(t)$$

where $\mathbf{w}(t)$ is a white noise of variance σ^2 which is uncorrelated with all past variables $\{\mathbf{y}(s) \mid s < t\}$ while $\frac{1}{a(z)}$ leads to an *anticausal AR representation*:

$$\mathbf{y}(t) = \sum_{k=1}^{n} a_k \mathbf{y}(t+k) + \bar{\mathbf{w}}(t) \tag{3.33}$$

where $\bar{\mathbf{w}}(t)$ is a white noise of same variance σ^2 which is uncorrelated with all *future* variables $\{\mathbf{y}(s) \mid s > t\}$. The two white processes are related by a transfer function

$$\bar{\mathbf{w}}(t) = \frac{a(z)}{a(z^{-1})} \mathbf{w}(t)$$

which has unitary values on the unit circle; in fact is a rational *inner function* see e.g. [10]. □

The recursion (3.33) is run backwards in time starting at some end point $t = N$, assuming that we have available a suitably long sequence of future measurements.
In our problem $\bar{\boldsymbol{\varphi}}(t)$ is a function of future measurements from time $t+1$ on and here, to estimate the initial conditions we shall apply a backward version of the conditional generalized Kalman filter, which is described in the theorem below.

Theorem 3.5.1. *Let $\bar{\mathbf{y}}^t$ denote the vector made with the available future components of*

y *at time t ordered with respect to a decreasing time index, so that*

$$\bar{\Phi}_t := \begin{bmatrix} \bar{\boldsymbol{\varphi}}(N)^\top \\ \cdots \\ \bar{\boldsymbol{\varphi}}(t)^\top \end{bmatrix} \tag{3.34}$$

which yields the pseudo-linear representation

$$\bar{\mathbf{y}}^t = \bar{\Phi}_t \mathbf{x} + \bar{\mathbf{w}}^t \tag{3.35}$$

then the conditional mean and the normalized conditional variance of the random parameter vector **x given the future data from time** *t:*

$$\bar{\mathbf{x}}(t) := \mathbb{E}\,[\mathbf{x} \mid \bar{\mathbf{y}}^t\,], \qquad \bar{P}(t) := \frac{1}{\sigma^2}\,\mathrm{Var}\,[\,\mathbf{x} \mid \bar{\mathbf{y}}^t\,] \tag{3.36}$$

satisfy the backward recursion

$$\bar{\mathbf{x}}(t-1) = \bar{\mathbf{x}}(t) + \bar{k}(t-1)[\,\mathbf{y}(t-1) - \bar{\boldsymbol{\varphi}}(t-1)^\top \bar{\mathbf{x}}(t)\,] \tag{3.37}$$

where the backward gain vector $\bar{k}(t-1)$ *is given by*

$$\bar{k}(t-1) = \bar{P}(t)\bar{\boldsymbol{\varphi}}(t-1)\big[\bar{\boldsymbol{\varphi}}(t-1)^\top \bar{P}(t)\bar{\boldsymbol{\varphi}}(t-1) + 1\big]^{-1} \tag{3.38}$$

while the backward covariance matrix $\bar{P}(t)$ *satisfies*

$$\bar{P}(t-1) = \bar{P}(t) - \bar{P}(t)\bar{\boldsymbol{\varphi}}(t-1)\big[\bar{\boldsymbol{\varphi}}(t-1)^\top \bar{P}(t)\bar{\boldsymbol{\varphi}}(t-1) + 1\big]^{-1}\,\bar{\boldsymbol{\varphi}}(t-1)^\top \bar{P}(t)\,. \tag{3.39}$$

The recursion

$$\bar{\lambda}^2(t-1) = \frac{t}{t+1}\bar{\lambda}^2(t) + \frac{1}{t+1}[\mathbf{y}(t-1) - \bar{\boldsymbol{\varphi}}(t-1)^\top \bar{\mathbf{x}}(t-1)\,]^2 \tag{3.40}$$

describes the estimate of the conditional variance of the backward innovation process $\bar{\mathbf{e}}(t) := \mathbf{y}(t) - \bar{\boldsymbol{\varphi}}(t)^\top \bar{\mathbf{x}}(t)$*. The corresponding estimate of* σ^2 *is expressed by the formula*

$$\bar{\sigma}^2(t) = \frac{\bar{\lambda}^2(t)}{\bar{\boldsymbol{\varphi}}(t)^\top \bar{P}(t)\bar{\boldsymbol{\varphi}}(t) + 1}\,. \tag{3.41}$$

This algorithm computes recursively in the backward direction, the conditional mean, the normalized backward *conditional variance* $\bar{P}(t)$ *of* $\mathbf{x}(t)$ *given the future history of*

y *and the conditional variance* $\bar{\sigma}^2(t)$ *of the* backward innovation process

Proof. The algorithm can be derived by following dual steps of those outlined in the appendix A. □

Remark 3.5.1. Of course the terminal conditions $\bar{x}(N)$ and $\bar{P}(N)$ are not known but could in principle be estimated by a *forward* conditional Kalman filter as described in the appendix. This algorithm of course depends on the proper initial conditions $\mathbf{x}_0$ and P_0 however, as we have shown, under certain rather general structural conditions on the signals of the true model generating the data, one can show that the Kalman filter tends to forget the initial conditions and its behavior is quite independent of them, of course provided they are chosen not too far from the true ones. The same will then be true also for the backward algorithm. One could *couple the two algorithms* and first run a forward filter to get preliminary estimates of $\mathbf{x}_N$ and P_N to be used as starting data for the backward filter but this procedure is seldom needed.

4

A Numerical Example

After having computed estimates of the hyperparameters of our ARX model by the backward procedure just described, the issue we are after is to compare the performance of Empirical Bayes with Bayesian estimators for the specific model class under study.

We shall simulate the simple autoregressive model

$$\mathbf{y}(t) = a_1\mathbf{y}(t-1) + a_2\mathbf{y}(t-2) + \mathbf{w}(t) \tag{4.1}$$

where $a_1 = 1.5$, $a_2 = -0.7$ and $\mathbf{w}$ is Gaussian i.i.d. with unit variance which generates an ergodic stationary process. We run the conditional Kalman Filter started at various initial conditions P_0, stopping the simulation at $N_1 = 50$, $N_2 = 100$, $N_3 = 200$ and compute the corresponding estimates of P_0 using the formula (3.27).
With these estimates compute the scalar variances and MSE of the Empirical Bayes and compare them with the Bayes MSE which can be computed using the differences of the estimates $\hat{\mathbf{x}}(N) - \theta_0$ where θ_0 is the true parameter+ the estimated Variance. We can use the formula

$$\text{Bayes MSE}(N) = \|\mathbb{E}\,\hat{\mathbf{x}}(N) - \theta_0\|^2 + \hat{\sigma}^2(N)\,\text{Tr}\;P(N) \tag{4.2}$$

(implicitly using the conditional expected value given the past data instead of the plain expectation) and compare this with the MSE of the Empirical Bayes, equal to the variance (3.20) which, after substituting the estimate of the prior

Initial P_0	$\hat{P}_0(N)$	N	Emp. Bayes Estimates ($\hat{\theta}_{EB}(N)$)	Emp. Bayes MSE	Bayes Estimates ($\hat{x}_B(N)$)	Bayes MSE
$\begin{bmatrix} .01 & 0 \\ 0 & .01 \end{bmatrix}$	$\begin{bmatrix} .01201 & -.00008 \\ -.00008 & .01194 \end{bmatrix}$	50	$\begin{bmatrix} 1.45760 \\ -.80157 \end{bmatrix}$	0.03674	$\begin{bmatrix} .72873 \\ -.16379 \end{bmatrix}$	0.90160
$\begin{bmatrix} .01 & 0 \\ 0 & .01 \end{bmatrix}$	$\begin{bmatrix} .01442 & -.00009 \\ -.00009 & .01434 \end{bmatrix}$	100	$\begin{bmatrix} 1.49922 \\ -.73953 \end{bmatrix}$	0.04079	$\begin{bmatrix} .98814 \\ -.26275 \end{bmatrix}$	0.46583
$\begin{bmatrix} .01 & 0 \\ 0 & .01 \end{bmatrix}$	$\begin{bmatrix} .01417 & -.00009 \\ -.00009 & .01409 \end{bmatrix}$	200	$\begin{bmatrix} 1.48003 \\ -.72652 \end{bmatrix}$	0.03201	$\begin{bmatrix} 1.12962 \\ -.40115 \end{bmatrix}$	0.23242
$\begin{bmatrix} .08 & 0 \\ 0 & .08 \end{bmatrix}$	$\begin{bmatrix} .63853 & -.03381 \\ -.03381 & .60759 \end{bmatrix}$	50	$\begin{bmatrix} 1.45760 \\ -.80157 \end{bmatrix}$	1.12862	$\begin{bmatrix} 1.19486 \\ -.55851 \end{bmatrix}$	0.15418
$\begin{bmatrix} .08 & 0 \\ 0 & .08 \end{bmatrix}$	$\begin{bmatrix} .60511 & -.03204 \\ -.03204 & .57579 \end{bmatrix}$	100	$\begin{bmatrix} 1.49922 \\ -.73953 \end{bmatrix}$	1.22632	$\begin{bmatrix} 1.40620 \\ -.64994 \end{bmatrix}$	0.07475
$\begin{bmatrix} .08 & 0 \\ 0 & .08 \end{bmatrix}$	$\begin{bmatrix} .47637 & -.02522 \\ -.02522 & .45329 \end{bmatrix}$	200	$\begin{bmatrix} 1.48003 \\ -.72652 \end{bmatrix}$	1.00381	$\begin{bmatrix} 1.42646 \\ -.67597 \end{bmatrix}$	0.03051

Table 4.1: Results of simulations with fixed parameters.

namely $\hat{P}_0(N)$ in place of P_0 becomes

$$\text{EmpBayes MSE}(N) = \hat{\sigma}^2(N)\,\text{Tr}\,\{[\Phi_N^\top \Phi_N]^{-1} + \hat{P}_0(N)\}. \tag{4.3}$$

The results are reported in Table 4.1.

In the table we have reported simulation results corresponding to two initial conditions (a priori variance matrices) $P_{0,i} = \sigma_i^2 \Pi_i$ (with $\sigma_i^2 = 1$ for simplicity):

$$P_{0,1} = \begin{bmatrix} 0.01 & 0 \\ 0 & 0.01 \end{bmatrix}, \quad P_{0,2} = \begin{bmatrix} 0.08 & 0 \\ 0 & 0.08 \end{bmatrix},$$

which, once compared with the relative estimates, can be classified as "medium" and "large", respectively. For instance, in the case of medium one with $N = 200$ samples the estimated prior variance gives a standard deviation $\sigma \approx 0.1187$ for both parameters $[a_1, a_2]$ and hence all the values of the two parameters are in the range $a_1 \pm 3\sigma$ and $a_2 \pm 3\sigma$, respectively, that for us seem to be a not so large deviation. Applying the same reasoning for the case of large prior then the standard deviation becomes $\sigma \approx 0.69$ for a_1 and $\sigma \approx 0.673$ for a_2 which for us seem to be high. For each $P_{0,i}$ we have calculated the relative estimate $\hat{P}_0$ using formula (3.27). In the right columns, there are listed the estimates of the parameters and the MSE both for the Empirical Bayes and Bayes a posteriori procedure. As we can see, for medium true prior (initial conditions for (A.12)),

the EB estimates are weakly dependent on the number of samples (N) and almost consistent, indeed with a small bias. Conversely, the Bayes estimates are strongly dependent on the number of samples (N), the higher N the better the estimates. For large P_0 we observe the same behavior of the parameter estimates, but with an increased quality of the Bayes ones. It is quite evident that for medium prior variance matrix the Empirical Bayes MSE is definitely smaller than the Bayes a posteriori one, and consequently Empirical Bayes behaves better, as it is also evident looking at the parameter estimates. Observe the decrease of the MSEs with the increase of the number of samples (N) as expected from theory.
Finally, for a considerably larger initial prior, one can clearly observe that the MSE of the Bayes a posteriori estimate becomes quite smaller than that of the Empirical Bayes estimates. Also now the prior variance estimates are very clearly biased.

Next, we consider a more realistic example generating the data $\mathbf{y}(t)$ by simulating the model (4.1) assuming *slowly varying random parameters*. We describe the parameter variation by the following linear random model:

$$\begin{bmatrix} a_1(t+1) \\ a_2(t+1) \end{bmatrix} = \begin{bmatrix} 1.5 \\ -.7 \end{bmatrix} + \begin{bmatrix} .98 & 0 \\ 0 & .97 \end{bmatrix} \begin{bmatrix} a_1(t) - 1.5 \\ a_2(t) + 0.7 \end{bmatrix} + \lambda \begin{bmatrix} \mathbf{w}_1(t) & 0 \\ 0 & \mathbf{w}_2(t) \end{bmatrix}$$

where $\mathbf{w}_1$ and $\mathbf{w}_2$ are independent white noises with a small standard deviation say λ equal to 0.01 or 0.02. The simulated behaviors of the output, compared with the nominal system, are shown in Figure 4.1, where the data correspond to the same trajectory of the generating noise $\mathbf{w}$.

By repeating the calculations of the previous example, we obtain the results shown in the tables 4.2 and 4.3 which correspond to two different standard deviations λ of the parameter fluctuation noise. In this case the behavior of the MSEs and estimates are similar to those with fixed parameters, instead we can observe the increases of the bias in the prior variance estimates, i.e. $\hat{P}_0(N)$.

The wording "medium" and a "large" a priori variance in the simulations is based on a comparison refering to the range of the empirical standard deviations around the mean of the parameters. When these two classes of prior variance are suitably fixed one can obtain a good comparison between the two techniques studied above. Once a reasonable “medium and large" prior variance is assigned as initial matrices, one can clearly see that the Empirical Bayes gives a nice estimate of parameters along a small MSE with respect to the Bayes a posteriori.

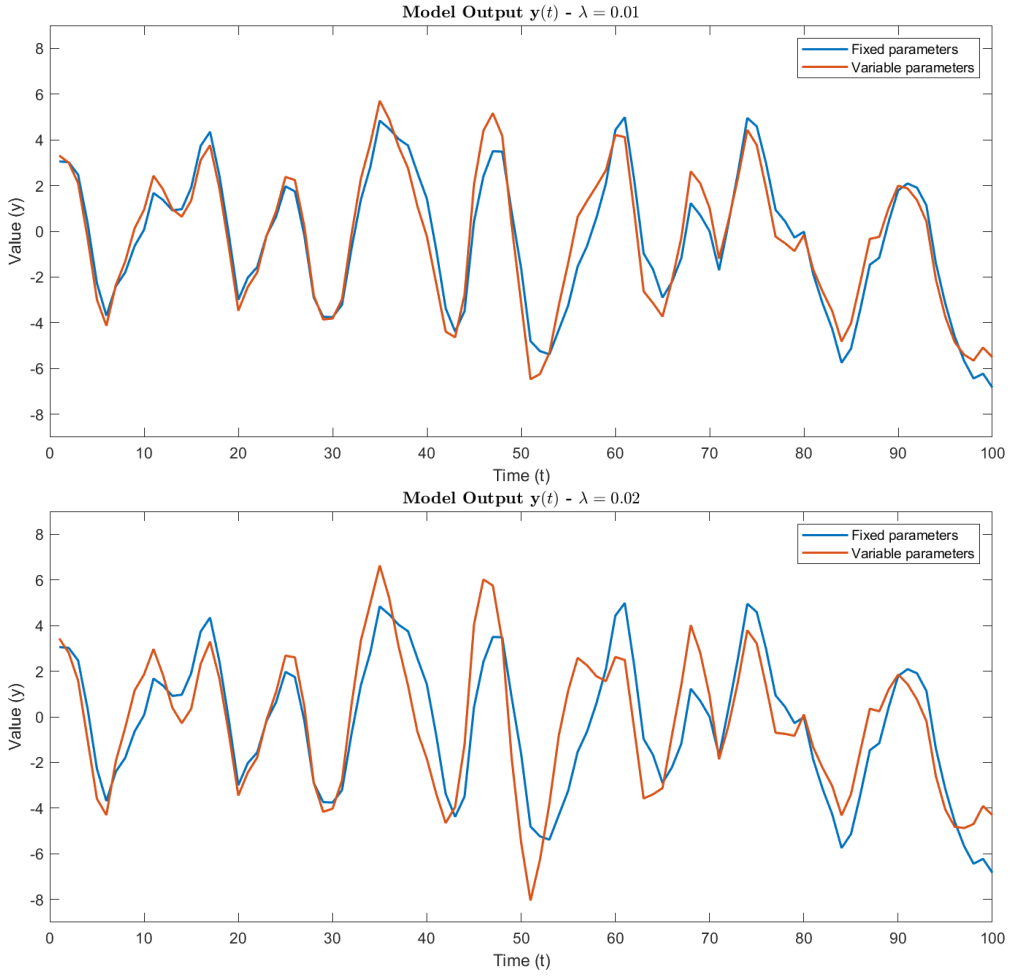

Figure 4.1: Outputs corresponding to N=100 and $\lambda = 0.01, 0.02$ for variable parameters. The plots with fixed parameters in the two graphics are identical.

In the case of randomly varying parameters there could be some strange fluctuations due to some non-linearity probably present in the data. Nonetheless, as number of samples increase the MSE decreases, in fact, as also the theory suggests.
For a relatively large prior variance the Empirical Bayes is no more a good alternative. The MSE becomes larger than the one of the Bayes a posteriori. Thus, with a large amount of data one should instead choose the Bayes a posteriori in order to have a more accurate estimation of the parameters.

From a numerical implementation point of view, when the two algorithms (i.e forward and backward Kalman filters) are coupled to find out the estimates of the hyperparameters, it is not necessary to write down two different codes to implement the recursive procedure, but it is enough to write down the forward

Initial P_0	$\hat{P}_0(N)$	**N**	**Emp. Bayes Estimates** $(\hat{\theta}_{EB}(N))$	**Emp. Bayes MSE**	**Bayes Estimates** $(\hat{x}_B(N))$	**Bayes MSE**
$\begin{bmatrix} .01 & 0 \\ 0 & .01 \end{bmatrix}$	$\begin{bmatrix} .01130 & -.00008 \\ -.00008 & .01110 \end{bmatrix}$	50	$\begin{bmatrix} 1.47650 \\ -.84354 \end{bmatrix}$	0.03319	$\begin{bmatrix} .78642 \\ -.23015 \end{bmatrix}$	0.74723
$\begin{bmatrix} .01 & 0 \\ 0 & .01 \end{bmatrix}$	$\begin{bmatrix} .01508 & -.00008 \\ -.00008 & .01484 \end{bmatrix}$	100	$\begin{bmatrix} 1.48029 \\ -.76864 \end{bmatrix}$	0.04286	$\begin{bmatrix} 1.01273 \\ -.33566 \end{bmatrix}$	0.38233
$\begin{bmatrix} .01 & 0 \\ 0 & .01 \end{bmatrix}$	$\begin{bmatrix} .01370 & -.00013 \\ -.00013 & .01356 \end{bmatrix}$	200	$\begin{bmatrix} 1.46703 \\ -.73076 \end{bmatrix}$	0.03106	$\begin{bmatrix} 1.12586 \\ -.41472 \end{bmatrix}$	0.22698
$\begin{bmatrix} .08 & 0 \\ 0 & .08 \end{bmatrix}$	$\begin{bmatrix} .48553 & -.02840 \\ -.02840 & .41492 \end{bmatrix}$	50	$\begin{bmatrix} 1.47650 \\ -.84354 \end{bmatrix}$	0.82156	$\begin{bmatrix} 1.33350 \\ -.70849 \end{bmatrix}$	0.09042
$\begin{bmatrix} .08 & 0 \\ 0 & .08 \end{bmatrix}$	$\begin{bmatrix} .63864 & -.02725 \\ -.02725 & .55624 \end{bmatrix}$	100	$\begin{bmatrix} 1.48029 \\ -.76864 \end{bmatrix}$	1.37825	$\begin{bmatrix} 1.40196 \\ -.69341 \end{bmatrix}$	0.06532
$\begin{bmatrix} .08 & 0 \\ 0 & .08 \end{bmatrix}$	$\begin{bmatrix} .45993 & -.03710 \\ -.03710 & .42031 \end{bmatrix}$	200	$\begin{bmatrix} 1.46703 \\ -.73076 \end{bmatrix}$	0.85930	$\begin{bmatrix} 1.41844 \\ -.68503 \end{bmatrix}$	0.02890

Table 4.2: Results of simulations with variable parameters, $\lambda = 0.01$.

Initial P_0	$\hat{P}_0(N)$	**N**	**Emp. Bayes Estimates** $(\hat{\theta}_{EB}(N))$	**Emp. Bayes MSE**	**Bayes Estimates** $(\hat{x}_B(N))$	**Bayes MSE**
$\begin{bmatrix} .01 & 0 \\ 0 & .01 \end{bmatrix}$	$\begin{bmatrix} .01126 & .00007 \\ .00007 & .01077 \end{bmatrix}$	50	$\begin{bmatrix} 1.49695 \\ -.87744 \end{bmatrix}$	0.03234	$\begin{bmatrix} .83624 \\ -.28188 \end{bmatrix}$	0.63180
$\begin{bmatrix} .01 & 0 \\ 0 & .01 \end{bmatrix}$	$\begin{bmatrix} .01463 & -.00003 \\ -.00003 & .01426 \end{bmatrix}$	100	$\begin{bmatrix} 1.45567 \\ -.79186 \end{bmatrix}$	0.04154	$\begin{bmatrix} 1.02877 \\ -.39871 \end{bmatrix}$	0.32363
$\begin{bmatrix} .01 & 0 \\ 0 & .01 \end{bmatrix}$	$\begin{bmatrix} .01625 & -.00008 \\ -.00008 & .01550 \end{bmatrix}$	200	$\begin{bmatrix} 1.47271 \\ -.73062 \end{bmatrix}$	0.03779	$\begin{bmatrix} 1.15014 \\ -.42892 \end{bmatrix}$	0.20237
$\begin{bmatrix} .08 & 0 \\ 0 & .08 \end{bmatrix}$	$\begin{bmatrix} .53880 & .02513 \\ .02513 & .34930 \end{bmatrix}$	50	$\begin{bmatrix} 1.49695 \\ -.87744 \end{bmatrix}$	0.83556	$\begin{bmatrix} 1.37114 \\ -.75324 \end{bmatrix}$	0.06840
$\begin{bmatrix} .08 & 0 \\ 0 & .08 \end{bmatrix}$	$\begin{bmatrix} .58419 & -.01039 \\ -.01039 & .46649 \end{bmatrix}$	100	$\begin{bmatrix} 1.45567 \\ -.79186 \end{bmatrix}$	1.24102	$\begin{bmatrix} 1.38857 \\ -.72718 \end{bmatrix}$	0.05562
$\begin{bmatrix} .08 & 0 \\ 0 & .08 \end{bmatrix}$	$\begin{bmatrix} .63437 & -.02433 \\ -.02433 & .39928 \end{bmatrix}$	200	$\begin{bmatrix} 1.47271 \\ -.73062 \end{bmatrix}$	1.08385	$\begin{bmatrix} 1.43353 \\ -.69226 \end{bmatrix}$	0.02675

Table 4.3: Results of simulations with variable parameters, $\lambda = 0.02$.

one and then use it also as backward just after inverted the order of stacking the past data in the matrix Φ_N and $\mathbf{y}$. Clearly, this holds using a single sequence of data of finite size.

5

Conclusions

In this work we investigated in which cases the Empirical Bayes can perform better than the usual Bayes a posteriori estimator for an ARX model. The results of the simple simulations of Section 4, both with fixed model parameters and with randomly varying parameters, seem to confirm the comparison of the two estimation principles as discussed in Section 3. The rough indication being that Empirical Bayes can perform better for small a priori variance and finite sample size.

Forward Kalman Filter for ARX Models

In this section we shall derive in a self-contained way, a simplified version of the so-called *Conditionally Gaussian Kalman filter* which is adapted to Bayesian parameter estimation of ARX models. The main idea was already in [13][7] but a rigorous justification is due to R. Liptser and is exposed in the book [11, Chap 12]. A related but more complicated version can be found in the paper [4].

We consider the ARX model (3.1) written as a "pseudo-linear regression"

$$\mathbf{y}(t) = \boldsymbol{\varphi}(t)^\top \theta + \mathbf{w}(t)\,, \qquad t \in \mathbb{Z}_+ \tag{A.1}$$

where we assign a Gaussian a priori distribution to the parameter which becomes a random p-dimensional vector $\mathbf{x}$:

$$\mathbf{x} \sim \mathcal{N}(\theta, \Sigma)$$

with θ some "nominal" mean value and Σ a variance matrix which is usually unknown. The noise $\mathbf{w}(t)$ is also assumed Gaussian i.i.d. with variance σ^2, independent of $\mathbf{x}$ for all t.

For simplicity we shall initially assume that $\boldsymbol{\varphi}(t)$ is only a function of $\mathbf{y}^{t-1}$ so that (A.1) reduces to a purely Auto Regressive model. At the end we shall consider 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 (A.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

$$\mathbf{y}(t) = h(\,\mathbf{x}^t,\, \mathbf{w}^t)\,; \qquad t \geq 0$$

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 A.0.1. *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 $(\,\mathbf{y}^t,\, \mathbf{x}\,)$.*

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 $\{\mathbf{z}_\alpha\,;\, \alpha \in A\}$.

By stacking the system equations (A.1) ordered for increasing time $t = 0, 1, \dots$ we obtain a relation among random vectors

$$\mathbf{y}^t = \begin{bmatrix} \boldsymbol{\varphi}(1)^\top \\ \dots \\ \boldsymbol{\varphi}(t)^\top \end{bmatrix} \mathbf{x} + \mathbf{w}^t \;:=\; \boldsymbol{\Phi}_t\, \mathbf{x} + \mathbf{w}^t \tag{A.2}$$

where the matrix $\boldsymbol{\Phi}_t$ is a function of the initial conditions and past outputs up to time $t-1$.

Theorem A.0.1. *Assume t is large enough and that $\boldsymbol{\Phi}_t$ has almost surely a left inverse $\boldsymbol{\Phi}_t^{-L}$. Then, the random variables $(\,\mathbf{y}(t+1),\, \mathbf{x}\,)$ are jointly conditionally Gaussian given $\mathbf{y}^t$.*

Proof. We shall first show that $p(\mathbf{x} = \theta \mid \mathbf{y}^t)$ is a conditionally Gaussian distribution. Left-multiply (A.2) by $\boldsymbol{\Phi}_t^{-L}$ (which only depends on $\mathbf{y}^{t-1}$) to get

$$\mathbf{x} = \boldsymbol{\Phi}_t^{-L}\mathbf{y}^t - \boldsymbol{\Phi}_t^{-L}\mathbf{w}^t$$

which shows that the conditional distribution of $\mathbf{x}$ is actually Gaussian given $\mathbf{y}^t$,

[1] with (conditional) mean vector $\mathbf{\Phi}_t^{-L}\mathbf{y}^t$ and conditional variance equal to $\sigma^2\mathbf{\Phi}_t^{-L}\,[\,\mathbf{\Phi}_t^{-L}\,]^\top$.
Then the statement follows from Bayes rule

$$p(y(t+1),\, x \mid \mathbf{y}^t = y^t) = p(y(t+1) \mid \mathbf{x} = x,\, \mathbf{y}^t = y^t)\, p(x \mid \mathbf{y}^t = y^t)$$

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 $\operatorname{var}\{\mathbf{w}(t+1)\}$. 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 (A.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 3.1 with a ridge penalty term related to the a priori variance of $\mathbf{x}$. We shall instead describe a *Bayesian recursive solution* which is much in the same spirit of the algorithm developed in Section 3.5. Consider then the following

Problem A.0.1. *Find a recursive updating algorithm to compute the conditional mean and conditional variance of the random parameter vector* $\mathbf{x}$*:*

$$\hat{\mathbf{x}}(t) := \mathbb{E}\,[\,\mathbf{x} \mid \mathbf{y}^t\,]\,, \qquad \Sigma(t) := \operatorname{Var}\,[\,\mathbf{x} \mid \mathbf{y}^t\,] \tag{A.3}$$

We shall rely on Theorem A.0.1 and on the full rank assumption of $\mathbf{\Phi}_t$. 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 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 iterated Gaussian conditional expectation formula yields

$$\begin{aligned}\mathbb{E}\,[\,\mathbf{x} \mid \mathbf{y}_1, \mathbf{y}_2] = \mathbb{E}_{\mathbf{y}_1}[\,\mathbf{x}\,] + \operatorname{Cov}\{\mathbf{x}, \mathbf{y}_2 \mid \mathbf{y}_1\}\, \operatorname{Var}\{\mathbf{y}_2 \mid \mathbf{y}_1\}^{-1} \\ \times \operatorname{Cov}\{\mathbf{y}_2, \mathbf{x} \mid \mathbf{y}_1\}\, \{\mathbf{y}_2 - \mathbb{E}\,[\,\mathbf{y}_2 \mid \mathbf{y}_1]\,\}\end{aligned} \tag{A.4}$$

[1] Since here $\mathbf{x}$ depends on the choice of the left inverse, more correctly one should say: any random vector $\mathbf{x}$ satisfying (A.2).

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 $t+1$

$$\hat{\mathbf{x}}(t+1) := \mathbb{E}\,[\mathbf{x} \mid \mathbf{y}(t+1),\, \mathbf{y}^t] = \mathbb{E}_{\mathbf{y}^t}[\,\mathbf{x} \mid \mathbf{y}(t+1)\,]$$

where the operator $\mathbb{E}_{\mathbf{y}^t}$ is as defined above. Then applying formula (A.4) we obtain

$$\begin{aligned}\mathbb{E}\,[\mathbf{x} \mid \mathbf{y}(t+1),\, \mathbf{y}^t] = \mathbb{E}_{\mathbf{y}^t}[\,\mathbf{x}\,] + \mathrm{Cov}\{\mathbf{x}, \mathbf{y}(t+1) \mid \mathbf{y}^t\}\,\mathrm{Var}\{\mathbf{y}(t+1) \mid \mathbf{y}^t\}^{-1} \\ \times \mathrm{Cov}\{\mathbf{y}(t+1), \mathbf{x} \mid \mathbf{y}^t\}\,\{\mathbf{y}(t+1) - \mathbb{E}\,[\,\mathbf{y}(t+1) \mid \mathbf{y}^t]\,\}\end{aligned} \tag{A.5}$$

where

$$\mathbb{E}_{\mathbf{y}^t}[\,\mathbf{x}\,] = \hat{\mathbf{x}}(t)\,, \qquad \mathbb{E}\,[\,\mathbf{y}(t+1) \mid \mathbf{y}^t] = \boldsymbol{\varphi}(t+1)^\top \hat{\mathbf{x}}(t)$$

the last equality following since $\mathbf{w}(t+1)$ and $\mathbf{y}^t$ are independent (Lemma A.0.1). The last equation describes the (one step ahead) predictor of $\mathbf{y}(t+1)$ given $\mathbf{y}^t$ which is commonly denoted $\hat{\mathbf{y}}(t+1 \mid t)$. The estimator (A.5) is a linear function of the (one step ahead) **prediction error**

$$\mathbf{e}(t) := \mathbf{y}(t) - \hat{\mathbf{y}}(t \mid t-1) = \boldsymbol{\varphi}(t)^\top[\,\mathbf{x} - \hat{\mathbf{x}}(t-1)\,] + \mathbf{w}(t) \tag{A.6}$$

which is a process with uncorrelated (and hence independent) variables by the orthogonality principle. Its conditional variance is just

$$\begin{aligned}\mathrm{var}[\,\mathbf{e}(t) \mid \mathbf{y}^{t-1}\,] = \mathrm{var}[\,\mathbf{y}(t) \mid \mathbf{y}^{t-1}] = \mathbb{E}\,\{\,[\boldsymbol{\varphi}(t)^\top(\,\mathbf{x} - \hat{\mathbf{x}}(t-1)\,) + \mathbf{w}(t)\,]^2 \mid \mathbf{y}^{t-1}\} \\ = \boldsymbol{\varphi}(t)^\top \Sigma(t-1)\boldsymbol{\varphi}(t) + \sigma^2\end{aligned}$$

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 called the **innovation process** of $\{\mathbf{y}(t)\}$. For the covariance matrices in (A.4) we obtain

$$\begin{aligned}&\mathrm{Cov}\{\mathbf{x}, \mathbf{y}(t+1) \mid \mathbf{y}^t\} = \mathbb{E}\,\{[\,\mathbf{x} - \hat{\mathbf{x}}(t)\,][\,\mathbf{y}(t+1) - \hat{\mathbf{y}}(t+1 \mid t)\,] \mid \mathbf{y}^t\} = \\ &= \mathbb{E}\,\{[\,\mathbf{x} - \hat{\mathbf{x}}(t)\,][\,\mathbf{x} - \hat{\mathbf{x}}(t)\,]^\top \boldsymbol{\varphi}(t+1) \mid \mathbf{y}^t\} + \mathbb{E}\,\{[\,\mathbf{x} - \hat{\mathbf{x}}(t)\,]\mathbf{w}(t+1) \mid \mathbf{y}^t\} = \\ &= \mathbb{E}\,\{[\,\mathbf{x} - \hat{\mathbf{x}}(t)\,][\,\mathbf{x} - \hat{\mathbf{x}}(t)\,]^\top \mid \mathbf{y}^t\}\boldsymbol{\varphi}(t+1) = \Sigma(t)\boldsymbol{\varphi}(t+1)\,.\end{aligned}$$

The last equality follows from the independence of $\mathbf{x}$ and $\mathbf{w}(t)$ and Lemma A.0.1.

The conditional variance $\mathrm{var}[\,\mathbf{y}(t+1) \mid \mathbf{y}^t\,]$ has been computed above for time t instead of $t+1$. The updating formula (A.5) can therefore be written as

$$\hat{\mathbf{x}}(t+1) = \hat{\mathbf{x}}(t) + k(t+1)[\,\mathbf{y}(t+1) - \boldsymbol{\varphi}(t+1)^\top \hat{\mathbf{x}}(t)\,] \tag{A.7}$$

where the *gain vector* $k(t+1)$ is given by

$$k(t+1) = \Sigma(t)\boldsymbol{\varphi}(t+1)\left[\boldsymbol{\varphi}(t+1)^\top \Sigma(t)\boldsymbol{\varphi}(t+1) + \sigma^2\right]^{-1} \tag{A.8}$$

The recursion is driven by the innovation $\mathbf{e}(t+1)$.The initial condition can be taken as $\hat{\mathbf{x}}(0) = \mathbb{E}\,\mathbf{x}$.
We still need an updating equation for $\Sigma(t)$. Using again the iterated conditioning formula (A.4) and (A.5) we get

$$\begin{aligned}
&\Sigma(t+1) = \mathrm{Var}[\,\mathbf{x} \mid \mathbf{y}^{t+1}] = \\
&= \mathrm{Var}[\,\mathbf{x} \mid \mathbf{y}^t\,] - \mathrm{Cov}\{\mathbf{x}, \mathbf{y}(t+1) \mid \mathbf{y}^t\}\,\mathrm{var}[\,\mathbf{y}(t+1) \mid \mathbf{y}^t]^{-1}\,\mathrm{Cov}\{\mathbf{y}(t+1), \mathbf{x} \mid \mathbf{y}^t\} = \\
&= \Sigma(t) - \Sigma(t)\boldsymbol{\varphi}(t+1)\left[\boldsymbol{\varphi}(t+1)^\top \Sigma(t)\boldsymbol{\varphi}(t+1) + \sigma^2\right]^{-1} \boldsymbol{\varphi}(t+1)^\top \Sigma(t)
\end{aligned}$$

now with initial condition the a priori covariance $\Sigma(0) = \mathrm{Var}[\,\mathbf{x}\,] = P$.

The reader could easily get a *normalized conditional Kalman filter algorithm* as described in Theorem A.0.2 below by *normalization* of the Covariance matrix dividing by σ^2.

In conclusion, we have

Theorem A.0.2. *Letting $\mathbf{y}^t$ denote the subvector made with the first t components of $\mathbf{y}$, the conditional mean and the normalized conditional variance of the random parameter vector $\mathbf{x}$:*

$$\hat{\mathbf{x}}(t) := \mathbb{E}\,[\mathbf{x} \mid \mathbf{y}^t\,], \qquad P(t) := \frac{1}{\sigma^2}\,\mathrm{Var}\,[\,\mathbf{x} \mid \mathbf{y}^t\,] \tag{A.9}$$

satisfy the recursion

$$\hat{\mathbf{x}}(t+1) = \hat{\mathbf{x}}(t) + k(t+1)[\,\mathbf{y}(t+1) - \boldsymbol{\varphi}(t+1)^\top \hat{\mathbf{x}}(t)\,] \tag{A.10}$$

where the gain vector $k(t+1)$ *is given by*

$$k(t+1) = P(t)\boldsymbol{\varphi}(t+1)\left[\boldsymbol{\varphi}(t+1)^\top P(t)\boldsymbol{\varphi}(t+1) + 1\right]^{-1} \tag{A.11}$$

while the matrix $P(t)$ satisfies

$$P(t+1) = P(t) - P(t)\boldsymbol{\varphi}(t+1)\left[\boldsymbol{\varphi}(t+1)^\top P(t)\boldsymbol{\varphi}(t+1) + 1\right]^{-1} \boldsymbol{\varphi}(t+1)^\top P(t)\,. \quad \text{(A.12)}$$

The recursion

$$\lambda^2(t+1) = \frac{t}{t+1}\lambda^2(t) + \frac{1}{t+1}[\mathbf{y}(t+1) - \boldsymbol{\varphi}(t+1)^\top \hat{\mathbf{x}}(t+1)\,]^2 \quad \text{(A.13)}$$

describes the estimate of the conditional variance of the (a posteriori) innovation process $\hat{\mathbf{e}}(t) := \mathbf{y}(t) - \boldsymbol{\varphi}(t)^\top \hat{\mathbf{x}}(t)$. *The corresponding estimate of σ^2 is expressed by the formula*

$$\hat{\sigma}^2(t) = \frac{\lambda^2(t)}{\boldsymbol{\varphi}(t)^\top P(t)\boldsymbol{\varphi}(t) + 1}\,. \quad \text{(A.14)}$$

This algorithm computes recursively the conditional mean, the normalized conditional variance $P(t)$ of $\mathbf{x}(t)$ and the conditional variance $\hat{\sigma}^2(t)$ of the innovation process

Note that 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 (3.1) 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 could 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.

Corollary A.0.1. *Consider the ARX model* (3.1) *with a Gaussian noise* $\mathbf{w}$ *independent of the input process* $\mathbf{u}$. *Then the estimator $\hat{\mathbf{x}}(t)$ which minimizes the conditional error variance $\Sigma(t) := \mathrm{Var}[\,\mathbf{x} \mid \mathbf{z}^t\,]$ evolves in time according to the same recursion, that is*

$$\hat{\mathbf{x}}(t+1) = \hat{\mathbf{x}}(t) + k(t+1)[\,\mathbf{y}(t+1) - \boldsymbol{\varphi}(t+1)^\top \hat{\mathbf{x}}(t)\,] \quad \text{(A.15)}$$

where the gain vector $k(t+1)$ *is given by*

$$k(t+1) = \Sigma(t)\boldsymbol{\varphi}(t+1)\left[\boldsymbol{\varphi}(t+1)^\top \Sigma(t)\boldsymbol{\varphi}(t+1) + \sigma^2\right]^{-1} \quad \text{(A.16)}$$

The process $\mathbf{e}(t+1) := \mathbf{y}(t+1) - \boldsymbol{\varphi}(t+1)^\top \hat{\mathbf{x}}(t)$ *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* $\hat{\mathbf{x}}(0) = \mathbb{E}\,\mathbf{x}$.
The conditional error variance $\Sigma(t)$ *can be updated by the same matrix recursion as above, namely*

$$\Sigma(t+1) = \Sigma(t) - \Sigma(t)\boldsymbol{\varphi}(t+1)\left[\boldsymbol{\varphi}(t+1)^\top \Sigma(t)\boldsymbol{\varphi}(t+1) + \sigma^2\right]^{-1} \boldsymbol{\varphi}(t+1)^\top \Sigma(t) \quad \text{(A.17)}$$

with initial condition the a priori variance $\Sigma(0) = P_0$.

To this end we rewrite the Bayesian model (3.10) componentwise assuming that we have available a suitably long sequence of measurements

$$\mathbf{y}(t) = \boldsymbol{\varphi}(t)^\top \mathbf{x} + \mathbf{w}(t)\,, \qquad t = 1, 2, \ldots, \quad \text{(A.18)}$$

In our problem $\boldsymbol{\varphi}(t)$ is a function of past measurements up to time $t-1$ and hence the classical version of the Kalman filter is not applicable. Under our assumption of Gaussian distributions however we may resort to a generalized Kalman filter algorithm which minimizes the *conditional error variance* given past data, and has a similar structure of the classical Kalman filter. A derivation of the formula is provided in Appendix **??**. We shall refer to [11, Chap. 12] and [4] for a complete treatment of this subject. It turns out that conditioning, formally allows to treat the column vectors $\boldsymbol{\varphi}(t)$ as being deterministic.

Proof. Omit. □

Remark A.0.1. Of course $x_0 = \mathbb{E}\,\mathbf{x}$ and $P_0 = \mathrm{Var}[\,\mathbf{x}\,]$ are not known to the analyst; however under certain rather general structural conditions on the true model generating the data, one can show that the Kalman filter tends to forget the initial conditions and its behavior is quite independent of them, of course provided they are chosen not too far from the true ones.

B

Stationarity and Ergodicity

Let us pretend that we have an infinite sequence of random 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 B.0.1. *A stochastic process* $\{\mathbf{y}(t)\}$ *is* **stationary** *(in the strict sense) if all PDF's relative to* $\mathbf{y}(t_1)$, $\mathbf{y}(t_2), \ldots \mathbf{y}(t_n)$ *say* $F_n(x_1, \ldots, x_n, t_1, \ldots, t_n)$ *are invariant for temporal translation, that is for every* n *it must hold that,*

$$F_n(x_1, \ldots, x_n, t_1 + \Delta, \ldots, t_n + \Delta) = F_n(x_1, \ldots, x_n, t_1, \ldots, t_n) \quad ,$$

(same function of $x_1, \ldots, x_n$, $t_1, \ldots, t_n$*), 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:
  $$\Sigma(t_1, t_2) := \mathbb{E}\ \left[\mathbf{y}(t_1) - \mu(t_1)\right]\ \left[\mathbf{y}(t_2) - \mu(t_2)\right]^\top$$
  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 B.0.1 (Birkhoff Ergodic Theorem)**.** *Let* $\{\mathbf{y}(t)\}$ *be a strictly stationary process. The limit*

$$\bar{\mathbf{z}} := \lim_{T\to\infty} \frac{1}{T} \sum_{k=1}^{T} f_k(\mathbf{y}) \tag{B.1}$$

exists with probability one for all functions f *such that* $\mathbb{E}\,|f(\mathbf{y})| < \infty$.

The limit can either be random or a deterministic constant. If it is random it must be a "very special" random variable. These are called *invariant random variables*. We shall not investigate them. to zero

Note in fact 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 $\bar{\mathbf{z}}_T$ in the second members of (B.1) is constant and hence converges as $T \to \infty$ so that one finds

$$\mathbb{E}\,\bar{\mathbf{z}} = \mathbb{E}\, f(\mathbf{y})\,.$$

Corollary B.0.1. *If* $\{\mathbf{y}(t)\}$ *is ergodic*

$$\lim_{T\to\infty} \frac{1}{T} \sum_{k=1}^{T} f_k(\mathbf{y}) = \mathbb{E}\; f(\mathbf{y}) \tag{B.2}$$

with probability one whatever may be $f(\mathbf{y})$ *having finite expectation.*

Proof (Sketch): In fact the L^1 norm of $\bar{\mathbf{z}}_T - \mathbb{E}\,(\mathbf{y})$ converges to zero 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 B.0.1. *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 B.0.2 (Kolmogorov)**.** *Every i.i.d. process having finite expectation is ergodic.*

The following is an important consequence.

Corollary B.0.2. *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

$$\mathbf{z}(t) := \sum_{-\infty}^{+\infty} c_k \, \mathbf{e}(t+k) = \sum_{-\infty}^{+\infty} c_{-k} \, \mathbf{e}(t-k)\,; \qquad t \in \mathbb{Z} \tag{B.3}$$

form an ergodic process.
Convergence of the sum follows from Cauchy-Schwartz inequality

$$\mathbb{E}\,|\sum_{-N}^{+N} c_k \, \mathbf{e}(t+k)| \le \sum_{-N}^{+N} |c_k|^2 \; \mathbb{E}\,|\mathbf{e}(t+k)|^2 = \sum_{-N}^{+N} |c_k|^2 \, \sigma_{\mathbf{e}}^2 \,.$$

References

[1] A. Aravkin et al. "On the estimation of hyperparameters for Empirical Bayes estimators: Maximum Marginal Likelihood vs Minimum MSE". In: *Proc. of the 16th IFAC Symposium on System Identification*. Brussels, Belgium, 2012, pp. 125–132.

[2] P.E. Caines. *Linear Stochastic Systems*. New York: Wiley, 1988.

[3] G. Casella. "An Introduction to Empirical Bayes Data Analysis". In: *The American Statistician* 39 (1985), pp. 83–87.

[4] H.F. Chen, P.R. Kumar, and J.H van Schuppen. "On Kalman filtering for conditionally Gaussian systems with random matrices". In: *System and Control Letters)* (1989), pp. 397–404.

[5] B. Efron. "Two modeling strategies for Empirical Bayes Estimation". In: *Statistical Science* 29.2 (2014), pp. 285–301.

[6] E. J. Hannan and M. Deistler. *The Statistical Theory of Linear Systems*. New York: John Wiley, 1988.

[7] Y. C. Ho and R. C. K. Lee. "A Bayesian approach to problems in stochastic estimation and control". In: *IEEE Trans. Autom. Control* 9 (1964), pp. 333–339.

[8] W. James and Charles Stein. "Estimation with Quadratic Loss". In: *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. Berkeley, Calif.: University of California Press, 1961, pp. 361–379. URL: `http://projecteuclid.org/euclid.bsmsp/1200512173`.

[9] E. L. Lehmann and G. Casella. *Theory of Point Estimation, Second Ed.* Springer Texts in Statistics, 1998.

[10] A. Lindquist and G. Picci. *Linear Stochastic Systems: A Geometric Approach to Modeling Estimation and Identification*. Springer Verlag, 2015.

[11] R.S. Liptser and A. N. Shiryaev. *Statistics of Random Processes vol. 2*. Springer-Verlag, 1977.

[12] L. Ljung. *System Identification; theory for the user*. Upper Saddle River, N.J.: Prentice Hall, 1999.

[13] D. Mayne. "Optimal non-stationary estimation of the parameters of a linear system with Gaussian inputs". In: *J. Electron. Control* 14 (1963).

[14] S. Petrone et al. "Empirical Bayes methods in classical and Bayesian inference". In: *Metron* 72 (2014), pp. 201–215.

[15] G. Picci and Bin Zhu. "Empirical Bayes identification of stationary processes and approximation of Toeplitz spectra". In: *Automatica* 142: 110362 (2022).

[16] G. C. Reinsel. "Mean Squared Error Properties of Empirical Bayes Estimators in a Multivariate Random Effects General Linear Model". In: *Journal of the American Statistical Association* 80 (391) (1985), pp. 642–650.

[17] A. N. Shiryaev. *Probability*. Second. Springer, 1995.

[18] T. Söderström and P. Stoica. *System Identification*. New York: Prentice Hall, 1989.

[19] M. Yuan, ChongLi Wan, and LaiSheng Wei. "Superiority of empirical Bayes estimator of the mean vector in multivariate normal distribution". In: *Science China Mathematics* 59 (2016), pp. 1175–1186.