Lecture Notes for a course on System Identification

Kristiaan Pelckmans

Contents

0.4pt0.4pt 0pt0 pt

Abstract

System identification in a narrow sense is concerned with tasks of parameter estimation based on observations originating from a dynamical system. System identification in a broad sense deals with many subtleties coming up when designing, conducting and interpreting results from such an experiment. The purpose of this text is to survey the collection of ideas, insights, theorems, intuitions, language and practical experiences which constitute the art of system identification. In this course we will guide you along the basic methods and concepts, as well as show you the many connections to related fields of applied science.

Part I

Basics of System Identification

Chapter 1

Aim

> ”Say we study an unknown dynamic system. How to design, conduct, process and interpret the results from an experiment applied to this system such that we will get an accurate model of its internal working?”

This question expresses the goal of system identification techniques. The actual answer to this question lies in how the different terms are translated mathematically. This course intends to illustrate the spectrum of choices one can make here. In order to bootstrap this course, let us give a working (’intuitive’) definition of those:

’Unknown’: Here we assume that a precise description of the system is not available. This might be as (i) the system is too complex to derive completely from (physical) laws; (ii) the system might behave different from what would be expected from a theoretical description (e.g. due to aging, unaccounted effects, or conversion from continuous to discrete time); (iii) we only need to have a ’good’ approximation of the system which serves our purpose well.

’Dynamic System’: The actual behavior relating input signal to output signal, often using a physical process. The keyword ’dynamic’ is understood here as follows: ’output values depend on present as well as past inputs given to the system’.

’Model’: A mathematical representation of the studied system. Many different flavors of models exists (see next chapter): this course will study mathematical models relating input and output signals using equations. (Logic, software, language or intuition can be used alternatively).

’Experiment’: A datum collected when feeding the system with a predesigned input signal. This book will often be concerned with signals taking continuous values in $\mathbb{R}$. (Alternatively, values of signals can be binary, categorical, strictly positive, or taking elements in other structured sets).

’Design’: How to choose the input signals in order to optimize the result of the overall analysis?

’Process’: In which shape should the signals be (pre-)processed before submitting to analysis?

’Interpret’: In what sense is the identified model accurate and reliable, and which results might be due to unknown disturbances, noisy measurements or artificial model structures?

In a specific sense, system identification is concerned with coming with an accurate model given the input-output signals recorded during working of the studied system.

Hence, it becomes plain that system identification is closely related to other fields of mathematical modeling. We mention here the various domains concerned with parameter estimation including statistical inference, adaptive filtering and machine learning. Historically, system identification originates from an engineering need to form models of dynamical systems: it then comes as no surprise that traditionally emphasis is laid on numerical issues as well on system-theoretical concerns.

Progress in the field has much been reinforced by introducing good software to execute the various algorithms. This makes the methodology semi-automatic: that is a user needs still have a conceptual overview on what is to be done, but the available software tools take care of most technical details. In this course, the use of the MATLAB System Identification toolbox is discussed in some detail.

System Identification as a field came only in existence in the 60s, while its roots can be traced back to the Least Squares techniques, other techniques of statistical inference. The field however originated from an engineering need to have good models for use in automatic control. 'Modern' system identification is often attributed to the work of Åström and Bohlin [1], while contemporary system identification is often associated to the work as reviewed in Ljung [4] or Söderström and Stoica [5]. The latter work will be the basis of this course.

1.1 Systems & Models

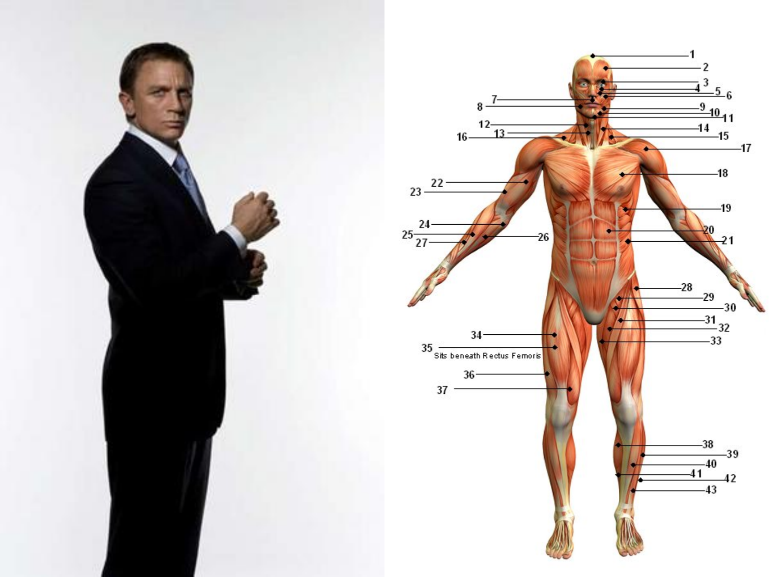
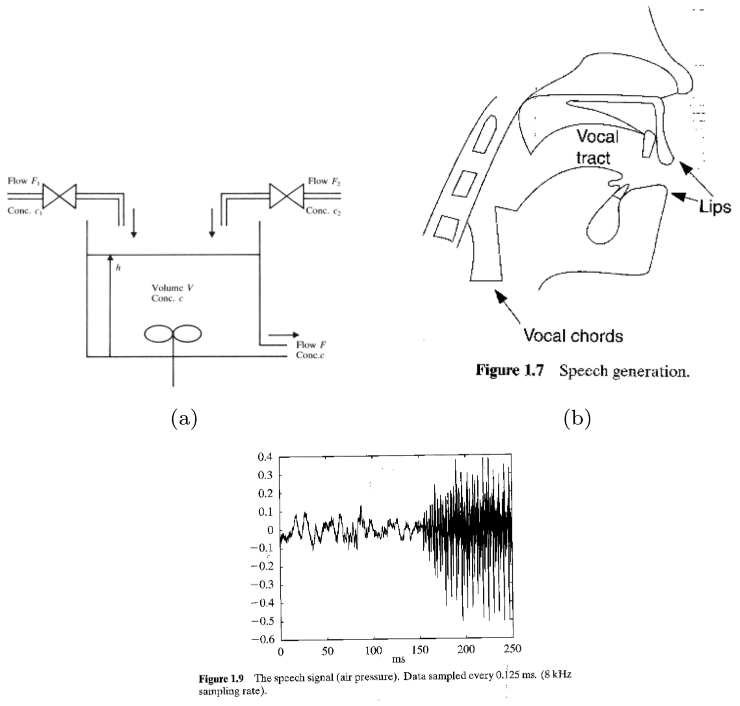


Figure 1.1: As a medical doctor you get to study the human body. As a doctor you work with an internal representation (=model) of how the body (=system) works.

Definition 1 (System or Model) *The overall behaviors you get to study is referred to as the* system*. The internal (mental) representation you as an researcher use in order to study the system, is called the* model*.*



(a)

(b)

(c)

Figure 1.2: (a) Schematic representation of a stirred thank process. (b) representation of a speech apparatus. (c) Example of a generated acoustic signal with two different filters shaped by intention ('k' versus 'a')

Let us begin by describing some of the systems which are being studied in the context of this course.

Stirred Thank: The following is a prototypical example in the context of process control. Consider a biochemical reactor, where two different substances go in via respective pipelines. Both inflows comes at a certain flow-rate and have a certain concentration, either of which can be controlled by setting valves. Then the substances interacts inside the stirred tank, and the yield is tapped from the tank. Maybe the aim of such process is to maximize the concentration of the yield at certain instances. A mathematical approach to such automatic control however requires a mathematical description of the process of interest. That is, we need to set up equations relating the setting of the valves and the output. Such model could be identified by experimenting on the process and compiling the observed results into an appropriate model.

Speech: Consider the apparatus used to generate speech in the human. In an abstract fashion, this can be seen as a white noise signal generated by the glottis. Then the mouth are used to filter this noise into structured signals which are perceived by an audience as meaningful. Hence, this apparatus can be abstracted into a model with unknown white noise input, a dynamical system shaped by intention, and an output which can be observed. Identification of the filter (dynamical system) can for example be used to make a artificial speech.

Lateron in this course, you will be asked to study how techniques of system identification can be applied in a range of different applications.

Industrial: The prototypical example of an engineering system is an industrial plant which is fed by an inflow of raw material, and some complicated process converts it into the desired yield. Often the internal mechanism of the studied process can be worked out in some detail. Nevertheless, it might be more useful to come up with a simlpler model relating input-signals to output-signals directly, as it is often (i) easier (cheaper) to develop, (ii) is directly tuned to our need, and (iii) makes abstraction of irrelevant mechanisms in the process, and (iv) might better handle the unforeseen disturbances.

Figure 1.3: Illustrative examples: (a) A petrochimical plant. (b) An application of a model for acoustic signals. (c) An example of an econometric system. (d) Signals arising from TV can be seen as coming from a system.

Acoustic: The processing of acoustical signals can be studied in the present context. Let us for example study the room which converts an acoustic signal (say a music signal) into an acoustic signal augmented with echo. It is then often of interest to compensate the signal sent into the room for this effect, so as to 'clean' the perceived signal by the audience. In this example, the room is conceived as the dynamical system, and it is of interest to derive a model based on acoustic signals going into the room, and the consequent signals perceived by an audience.

Econometric: The following example is found in a financial context. Consider the records of the currency exchange rates. This multivariate time-series is assumed to be driven by political, socio-economic or cultural effects. A crude way to model such non-measurable effects is as white noise. Then the interesting bit is how the exchange rates are interrelated: how for example a injection of resources in one market might alter other markets as well.

Multimedial: Finally, consider the sequence of images used to constitute a cartoon on TV say. Again, consider the system driven by signals roughly modeling meaning, and outputting the values projected in the different pixels. It is clear that the signals of neighboring pixels are inter-related, and that the input signal is not as high-dimensional as the signals projected on the screen.

1.2 The System Identification Procedure

Let us sketch a prototypical example of the different steps taken in a successful system identification experiment. The practical way to go ahead is typically according to the following steps, each one raising their own challenges:

1. Description of the task. What is a final desideratum of a model? For what purpose is it to be used? How will we decide at the end of the day if the identified model is satisfactory? On which properties to we have to focus during the identification experiments?

2. Look at initial Data. What sort of effects are of crucial importance to capture? What are the challenges present in the task at hand. Think about useful graphs displaying the data. Which phenomena in those graphs are worth pinpointing?

3. Nonparametric analysis. If possible, do some initial experiments: apply an pulse or step to the system, and look at the outcome. Perhaps a correlation or a spectral analysis are possible as well. Look at where random effects come in. If exactly the same experiment is repeated another day, how would the result differ? Is it possible to get an idea of the form of the disturbances?

4. Design Experiment. Now that we have acquired some expertise of the task at hand, it is time to set up the large identification experiment. At first, enumerate the main challenges for identification, and formalize where to focus on during the experiment. Then design an experiment so as to maximize the information which can be extracted from observations made during the experiment. For example. make sure all the dynamics of the studied system are sufficiently excited. On the other hand, it is often paramount to make sure that the system remains in the useful 'operation mode' throughout the experiment. That is, it is no use to inject the system with signals which do not apply in situations where/when the model is to be used.

5. Identify model. What is a good model structure? What are the parameters which explain the behavior of the system during the experiment.

6. Refine Analysis: It is ok to start off with a hopelessly naive model structure. But it is then paramount to refine the model structure and the subsequent parameter estimation in order to compensate for the effects which could not be expressed in the first place. It is for example common practice to increase the order of the dynamical model. Is the noise of the model reflecting the structure we observe in the first place, or do we need more flexible noise models? Which effects do we see in the data but are not captured by the model: time-varying, nonlinear, aging, saturation,... ?

7. Verify Model: is the model adequate for the purpose at hand? Does the model result in satisfactory results as written down at the beginning? Is it better than a naive approach? Is the model accurately extracting or explaining the important effects? For example, analyze the residuals left over after subtracting the modeled behavior from the observed data. Does it still contain useful information, or is it white? Implement the model for the intended purpose, thus it work satisfactory?

A flowchart of the typical system identification experiment is shown in Fig. (1.2.a). The identification procedure can be compared to an impedance meter (see Fig. (1.2.b)) which measures the impedance of a system by comparing the current measured at input with the corresponding voltage at the output line of a system. Similarly, system identification tries to figure out the dynamics of a system by relating input signal to corresponding output, i.e. from observed input-output behavior.

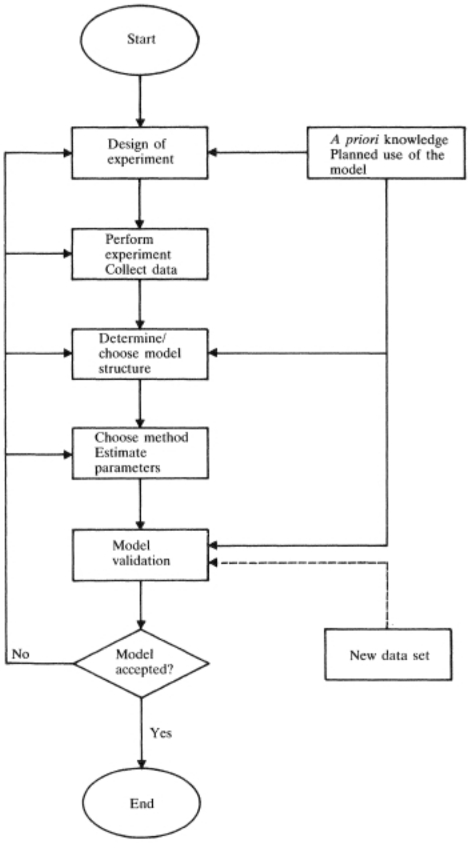

Start
Design of experiment
A priori knowledge Planned use of the model
Perform experiment Collect data
Determine/ choose model structure
Choose method Estimate parameters
Model validation
New data set
No
Model accepted?
Yes
End

FIGURE 1.3 Schematic flowchart of system identification.

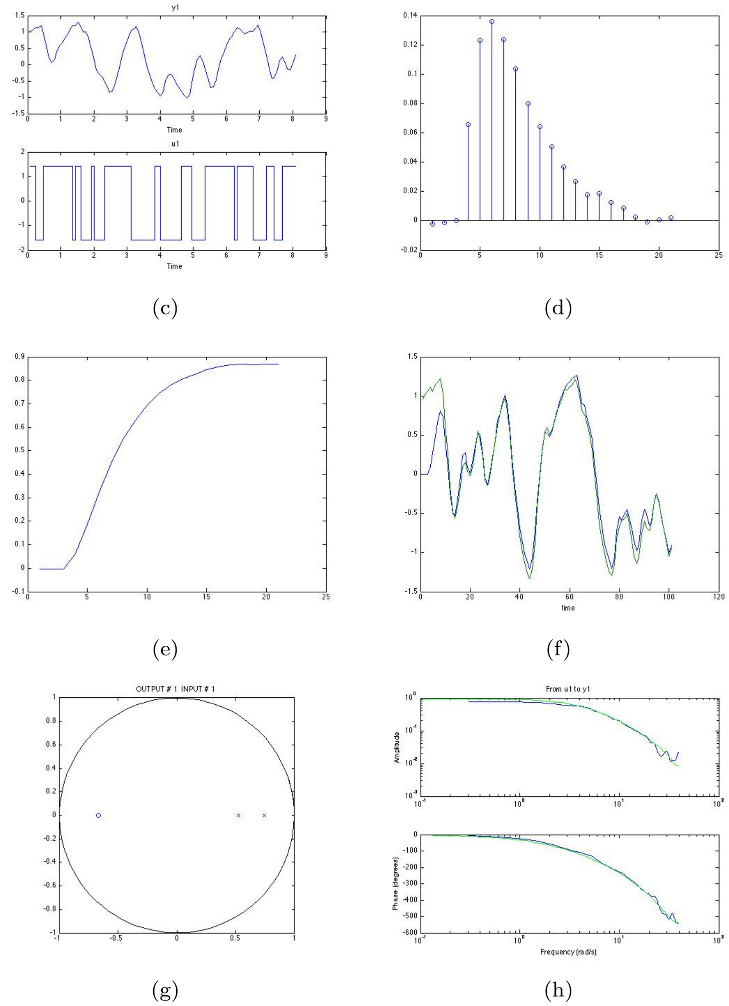
(a)

(b)

1.3 A simple Example

The code of a simple identification experiment in MATLAB is given. The task is to identify a model for a hairdryer. This system fans air through a tube which is heated at the inlet. The input signal $u(t)$ reflects the power of the heating device. The output signal $y(t)$ reflects the temperature of the air coming out. The data used for identification is displayed as follows.

```
>> load dryer2
>> z2 = [y2(1:300) u2(1:300)];
>> idplot(z2, 200:300, 0.08)
```



(c) (d)

(e) (f)

(g) (h)

Figure 1.4:

At first, the structural properties of the system are displayed using nonparametric tools as follows

```
>> z2 = dtrend(z2);
>> ir = cra(z2);
>> stepr = cumsum(ir);
>> plot(stepr)
```

Inspection of those properties suggest the following parametric model relating input to output signals.

$$y(t) + a_1 y(t-1) + a_2 y(t-2) = b_1 u(t-3) + b_2 u(t-4) \tag{1.1}$$

Here $\{a_1, a_2, b_1, b_2\}$ are to be estimated based on the collected data as given before.

```
>> model = arx(z2, [2 2 3]);
>> model = sett(model,0.08);
>> u = dtrend(u2(800:900));
>> y = dtrend(y2(800:900));
>> yh = idsim(u,model);
>> plot([yh y]);
```

The dynamics of this estimated model are characterized as the poles and zeros of the system. This is given as

```
>> zpth = th2zp(model);
>> zpplot(zpth);
```

The transfer function corresponding to the estimated model, and derived from the non-parametric method is compared as follows

```
>> gth = th2ff(model);
>> gs = spa(z2); gs = sett(gs,0.08);
>> bodeplot([gs gth]);
```

Many open questions on the studied system remain after this simple analysis. For example 'Is the estimated model accurate enough (model validation)?', 'Is the model structure as given in eq. (1.1) appropriate?' 'If we can freely choose the input signals, what would the optimal inputs be (experiment design)?', ...

1.4 Notation

> Hardly any non-mathematician will admit that mathematics has a cultural and aesthetic appeal, that it has anything to do with beauty, power or emotion. I categorically deny such cold and rigid view. [N.Wiener, 1956]

Mathematical manuscripts tend to scare people away because of their intrinsic technical notation. However, mathematical notation is carefully crafted over the years to express ideas, inventions, truths, intuitions and reasonings of exact sciences better than any spoken language could do. Similar to spoken languages, mathematical notation comes in many different dialects, obstructing the fluent interaction between different groups of researchers. Dialects may differ in subtle issues as e.g.

the use of capital symbols for certain quantities or in indexing systems, or in capital difference as the use of operators versus explicit representations of transforms. As for any compiler, ideas might not work out properly if the syntax is not as intended. As such, it really pays off to get familiar with different notational conventions.

In this course we adapt the following notation.

(Constant): A constant quantity will be denoted as a lower-case letter in an equation. For example scalars (e.g. c, a, b), indices (as e.g. $i, j, k, n, \dots$), functions (as e.g. $f, g, \dots$) and so on.

(Vector): A vector - an array of scalars taking value in $\mathbb{R}^d$ - is denoted as a boldface lowercase letter (e.g. $\mathbf{a}, \mathbf{b}, \mathbf{c}, \dots$).

(Matrix): A matrix - a tableau of scalars taking values in $\mathbb{R}^{n\times d}$ - is denoted as a boldface capital letters (e.g. $\mathbf{A}, \mathbf{B}, \mathbf{M},$)

(Random variable): Random quantities will be noted as capital letters in equations. Random quantities need a proper stochastic setting to be well-defined (see chapter 4), in practice they can be recognized as they can take different values (realizations).

(Set): Sets of elements are denoted using curled brackets: e.g. $\{a, b, c\}$ is a set of three constant values. Sets are referred to using mathbb letters, e.g. $\mathbb{R}, \mathbb{N}, \mathbb{C}, \dots$.

(Operator): An operator is a mapping of a function to another. An operator is denoted as a calligraphic letter, e.g. $\mathcal{L} : \{f\} \to \{f\}$ is a mapping from the set of functions into the same set.

(Reserved): In order to ease notational load, a number of symbols connect (in this course) to a given meaning. An overview of those is given in Table (1.1).

Symbol	Type	Meaning
$\mathbb{R}, \mathbb{N}, \mathbb{C}$	Set	Set of real, integer or complex values.
n	Scalar	Number of observations.
θ	Vector	Unknown parameters of the problem to be estimated.
τ	Scalar	Delay of the model/system.
$i, j = 1, \dots, n$	Index	Index ranging over the n observations.
$\mathbf{i}$	Imaginary number	$\mathbf{i} = \sqrt{-1}$
d	Scalar	Order of Input Response, or dimension.
$u, \mathbf{u}$	function and vector	Continuous and discrete input signal.
$y, \mathbf{y}$	function and vector	Continuous and discrete output signal.
$\hat{\cdot}, \cdot_n$	Operator	An estimate, that is $\hat{f}$ and f_n are both an estimate of the quantity f.

Table 1.1: Enumeration of symbols taking a fixed meaning in this text.

Bibliography

[1] K.J. Åström and P. Eykhoff. *System Identification – A Survey.* Automatica 7(2), pp. 123–162, 1971.

[2] B.D.O. Anderson and J.B. Moore. *Optimal Filtering.* Prentice-Hall, 1979.

[3] G. Box and G. Jenkins. *Time Series Analysis: Forecasting and Control* . Prentice-Hall, 1987.

[4] L. Ljung. *System Identification, Theory for the User.* Prentice Hall, 1987.

[5] T. Söderström. and P. Stoica. *System Identification.* Prentice-Hall Englewood Cliffs, 1989.

[6] K. Åström and B. Wittenmark. *Adaptive Control.* Addisson-Wesley, 1994.

[7] P.J.. Brockwell and R.A. .Davis. *Time series: theory and methods.* Springer, 1987.

[8] P. Overschee and B.. Moor. *Subspace Identification for Linear Systems: Theory & Implementation.* Kluwer, 1996.

[9] P. Stoica and R. Moses. *Spectral analysis of signals.* Prentice-Hall, 2005.

Chapter 2

Least Squares Rules

> "Given a set of observations, which model parameters gives a model which approximates those up to the smallest sum of squared residuals?"

Least squares estimation serves as the blueprint and touchstone of most estimation techniques, and ideas should be mastered by any student. The purpose of this chapter is to survey results, to review the geometric insights and to elaborate on the numerical techniques used. In order to keep the exposition as simple as possible (but no simpler), we suppress for the moment the 'dynamic' nature of the studied systems, and focus on the static estimation problem.

2.1 Least Squares (OLS) estimates

2.1.1 Models which are Linear in the Parameters

This chapter studies a classical estimators for unknown parameters which occur linearly in a model structure. Such model structure will be referred to as Linear In the Parameters, abbreviated as LIP. At first we will give some examples in order to get an intuition about this class. Later sections then discuss how those parameters can be estimated using a least square argument. It is important to keep in the back of your mind that such least squares is not bound to LIP models, but then one ends up in general with less convenient (numerical, theoretical and conceptual) results.

Definition 2 (LIP) *A model for $\{y_i\}_i$ is linear in the unknowns $\{\theta_j\}_{j=1}^d$ if for each $y_i : i = 1, \dots, n$, one has given values $\{x_{ij}\}_{j=1}^d$ such that*

$$y_i = \sum_{j=1}^{d} x_{ij}\theta_j + e_i, \tag{2.1}$$

and the terms $\{e_i\}$ are assumed to be chosen independently of $\{x_{ij}\}$ and i, j. Intuitively, this means that $\{e_i\}$ are residuals which should be as small as possible. Such model can be summarized schematically as

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_{11} & \dots & x_{1d} \\ \vdots & & \vdots \\ x_{n1} & \dots & x_{nd} \end{bmatrix} \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_d \end{bmatrix} \tag{2.2}$$

or in matrix notation

$$\mathbf{y} = \mathbf{X}\theta + \mathbf{e}. \tag{2.3}$$

where the matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$*, the vectors* $\theta \in \mathbb{R}^d$ *and* $\mathbf{e} \in \mathbb{R}^n$ *are as in eq. (2.2). In case* d *is 'small' (compared to* n*), one refers to them as the* parameter *vector.*

If the input-output data we try to model can be captured in this form, the resulting problems , algorithms, analysis and interpretations become rather convenient. So the first step in any modeling task is to try to phrase the model formally in the LIP shape. Later chapters will study also problems who do not admit such parameterization. However, the line which models admit such parameterizations and which do not is not always intuitively clear. We support this claim with some important examples.

Example 1 (Constant Model) *Perhaps the simplest example of a linear model is*

$$y_t = \theta + e_t. \tag{2.4}$$

where $\theta \in \mathbb{R}$ *is the single parameter to estimate. This can be written as in eq. (2.1) as*

$$y_t = \mathbf{x}^T\theta + e_t. \tag{2.5}$$

where $\mathbf{x}_t = 1 \in \mathbb{R}$*, or in vector form as*

$$\mathbf{y} = 1_n\theta + \mathbf{e}, \tag{2.6}$$

where $\mathbf{y} = (y_1, \ldots, y_n)^T \in \mathbb{R}^n$*,* $\mathbf{e} = (e_1, \ldots, e_n)^T \in \mathbb{R}^n$ *and* $1_n = (1, \ldots, 1)^T \in \mathbb{R}^n$*. Hence, the 'inputs' take the constant value 1. This thinking is also used in order to express a model of* d *inputs* $\mathbf{X}_t = \mathbf{x}_t^T \in \mathbb{R}^d$ *for all* $t = 1, \ldots, n$ *with a constant intercept term* θ_0 *given as*

$$y_t = \mathbf{x}_t^T\theta + \theta_0 + e_t. \tag{2.7}$$

In matrix form one has

$$\mathbf{y} = \mathbf{X}'\theta' + \mathbf{e}, \tag{2.8}$$

where now $\theta' = (\theta^T\theta_0)^T \in \mathbb{R}^{d+1}$ *and* $\mathbf{X}' = \begin{bmatrix}\mathbf{X} & 1_n\end{bmatrix}$*.*

Example 2 (Polynomial Trend) *Assume the output has a polynomial trend of order smaller than* $m > 0$*, then it is good practice to consider the model*

$$y_t = \sum_{j=1}^{d} x_{tj}\theta_j + \sum_{k=0}^{m} t^k\theta'_k + e_t. = \mathbf{x}_t^T\theta + \mathbf{z}^T(t)\theta' + e_t, \tag{2.9}$$

where $\mathbf{z}(t) = (1, t, t^2, \ldots, t^m)^T \in \mathbb{R}^{m+1}$ *and* $\theta' = (\theta'_0, \theta'_1 \ldots, \theta_m)^T \in \mathbb{R}^{m+1}$*. Again, in matrix notation one has*

$$\mathbf{y} = \mathbf{X}'\theta' + \mathbf{e}. \tag{2.10}$$

where $\mathbf{X}'_t = (\mathbf{x}_t^T, 1, t, \ldots, t^m)$ *and* $\theta' = (\theta^T, \theta'_0, \theta'_1, \ldots, \theta'_m)^T \in \mathbb{R}^{d+m+1}$*.*

Example 3 (A Weighted sum of Exponentials) *It is crucial to understand that models which are linear in the parameters are not necessary linear in the covariates. For example, consider a nonlinear function*

$$y_t = f(\mathbf{x}_t) + e_t, \tag{2.11}$$

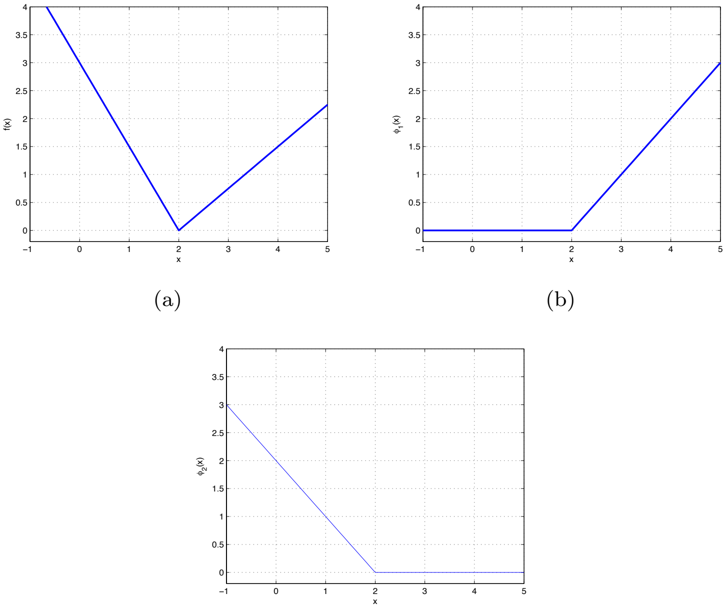
where $f : \mathbb{R}^d \to \mathbb{R}$ is any function. Then one can find an arbitrary good approximation to this model

$$y_t = \sum_{k=1}^{m} \varphi_k(\mathbf{x}_t)\theta_k + e_t = \varphi(\mathbf{x}_t)\theta + e_t, \tag{2.12}$$

where $\{\phi_1, \ldots, \phi_m\} \subset \{f : \mathbb{R}^d \to \mathbb{R}\}$ are an appropriate set of basis functions*. Then $\varphi(\mathbf{x}) = (\varphi_1(\mathbf{x}), \ldots, \varphi_m(\mathbf{x}))^T \in \mathbb{R}^m$. There are ample ways on which form of basis functions to consider. A method which often works is to work with exponential functions defined for any $k = 1, \ldots, m$ as*

$$\varphi_k(\mathbf{x}) = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}_k\|}{\sigma_k}\right), \tag{2.13}$$

where $\sigma_k > 0$ and $\mathbf{x}_k \in \mathbb{R}^d$ is chosen suitably. Specific examples of basis functions are the orthogonal polynomials (e.g. Chebychev, Legendre or Laguerre polynomials). More involved sets lead to methods as wavelets, splines, orthogonal functions, or kernel methods.



(a)

(b)

(c)

Figure 2.1: A Simple Example of a representation of a function $f(x)$ in panel (a) as the sum of two basis functions ϕ_1 (b) and ϕ_2 (c).

Example 4 (Dictionaries) *Elaborating on the previous example, it is often useful to model*

$$y_t = f(\mathbf{x}_t) + e_t, \tag{2.14}$$

as

$$y_t = \sum_{k=1}^{m} f_k(\mathbf{x}_t)\theta_k + e_t, \tag{2.15}$$

where the set $\{f_1, \ldots, f_m\}$ *is assumed to contain the unknown function* $f : \mathbb{R}^d \to \mathbb{R}$ *up to a scaling. If this is indeed the case, and the* $f \propto f_j$ *then eq. (2.14) can be represented as*

$$y_t = F_m(\mathbf{x}_t)\mathbf{e}_k a + e_t, \tag{2.16}$$

where $a \neq 0$ *is a constant,* $\mathbf{e}_k \in \{0,1\}^m$ *is the kth unit vector, that is* $\mathbf{e}_k = (0, \ldots, 1 \ldots, 0)^T$ *with unit on the kth position, and zero elsewhere.* F_m *denotes the dictionary, it is* $F_m(\mathbf{x}) = (f_1(\mathbf{x}), \ldots, f_m(\mathbf{x}))^T \in \mathbb{R}^m$ *for all* $\mathbf{x} \in \mathbb{R}^d$*.*

In practice, if the model which is proposed to use for the identification experiment can be written as an expression which is linear in the parameters, then the subsequent analysis, inference and interpretation are often rather straightforward. The first challenge for successful identification is hence to phrase the problem in this form. It is however not always possible to phrase mathematical models in this form as indicated in the following example.

Example 5 (Nonlinear in the Parameters) *Consider the following model for observations* $\{y_t\}_{t=1}^n$

$$y_t = a\sin(bt + c), \tag{2.17}$$

where (a, b, c) *are unknown parameters. Then it is seen that the model is linear in* a*, but not in* b, c*. A way to circumvent this is to come up with plausible values* $\{b_1, \ldots, b_m\}$ *for* b*, and* $\{b_1, \ldots, b_m\}$ *for* c*, and to represent the model as*

$$y_t = \sum_{i,j=1}^{m} a_{i,j}\sin(b_i t + c_j), \tag{2.18}$$

where the model (2.17) is recovered when $a_{i,j} = a$ *when* $b_i = b$ *and* $c_j = c$*, and is zero otherwise.*

Example 6 (Quadratic in the Parameters) *Consider the following model for observations* $\{(y_t, x_t)\}_{t=1}^n$

$$y_t = ax_t + e_t + be_{t-1}, \tag{2.19}$$

where $\{e_t\}_t$ *are unobserved noise terms. Then we have cross-terms* $\{be_{t-1}\}$ *of unknown quantities, and the model falls not within the scope of models which are linear in the parameters.*

Other examples are often found in the context of grey-box models where theoretical study (often expressed in terms of PDEs) decide where to put the parameters. Nonlinear modeling also provide a fertile environment where models which are nonlinear in the parameters thrive. One could for example think of systems where nonlinear feedback occurs.

2.1.2 Ordinary Least Squares (OLS) estimates

Example 7 (Univariate LS) *Suppose we have given n samples $\{(x_i, y_i)\}_{i=1}^n$, where $x_i, y_i \in \mathbb{R}$. We suspect that they are strongly correlated in the sense that there is an unknown parameter $\theta \in \mathbb{R}$ such that $y_i \approx \theta x_i$ for all $i = 1, \ldots, n$. We hope to find a good approximation to θ by solving the following problem*

$$\hat{\theta} = \operatorname*{argmin}_{\theta \in \mathbb{R}} \sum_{i=1}^n (y_i - \theta x_i)^2. \tag{2.20}$$

At this stage, we have done the most work - i.e. converting the practical problem in a mathematical one - and what is left is mere technical. In this particular case, the solution is rather straightforward: first note that eq. (2.20) requires us to solve an optimization problem with (i) optimization criterion $J_n(\theta) = \sum_{i=1}^n (y_i - \theta x_i)^2$, and (ii) $\theta \in \mathbb{R}$ the variable to optimize over. It is easily checked that in general there is only a single optimum, and this one is characterized by the place where $\dfrac{\partial J_n(\theta)}{\partial \theta} = 0$. Working this one out gives

$$-2 \sum_{i=1}^n x_i y_i + 2 \sum_{i=1}^n x_i x_i \hat{\theta} = 0, \tag{2.21}$$

or

$$\hat{\theta} = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}, \tag{2.22}$$

That is, in case $\sum_{i=1}^n x_i^2 \neq 0$! This is a trivial remark in this case, but in general such conditions will play a paramount role in estimation problems.

Example 8 (Average) *Consider the simpler problem where we are after a variable θ which is 'close' to all datasamples $\{y_i\}_{i=1}^n$ taking values in $\mathbb{R}$. Again, we may formalize this as*

$$\hat{\theta} = \operatorname*{argmin}_{\theta \in \mathbb{R}} \sum_{i=1}^n (y_i - \theta)^2. \tag{2.23}$$

Do check that the solution is given as

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^n y_i. \tag{2.24}$$

In other words, the sample average is the optimal least squares approximation of a bunch of samples. This is no coincidence: we will explore the relation between sample averages, means and least squares optimal estimators in depth in later chapters. Note that in this particular case, there is no caveat to the solution, except for the trivial condition that $n > 0$.

The extension to more than one parameter is much easier by using matrix representations.

Example 9 (Bivariate Example) *Assume that we have a set of $n > 0$ couples $\{(x_{i,1}, x_{i,2}, y_i)\}_{i=1}^n$ to our disposition, where $x_{i,1}, x_{i,2}, y \in \mathbb{R}$. Assume that we try to 'fit a model'*

$$\theta_1 x_{i,1} + \theta_2 x_{i,2} + e_i = y_i, \tag{2.25}$$

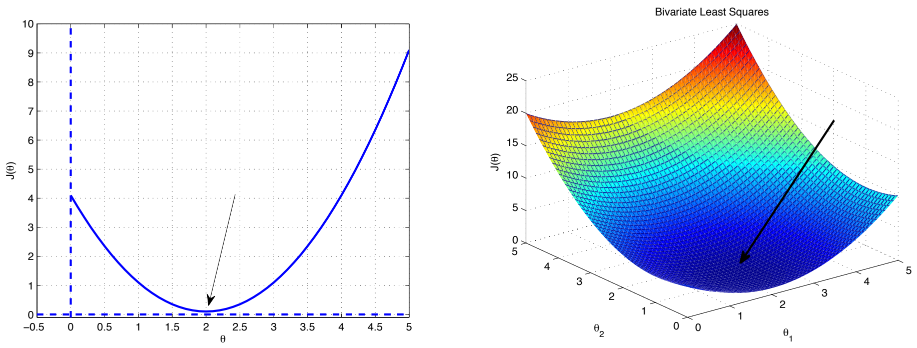


Figure 2.2: Illustration of the squared loss function in function of θ. The arrow indicates $\hat{\theta}$ where the minimum to $J(\theta)$ is achieved. Panel (a) shows the univariate case, or $\theta \in \mathbb{R}$ as in Example 2. Panel (b) shows the bivariate case, or $\theta \in \mathbb{R}^2$ as in Example 3.

where the unknown residuals $\{e_i\}_{i=1}^n$ are thought to be 'small' in some sense. The Least Squares estimation problem is then written as

$$(\hat{\theta}_1, \hat{\theta}_1) = \operatorname*{argmin}_{\theta_1, \theta_2 \in \mathbb{R}} \sum_{i=1}^{n} (y_i - \theta_1 x_{i1} - \theta_2 x_{i2})^2. \tag{2.26}$$

This can be written out in matrix notation as follows. Let us introduce the matrix and vectors $\mathbf{X}_2 \in \mathbb{R}^{n \times 2}$, $\mathbf{y}, \mathbf{e} \in \mathbb{R}^n$ and $\theta \in \mathbb{R}^2$ as

$$\mathbf{X}_2 = \begin{bmatrix} x_{1,1} & x_{1,2} \\ x_{2,1} & x_{2,2} \\ \vdots & \vdots \\ x_{n,1} & x_{n,2} \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{e} = \begin{bmatrix} e_1 \\ y_2 \\ \vdots \\ e_n \end{bmatrix}, \quad \theta = \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}. \tag{2.27}$$

Then the model (2.25) can be written as

$$\mathbf{X}_2 \theta + \mathbf{e} = \mathbf{y}. \tag{2.28}$$

and the Least Squares estimation problem (2.26) becomes

$$\hat{\theta} = \operatorname*{argmin}_{\theta \in \mathbb{R}^2} J_n(\theta) = (\mathbf{X}_2 \theta - \mathbf{y})^T (\mathbf{X}_2 \theta - \mathbf{y}) \tag{2.29}$$

where the estimate $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2)^T \in \mathbb{R}^2$ is assumed to be unique. Taking the derivative of $J_n(\theta)$ and equating it to zero (why?) gives the following set of linear equations characterizing the solution:

$$(\mathbf{X}_2^T \mathbf{X}_2)\theta = \mathbf{X}^T \mathbf{y}. \tag{2.30}$$

This set of linear equations has a unique solution in case the matrix $(\mathbf{X}_2^T \mathbf{X}_2)$ is sufficiently 'informative'. In order to formalize this notion let us first consider some examples:

1. *Assume $x_{i,2} = 0$ for all $i = 1, \ldots, n$.*

2. *Assume $x_{i,2} = x_{i,1}$ for all $i = 1, \dots, n$*

3. *Assume $x_{i,2} = ax_{i,1}$ for all $i = 1, \dots, n$, for a constant $a \in \mathbb{R}$.*

How does the matrix

$$(\mathbf{X}_2^T \mathbf{X}_2) = \begin{bmatrix} \sum_{i=1}^n x_{i,1}x_{i,1} & \sum_{i=1}^n x_{i,1}x_{i,2} \\ \sum_{i=1}^n x_{i,2}x_{i,1} & \sum_{i=1}^n x_{i,2}x_{i,2} \end{bmatrix}, \tag{2.31}$$

look like? Why does (2.30) give an infinite set of possible solutions in that case?

This reasoning brings us immediately to the more general case of $d \geq 1$ covariates. Consider the model which is *linear in the parameters*

$$y_i = \sum_{j=1}^d \theta_j x_{i,j} + e_i, \ \forall i = 1, \dots, n. \tag{2.32}$$

Defining

$$\mathbf{X}_d = \begin{bmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_n \end{bmatrix} = \begin{bmatrix} x_{1,1} & \dots & x_{1,d} \\ \vdots & & \vdots \\ x_{n,1} & \dots & x_{n,d} \end{bmatrix}, \tag{2.33}$$

or $\mathbf{X}_d \in \mathbb{R}^{n \times d}$ with $\mathbf{X}_{d,\ i,j} = x_{i,j}$. Note the orientation (i.e. the transposes) of the matrix as different texts use often a different convention. Equivalently, one may define

$$\mathbf{y} = \mathbf{X}_d \theta + \mathbf{e}, \tag{2.34}$$

where $\theta = (\theta_1, \dots, \theta_d)^T \in \mathbb{R}^d$ and $\mathbf{y} = (y_1, \dots, y_n)^T \in \mathbb{R}^n$ and $\mathbf{e} = (e_1, \dots, e_n)^T \in \mathbb{R}^n$. The Least Squares estimation problem solves as before

$$\min_{\theta \in \mathbb{R}^d} J_n(\theta) = (\mathbf{X}_d\theta - \mathbf{y})^T (\mathbf{X}_d\theta - \mathbf{y}) \tag{2.35}$$

where the estimate is now $\hat{\theta} = (\hat{\theta}_1, \dots, \hat{\theta}_d)^T \in \mathbb{R}^d$. Equating the derivative to zero gives a characterization of a solution θ in terms of a set of linear equations as

$$(\mathbf{X}_d^T \mathbf{X}_d)\theta = \mathbf{X}_d^T \mathbf{y}, \tag{2.36}$$

this set of equations is referred to as *the normal equations* associated to (2.35). Now it turns out that the condition for uniqueness of the solution to this set goes as follows.

Lemma 1 *Let $n, d > 0$, and given observations $\{(x_{i,1}, \dots, x_{i,d}, y_i)\}_{i=1}^n$ satisfying the model (2.34) for a vector $\mathbf{e} = (e_1, \dots, e_n)^T \in \mathbb{R}^n$. The solutions $\{\theta\}$ to the optimization problem (2.35) are characterized by the normal equations (2.36). This set contains a single solution if and only if (iff) there exists no $\mathbf{w} \in \mathbb{R}^d$ with $\|\mathbf{w}\|_2 = 1$ such that $(\mathbf{X}_d^T \mathbf{X}_d)\mathbf{w} = 0_d$.*

Proof: At first, assume there exists a $\mathbf{w} \in \mathbb{R}^d$ with $\|\mathbf{w}\|_2 = 1$ such that $(\mathbf{X}_d^T \mathbf{X}_d)\mathbf{w} = 0_d$. Then it is not too difficult to derive that there has to be many different solutions to (2.35). Specifically, let θ be a solution to the problem (2.35), then so is $\theta + a\mathbf{w}$ for any $a \in \mathbb{R}$.

Conversely, suppose there exists two different solutions, say θ and θ', then $\mathbf{w} \neq 0_d$ is such that $(\mathbf{X}_d^T \mathbf{X}_d)\mathbf{w} = 0_d$. This proofs the Lemma.

□

It is interesting to derive what the minimal value $J(\hat\theta)$ will be when the optimum is achieved. This quantity will play an important role in later chapters on statistical interpretation of the result, and on model selection. Let's first consider a simple example:

Example 10 (Average, Ct'd) *Consider the again the case where we are after a variable θ which is 'close' to all datasamples $\{y_i\}_{i=1}^n$ taking values in $\mathbb{R}$, or*

$$\hat\theta = \operatorname*{argmin}_{\theta\in\mathbb{R}} J(\theta) = \sum_{i=1}^n (y_i-\theta)^2. \tag{2.37}$$

The solution is characterized as $\hat\theta = \frac{1}{n}\sum_{i=1}^n y_i$. Then the achieved minimal value $J(\hat\theta)$ equals

$$J(\hat\theta) = \sum_{i=1}^n (y_i-\hat\theta)^2 = \sum_{i=1}^n y_i^2 - 2\sum_{i=1}^n y_i\hat\theta + \hat\theta^2 = \sum_{i=1}^n y_i^2 - \left(\sum_{i=1}^n y_i\right)^2. \tag{2.38}$$

as verified by straightforward calculation.

In general, the value $J(\hat\theta)$ is expressed as follows.

2.1.3 Ridge Regression

What to do in case multiple solutions exists? It turns out that there exists two essentially different approaches which become almost as elementary as the OLS estimator itself. we consider again the estimators solving

$$\Omega = \operatorname*{argmin}_{\theta\in\mathbb{R}^d} J_n(\theta) = (\mathbf{X}_d\theta-\mathbf{y})^T(\mathbf{X}_d\theta-\mathbf{y}), \tag{2.39}$$

where $\Omega\subset\mathbb{R}^d$ is now a set.

- Select the smallest solution amongst the set of all solutions Ω.
- Modify the objective such that there is always a unique solution.

The first approach is very much a procedural approach, and details will be given in the section about numerical tools. It is noteworthy that such approach is implemented through the use of the pseudo-inverse.

The second approach follows a more general path. In its simplest form the modified optimization problem becomes

$$\min_{\theta\in\mathbb{R}^d} J_n^\gamma(\theta) = (\mathbf{X}_d\theta-\mathbf{y})^T(\mathbf{X}_d\theta-\mathbf{y}) + \gamma\theta^T\theta, \tag{2.40}$$

where $\gamma\geq 0$ regulates the choice of how the terms (i) $\|\mathbf{X}_d\theta-\mathbf{y}\|_2^2$ and (ii) $\|\theta\|_2^2$ are traded off. If γ is chosen large, one emphasizes 'small' solutions, while the corresponding first term (i) might be suboptimal. In case $\gamma\approx 0$ one enforces the first term to be minimal, while imposing a preference on all vectors $\{\theta\}$ minimizing this term. It is easy to see that in case $\gamma>0$ there is only a single solution to (2.40). Indeed equating the derivative of (2.40) to zero would give the following characterization of a solution $\theta\in\mathbb{R}^d$

$$(\mathbf{X}_d^T\mathbf{X}_d+\gamma I_d)\theta = \mathbf{X}_d^T\mathbf{y}, \tag{2.41}$$

and it becomes clear that no $\mathbf{w} \in \mathbb{R}^d$ with $\|\mathbf{w}\|_2 = 1$ exist such that $(\mathbf{X}_d^T \mathbf{X}_d + \gamma I_d)\mathbf{w} = 0_d$. in case there is only a single θ which achieves the minimum to $\|\mathbf{X}_d\theta - \mathbf{y}\|_2^2$, a nonzero γ would give a slightly different solution to (2.40), as opposed to this θ. It would be up to the user to control this difference, while still ensuring uniqueness of the solution when desired.

Recently, a related approach came into attention. Rather than adding a small jitter term $\theta^T\theta$, it is often advantageous to use a jitter term $\|\theta\|_1 = \sum_{j=1}^d |\theta_j|$. The objective then becomes

$$\min_{\theta\in\mathbb{R}^d} J_n^\gamma(\theta) = (\mathbf{X}_d\theta - \mathbf{y})^T(\mathbf{X}_d\theta - \mathbf{y}) + \gamma\|\theta\|_1. \tag{2.42}$$

The solution to this problem can be computed efficiently using tools of numerical optimization as surveyed in Chapter 15. Why to prefer (2.42) over (2.40)? Denote the estimates resulting from solving (2.40) as $\hat{\theta}_2$, and the estimates based on the same $\mathbf{X}, \mathbf{y}$ obtained by solving (2.42) as $\hat{\theta}_1$. Then the main insight is that the latter will often contain zero values in the vector $\hat{\theta}_1$. Those indicate often useful information on the problem at hand. For example, they could be used for selecting relevant inputs, orders or delays. Solution $\hat{\theta}_2$ in contrast will rarely contain zero parameters. But then, it is numerically easier to solve (2.40) and to characterize theoretically the optimum.

2.2 Numerical Tools

The above techniques have become indispensable tools for researchers involved with processing data. Their solutions are characterized in terms of certain matrix relations. The actual power of such is given by the available tools which can be used to solve this problems numerically in an efficient and robust way. This section gives a brief overview of how this works.

2.2.1 Solving Sets of Linear Equations

A central problem is how a set of linear equations can be solved. That is, given coefficients $\{a_{ij}\}_{i=1,\dots,d,j=1,\dots,d'}$ and $\{b_i\}_{i=1}^d$, find scalars $\{\theta_i \in \mathbb{R}\}_{i=1}^{d'}$ such that

$$\begin{cases} a_{11}\theta_1 + \dots a_{1d'}\theta_{d'} = b_1 \\ \vdots \\ a_{d1}\theta_1 + \dots a_{dd'}\theta_1 = b_d. \end{cases} \tag{2.43}$$

This set of linear equations can be represented in terms of matrices as $\mathbf{b} = (b_1, \dots, b_d)^T \in \mathbb{R}^d$ and

$$\mathbf{A} = \begin{bmatrix} a_{11} & \dots & a_{1d'} \\ \vdots & & \vdots \\ a_{d1} & \dots & a_{dd'} \end{bmatrix}. \tag{2.44}$$

The set of linear equations is then written shortly as

$$\mathbf{A}\theta = \mathbf{b}. \tag{2.45}$$

Now we discriminate between 3 different cases:

$d < d'$ Then the matrix $\mathbf{A}$ looks *fat*, and the system is underdetermined. That is, there are an infinite set of possible solutions: there are not enough equality conditions in order to favor a single solution.

$d > d'$ Then the matrix $\mathbf{A}$ looks *tall*, and the system is in general overdetermined. That is, there is in general no solution vector $\theta = (\theta_1, \ldots, \theta_{d'})^T \in \mathbb{R}^{d'}$ which satisfies all equations simultaneously. Note that in certain (restrictive) conditions on the equality constraints, it is possible for a solution to exist.

$d = d'$ This implies that $\mathbf{A} \in \mathbb{R}^{d\times d}$ is square. In general, there is exactly one vector $\theta = (\theta_1, \ldots, \theta_d) \in \mathbb{R}^d$ which obeys all the equality constraints at once. In some cases this solution is however not unique.

As explained in the previous section, a vector $\theta \in \mathbb{R}^{d'}$ can satisfy more than d' equalities (i.e. $d > d'$) only when at least one of the equalities can be written as a linear combination of the other equalities. Numerical solutions to solve this system of equations include the Gauss Elimination or the Gauss-Newton algorithms. It is found that both theoretical as well as practical advantages are achieved when using a Conjugate Gradient Algorithm (CGA). Plenty of details of such schemes can be found in standard textbooks on numerical analysis and optimization algorithms, see e.g. [].

In case that there is no exact solution to the set of equality constraints, one can settle for the next best thing: the best approximate solution. If 'best' is formalized in terms of least squares norm of the errors needed to make the equalities hold approximatively, one gets

$$\min_\theta \sum_{i=1}^{d} \left(\sum_{j=1}^{d} a_{ij}\theta_j - b_i \right)^2 = \min_\theta \|\mathbf{A}\theta - \mathbf{b}\|_2^2. \tag{2.46}$$

which can again be solved as ... an OLS problem, where the solution in turn is given by solving according normal equations $\mathbf{A}^T\mathbf{A}\theta = \mathbf{A}^T\mathbf{b}$ of size d'.

A crucial property of a matrix is its rank, defined as follows.

Definition 3 (Rank) *A matrix $\mathbf{A} \in \mathbb{C}^{n\times d}$ with $n \geq d$ is rank-deficient if there exists a nonzero vector $\mathbf{x} \in \mathbb{C}^d$ such that*

$$\mathbf{A}\mathbf{x} = 0_n \tag{2.47}$$

where $0_n \in \mathbb{R}^n$ denotes the all-zero vector. Then the rank of a matrix is defined as the number of nonzero linear independent vectors $\{\mathbf{x}_1, \ldots, \mathbf{x}_r\} \subset \mathbb{R}^d$ which have that $\mathbf{A}\mathbf{x}_i \neq 0_n$, or

$$\operatorname{rank}(\mathbf{A}) = \max \left| \{\mathbf{x}_i \in \mathbb{R}^d \ \ s.t. \ \ \mathbf{A}\mathbf{x}_i \neq 0_n, \mathbf{x}_i^T\mathbf{x} = \delta_{i-j}, \ \forall i,j = 1, \ldots, r\} \right| \leq \min(n,d). \tag{2.48}$$

2.2.2 Eigenvalue Decompositions

The previous approaches are mostly procedural, i.e. when implementing them the solution is computed under suitable conditions. However, a more fundamental approach is based on characterizing the properties of a matrix. In order to achieve this, the notion of a n Eigen Value Decomposition (EVD) is needed.

Definition 4 (Eigenpair) *Given a matrix $\mathbf{A} \in \mathbb{C}^{d' \times d}$ which can contain complex values. Then a vector $\mathbf{x} \in \mathbb{R}^d$ with $\|\mathbf{x}\|_2 = 1$ and corresponding value $\lambda \in \mathbb{C}$ constitutes an eigenpair $(\mathbf{x}, \lambda) \in \mathbb{C}^d \times \mathbb{C}$ if they satisfy*

$$\mathbf{A}\mathbf{x} = \lambda \mathbf{x}, \tag{2.49}$$

That is, if the matrix $\mathbf{A}$ *applied* to the vector $\mathbf{x}$ transforms into a rescaled version the same vector. It is intuitively clear that working with such eigenpairs simplifies an analysis since it reduces to working with scalars instead. Suppose we have d' such eigenpairs $\{(\mathbf{x}_i, \lambda_i)\}_{i=1}^n$, then those can be represented in matrix formulation as

$$\mathbf{A}\left[\mathbf{x}_1, \dots, \mathbf{x}_{d'}\right] = \left[\lambda_1 \mathbf{x}_1, \dots, \lambda_{d'} \mathbf{x}_{d'}\right] = \left[\mathbf{x}_1, \dots, \mathbf{x}_{d'}\right] \operatorname{diag}(\lambda_1, \dots, \lambda_{d'}), \tag{2.50}$$

or

$$\mathbf{A}\mathbf{X} = \mathbf{X}\Lambda. \tag{2.51}$$

where $\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_{d'}) \in \mathbb{C}^{d' \times d'}$ is a diagonal matrix, and $\mathbf{X} = \left[\mathbf{x}_1, \dots, \mathbf{x}_{d'}\right] = \left[\mathbf{x}_1, \dots, \mathbf{x}_{d'}\right] \in \mathbb{R}^{d \times d'}$.

The eigenvalues become have a special form when the matrix $\mathbf{A}$ has special structure. The principal example occurs when $\mathbf{A} \in \mathbb{C}^{d \times d}$ and $\mathbf{A} = \mathbf{A}^*$, i.e. the matrix is Hermitian. In case $\mathbf{A} \in \mathbb{R}^{d \times d}$, this means that $\mathbf{A} = \mathbf{A}^T$ is squared and symmetric. In both above cases we have that

(Real) All eigenvalues $\{\lambda_i\}_{i=1}^d$ are real valued.

(Ortho) All eigenvectors are orthonormal to each other, or $\mathbf{x}_i^T \mathbf{x}_j = \delta_{i-j}$.

Such orthonormal matrices are often represented using the symbol $\mathbf{U}$, here for example we have that $\mathbf{X} = \mathbf{U}$. The last property means that $\mathbf{U}^T\mathbf{U} = \mathbf{I}_d$, where $\mathbf{I}_d = \operatorname{diag}(1, \dots, 1) \in \{0,1\}^{d \times d}$ is the identity matrix of dimension d. But it also implies that $\mathbf{U}\mathbf{U}^T = \mathbf{U}(\mathbf{U}^T\mathbf{U})\mathbf{U}^T = (\mathbf{U}\mathbf{U}^T)^2$. Then, the only full-rank matrix $\mathbf{C} \in \mathbb{R}^{d \times d}$ which satisfies the problem $\mathbf{C}\mathbf{C} = \mathbf{C}$ is $\mathbf{I}_d$, such that we have also that $\mathbf{U}\mathbf{U}^T = I_d$. As such we can write

$$\mathbf{U}^T\mathbf{A}\mathbf{U} = \Lambda. \tag{2.52}$$

That is, the matrix of eigenvectors of a symmetric matrix *diagonalizes* the matrix. The proof of the above facts are far from trivial, both w.r.t. existence of such eigenpairs as well as concerning the properties of the decomposition, and we refer e.g. to [9], Appendix A for more information and pointers to relevant literature. Then we define the concepts of definiteness of a matrix as follows.

Definition 5 (Positive Definite Matrices) *A square matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$ is called Positive Definite (PD) iff one has for all vectors $\mathbf{x} \in \mathbb{R}^d$ that*

$$\mathbf{x}^T\mathbf{A}\mathbf{x} > 0. \tag{2.53}$$

A matrix $\mathbf{A} = \mathbf{A}^*$ is called Positive Semi-Definite (PSD) iff one has for all vectors $\mathbf{v} \in \mathbb{R}^d$ that

$$\mathbf{x}^T\mathbf{A}\mathbf{x} \geq 0. \tag{2.54}$$

The notation $\mathbf{A} \succeq 0$ denotes the $\mathbf{A}$ is PSD, and $\mathbf{A} \succ 0$ means that $\mathbf{A}$ is PD.

In the same vein, one defines negative definite, negative semi-definite and non-definite matrices. It turns out that such properties of a squared matrix captures quite well how different matrices behave in certain cases.

Lemma 2 (A PD Matrix) *A matrix $\mathbf{A} = \mathbf{A}^* \in \mathbb{C}^{d\times d}$ is positive definite if any of the following conditions hold:*

(i) If all eigenvalues $\{\lambda_i\}$ are strictly positive.

(ii) If there exist a matrix $\mathbf{C} \in \mathbb{R}^{n\times d}$ of rank $\operatorname{rank}(\mathbf{A})$ where

$$\mathbf{A} = \mathbf{C}^T\mathbf{C}. \tag{2.55}$$

(iii) If the determinant of any submatrix of $\mathbf{A}$ is larger than zero. A submatrix of $\mathbf{A}$ is obtained by deleting $k < d$ rows and corresponding columns of $\mathbf{A}$.

This decomposition does not only characterize the properties of a matrix, but is as well optimal in a certain sense.

Lemma 3 (Rayleigh Coefficient) *Let $\lambda_1 \geq \cdots \geq \lambda_d$ be the ordered eigenvalues of a matrix $\mathbf{A} = \mathbf{A}^T \in \mathbb{R}^{d\times d}$. Then*

$$\lambda_n = \min_{\mathbf{x}} \frac{\mathbf{x}^T\mathbf{A}\mathbf{x}}{\mathbf{x}^T\mathbf{x}}, \tag{2.56}$$

and this minimum is achieved when $\mathbf{x} \propto \mathbf{x}_1$, that is, is proportional to an eigenvector corresponding to a minimal eigenvalue.

$$\lambda_n = \max \min_{\mathbf{x}} \frac{\mathbf{x}^T\mathbf{A}\mathbf{x}}{\mathbf{x}^T\mathbf{x}}, \tag{2.57}$$

Moreover, one has for all $i = 1, \ldots, d$ that

$$\lambda_i = \max_{\mathbf{W}\in\mathbb{R}^{d\times(d-i)}} \min_{\mathbf{x}:\mathbf{W}^T\mathbf{x}=0} \frac{\mathbf{x}^T\mathbf{A}\mathbf{x}}{\mathbf{x}^T\mathbf{x}} = \min_{\mathbf{W}\in\mathbb{R}^{d\times(i-1)}} \max_{\mathbf{x}:\mathbf{W}^T\mathbf{x}=0} \frac{\mathbf{x}^T\mathbf{A}\mathbf{x}}{\mathbf{x}^T\mathbf{x}}, \tag{2.58}$$

which is known as the Courant-Fischer-Weyl min-max principle.

This is intuitively seen as

$$\lambda_i = \mathbf{x}_i^T\mathbf{A}\mathbf{x}_i = \frac{\mathbf{x}_i^T\mathbf{A}\mathbf{x}_i}{\mathbf{x}_i^T\mathbf{x}_i},\ \mathbf{x}_i\mathbf{x}_j^T,\ \forall j \neq i, \tag{2.59}$$

for all $i = 1, \ldots, d$, by definition of an eigenpair. Equation (2.58) implies that the eigenvectors are also endowed with an optimality property.

2.2.3 Singular Value Decompositions

While the EVD is usually used in case $\mathbf{A} = \mathbf{A}^*$ is Hermitian and PSD, the related Singular Vector Decomposition is used when $\mathbf{A} \in \mathbb{C}^{n\times d}$ is rectangular.

Definition 6 (Singular Value Decomposition) *Given a matrix $\mathbf{A} \in \mathbb{C}^{n\times d}$, the Singular Value Decomposition (SVD) is given as*

$$\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^*, \tag{2.60}$$

where $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_n) \in \mathbb{C}^{n\times n}$ and $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_d) \in \mathbb{C}^{d\times d}$ are both unitary matrices, such that $\mathbf{U}^T\mathbf{U} = \mathbf{U}\mathbf{U}^T = \mathbf{I}_n$ and $\mathbf{V}^T\mathbf{V} = \mathbf{V}\mathbf{V}^T = \mathbf{I}_d$. The matrix $\Sigma \in \mathbb{R}^{n\times d}$ which is all zero except for the elements $\Sigma_{ii} = \sigma_i$ for $i = 1, \ldots, \min(n, d)$. Here, $\{\sigma_i\}$ are the singular vectors, and the corresponding vectors $\{\mathbf{u}_i\} \subset \mathbb{C}^n$ and $\{\mathbf{v}_i\} \subset \mathbb{C}^d$ are called the left- and right singular vectors respectively.

The fundamental result then goes ass follows.

Lemma 4 (Existence and Uniqueness) *Given a matrix $\mathbf{A} \in \mathbb{C}^{n\times d}$, the Singular Value Decomposition (SVD) always exists and is unique up to linear transformations of the singular vectors corresponding to equal singular values.*

This implies that

$$\operatorname{rank}(\mathbf{A}) = \left|\{\sigma_i \neq 0\}\right|, \tag{2.61}$$

that is, the rank of a matrix equals the number of nonzero singular values of that matrix. The intuition behind this result is that the transformations $\mathbf{U}$ and $\mathbf{V}$ do not change the rank of a matrix, and the rank of Σ equals by definition the number of non-zero 'diagonal' elements. Similarly, the 'best' rank r approximation of a matrix $\mathbf{A}$ can be computed explicitly in terms of the SVD. Formally,

$$\mathbf{B} = \operatorname*{argmin}_{\mathbf{B}\in\mathbb{C}^{n\times d}} \|\mathbf{A} - \mathbf{B}\|_F \quad \text{s.t.} \quad \operatorname{rank}(\mathbf{B}) = r. \tag{2.62}$$

where the *Frobenius* norm of a matrix $\mathbf{A}$ is defined as $\|\mathbf{A}\|_F = \operatorname{tr}(\mathbf{A}^T\mathbf{A})$. For simplicity, assume that the singular values which are not equal to zero are distinct, and sort them as $\sigma_{(1)} > \ldots \sigma_{(d')} \geq 0$ where $\min(n, d) \geq d' > r$. This notation is often used: $a_1, \ldots, a_n$ denoted a sequence of numbers, and $a_{(1)}, \ldots, a_{(n)}$ denotes the corresponding sorted sequence of numbers. The unique matrix optimizing this problem is given as

$$\hat{\mathbf{B}} = \sum_{i=1}^{r} \sigma_{(i)}\mathbf{u}_{(i)}\mathbf{v}^*_{(i)}. \tag{2.63}$$

where $\mathbf{u}_{(i)}, \mathbf{v}_{(i)}$ are the left- and right singular vector corresponding to eigenvalue $\sigma_{(i)}$. In matrix notation this becomes

$$\hat{\mathbf{B}} = \begin{bmatrix}\mathbf{U}_1 & \mathbf{U}_2\end{bmatrix} \begin{bmatrix}\Sigma^r & 0\\ 0 & 0\end{bmatrix} \begin{bmatrix}\mathbf{V}_1^*\\ \mathbf{V}_1^*\end{bmatrix} = \mathbf{U}\Sigma_{(r)}\mathbf{V}^*, \tag{2.64}$$

where Σ^r denote the matrix consisting of the first r rows and columns of Σ, and $\Sigma_{(r)} \in \mathbb{R}^{n\times d}$ equals Σ except for the singular values $\sigma_{(r+1)}, \sigma_{(r+2)}, \ldots$ which are set to zero. This result appeals again to the min-max result of the EVD. That is, the EVD and SVD decomposition are related as

Proposition 1 (SVD - EVD) *Let $\mathbf{A} \in \mathbb{C}^{n\times d}$, let then $\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^*$ be the SVD, then*

$$\mathbf{A}^T\mathbf{A} = \mathbf{V}^*\Sigma^T\mathbf{U}\mathbf{U}^T\Sigma\mathbf{V} = \mathbf{V}^*(\Sigma^T\Sigma)\mathbf{V}. \tag{2.65}$$

Let $\mathbf{XA} = \Lambda\mathbf{X}$ be the EVD of the PSD Hermitian matrix $\mathbf{A}^T\mathbf{A}$. Then $\Sigma^T\Sigma = \Lambda$ and $\mathbf{X} = \mathbf{V}$. That is $\sigma_i^2 = \lambda_i$ for all $i = 1, \dots, \min(d, n)$ and $\lambda_i = 0$ otherwise. Similarly,

$$\mathbf{AA}^T = \mathbf{U}\Sigma\mathbf{V}^*\mathbf{V}\Sigma^T\mathbf{U}^* = \mathbf{U}(\Sigma\Sigma^T)\mathbf{U}. \tag{2.66}$$

and $\mathbf{V}$ as such contains the eigenvectors of the outer-product $\mathbf{AA}^T$, and the subsequent eigenvalues are $\lambda_i = \sigma_i^2$ for all $i = 1, \dots, \min(d, n)$ and $\lambda_i = 0$ otherwise

2.2.4 Other Matrix Decompositions

There exist a plethora of other matrix decompositions, each with its own properties. For the sake of this course the QR-decomposition is given. Let $\mathbf{A} = \mathbf{A}^T \in \mathbb{R}^{d\times d}$ be a symmetric positive definite matrix (i.e. without zero singular values). Then we can decompose the matrix $\mathbf{A}$ uniquely as the product of an uppertriangular matrix $\mathbf{R} \in \mathbb{R}^{d\times d}$ and a unitary matrix $\mathbf{Q} \in \mathbb{R}^{d\times d}$ or

$$\mathbf{A} = \begin{bmatrix} q_{11} & \dots & q_{1d} \\ \vdots & & \vdots \\ q_{d1} & \dots & q_{dd} \end{bmatrix} \begin{bmatrix} r_{11} & \dots & r_{1d} \\ & \ddots & \vdots \\ 0 & & r_{dd} \end{bmatrix} = \mathbf{QR}, \tag{2.67}$$

where $\mathbf{QQ}^T = \mathbf{Q}^T\mathbf{Q} = \mathbf{I}_d$.

2.2.5 Indirect Methods

Let us return to the question how to solve the normal equations (2.36) associated to a least squares problem, given as

$$(\mathbf{X}^T\mathbf{X})\theta = \mathbf{X}^T\mathbf{y}. \tag{2.68}$$

Rather than solving the normal equations directly by using numerical techniques, one often resorts to solving related systems. For example in order to achieve numerical stable results, to speed up multiple estimation problems. Some common approaches go as follows

QR: If $\mathbf{A}\theta = \mathbf{b}$ is a set of linear equations where $\mathbf{A}$ is upper-triangular, then the solution θ can be found using a simple backwards substitution algorithm. But the normal equations can be phrased in this form using the QR decomposition of the covariance matrix as $(\mathbf{X}^T\mathbf{X}) = \mathbf{QR}$ where $\mathbf{Q}$ is orthonormal (unitary) and $\mathbf{R}$ is upper-triangular. Hence

$$(\mathbf{X}^T\mathbf{X})\theta = (\mathbf{X}^T\mathbf{y}) \Leftrightarrow \mathbf{Q}^T(\mathbf{X}^T\mathbf{X})\theta = \mathbf{Q}^T(\mathbf{X}^T\mathbf{y}) \Leftrightarrow \mathbf{R}\theta = \mathbf{b}, \tag{2.69}$$

where $\mathbf{b} = \mathbf{Q}^T(\mathbf{X}^T\mathbf{y})$. Hence the solution θ can then be found by backwards substitution. The QR decomposition of the matrix $(\mathbf{X}^T\mathbf{X})$ can be found using a Gram-Schmid algorithm, or using Householder or Givens rotations. Such approaches have excellent numerical robustness properties.

SVD: Given the SVD of the matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$ as $\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^T$, and assume that all the singular values $\{\sigma_i > 0\}$ are strictly positive. Then the solution θ_n to (2.68) is given as

$$(\mathbf{X}^T\mathbf{X})\theta = (\mathbf{X}^T\mathbf{y}) \Leftrightarrow (\mathbf{V}\Sigma^T\mathbf{U}^T\mathbf{U}\Sigma\mathbf{V}^T)\theta = \mathbf{V}\Sigma^T\mathbf{U}^T\mathbf{y} \Leftrightarrow \theta = \mathbf{V}\Sigma^{-1}\mathbf{U}^T\mathbf{y}, \tag{2.70}$$

where $\Sigma^{-1} = \operatorname{diag}(\sigma_1^{-1}, \dots, \sigma_d^{-1}) \in \mathbb{R}^{d\times d}$ and the inverses exist by assumption.

†: In case the matrix $\mathbf{X}$ is not full rank, one has to modify the reasoning somewhat. That is, it is not guaranteed that there exists a solution θ_n to the normal equations of eq. (2.68). And in case that a solution θ_n exists, it will not be unique: assume that $\mathbf{a} \in \mathbb{R}^d$ is a nonzero vector such that $\mathbf{X}\mathbf{a} = 0$, then $\theta_n + \mathbf{a}$ solves the normal equations as well as

$$(\mathbf{X}^T\mathbf{X})(\theta_n + a_n) = (\mathbf{X}^T\mathbf{X})\theta_n = \mathbf{X}^T\mathbf{y}. \tag{2.71}$$

So it makes sense in case $\mathbf{X}$ is rank-deficient to look for a solution θ_n which is solves the normal equations as good as possible, while taking the lowest norm of all equivalent solutions. From properties of the SVD we have that any vector $\theta \in \mathbb{R}^d$ solving the problem as well as possible is given as

$$\theta = \sum_{i=1}^{r} \mathbf{v}_{(i)}\sigma_{(i)}^{-1}\mathbf{u}_{(i)}^T\mathbf{y} + \sum_{j=1}^{d-r} a_j\mathbf{v}_{(r+j)}a_j\mathbf{u}_{(r+j)}^T\mathbf{y}, \tag{2.72}$$

where $\{\sigma_{(1)}, \ldots, \sigma_{(r)}\}$ denote the r non-zero singular values. The smallest solution θ in this set is obviously the one where $a_1 = \cdots = a_{d-r} = 0$, or

$$\theta_n = \sum_{i=1}^{r} \mathbf{v}_{(i)}\sigma_{(i)}^{-1}\mathbf{u}_{(i)}^T\mathbf{y}. \tag{2.73}$$

Note that this is not quite the same as the motivation behind ridge regression where we want to find the solution trading the smallest norm requirement with the least squares objective.

From a practical perspective, the last technique is often used in order to get the best numerically stable technique. In common software packages as MATLAB, solving of the normal equations can be done using different commands. The most naive one is as follows:

```
>> theta = inv(X’*X) * (X’y)
```

But since this requires the involved inversion of a square matrix, a better approach is

```
>> theta = (X’*X) \ (X’y)
```

which solves the set of normal equations. This approach is also to be depreciated as it requires the software to compute the matrix $(\mathbf{X}^T\mathbf{X})$ explicitly, introducing numerical issues as a matrix-matrix product is known to increase rounding errors. The better way is

```
>> theta = pinv(X)*Y
```

MATLAB implements such technique using the shorthand notation

```
>> theta = X \ Y
```

2.3 Orthogonal Projections

Let us put our geometric glasses on, and consider the calculation with vectors and vector spaces. A vector space $\mathcal{A} \subset \mathbb{R}^m$ generated by a matrix $\mathbf{A} \in \mathbb{R}^{n\times m}$ is defined as

$$\mathcal{A} = \left\{ \mathbf{a} \;\middle|\; \mathbf{a} = \sum_{j=1}^{m} \mathbf{w}_j\mathbf{A}^j = \mathbf{A}\mathbf{w}, \forall \mathbf{w} \in \mathbb{R}^m \right\}. \tag{2.74}$$

Consider the following geometric problem: "Given a vector $\mathbf{x} \in \mathbb{R}^n$, and a linear space $\mathcal{A}$, extract from $\mathbf{x}$ the contribution lying in $\mathcal{A}$." Mathematically, this question is phrased as a vector which can be written as $\hat{\mathbf{x}} = \mathbf{A}\mathbf{w}$, where

$$(\mathbf{x} - \mathbf{A}\mathbf{w})^T \mathbf{A} = 0, \tag{2.75}$$

saying that "the remainder $\mathbf{x} - \hat{\mathbf{x}}$ cannot contain any component that correlate with $\mathbf{A}$ any longer". The projection $\mathbf{A}\mathbf{w}$ for this solution becomes as such

$$\mathbf{A}\mathbf{w} = \mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}\mathbf{x}, \tag{2.76}$$

that is, if the matrix $(\mathbf{A}^T\mathbf{A})$ can be inverted. Yet in other words, we can write that the projection $\hat{\mathbf{x}}$ of the vector $\mathbf{x}$ onto the space spanned by the matrix $\mathbf{A}$ can be written as

$$\hat{\mathbf{x}} = \Pi_{\mathbf{A}}\mathbf{x} = \mathbf{A}\left((\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}\mathbf{x}\right) = (\mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A})\mathbf{x}, \tag{2.77}$$

and the matrix $\Pi_{\mathbf{A}} = (\mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A})$ is called the projection matrix. Examples are

- The identity projection $\Pi_{\mathbf{A}} = I_n$, projecting any vector on itself.
- The coordinate projection $\Pi_{\mathbf{A}} = \operatorname{diag}(1, 0, \ldots, 0)$, projecting any vector onto its first coordinate.
- Let $\Pi_{\mathbf{w}} = \frac{1}{\mathbf{w}^T\mathbf{w}}(\mathbf{w}\mathbf{w}^T)$ for any nonzero vector $\mathbf{w} \in \mathbb{R}^n$, then $\Pi_{\mathbf{w}}$ projects any vector orthogonal onto $\mathbf{w}$.
- In general, since we have to have that $\Pi_{\mathbf{A}}\Pi_{\mathbf{A}}\mathbf{x} = \Pi_{\mathbf{A}}\mathbf{x}$ for all $\mathbf{x} \in \mathbb{R}^n$ (idempotent property), a projection matrix $\Pi_{\mathbf{A}}$ has eigenvalues either 1 or zero.

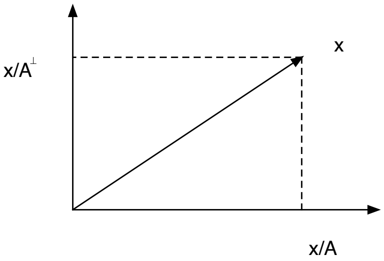


Figure 2.3: Orthogonal projection of vector $\mathbf{x}$ on the space $\mathcal{A}$, spanned by vector $\mathbf{a}$;

2.3.1 Principal Component Analysis

Principal Component Analysis (PCA) aims at uncovering the structure hidden in data. Specifically, given a number of samples $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^n$ - each $\mathbf{x}_i \in \mathbb{R}^d$ - PCA tries to come up with a shorter description of this set using less than d features. In that sense, it tries to 'compress' data, but it turns out that this very method shows up using other motivations as well. It is unlike a Least Squares estimate as there is no reference to a label or an output, and it is sometimes referred to as an *unsupervised* technique, [**?**]. The aim is translated mathematically as finding that direction

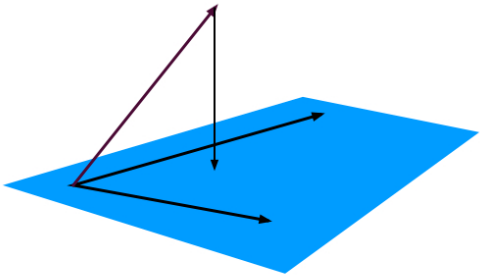


Figure 2.4: Schematic Illustration of an orthogonal projection of the angled upwards directed vector on the plane spanned by the two vectors in the horizontal plane.

$\mathbf{w} \in \mathbb{R}^d$ that explains most of the variance of the given data. 'Explains variance' of a vector is encoded as the criterion $\sum_{i=1}^n (\mathbf{x}_i^T \mathbf{w})^2$. Note that multiplication of the norm of the vector $\mathbf{w}$ gives a proportional gain in the 'explained variance'. As such it makes sense to fix $\|\mathbf{w}\|_2 = 1$, or $\mathbf{w}^T\mathbf{w} = 1$, in order to avoid that we have to deal with infinite values.

This problem is formalized as the following optimization problem. Let $\mathbf{x}_i \in \mathbb{R}^d$ be the observation made at instant i, and let $\mathbf{w} \in \mathbb{R}^d$ and let $z_i \in \mathbb{R}$ be the latent value representing $\mathbf{x}_i$ in a one-dimensional subspace. Then the problem becomes

$$\min_{\mathbf{w}\in\mathbb{R}^d, \{\mathbf{z}_i\}_i} \sum_{i=1}^n \|\mathbf{x}_i - \mathbf{w}\mathbf{z}_i\|_2^2 \text{ s.t. } \mathbf{w}^T\mathbf{w} = 1. \tag{2.78}$$

In order to work out how to solve this optimization problem, let us again define the matrix $\mathbf{X}_n \in \mathbb{R}^{n\times d}$ stacking up all the observations in $\mathcal{D}$, and the matrix $\mathbf{z}_n \in \mathbb{R}^n$ stacking up all the corresponding latent values, or

$$\mathbf{X}_n = \begin{bmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix}, \quad \mathbf{z}_n = \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{bmatrix}. \tag{2.79}$$

Then the problem eq. (2.78) can be rewritten as

$$\min_{\mathbf{w}\in\mathbb{R}^d, \mathbf{z}_n\in\mathbb{R}^n} J_n(\mathbf{z}_n, \mathbf{w}) = \operatorname{tr}\left(\mathbf{X}_n - \mathbf{z}_n\mathbf{w}^T\right)\left(\mathbf{X}_n - \mathbf{z}_n\mathbf{w}^T\right)^T \text{ s.t. } \mathbf{w}^T\mathbf{w} = 1, \tag{2.80}$$

where $\operatorname{tr}\mathbf{G} = \sum_{i=1}^n \mathbf{G}_{ii}$ where $\mathbf{G} \in \mathbb{R}^{n\times n}$. Suppose that $\mathbf{w}$ satisfying $\mathbf{w}^T\mathbf{w} = 1$ were known, then we can find the corresponding optimal $\hat{\mathbf{z}}_n(\mathbf{w})$ as a simple least squares problem: as classically we derive the objective to eq. (2.80) and equate it to zero, giving the condition for any $i = 1, \ldots, n$ that

$$\frac{\partial J_n(\hat{\mathbf{z}}_{n,i}(\mathbf{w}), \mathbf{w})}{\partial \hat{\mathbf{z}}_{n,i}(\mathbf{w})} = 0 \Leftrightarrow -2\sum_{i=1}^n \left(\mathbf{x}_i - \mathbf{w}\hat{\mathbf{z}}_{n,i}(\mathbf{w})\right)^T \mathbf{w} = 0, \tag{2.81}$$

or

$$\hat{\mathbf{z}}_n(\mathbf{w}) = \frac{1}{(\mathbf{w}^T\mathbf{w})}\mathbf{X}_n\mathbf{w}. \tag{2.82}$$

Now having this closed form solution for $\hat{\mathbf{z}}_n$ corresponding to a $\mathbf{w}$, one may invert the reasoning and try to find this $\mathbf{w}$ satisfying the constraint $\mathbf{w}^T\mathbf{w} = 1$ and optimizing the objective $J_n(\hat{\mathbf{z}}_n(\mathbf{w}), \mathbf{w})$ as

$$\min_{\mathbf{w}\in\mathbb{R}^d} J'_n(\mathbf{w}) = J_n(\hat{\mathbf{z}}_n(\mathbf{w}), \mathbf{w}) = \operatorname{tr}\left(\mathbf{X}_n - \hat{\mathbf{z}}_n(\mathbf{w})\mathbf{w}^T\right)\left(\mathbf{X}_n - \hat{\mathbf{z}}_n(\mathbf{w})\mathbf{w}^T\right)^T. \ \text{ s.t. } \ \mathbf{w}^T\mathbf{w} = 1. \tag{2.83}$$

Working out the terms and using the normal equations (2.82) gives the equivalent optimization problem

$$\min_{\mathbf{w}\in\mathbb{R}^d} J'_n(\mathbf{w}) = \left\|\mathbf{X}_n - (\mathbf{w}\mathbf{w}^T)\mathbf{X}_n\right\|_F \ \text{ s.t. } \ \mathbf{w}^T\mathbf{w} = 1. \tag{2.84}$$

where the Frobenius norm $\|\cdot\|_F^2$ is defined for any matrix $\mathbf{G} \in \mathbb{R}^{n\times d}$ as

$$\|\mathbf{G}\|_F = \operatorname{tr}\mathbf{G}\mathbf{G}^T = \sum_{i=1}^n \mathbf{G}_i^T\mathbf{G}_i = \sum_{ij}\mathbf{G}_{ij}^2, \tag{2.85}$$

and where $\mathbf{G}_i$ denotes the ith row of $\mathbf{G}$. It is useful to interpret this formula. It is easy to see that the matrix $\Pi_{\mathbf{w}} = (\mathbf{w}\mathbf{w}^T)$ as a projection matrix as $(\mathbf{w}^T\mathbf{w})^{-1} = 1$ by construction, and as such we look for the best projection such that $\Pi\mathbf{X}_n$ is as close as possible to $\mathbf{X}_n$ using a Euclidean norm. To solve this optimization problem, let us rewrite eq. (2.84) in terms of the arbitrary vector $\mathbf{v} \in \mathbb{R}^d$ such that $\mathbf{w} = \frac{\mathbf{v}}{\|\mathbf{v}\|_2}$ has norm 1 by construction. We take care of this rescaling by dividing the objective through $\mathbf{v}^T\mathbf{v}$. Recall that linearity of the trace implies $\operatorname{tr}\mathbf{G}\mathbf{G}^T = \operatorname{tr}\mathbf{G}^T\mathbf{G}$. Since $(\mathbf{w}\mathbf{w})^T(\mathbf{w}\mathbf{w}) = (\mathbf{w}\mathbf{w})$ (idempotent), one has

$$\min_{\mathbf{v}\in\mathbb{R}^d} J'_n(\mathbf{v}) = \min_{\mathbf{v}\in\mathbb{R}^d} \frac{\mathbf{v}^T\mathbf{v} - \mathbf{v}^T(\mathbf{X}_n^T\mathbf{X}_n)\mathbf{v}}{\mathbf{v}^T\mathbf{v}} = 1 - \max_{\mathbf{v}\in\mathbb{R}^d}\frac{\mathbf{v}^T(\mathbf{X}_n^T\mathbf{X}_n)\mathbf{v}}{\mathbf{v}^T\mathbf{v}}, \tag{2.86}$$

and $\mathbf{w}$ solving eq. (2.78) is given as $\mathbf{w} = \frac{\mathbf{v}}{\|\mathbf{v}\|}$.

Now luckily enough maximization of $\frac{\mathbf{v}^T(\mathbf{X}_n^T\mathbf{X}_n)\mathbf{v}}{\mathbf{v}^T\mathbf{v}}$ is a wellknown problem, studied for decades in analyses and numerical algebra as the problem of maximizing the Rayleigh coefficient. From this we know not only how the maximum is found, but how all local maxima can be found. Equating the derivative of the Rayleigh coefficient to zero gives the conditions

$$\lambda(\mathbf{v}) = \frac{\mathbf{v}^T(\mathbf{X}_n^T\mathbf{X}_n)\mathbf{v}}{\mathbf{v}^T\mathbf{v}} \Leftrightarrow \lambda(\mathbf{v})(\mathbf{v}^T\mathbf{v}) = \mathbf{v}^T(\mathbf{X}_n^T\mathbf{X}_n)\mathbf{v}. \tag{2.87}$$

Now deriving to $\mathbf{v}$ and equating to zero gives the conditions

$$\lambda(\mathbf{v})\mathbf{v} = (\mathbf{X}_n^T\mathbf{X}_n)\mathbf{v}, \tag{2.88}$$

and we know that the d orthogonal solutions $\{\mathbf{v}_i\}_i$ and corresponding coefficients $\{\lambda(\mathbf{v}_i)\}$ are given by the eigenvectors and eigenvalues of the matrix $\mathbf{X}_n^T\mathbf{X}_n$, such that $\mathbf{v}_i^T\mathbf{v}_j = \delta_{ij}$ (i.e. is one if $i = j$, and zero otherwise), and $\mathbf{v}_i^T(\mathbf{X}_n^T\mathbf{X}_n)\mathbf{v}_j = \lambda_i(\mathbf{v}_i)\delta_{ij}$. We will use the notation that $\{(\lambda_i(\mathbf{X}_n^T\mathbf{X}_n), \mathbf{v}_i(\mathbf{X}_n^T\mathbf{X}_n))\}$ to denote this set. In fact, the relation PCA - eigenvalue decomposition is so close that they are often considered to be one and the same. That is if an algorithm performs an eigenvalue decomposition at a certain stage of a certain matrix, one may often think of it as a PCA of this matrix thereby helping intuition.

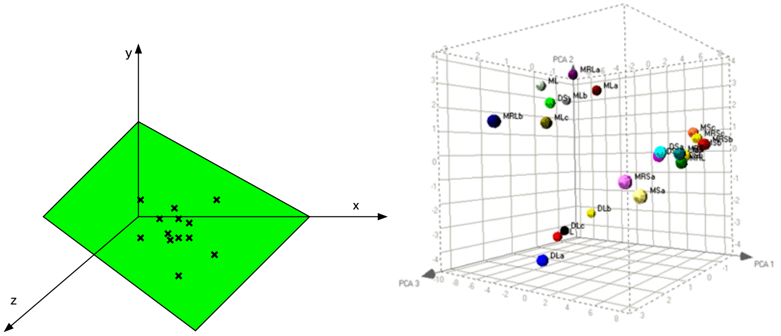


Figure 2.5: (a) An example of $n = 13$ and $d = 2$, where all the samples 'x' lie in a two-dimensional linear subspace denoted as the filled rectangle. PCA can be used to recover this subspace from the data matrix $\mathbf{X} \in \mathbb{R}^{13\times 3}$. (b) An example of the results of a PCA analysis on 2000 of expression levels observed in 23 experiments. The 3 axes correspond with the 3 principal components of the matrix $\mathbf{X} \in \mathbb{R}^{23\times 2000}$.

Chapter 3

Flavors of Dynamic Systems and Models

> ”How can we mathematically represent and characterize a dynamical system? And what is a good representation for a certain task?”

This chapter is a survey of some important results in systems theory, and will serve to fix ideas, notation and language to be used in later chapters. Models come in many different flavors. We will mainly be interested in models relating input to output signals using formal equations. But keep in mind that this is but one choice. In some cases it is good practice to express models using predicate logic, software code, graphs and pictures, human language or even intuitions. At the end a model is only an internal representation of the actual system, and is as good proportionally to how well it serve its final purpose.

3.1 General Characterization of Dynamic Systems

In the context of this course, we are interested in dynamic systems, i.e. systems which have a nontrivial dependence on its past (future) signals. To make the ideas in this chapter easier to present, we will restrict ourself to systems accepting a univariate input signal $\{u_t\}_t \subset \mathbb{R}$, and outputting a univariate signal $\{y_t\}_t \subset \mathbb{R}$. Such systems are called SISO ('Single Input, Single Output'). It is not to hard then to figure out what MIMO, MISO, TITO stands for, is it?

Dynamic models come in different forms, but arguably the largest watershed is drawn between models represented in continuous-time, or in discrete-time. While in the physical sciences one often builds up models based on first-order principles expressed as continuous Ordinary Differential Equations (ODEs) or Partial Differential equations (PDEs), in the context of system identification one typically chooses discrete representations.

(Digital): The system's input-output behavior is recorded as numerical values, The keyword 'digital' suggests often a discrete nature either in time scale, or in the values of the signals.

(Parameters): The unknown quantities in the assumed model structure which is used to describe the system of interest. Those parameters will be estimated based on the collected input-output data of

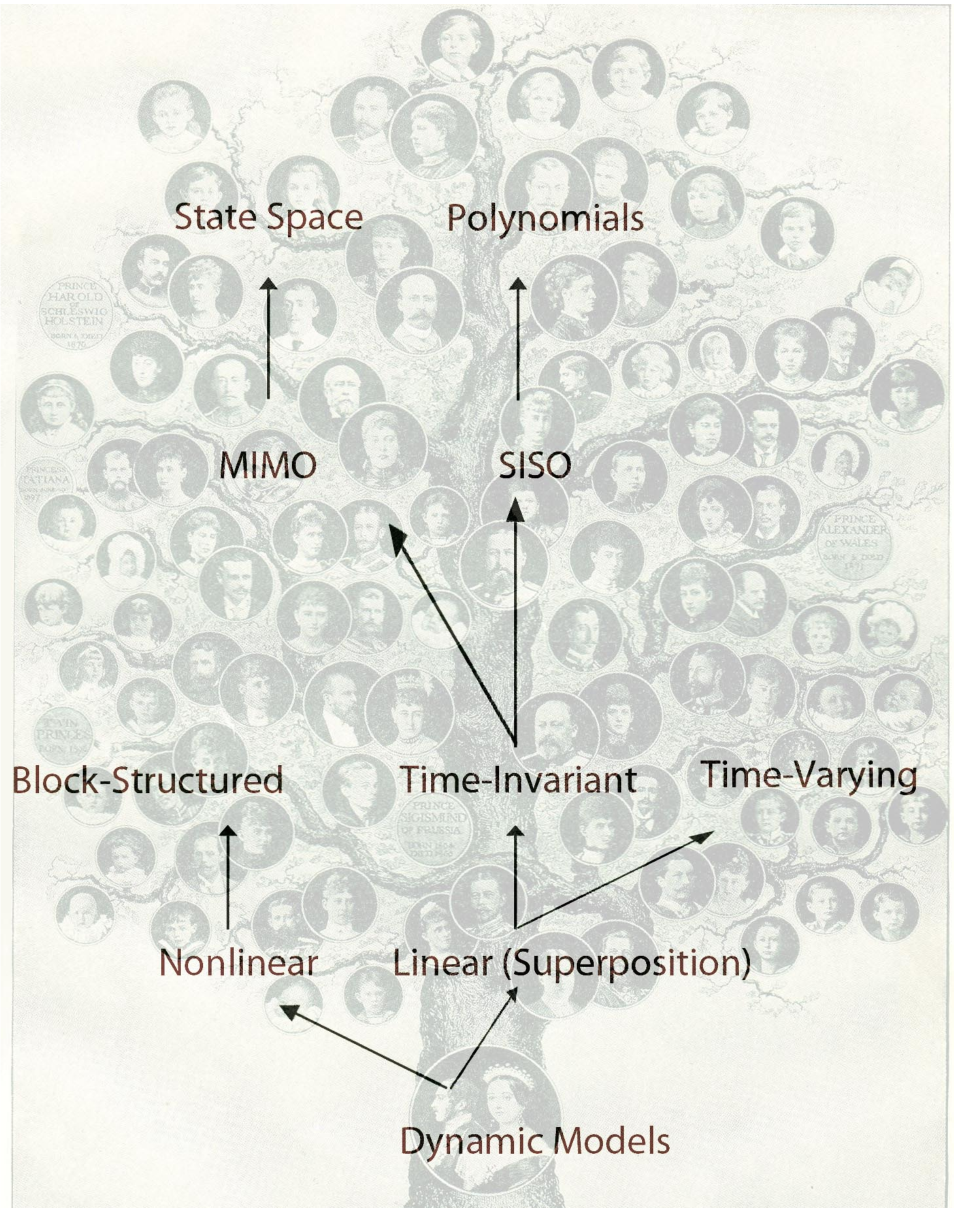


Figure 3.1: A Taxonomy of Systems

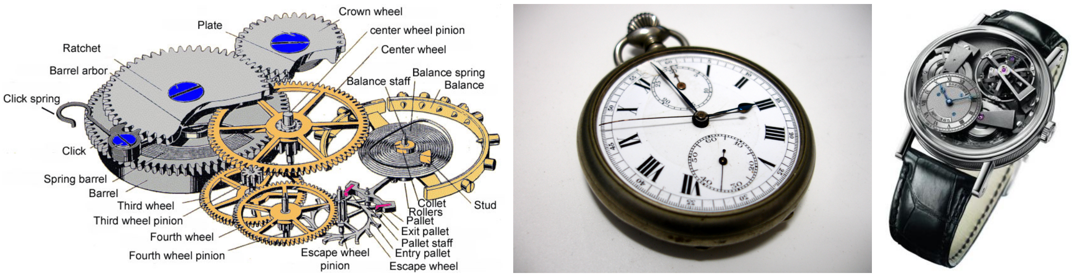
the systems. In general, the there are only a small, constant number of such parameters (as e.g. compared to n) to be estimated.

(White-box): Sometimes a model can be formalized in terms of physical laws, chemical relations, or other theoretical considerations for the studied system. Such a model is called white-box, as the intern model description directly appeals to the intern mechanism supposed to underly the system. As such, the system explains (in a sense) *why* it operates as is seen in input-output behavior.

(Black-box): A black-box model does not directly appeal to the actual mechanisms which are supposed to underly the studied system. A black-box model merely intends to make good predictions of the (future) system behavior: the internal description merely serves to relate input signals to output signals.

(Grey-box): Grey-box models are a mixture between black-box and white-box models. Some of the internal working of the system can be assumed to be dictated by laws, The complete input-output behavior is only known however up to some gaps. Those blanks can then be modeled using black-box techniques.

In the context of this course, mainly black-box models are considered. Section 5.3 will discuss how techniques can be extended to white- and grey-box models.



(a) (b) (c)

Figure 3.2: Metaphorical representation of a (a) white-box model where you see the internal working, (b) black-box model where you only see the output of the system, and (c) grey-box model where both are present to some degree.

The conversion of continuos time-models into an (approximate) discrete-time counterpart falls within the area of systems theory. This overall challenge is historically and research-wise related to numerical solvers used for finding solutions for sets of differential equations. Again, this book will adopt the convention to denote continuous signals using its 'argument form', e.g. $u(t)$. Its discrete counterpart are denoted as u_t where now t becomes an index taking integer values, i.e. $t = 1, 2, 3, \ldots$.

3.2 LTI Systems

The following definitions help us to narrow down the class of models studied here. The central notion is the idea of an impulse response, intuitively the timely effect observed at the output of a system when injecting a pulse signal into the system. An example is given in Fig. (3.3).

Figure 3.3: An intuitive illustration of an impulse response: a sound beep ripples an acoustic signal through a room.

Definition 7 (Time-invariant System) *A system is called* time-invariant *if its response to a certain input signal does not depend on absolute time.*

Formally, the properties of the model (e.g. the orders, parameters or sampling time) do not depend on the precise index t. An example of a time-invariant model is $y_t = \theta + e_t$. If this model were time-varying it would be denoted as $y_t = \theta_t + e_t$.

Definition 8 (Causal System) *A system is called* causal *if an output response at a certain instant depends only on the inputs up to that time. That is, if a useful prediction can be made based on the past signals only.*

Definition 9 (Superposition Principle) *The superposition principle states that for any linear system the net response at a given place and time caused by a linear combination of different input signals equals the same linear combination of the output responses which would have been caused by individual application of each input signal.*

Theorem 1 (Continuous Impulse Response (IR) representation) *Given a causal LTI system $\mathcal{S}$, then its mapping from any continuous input signal $\{u(s)\}_{s\leq t}$ to a corresponding output $y(t)$, for any such t, can be represented using a fixed function $h : \mathbb{R}_+ \to \mathbb{R}$ as*

$$y(t) = \int_{\tau=0}^{\infty} h(\tau)u(t-\tau)d\tau. \tag{3.1}$$

The function h is called the continuous Impulse Response (IR) function.

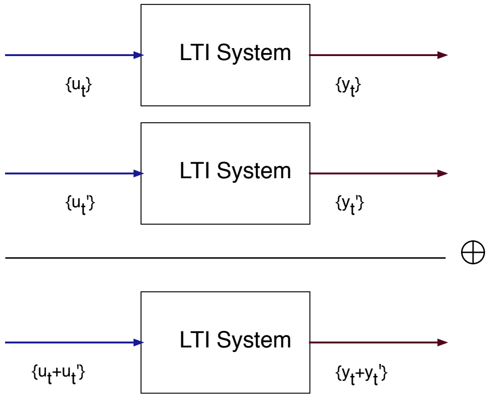


Figure 3.4: Illustration of the superposition principle as in Def. (9).

Theorem 2 (Discrete Impulse Response (IR) representation) *Given a causal LTI system $\mathcal{S}$, then its mapping from any input signal $\{\dots, u_{t-1}, u_t\}$ to a corresponding output y_t can be represented using a fixed sequence $\{h_0, h_1, \dots, h_d\}$ as*

$$y_t = \sum_{\tau=0}^{d} h_\tau u_{t-\tau}, \ \forall t = 1, \dots. \tag{3.2}$$

with order d which is possibly infinite. The (infinite) vector $\mathbf{h} = (h_0, h_1, \dots, h_d)^T \in \mathbb{R}^d$ is called the discrete Impulse Response (IR) vector.

This step from continuous representation (or 'model') to a discrete representation ('model') is intuitively seen as follows: As working assumption we take that we sample the (continuous) time as every other $\Delta > 0$ period, such that the sample after a time Δ correspond with the zeroth sample. Formally, this is written as the relation for every $t' = \dots, -2, -1, 0, 1, 2, \dots$ as

$$u_{t'} = u(t'\Delta + \tau), \ \forall 0 \leq \tau < \Delta, \tag{3.3}$$

and $u_0 = u(0)$. Note the use of the symbol t and t' for differentiating between continuous time and discrete index. In the rest van this text we will use t in both cases as its meaning is almost always clear from its context.

$$\begin{aligned}
y(t'\Delta) &= \int_{\tau=0}^{\infty} h(\tau) u(t'\Delta - \tau) d\tau \\
&= \sum_{\tau'=1}^{\infty} \int_{\tau=(\tau'-1)\Delta}^{\tau'\Delta} h(\tau) u(t'\Delta - \tau) d\tau \\
&= \sum_{\tau'=1}^{\infty} \left(\int_{\tau=(\tau'-1)\Delta}^{\tau'\Delta} h(\tau) d\tau \right) u_{t'-\tau'} \\
&= \sum_{\tau'=1}^{\infty} h_{\tau'} u_{t'-\tau'}, \qquad (3.4)
\end{aligned}$$

where we define for any $\rho = 1, 2, 3, \ldots$ that

$$h_{\tau'} = \int_{\tau=(\tau'-1)\Delta}^{\tau'\Delta} h(\tau) d\tau. \tag{3.5}$$

Again we will use τ in general to denote both the displacement in the continuous case (i.e. τ), as the distant in the discrete case (i.e. τ').

Observe that no approximation need to be made if one assumes eq. (3.5), and that it is sufficient in the sense that $\{h_\tau\}_\tau$ fully specifies the (continuous) response to the input signal. Even if eq. (3.5) does not hold, $\{h_\tau\}_\tau$ might still give a good discrete to what is happening in continuous time, provided the signal $u(t)$ does not change too much during intersample intervals. The study of different sampling schemes and the influence on subsequent analysis steps goes to the heart of digital control theory and signal processing, and has its obvious relevance in the design of A/D convertors.

The following two examples are prototypical.

Example 11 (A First Order Example) *At first an example is given of a continuous time first order linear system. Assume a system is modeled as a first order differential equation, or*

$$T\frac{dy(t)}{dt} + y(t) = Ku(t - \tau), \tag{3.6}$$

for a time delay $\tau \geq 0$, a gain $K > 0$ and time constant $T > 0$. Then the impulse response can be computed by equating the signal $u(t)$ to an impulse $\delta(t)$, and to solve for the corresponding $y(t)$. Similarly, one can compute the solution when the system is excited with the step signal $\{u(t) = 1(t \geq 0), t = -\infty, \ldots, \infty\}$. This solution is given in Fig. (3.5).

Conversely, one can determine the parameters τ, K, T by looking at the step response. Figure (3.5) demonstrates a graphical method for determining the parameters K, T, τ from the step response:

- *The gain K is given by the final value.*
- *By fitting the steepest tangent, T and τ can be obtained. The slope of this tangent is $\frac{K}{T}$*
- *This tangent crosses the time-axis at τ, the time delay.*

Example 12 (Step Response of a damped Oscillator) *Consider a second-order continuous system characterized by a differential equation, or*

$$\frac{d^2y(t)}{dt^2} + 2\xi\omega_0\frac{dy(t)}{dt} + \omega_0^2 y(t) = K\omega_0^2 u(t). \tag{3.7}$$

After some calculations the solution when of the differential equation when exited with a step input is found to be

$$y(t) = K\left(1 - \frac{e^{-\xi\omega_0 t}}{\sqrt{1-\xi^2}} \sin\left(\omega_0 t\sqrt{1-\xi^2} + \tau\right)\right), \tag{3.8}$$

where the time delay is $\tau = \arccos\xi$. Here the gain is parametrized by K, and the timescale is parametrized by ω_0. The parameter ξ regulates the damping of the oscillation of the system, hence the name. This step response is illustrated in Fig. (3.6) for different values of $\xi = 0.1, 0.2, 0.5, 0.7, 0.99$. In the example we fix the gain as $K = 1$, and the timescale as $\omega_0 = 1$.

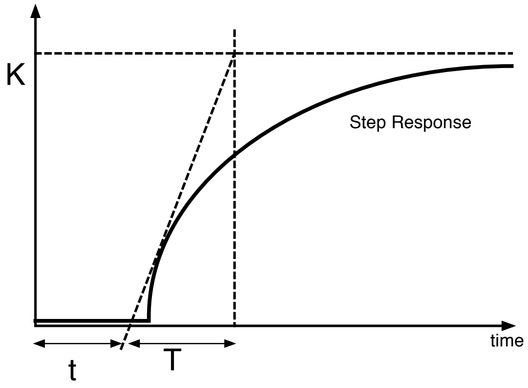


Figure 3.5: Example of a first-order system given in eq. (3.6)

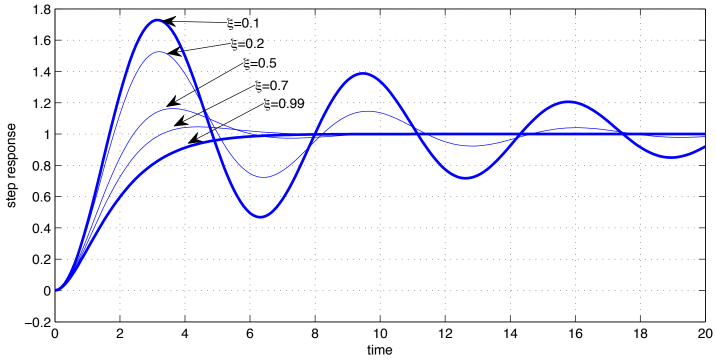


Figure 3.6: Example of a step response of a second-order system given in eq. (3.7) In this example $K = 1$ and $\omega_0 = 1$.

Chapter 13 discusses how one can extend this model class to account for nonlinear effects. That is, how identification techniques can be applied when the Linear superposition property is not valid any longer.

3.2.1 Transforms

In order to get insight into why, how and what a certain model is capable of, it turns out to be quite useful to express the dynamics using different languages. That is, the model dynamics are transformed to various descriptions. This amounts in general to the theory and practice of transforms.

Sinusoid Response and the Frequency Function

Let us elaborate on the simple case where the input of a system is a cosine function:

$$u_t = \cos(\omega t),\ \forall t = \ldots, -2, -1, 0, 1, 2, \ldots. \tag{3.9}$$

It will be convenient to rewrite this as

$$u_t = \Re e^{i\omega t},\ \forall t = -2, -1, 0, 1, 2, \ldots. \tag{3.10}$$

with $\Re\cdot$ denoting the 'real part'. This follows from Fermat's equality that $e^{i\omega_0 t} = \cos(i\omega_0 t) + i\sin(i\omega_0 t)$ as depicted in Fig. (3.7).

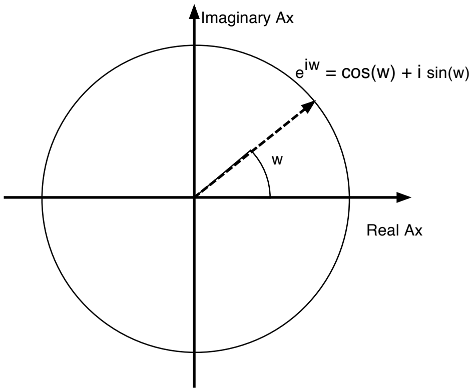


Figure 3.7: The Complex unit-circle.

Then given a (discrete) system $\{h_\tau\}_\tau$ we have for any $t = \dots, -1, 0, 1, \dots$ that

$$\begin{aligned} y_t &= \sum_{\tau=1}^{\infty} h_\tau \cos(\omega t - \tau) \\ &= \sum_{\tau=1}^{\infty} h_\tau \Re e^{i\omega(t-\tau)} \\ &= \Re \sum_{\tau=1}^{\infty} h_\tau e^{i\omega(t-\tau)} \\ &= \Re \left(e^{i\omega t} \sum_{\tau=1}^{\infty} h_\tau e^{-i\omega\tau} \right), \end{aligned} \tag{3.11}$$

the second equality follows as h_τ is real. Using the definition

$$H\left(e^{i\omega}\right) = \sum_{\tau=1}^{\infty} h_\tau e^{-i\omega\tau}, \tag{3.12}$$

one can write further

$$\begin{aligned} y_t &= \Re\left(e^{i\omega t} H\left(e^{i\omega}\right)\right) \\ &= \left|H\left(e^{i\omega}\right)\right| \cos\left(\omega t + \varphi\right), \end{aligned} \tag{3.13}$$

where

$$\varphi = \arg H\left(e^{i\omega}\right). \tag{3.14}$$

And the absolute value and the argument of a complex value $x + iy \in \mathbb{C}$ are defined as

$$|x + iy| = \sqrt{x^2 + y^2}, \tag{3.15}$$

and

$$\arg(x + iy) = \begin{cases} 2\arctan \frac{y}{\sqrt{x^2+y^2}+x} & x > 0, y \neq 0 \\ \pi & x < 0, y = 0 \\ \infty & \text{else.} \end{cases} \tag{3.16}$$

This follows from the rules of complex calculus. This derivation is paramount into understanding how LTIs work. The engineering view is that an LTI H is a mapping from a sinusoid with frequency $-\pi \leq \omega \leq \pi$, to a corresponding one with the same frequency, but with amplitude magnified by $\left|H\left(e^{i\omega}\right)\right|$, and phase increased by $\arg H\left(e^{i\omega}\right)$. For that reason, the function $H\left(e^{i\omega}\right)$ is denoted as the *frequency function* or *transfer function* of the LTI H. The method of transforms - coming in the form of Laplace, z- or Fourier transforms is then all about the concept of $H\left(e^{i\omega}\right)$, as will be elaborated in the remainder of this section.

History has provided us with many graphical tools to characterize LTIs, amongst which

(Bode) Represents the amplitudes $\left|H\left(e^{i\omega}\right)\right|$ and phases $\arg H\left(e^{i\omega}\right)$ as a function of the frequency $-\pi \leq \omega \leq \pi$.

(Nyquist) Represents for any $-\pi \leq \omega \leq \pi$ the complex numbers

$$\left(\left|H\left(e^{i\omega}\right)\right|, \arg H\left(e^{i\omega}\right)\right) \tag{3.17}$$

as curves in the 2D graph.

These concepts are often studied in the continuous-time case, but their basic properties carry over to the discrete-time case as well.

Now, those reasonings motivate us to decompose the given signals into contributions of sinusoids with various phases and amplitudes. Indeed if we know this decomposition it is straightforward to characterize the system from observed input- and output signals. Let us consider first the case where this decomposition is performed on a input-signal $\{u_t\}_{t=1}^n$ of finite length (!). Now define the function $\mathcal{U}_n : \mathbb{R} \to \mathbb{C}$ for any $-\pi \leq \omega \leq \pi$ as

$$\mathcal{U}_n(\omega) = \frac{1}{\sqrt{n}} \sum_{t=1}^n u_t e^{-i\omega t}. \tag{3.18}$$

The values obtained for $\omega = \frac{1}{n}2\pi k$ for $k = 1, \ldots, n$ form the Discrete Fourier Transform of the finite sequence $\{u_t\}_{t=1}^n$. We can reconstruct the original sequence $\{u_t\}_{t=1}^n$ from $\{\mathcal{U}(\frac{i2\pi k}{n})\}_{k=1}^n$ as

$$u_t = \frac{1}{\sqrt{n}} \sum_{k=1}^n \mathcal{U}_n\left(\frac{2\pi k}{n}\right) e^{i2\pi kt}, \tag{3.19}$$

for any $t = 1, \ldots, n$. Indeed

$$\frac{1}{\sqrt{n}} \sum_{k=1}^n \mathcal{U}_n\left(\frac{2\pi k}{n}\right) e^{i2\pi kt} = \frac{1}{n} \sum_{k=1}^n \sum_{s=1}^n u_s \exp\left(\frac{2i\pi k(t-s)}{n}\right) = \frac{1}{n} \sum_{s=1}^n u_s n \delta_{t-s} = u_t, \tag{3.20}$$

where

$$\delta_{k-s} = \frac{1}{n} \sum_{k=1}^n \exp\left(\frac{i2\pi rk}{n}\right) = \begin{cases} 1 & r = 0 \\ 0 & r \neq 0. \end{cases}. \tag{3.21}$$

This follows in turn as the different functions $\{\exp(i2\pi k/n)\}_{k=1}^n$ are carefully crafted as orthonormal functions.

As such, we have found a good decomposition of the signal $\{u_t\}$ in its frequency components $\{\mathcal{U}_n(\omega_k)\}_k$. It is instructive to consider the 'energy' of the different frequencies. This notion is captured by the *periodogram*, which is a function defined for any $-\pi \leq \omega \leq \pi$ as

$$\left|\mathcal{U}_n(\omega_k)\right|^2, \ \forall \omega_k. \tag{3.22}$$

Parseval's equality gives then that

$$\sum_{k=1}^n \left|\mathcal{U}_n\left(\frac{2\pi k}{n}\right)\right|^2 = \sum_{t=1}^n u_t^2. \tag{3.23}$$

3.3 Useful Representations

3.3.1 Disturbances

Again, we will focus on the LTI model

$$y_t = \sum_{\tau=0}^{\infty} h_\tau u_{t-\tau} = H(q^{-1})u_t. \tag{3.24}$$

A disturbance refers to the collection of effects such that the system at hand is not entirely described as in eq. (3.24). Herefor, the model of eq. (3.24) is extended as

$$y_t = H(q^{-1})u_t + d_t, \tag{3.25}$$

where $\{d_t\}_t \subset \mathbb{R}$ denote the disturbances. It is often realistic to have such terms as in practice one may be faced with data which is perturbed by

(Random Noise): The observations made during the experiment are often due to stochastic signals influencing the system during operation. Effects are roughly called stochastic if they differ from experiment to experiment. That is, if they result in different signals even when the experiment is repeated under identical circumstances. Noise terms may be due to inaccuracies in the measurement devices, or to external influences which are not directly measured, and which are beyond control of the designed experiment. Often, the stochastic framework and statistical models provide a reasonable description of such effects.

(Under-modeling): In case the system does not quite fit the model structure which is chosen, disturbance terms might reflect the 'residual' dynamics of the system. Those are often present as a model is a useful abstraction of the studied system, and a complete modeling is in many real-world cases beyond reach.

(Nonlinear Effects): A model satisfying the superposition principle is often accurate (convenient) enough for our purposes. However, most systems reflect some sort of deviation of this mathematical principle, in general denoted as a 'nonlinear effects'. In process industry for example, saturation effects often occur. But in case the system remains at more or less the same operation regime the system can be expected to behave linear. A disturbance term can absorb the occasional nonlinear effects nonetheless.

(Time Varying): In the same spirit as the nonlinear effects, any real system displays non time-invariant effects. But if one remains more or less in the same operation regime and avoids structural changes during the experiment, a disturbance term might get rid of minor time-varying effects. A particular time-varying effect is due to aging of the experimental setup, which can be avoided by collecting data in a relatively short timespan.

3.3.2 Polynomial Representations

Consider the SISO case. Then any LTI can be represented as

$$\begin{aligned} y_t = \sum_{\tau=0}^{\infty} h_\tau u_{t-\tau} + \sum_{\tau=0}^{\infty} g_\tau e_{t-\tau} \\ = (1 + h_1 q^{-1} + \dots + h_\tau q^{-\tau} + \dots) u_t + (1 + g_1 q^{-1} + \dots + g_\tau q^{-\tau} + \dots) e_t \\ = H(q^{-1}) u_t + G(q^{-1}) e_t. \quad (3.26) \end{aligned}$$

with suitable polynomials $H(z)$ and $G(z)$. The former characterizes the dynamics of the (observed) input signals, the latter captures the dynamics of the disturbances to the system. Now there exists a wide spectrum of models which parameterize this general description in some convenient, appropriate way. A few are enumerated here:

FIR(m_b):

$$y_t = b_0 u_t + b_1 u_{t-1} + \dots + b_{m_b} u_{t-m_b} + e_t = B(q^{-1}) u_t + e_t, \quad (3.27)$$

where

$$B(q^{-1}) = b_0 + b_1 q^{-1} + \dots + b_{m_b} q^{-m_b}. \quad (3.28)$$

ARX(m_a, m_b): An Auto-Regressive model with eXogenous inputs is specified as

$$A(q^{-1}) y_t = B(q^{-1}) u_t + e_t, \quad (3.29)$$

where

$$A(q^{-1}) = 1 + a_1 q^{-1} + \dots + a_{m_a} q^{-m_a}. \quad (3.30)$$

and $B(q^{-1})$ is given as in eq. (3.28). This implies that the model equals

$$y_t = \frac{B(q^{-1})}{A(q^{-1})} u_t + \frac{1}{A(q^{-1})} e_t, \quad (3.31)$$

and the noise influences the outcome in a nontrivial way. This is typical for situations where the disturbances come into the dynamical systems at earlier stages, i.e. the noise shares some important aspects of the dynamics with the influence of an input. Its appeal in practice comes however from a different phenomenon. This model fits straightforwardly a model description which is linear in the parameters.

$$y_t = \phi_t^T \theta, \quad (3.32)$$

where

$$\begin{cases} \phi_t = (-y_{t-1}, \dots, y_{t-m_a}, u_t, \dots, u_{t-m_b})^T \in \mathbb{R}^{m_a+m_b+1} \\ \theta = (a_1, \dots, a_{m_a}, b_0, b_1, \dots, b_{m_b})^T \in \mathbb{R}^{m_a+m_b+1}. \end{cases} \quad (3.33)$$

ARMAX(m_a, m_b, m_c): An Auto-Regressive model with eXogenous inputs and Moving Average model for the disturbances is given as

$$A(q^{-1}) y_t = B(q^{-1}) u_t + C(q^{-1}) e_t, \quad (3.34)$$

where

$$C(q^{-1}) = 1 + c_1 q^{-1} + \dots + c_{m_c} q^{-m_c}. \quad (3.35)$$

and $A(q^{-1})$ is as in eq. (3.30) and $B(q^{-1})$ is given as in eq. (3.28). This model fits the description of eq. (3.26) as

$$y_t = \frac{B(q^{-1})}{A(q^{-1})} u_t + \frac{C(q^{-1})}{A(q^{-1})} e_t, \tag{3.36}$$

where the dynamics of the noise are parametrized more flexible than in the ARX(m_a, m_b) model.

OE(m_a, m_b): The Output Error model of order m_a, m_b is given as

$$y_t = \frac{B(q^{-1})}{A(q^{-1}} u_t + e_t, \tag{3.37}$$

where $A(q^{-1})$ is as in eq. (3.30) and $B(q^{-1})$ is given as in eq. (3.28). This model is often used in case the noise comes only in at the end-stages of the process to be modeled: it does not share many dynamics with the input.

$(m_a, m_b, m_c, m_d, m_f)$: The general fractional representation of a polynomial model is refered to as a Box-Jenkins (BJ) model structure of orders m_a, m_b, m_c, m_d, m_f defined as

$$A(q^{-1} y_t = \frac{B(q^{-1})}{F(q^{-1}} u_t + \frac{C(q^{-1})}{D(q^{-1}} e_t. \tag{3.38}$$

where

$$\begin{cases} D(q^{-1}) = 1 + d_1 q^{-1} + \cdots + d_{m_d} q^{-m_d} \\ F(q^{-1}) = 1 + f_1 q^{-1} + \cdots + f_{m_f} q^{-m_f}, \end{cases} \tag{3.39}$$

and $A(q^{-1}), B(q^{-1}), C(q^{-1})$ are as defined above. It should be stressed that its not often useful to use this model structure in its general form. On the contrary, it is good practice to reduce it by setting one or more of the polynomials to unity.

3.3.3 Models for Timeseries

W enow study some common model structured useful for characterizing timeseries $\{y_t\}$. The input signals to such systems are not observed, and are commonly assumed to be noise signals. In general, we consider the model

$$y_t = \sum_{\tau=0}^{\infty} g_\tau e_{t-\tau} = (1 + g_1 q^{-1} + \cdots + g_\tau q^{-\tau} + \dots) e_t = G(q^{-1}) e_t. \tag{3.40}$$

MA(m): A Moving Average model of order m:

$$y_t = e_t + c_1 e_{t-1} + \cdots + c_m e_{t-m} = C(q^{-1}) e_t. \tag{3.41}$$

Such an all-zero model is useful to model signals with power spectra which have sharp valleys toward zero.

AR(m): An Auto-Regressive model of order m:

$$y_t + a_1 y_{t-1} + \cdots + a_m y_{t-m} = A(q^{-1}) y_t = e_t. \tag{3.42}$$

Equivalently, one writes

$$y_t = A^{-1}(q^{-1})e_t. \tag{3.43}$$

Such an all-pole model is useful to model signals with power spectra which have sharp upward peaks.

ARMA(m_a, m_c): An Auto-Regressive Moving Average model of orders m_a and m_c:

$$A(q^{-1})y_t = C(q^{-1})e_t, \tag{3.44}$$

where

$$\begin{cases} A(q^{-1}) = 1 + a_1 q^{-1} + \cdots + a_{m_a} q^{-m_a} \\ C(q^{-1}) = 1 + c_1 q^{-1} + \cdots + c_{m_c} q^{-m_c}. \end{cases} \tag{3.45}$$

Equivalently, one has

$$y_t = \frac{C(q^{-1})}{A(q^{-1})} e_t, \tag{3.46}$$

and this model as such uses a fractional noise model.

ARIMA(d, m_a, m_c): An Auto-Regressive Integrated Moving Average model of orders m_a and m_c:

$$(1 - q^{-1})^d A(q^{-1})y_t = C(q^{-1})e_t, \tag{3.47}$$

with $A(q^{-1})$ and $C(q^{-1})$ the polynomials as defined in (3.45). In case $d = 1$, this is equivalent to the model

$$A(q^{-1})(y_t - y_{t-1}) = C(q^{-1})e_t, \tag{3.48}$$

hence explaining the naming convention.

3.3.4 Stability and Minimal Phase

A system $\mathcal{S}$ is called Bounded-Input, Bounded Output (BIBO) stable if any input signal of bounded input can only imply an output which is bounded.

Definition 10 (Minimum Phase) *A polynomial $A(z) = 1 + a_1 z + \cdots + a_m z^m$ with real-valued coefficients $\{a_1, \ldots, a_m\}$ is called minimum phase if it has all its zeros $\{z \in \mathbb{C}\}$ strictly inside the unit circle. This naming convention is as for any other polynomial with real-valued coefficients and such that $|B(e^{\mathbf{i}\omega})| = |A(e^{\mathbf{i}\omega})|$ for all ω has larger or equal phase lag $-\arg B(e^{\mathbf{i}\omega})$. That implies in turn that a minimal phase system has zero delay, and is causally invertible. That is the inverse system is BIBO stable.*

This nomenclature is used as polynomial with minimal phase does imply the smallest phase-lag of all polynomials sharing the same characteristics in terms of the magnitudes of the frequency responses. A model is in minimal phase if its zeros lie strictly inside the unit circle.

3.4 Simulation and Prediction

A model is often used in two different regimes in order to forecast what is going to happen next.

(Simulation): The most basic use of model is to simulate the system's response to various input scenarios. This simply means that an input sequence $\{u_t^*\}_{t=1}^n$ is chosen by the user, and and is applied to the model $H(q^{-1})$ in order to obtain the undisturbed output

$$y_t^* = H(q^{-1})u_t^*, \ \forall t = 1, \ldots, n. \tag{3.49}$$

This is the output produced by the model when there are no external disturbances which need to be taken into account.

(Prediction): Given the input-signals and the past output signals recorded before instant t, as well as the model, what will the outcome be at instance t? That is, in this case we have some knowledge about the disturbances which acted in the past on the system, and hence for the disturbance terms in the model.

Let us now see how the latter can be formalized. We shall start by discussing how a future value of v_t can be predicted in case it is described as

$$v_t = H(q^{-1})e_t = \sum_{\tau=0}^{\infty} h_\tau e_{t-\tau}. \tag{3.50}$$

For this equation to be meaningful we assume that H is stable, that is

$$\sum_{\tau=0}^{\infty} |h_\tau| < \infty. \tag{3.51}$$

A crucial property of eq. (3.50) which we will impose is that it should be invertible, that is, if v_s is known for all $s \leq t$, then we should be able to compute e_t as

$$e_t = \tilde{H}(q^{-1})v_t = \sum_{\tau=0}^{\infty} \tilde{h}_\tau v_{t-\tau} \tag{3.52}$$

with

$$\sum_{\tau=0}^{\infty} |\tilde{h}_\tau| < \infty. \tag{3.53}$$

The filters H and $\tilde{H}$ are related as follows. Conider the polynomial in $z \in \mathbb{C}$ defined as

$$H(z) = \sum_{\tau=0}^{\infty} h_\tau z^{-\tau}, \tag{3.54}$$

and assume that the inverse function $\frac{1}{H(z)}$ is analytic in $|z| \geq 1$, or

$$\frac{1}{H(z)} = \sum_{\tau=0}^{\infty} \bar{h}_\tau z^{-\tau}. \tag{3.55}$$

Define then the filter $H^{-1}(q^{-1})$ as

$$H^{-1}(q^{-1}) = \sum_{\tau=0}^{\infty} \bar{h}_\tau q^{-\tau}. \tag{3.56}$$

Then $H^{-1}(q^{-1}) = \tilde{H}(q^{-1})$. The proof of this result needs quite a few subtle reasonings. However, this result is quite powerful as it indicates that the properties of the filter $H(q^{-1})$ are similar to those of the function $H(z)$. All that is needed is that the function $\frac{1}{H(z)}$ be analytic in $|z| \geq 1$. That is, it has no poles on or outside the unit circle. We could also phrase the condition as $H(z)$ must have zeros on or outside the unit circle.

Example 13 (Moving Average) *Suppose that one has for all $t = -\infty, \ldots, \infty$ that*

$$v_t = e_t + ce_{t-1}, \tag{3.57}$$

That is

$$H(q^{-1}) = 1 + cq^{-1}, \tag{3.58}$$

that is the process $\{v_t\}_t$ is a MA(1) process of order 1. Then

$$H(z) = 1 + cz^{-1} = \frac{z+c}{z}, \tag{3.59}$$

$$H^{-1}(z) = \frac{1}{1+cz^{-1}} = \sum_{\tau=0}^{\infty} (-c)^\tau z^{-\tau}, \tag{3.60}$$

where we use the geometric series expansion. Then e_t can be computed from $\{v_t\}_{s\leq t}$ as

$$e_t = \sum_{\tau=0}^{\infty} (-c)^\tau v_{t-\tau}. \tag{3.61}$$

Suppose now that we have observed only $\{v_t\}_{s<t}$, and that we want to predict the value of v_t based on these observations. We have then that, since $H(z)$ is assumed to be monic, that

$$v_t = \sum_{\tau=0}^{\infty} h_\tau e_{t-\tau} = e_t + \sum_{\tau=1}^{\infty} h_\tau e_{t-\tau}. \tag{3.62}$$

Now, knowledge of v_s implies knowledge of e_s for all $s < t$ as $v_t = H(q^{-1})e_t$ or

$$e_t = H^{-1}(q^{-1})v_t, \tag{3.63}$$

and thus a prediction of v_t is made as

$$\hat{v}_{t|t-1} = (H(q^{-1}) - 1)H^{-1}(q^{-1})v_t, = (1 - H^{-1}(q^{-1})v_t. \tag{3.64}$$

This is in a sense the best one can do in case $\{e_t\}_t$ contains no information information which contains information to predict next values. That is, the information of e_t is not predictable based on $\{e_s\}_{s<t}$, nor of linear combinations of those (as e.g. $\{v_t\}_{s<t}$). Such terms are denoted as *innovations.*

3.5 Identifiability Issues

System identification is concerned with finding appropriate models from experimental input-output behavior of the system. Conversely, an essential limitation of tools of system identification is that they cannot be used to recover properties of the system which are not reflected by its input-output behavior. The formal way to characterize this notion is by using the following definition

Definition 11 (Globally Identifiability at θ^* of a Model Structure) *Consider a class of models parameterized by a vector θ, or $\mathcal{H} = \{H(z,\theta), G(z,\theta) : \forall\theta\}$. Identification then tries to recover θ using observations of $\{(u_t, y_t)\}_t$. where $y_t = H(z,\theta)u_t + G(z,\theta)v_t$ for all t. Then the model class $\mathcal{H}$ is globally identifiable at θ^* if and only if*

$$\theta = \theta^* \Leftrightarrow H(z,\theta) = H(z,\theta^*), G(z,\theta) = G(z,\theta^*),\ \forall z \in \mathbb{C}. \tag{3.65}$$

This property comes also in a global flavor.

Definition 12 (Globally Identifiability of a Model Structure) *Consider a class of models parameterized by a vector θ, or $\mathcal{H} = \{H(z,\theta), G(z,\theta) : \forall\theta\}$. The class $\mathcal{H}$ is globally identifiable if and only it is globally identifiable at any possible value θ.*

Or in other words, no two different sets of parameter $\theta \neq \theta'$ can express models having the same LTI. Both notions are used as asymptotically, we are only interested at global identifiability at the actual parameter θ_0, a property which is obviously implied by global identifiability. The following counterexample is informative.

Example 14 (Sequence of LTIs) *Consider two LTI models given as*

$$\begin{cases} y_t = a_1 \frac{1}{1-a_2q^{-1}} u_t \\ y_t = b_1(1-b_2q^{-1})u_t, \end{cases} \tag{3.66}$$

where $a_2 \neq b_2$. Then samples $\{(u_t, y_t)\}_t$ originating from joining the two systems sequentially would obey the relation

$$y_t = (b_1a_1)\frac{1-b_2q^{-1}}{1-a_2q^{-1}}u_t, \tag{3.67}$$

and the outcome is identifiable up to the term (a_1b_1). That is, merely looking at the input-output behavior, one cannot recover which part of the overall gain is to the first subsystem, and which is due to the second one.

In case $a_2 = b_2$, the overall model behaves as

$$y_t = (b_1a_1)u_t. \tag{3.68}$$

Conversely, if one does not know the orders of the subsystems, and whether they do share common factors, then an input behavior $\{(u_t, y_t)\}_t$ described as eq. (3.67) could as well be caused by the systems

$$\begin{cases} y_t = a_1 \frac{1}{(1-a_2q^{-1})(1-c_1q^{-1})(1-c_2q^{-1})} u_t \\ y_t = b_1(1-b_2q^{-1})(1-c_1q^{-1})(1-c_2q^{-1})u_t, \end{cases} \tag{3.69}$$

for all $c_1, c_2 \in \mathbb{C}$ where $|c_1| < 1$ and $|c_2| < 1$, as far as we know.

Hence a sequence of two LTIs is only identifiable up to the gains of the subsystem and under the assumption that the models do not share canceling factors.

The following example gives a flash-forward of the difficulties we will discuss in the context of state-space models

Example 15 (Identifiability of State Space Systems) *Consider a sequence of data* $(u_1, y_1), (u_2, y_2), \ldots, (u_t, y_t$ *which obey the difference equations for all* $t = 1, 2, \ldots, t, \ldots$

$$\begin{cases} \mathbf{x}_t = \mathbf{A}\mathbf{x}_{t-1} + \mathbf{b}u_t \\ y_t = \mathbf{c}^T\mathbf{x}_t \end{cases} \tag{3.70}$$

with $\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t, \ldots$ *which are elements of* $\mathbb{R}^d$ *a sequence of (unknown)* state *vectors Then the input-output behavior is determined up to a linear transformation of the system matrix* $\mathbf{A} \in \mathbb{R}^{d\times d}$. *That is, let* $\mathbf{G} \in \mathbb{R}^{d\times d}$ *be a matrix which is full rank such that its inverse* $\mathbf{G}^{-1}$ *exists, then the sequence* $(u_1, y_1), (u_2, y_2), \ldots, (u_t, y_t), \ldots$ *obeys as well the equations*

$$\begin{cases} \tilde{\mathbf{x}}_t = \mathbf{G}\mathbf{A}\mathbf{G}^{-1}\tilde{\mathbf{x}}_{t-1} + \mathbf{G}\mathbf{b}u_t \\ y_t = \mathbf{c}^T\mathbf{G}^{-1}\tilde{\mathbf{x}}_t \end{cases} \tag{3.71}$$

where now $\tilde{\mathbf{x}}_t = \mathbf{G}\mathbf{x}_t$. *We say that a state-space system is only identifiable up to a (full rank) linear transformation.*

3.5.1 Persistency of Exitation

As seen in the previous chapter, in order for the LS estimate to give unique solutions, we need to have that the associated sample covariance matrix is full rank. The notion of Persistency of Excitation (PE) is an interpretation of this condition when the LS estimate is applied for estimating the parameters of a dynamical model based on input- and output behavior. Let us first illustrate this with the following example

Example 16 (PE for a FIR(d) Model) *Let us consider signals* $\{u_t\}_t$ *and* $\{y_t\}$ *of length* n, *and suppose their relation can adequately be captured using the following model structure*

$$y_t = \sum_{\tau=1}^{d} h_{0,\tau} u_{t-\tau} + e_t. \tag{3.72}$$

where $\theta_0 = (h_{0,1}, \ldots, h_{0,d})^T \in \mathbb{R}^d$ *are unknown. Stacking all* $n - d$ *such equalities yields the linear system*

$$\begin{bmatrix} u_1 & \ldots & u_d \\ u_2 & & u_{d+1} \\ & \vdots & \\ u_{n-d+1} & \ldots & u_n \end{bmatrix} \begin{bmatrix} h_{0,1} \\ \vdots \\ h_{0,d} \end{bmatrix} = \begin{bmatrix} y_{d+1} \\ \vdots \\ y_n \end{bmatrix} + \begin{bmatrix} e_{d+1} \\ \vdots \\ e_n \end{bmatrix}, \tag{3.73}$$

or shortly $\Phi\theta_0 = \mathbf{y} + \mathbf{e}$, *using appropriate definitions of the matrix* Φ *and the vectors* $\mathbf{y}, \mathbf{e}$. *Then the LS estimate of those parameters is given as* θ_n *which solves the system*

$$\begin{bmatrix} \hat{r}_{uu}(0) & \ldots & \hat{r}_{uu}(d-1) \\ & \ddots & \\ \hat{r}_{uu}(d-1) & \ldots & \hat{r}_{uu}(0) \end{bmatrix} \theta_n = \hat{\mathbf{R}}_d\theta_n = \begin{bmatrix} \hat{r}_{uy}(1) \\ \ldots \\ \hat{r}_{uy}(d) \end{bmatrix} = \hat{\mathbf{r}}_d, \tag{3.74}$$

where $\hat{r}_{uu}(\tau) = \frac{1}{n}\sum_{t=1}^{n-\tau} u_t u_{t+\tau}$ *and* $\hat{r}_{uy}(\tau) = \frac{1}{n}\sum_{t=1}^{n-\tau} u_t y_{t+\tau}$*, and where* $\hat{\mathbf{R}}_d \in \mathbb{R}^{d\times d}$ *and* $\hat{\mathbf{r}}_d \in \mathbb{R}^d$ *are defined appropriately. Then, this set of equations has a unique solution if and only if* $\hat{\mathbf{R}}_d$ *is of full rank, e.g. invertible. This requires in turn that the input signals are sufficiently rich: for example if* $d > 1$ *and* $u_t = 1$ *for all* $t = 1, \ldots, n$*, this condition is obviously not satisfied.*

This intuition leads to the general definition of PE:

Definition 13 (PE) *A signal* $\{u_t\}_t$ *of infinite length is called Persistently Exciting (PE) of order* d *in case the following two conditions are satisfied.*

(i): For any $\tau = 0, \ldots, d-1$*, the limit*

$$r_{uu}(\tau) = \lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n-\tau} u_t u_{t+\tau}, \tag{3.75}$$

exists.

(ii): The matrix

$$\mathbf{R}_d = \begin{bmatrix} r_{uu}(0) & \ldots & r_{uu}(d-1) \\ & \ddots & \\ r_{uu}(d-1) & \ldots & r_{uu}(0) \end{bmatrix} \succ 0, \tag{3.76}$$

i.e., this matrix is positive definite, and hence is of full rank.

Example 17 (White Noise) *Let* $\{U_t\}_t$ *be a sequence of zero mean white noise with variance* $\sigma^2 > 0$*, then* $r_u(\tau) = \mathbb{E}[U_t U_{t-\tau}] = \sigma^2 \delta_\tau$*, and the matrix* $\mathbf{R}_d = \sigma^2 I_d$*, and is thus of full rank for any* d*. That means that a sequence of white noise* $\{U_t\}_t$ *is PE of any order.*

Example 18 (Step) *Let* $\{u_t\}$ *be a step function where the step is made at* $t = 1$*, i.e.* $u_t = I(t \geq 1)$*. Hence for finite* n *the matrix* $\hat{\mathbf{R}}_d$ *is full rank for* $d \leq 2$*, but when considering the limit for* $n \to \infty$*, the initial difference at step* $t = 0, 1$ *is averaged out, and the matrix* $\mathbf{R}_d$ *is only full rank for* $d = 1$*. Hence the step function is PE of order 1.*

Example 19 (Finite Impulse) *Let* $\{u_t\}$ *be a finite impulse function, such that* $u_t = 1$ *only if* $t = 0$*, zero otherwise. Then this signal is not PE of any order as* $\mathbf{R}_0 = 0$*. Note that the situation is entirely different when* $u_1 = a_n$ *with* $a_n \to \infty$ *when* $n \to \infty$*, as then the excitation in the initial steps is not necessarily averaged out.*

The notion of PE is not restricted to the use of FIR models, but one has that a model of order d is identifiable in case the input signal $\{u_t\}_t$ is PE of order d when using the IV or PEM method as will be introduced in later chapters. However, if the model can be written as a FIR model of order d, PE of order d is already sufficient for identifiability (uniqueness) of the solution.

Also observe that the concept of PE is useful for noisy systems. If the system is noiseless, one can obtain identifiability using less restrictive conditions on the input signal. Specifically, one does not have to consider the limiting behavior. For example, if the given signals $\{u_t\}_t$ and $\{y_t\}_t$ obey a

FIR system without noise, i.e. $y_t = \sum_{\tau=1}^d h_{0,\tau} u_{t-d}$, the parameters can be recovered *exactly* from the system

$$\begin{bmatrix} u_1 & \dots & u_d \\ u_2 & & u_{d+1} \\ \vdots & & \vdots \\ u_d & \dots & u_{2d-1} \end{bmatrix} \begin{bmatrix} h_{0,1} \\ \vdots \\ h_{0,d} \end{bmatrix} = \Phi_d \theta_0 = \begin{bmatrix} y_{d+1} \\ \vdots \\ y_{2d} \end{bmatrix} \tag{3.77}$$

in case the matrix $\Phi_d \in \mathbb{R}^{d \times d}$ were full rank.

Let us now exemplify PE conditions derived for more complex ways used to generate signals. The proofs as in the SI book, albeit simple, are not reproduced here.

Example 20 (PE for d distinct frequencies) *Let $\{U_t\}$ be a (multivariate) ergodic stochastic process. Assume the its spectral density (matrix) is positive (definite) in at least d distinct frequencies, then $\{U_t\}$ is PE of at least order d.*

Example 21 (frequency of a PE signal) *Let $\{u_t\}_t$ be a signal which is PE of at least order d, then its spectral density is nonzero in at least d different frequencies.*

Example 22 (PE for filtered signals) *Let $\{u_t\}$ be a signal which is PE of order d. Let $H(q^{-1})$ be an asymptotically stable, linear filter with k zeros on the unit circle, then the filtered signal $\{y_t = H(q^{-1})u_t\}$ is PE of order m with $d - k \leq m \leq d$.*

Example 23 (PE and zero filtering) *Let $\{U_t\}_t$ be a stationary stochastic process which is PE of order at least d. Define*

$$Z_t = \sum_{\tau=1}^d h_\tau U_{t-\tau}. \tag{3.78}$$

Then the condition that $\mathbb{E}[Z_t Z_t] = 0$ implies (if and only if) that $h_1 = \dots = h_d = 0$.

3.5.2 Input Signals

It is hence clear that a successful SI experiment relies on a good input signal. We already mentioned that a stochastic white noise sequence has good properties w.r.t. PE. Again, we use capital letters to denote random quantities (i.e. depending on some sort of sampling mechanism), and lower case letters indicate Some other examples are given here.

Example 24 (A PRBS) *A* Pseudo Random Binary Sequence *(PRBS) is a signal that shifts between two levels (typically ±a) in a deterministic fashion. Typically, such signals are realized by using a circuit with shift registers such that the outcome 'looks similar to white stochastic noise'. However, the essential difference is that when computing a PRBS again on a different occasion, the signal will be exactly the same. The signal is necessarily periodic, that is, it repeats itself after a given period. In most practical cases however, the period would be chosen such that it exceeds the number of samples, such that no artifacts come up in the analysis due to such property. When applying a PRBS, the user has to design the two levels, the period as well as the clock period. The clock period is the minimal time the signal varies its level. Typically, the clock period is taken equal to one sampling interval.*

Example 25 (A PRBS of period m) *Let $\{u_t\}_t$ be a signal which is generated as a pseudo-random binary sequence (PRBS) with period m, and magnitude a, that is one has*

1. *The signal is deterministic (e.g. not depending on any sampling mechanism).*
2. *The magnitudes of the signal are a, such that $u_t = \pm a$.*
3. *The autocovariances are almost zero for $\tau = 1, \dots, m-1$, specifically one has that $\hat{\mathbf{R}}_m^{-1}$ equals*

$$\hat{\mathbf{R}}_m^{-1} = \begin{bmatrix} a^2 & \frac{-a^2}{m} & \dots & \frac{-a^2}{m} \\ \frac{-a^2}{m} & a^2 & \dots & \frac{-a^2}{m} \\ & & \ddots & \\ \frac{-a^2}{m} & \frac{-a^2}{m} & \dots & a^2 \end{bmatrix} \tag{3.79}$$

4. *The signal repeats itself exactly after m time instances.*

Then this signal is PE of order exactly m.

Example 26 (An ARMA PRBS Process) *Let $\{e_t\}$ be a PRBS. Then this process filtered by an ARMA model gives $\{u_t\}_t$ such that*

$$A(q^{-1})u_t = B(q^{-1})e_t. \tag{3.80}$$

such that one may tune the properties of the filter by design of appropriate A, B polynomials

Example 27 (An ARMA Process) *Let $\{D_t\}$ be a white, zero mean stochastic process. Then this process filtered by an ARMA model gives $\{U_t\}_t$ such that*

$$A(q^{-1})U_t = B(q^{-1})D_t. \tag{3.81}$$

such that one may tune the properties of the filter by design of appropriate A, B polynomials

Then we have the property that

Example 28 (PE for an ARMA process) *A stochastic process following a nontrivial ARMA system is PE of any order.*

Another example which is often used is

Example 29 (A Sum of Sinusoids) *The following deterministic signal $\{u_t\}_t$ is often used:*

$$u_t = \sum_{j=1}^{m} a_j \sin\left(\omega_j t + \varphi_j\right), \tag{3.82}$$

with amplitudes $\mathbf{a} = (a_1, \dots, a_m)^T$, frequencies $\omega = (\omega_1, \dots, \omega_m)^T$ and phases $\varphi = (\varphi_1, \dots, \varphi_m)^T$. A term with $\omega_j = 0$ will give a constant contribution $a_j \sin(\varphi_j)$, a term with $\omega_j = \pi$ gives a contribution which will oscillate in two sampling intervals, or

$$a_j \sin(\omega_j(t+1) + \varphi_j) = -a_j \sin(\omega_j(t+1) + \varphi_j). \tag{3.83}$$

for any t.

finally, it is often useful to design the input signals such that the resulting identified model is adequate w.r.t. a certain frequency range. In most cases, the input signal must emphasize the low-frequency properties during the modeling. There are different ways of obtaining such inputs, including

- Standard Filtering. This can be done by pre-filtering the input signals such that the resulting signal has the desirable property in frequency domain. An example is given during the computer labs.

- Increasing the clock period. If keeping the input signal constant over an increased amount of time exceeding the sampling interval, it must be clear that in that way one reduces the rapid fluctuations (high-frequencies) present in the original signal. This reasoning makes it clear that if the given sampling interval of the case at hand is relatively large, there is not so much hope to recover the dynamics of the system corresponding to the high-frequencies.

- Decreasing the probability of changing level. Consider the case of a binary *Stochastic sequence* $\{U_t\}_t$ taking values in $\{-a, a\}$ for $a > 0$, which has the stochastic model for $0 < p \leq 1$ as

  $$U_t = \begin{cases} -U_{t-1} & \text{with probability} p \\ U_{t-1} & \text{else.} \end{cases} \tag{3.84}$$

  and $U_0 = a$. Then by increasing p, the signal reflects more rapid fluctuations. By decreasing p, the signal has a larger power in the low-frequency area.

Chapter 4

Nonparametric Techniques

"Which experiments reveal structural properties of the studied system?"

Now let us look at a converse problem. Here we do not look at the properties of the assumed model class, but we compute such properties based on experiments carried out on the studied system. In general, such methods are not tied to a specific (parameterized) model, but nevertheless embody a description of the system. Such method have the denominator 'non-parametric' or 'distribution-free'. They often come in the forms graphs, curves, tables or other intuitive representations, and give as such structural information of the system. Their use is often found in

(Preprocess) Indicate important effects present in the studied system.

(Model class) Suggest a suitable class of parametric models which can be used to capture such effects.

(Validate) Check whether the identified model behaves similarly than the actual system.

4.1 Transient Analysis

A first approach is to inject the studied system with a simple input as a pulse or a step, and to record the subsequent output of the system. This gives then an impression of the impulse response of the studied system. Let us look further into the pros and cons of this strategy. Formally, let the following input signal $\{u_t\}_t$ be injected to the system H

$$u_t = \begin{cases} K & t = 0 \\ 0 & \text{else.} \end{cases} \tag{4.1}$$

Then, if the system could be described exactly (i.e. without any effect of unknown disturbances) as $H(q^{-1}) = h_0 + \dots + h_\tau q^{-\tau} + \dots$, then the output of the system becomes

$$y_t = H(q)u_t = K \begin{cases} h_t & t \geq 0 \\ 0 & \text{else.} \end{cases} \tag{4.2}$$

So conversely, if one knows that the system follows very closely an LTI description $H(q^{-1}) = h_0 + \dots + h_\tau q^{-\tau} + \dots$, the different unknowns $\{h_\tau\}_\tau$ can be observed directly when injecting the studied system with a pulse signal as in eq. (4.1). The pros of this approach are that (i) it is simple to understand or to (ii) implement, while the model need not be specified further except for the LTI property. The downsides are of course that (i) this method breaks down when the LTI model fits not exactly the studied system. Since models serve merely as mathematical convenient approximations of the actual system, this is why this approach is in practice not often used. (ii) It cannot handle random effects very well. (iii) such experiment is not feasible in the practical setting at hand. As for this reason it is merely useful in practice to determine some structural properties of the system. For example consider again the first order system as in the previous example, then a graph of the impulse response indicates the applicable time-constants and gain of the system.

Similarly, consider the step input signal $\{u_t\}_t$ defined as

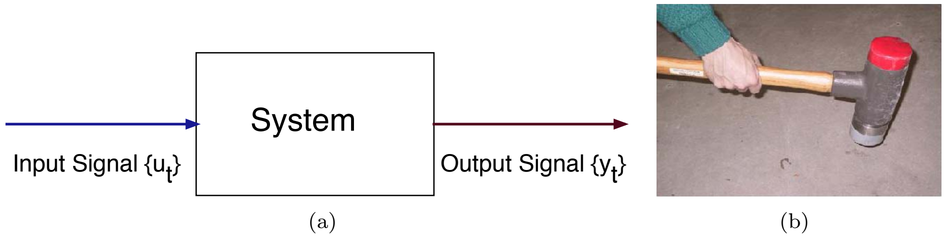
$$u_t = \begin{cases} K & t \geq 0 \\ 0 & \text{else.} \end{cases} \tag{4.3}$$

Then, if the system could be described exactly (i.e. without any effect of unknown disturbances) as $H(q^{-1}) = h_0 + \cdots + h_\tau q^{-\tau} + \dots$, then the output of the system becomes

$$y_t = H(q^{-1})u_t = K \begin{cases} \sum_{\tau=0}^{t} h_\tau & t \geq 0 \\ 0 & \text{else.} \end{cases} \tag{4.4}$$

or equivalently

$$y_t - y_{t-1} = K \begin{cases} h_\tau & t \geq 1 \\ 0 & \text{else.} \end{cases} \tag{4.5}$$



(a)

(b)

Figure 4.1: (a) A Block Representation of a system. (b) An impact hammer used for modal analysis of bridges and other large constructions.

4.2 Frequency Analysis

As seen in the previous chapter, an LTI is often characterized in terms of its reaction to signals with a certain frequency and phase. It is hence only natural to try to learn some properties of the studied system by injecting it with a signal having such a form. Specifically, let $\{u_t\}$ be defined for $t = \dots, -1, 0, 1, \dots$ as

$$u_t = a \sin(\omega t). \tag{4.6}$$

where $a > 0$ is the gain of the signal. Then as seen in the previous chapter the corresponding output of the system H is given as

$$y_t = K \sin(\omega t + \phi), \tag{4.7}$$

where

$$\begin{cases} K = a|H(e^{i\omega})| \\ \phi = \arg G(e^{i\omega}). \end{cases} \tag{4.8}$$

Note that normally the phase ϕ will be negative. By measuring the amplitude a, K and the phase ϕ for for given ω, one can find the complex variable $H(e^{i\omega})$ from (4.8). If repeating this procedure for a range of frequencies ω, one can obtain a graphical representation of $H(e^{i\omega})$. Such Bode plots (or Nyquist or related plots) are well suited for the design and analysis of automatic control systems. The procedure described above is rather sensitive to disturbances. This is not difficult to understand. If one has disturbance terms with Laplace transform $E(s)$, one gets

$$Y(s) = H(s)U(s) + E(s). \tag{4.9}$$

Then when injecting the system with a signal $\{u_t\}$ as in eq. (4.6) one gets the output signal $\{y_t\}$ where

$$y_t = K\sin(\omega t + \phi) + e_t, \tag{4.10}$$

and due to the presence of noise it will be difficult to extract good estimates of K and ϕ from those signals.

4.3 A Correlation Analysis

The above ideas are taken a step further into a correlation analysis. But instead of using simple input signals, the system is injected with a random signal $\{u_t\}_t$ which has zero mean or

$$\mathbb{E}[u_t] = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} u_t, \tag{4.11}$$

which has finite values. A formal definition of such white noise sequence is given in Chapter 4, but for now it is sufficient to let the expectation $\mathbb{E}[\cdot]$ denote an limit of an average, or $\mathbb{E}[u_t] = \lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n} u_t$. Then the output which is recorded at the output of the system

$$y_t = \sum_{\tau=0}^{\infty} h_\tau u_{t-\tau}. \tag{4.12}$$

When taking multiplying both sides by $u_{t'}$ for any t', and taking expectations one gets

$$\mathbb{E}[u_{t'}y_t] = \mathbb{E}\left[\sum_{\tau=0}^{\infty} h_\tau u_{t-\tau}u_{t'}\right] = \sum_{\tau=0}^{\infty} h_\tau \mathbb{E}\left[u_{t-\tau}u_{t'}\right]. \tag{4.13}$$

Summarizing this for all t, t' and canceling the cross-terms gives the linear system

$$\begin{bmatrix} r_{uy}(0) \\ r_{uy}(1) \\ \vdots \\ r_{uy}(\tau) \\ \vdots \end{bmatrix} = \begin{bmatrix} r_{uu}(0) & r_{uu}(1) & r_{uu}(2) & \dots & r_{uu}(\tau) & \dots \\ r_{uu}(1) & r_{uu}(0) & & & r_{uu}(\tau-1) & \dots \\ r_{uu}(2) & & & & & \\ \vdots & & & \ddots & & \\ r_{uu}(\tau) & r_{uu}(\tau-1) & & & r_{uu}(0) & \dots \\ \vdots & & & & & \end{bmatrix} \begin{bmatrix} h_0 \\ h_1 \\ \vdots \\ h_\tau \\ \vdots \end{bmatrix} \tag{4.14}$$

where

$$r_{uu}(\tau) = \mathbb{E}[u_t u_{t-\tau}] = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} u_t u_{t-\tau}. \tag{4.15}$$

and

$$r_{uy}(\tau) = \mathbb{E}[y_t u_{t-\tau}] = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} y_t u_{t-\tau}. \tag{4.16}$$

Since this limit cannot be computed explicitly in practice, one settles for working with the estimates

$$\hat{r}_{uu}(\tau) = \frac{1}{n} \sum_{i=1}^{n} u_t u_{t-\tau}. \tag{4.17}$$

and

$$\hat{r}_{uy}(\tau) = \frac{1}{n} \sum_{i=1}^{n} y_t u_{t-\tau}. \tag{4.18}$$

Secondly, rather than solving the infinite system (4.19), one solves the corresponding finite linear system for appropriate $m > 0$ given as

$$\begin{bmatrix} \hat{r}_{uy}(0) \\ \hat{r}_{uy}(1) \\ \vdots \\ \hat{r}_{uy}(m-1) \end{bmatrix} = \begin{bmatrix} \hat{r}_{uu}(0) & r_{uu}(1) & \hat{r}_{uu}(2) & \dots & \hat{r}_{uu}(m-1) \\ \hat{r}_{uu}(1) & r_{uu}(0) & & & \hat{r}_{uu}(m-2) \\ \hat{r}_{uu}(2) & & & & \\ \vdots & & & \ddots & \\ \hat{r}_{uu}(m-1) & r_{uu}(m-2) & & & \hat{r}_{uu}(0) \end{bmatrix} \begin{bmatrix} h_0 \\ h_1 \\ \vdots \\ h_M \end{bmatrix} \tag{4.19}$$

in order to get an idea about $\{h_\tau\}_{\tau=0}^{n}$. Those equations are known as Wiener-Hopf type of equations. This technique is related to the Least Squares estimate and the Prediction Error Method in Chapter 5.

4.4 Spectral Analysis

Now both the correlation technique and the frequency analysis method can be combined into a signal nonparametric approach as follows. The idea is to take the Discrete Fourier Transforms (DFT) of the involved signals, and find the transfer function relating them.

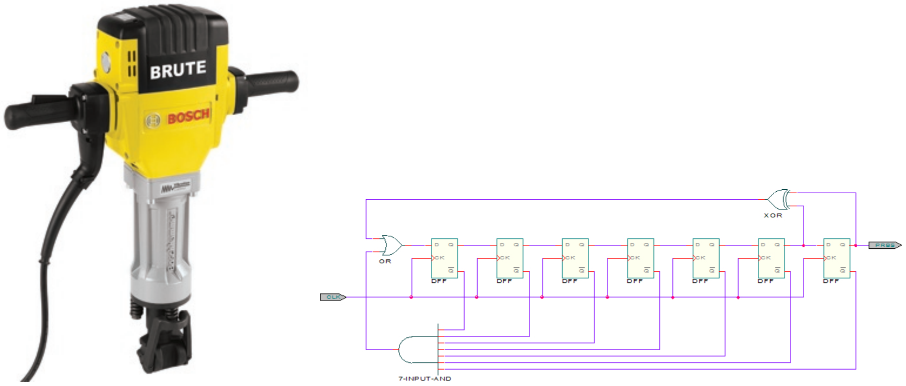
$$\begin{cases} \phi_{uu}(\omega) = \frac{1}{2\pi} \sum_{\tau=-\infty}^{\infty} r_{uu}(\tau) e^{-i\omega\tau} \\ \phi_{uy}(\omega) = \frac{1}{2\pi} \sum_{\tau=-\infty}^{\infty} r_{uy}(\tau) e^{-i\omega\tau}. \end{cases} \tag{4.20}$$

Then the previous chapter learns us that we have that for all ω one has

$$\phi_{uy}(\omega) = H(e^{-i\omega}) \phi_{uu}(\omega), \tag{4.21}$$

where

$$H(e^{-i\omega}) = \sum_{\tau=0}^{\infty} h_\tau e^{-i\tau\omega}. \tag{4.22}$$



(a) (b)

Figure 4.2: (a) A vibration Hammer (b) A circuit scheme for realizing a Pseudo Random Binary Sequence.

Consequently, a reasonable estimate for H would be

$$\hat{H}(e^{-i\omega}) = \frac{\hat{\phi}_{uy}(\omega)}{\hat{\phi}_{uu}(\omega)}, \tag{4.23}$$

where $\hat{\phi}$ are reasonable estimates of ϕ. A straightforward estimate would be

$$\hat{\phi}_{uy} = \frac{1}{2\pi} \sum_{\tau=-n}^{n} \hat{r}_{uy} e^{-i\tau\omega}, \tag{4.24}$$

and similarly for $\hat{\phi}_{uu}$. Working out $\hat{r}_{uy}$ gives

$$\hat{\phi}_{uy} = \frac{1}{2\pi n} \sum_{\tau=-n}^{n} \sum_{t=1-\min(\tau,0)}^{\max(\tau,0)} y_{t+\tau} u_t e^{-i\tau\omega}. \tag{4.25}$$

Then change of indices τ and t gives

$$\hat{\phi}_{uy} = \frac{1}{2\pi n} \sum_{s=1}^{n} \sum_{t=1}^{n} y_s u_t e^{-i(s-t)\omega} = \frac{1}{2\pi n} Y_n(\omega) U_n(-\omega), \tag{4.26}$$

where

$$\begin{cases} U_n(\omega) = \sum_{s=1}^{n} u_s e^{-is\omega} \\ Y_n(\omega) = \sum_{s=1}^{n} y_s e^{-is\omega}. \end{cases} \tag{4.27}$$

Those are the Discrete Fourier Transforms of the signals $\{u_t\}$ and $\{y_t\}$ (padded with zeros). For $\omega = 0, \frac{2\pi}{n}, \frac{4\pi}{n}, \ldots, \pi$ those can be estimated efficiently using the Fast Fourier Transform (FFT) algorithms. In a similar fashion one has

$$\hat{\phi}_{uu} = \frac{1}{2\pi n} U_n(\omega) U_n(-\omega) = \frac{1}{2\pi n} |U_n(\omega)|. \tag{4.28}$$

This estimate is called the *periodogram.* From the derivations above it follows that

$$\hat{H}(e^{-i\omega}) = \frac{Y_n(\omega)}{U_n(\omega)}. \tag{4.29}$$

This estimate is sometimes called the *empirical transfer function* estimate.

However the above estimate to the spectral densities and the transfer function will give poor results. For example, if u_t is a stochastic process, then the estimates eq. (4.28) and (4.26) do not converge in (the mean square sense) to the true spectrum as n, the number of datapoints tends to infinity. In particular, the estimate $\hat{\phi}_{uu}$ will on average behave as ϕ_{uu}, but its variance does not tend to zero as $n \to \infty$. One of the reasons for this behavior is that $\phi\phi_{uu}(\tau)$ will be quite inaccurate for large values for τ, but all covariance elements $\hat{r}_{uy}(\tau)$ are given the same weight in eq. (4.26) regardless of their accuracy. Another more subtle reason goes as follows. In eq. (4.26) $2n+1$ terms are summed. Even if the estimation error of each term goes to zero, there is no guarantee that the global sum goes to zero. These problems may be overcome if the terms of eq. (4.26) corresponding with large τ are weighted out. Thus, instead of eq. (4.26) the following improed estimate of the cross-spectrum can be used

$$\hat{\phi}'_{uy} = \frac{1}{2\pi} \sum_{\tau=-n}^{n} \hat{r}_{uy}(\tau) w(|\tau|) e^{-i\tau\omega}, \tag{4.30}$$

where $w : \mathbb{R} \to \mathbb{R}_+$ is a socalled *lag window.* It should $w(0) = 1$, and decreasing. Several forms of the lag window have been proposed in the literature. Some simple lag windows are presented in the following example.

Example 30 (Lag Windows) *The following lag windows are often used in the literature.*

- *Rectangular window:*

$$w_1(|\tau|) = \begin{cases} 1 & |\tau| \leq M \\ 0 & |\tau| > M \end{cases} \tag{4.31}$$

- *Bartlett window:*

$$w_2(|\tau|) = \begin{cases} 1 - \frac{|\tau|}{M} & |\tau| \leq M \\ 0 & |\tau| > M \end{cases} \tag{4.32}$$

- *Hamming and Tukey*

$$w_3(|\tau|) = \begin{cases} \frac{1}{2}(1 + \cos\frac{\pi\tau}{M}) & |\tau| \leq M \\ 0 & |\tau| > M. \end{cases} \tag{4.33}$$

Note that all the windows vanish for $|\tau| > M$. If the parameters M is chosen to be sufficiently large, the periodogram will not be smoothed very much. On the other hand a small M may mean that essential parts of the spectrum are smoothed out.It is not trivial to choose the parameter M. Roughly speaking M should be chosen according to trading off the following two objectives:

- *M should be small compared to n:*
- *$|\hat{r}_{uy}(\tau)| \ll \hat{r}_{uu}(0)$ for $\tau \geq M$ so as not to smooth out the parts of interest in the true spectrum.*

The use of a lag window is necessary to obtain a reasonable accuracy. On the other hand, sharp peaks in the spectrum might be smeared out. It may therefore not be possible to separate adjacent peaks. Thus the use of a lag window will give a limited frequency resolution. The effect of a lag window is illustrated in the following example.

4.5 Nonparameteric Techniques for Timeseries

Let us study a similar technique for estimating the transfer function of a timeseries. Here, we do not have an input sequence available, but we assume that the observed sequence $\{y_t\}_t$ is driven by unobserved white noise $\{e_t\}$. The previous chapter enumerates some common models for such systems. Here we will give a main nonparametric technique useful for recovering the underlying structure. Again such approaches are based on working with the covariances observed in the system.

4.5.1 Yule-Walker Correlation Analysis

Let's consider the equivalent of the correlation approach when timeseries are concerned. At first, assume the studied timeseries follows an AR(m) model as

$$y_t - a_1 y_{t-1} - \cdots - a_m y_{t-m} = e_t, \tag{4.34}$$

where $\{e_t\}$ is zero mean white noise such that

$$\begin{cases} \mathbb{E}[e_t] = 0 \\ \mathbb{E}[e_t e_{t-\tau}] = \sigma^2 \delta_\tau \end{cases} \tag{4.35}$$

with δ_τ equal to one if $\tau = 0$, and equals zero otherwise. Note that then $\{y_t\}$ can be seen as a linear combination of past values of the noise $\{e_s\}_{s \leq t}$. B multiplication of both sides of eq. (4.34) with a (delayed) value of the process $y_{t-\tau}$ for all $\tau = 0, 1, 2, \ldots$, and taking the expectation one gets

$$\mathbb{E}[y_{t-\tau}(y_t - a_1 y_{t-1} - \cdots - a_m y_{t-m})] = \mathbb{E}[y_{t-\tau} e_t], \tag{4.36}$$

or

$$r_y(\tau) - a_1 r_y(\tau - 1) - \cdots - a_m r_y \tau - m = \begin{cases} \sigma^2 & \text{if } \tau = 0 \\ 0 & \text{otherwise,} \end{cases} \tag{4.37}$$

where we defined as before

$$r_y(\tau) = \mathbb{E}[y_t y_{t-\tau}] = \lim_{n \to \infty} \frac{1}{n} \sum_{t=\tau+1}^{n} y_t y_{t-\tau}. \tag{4.38}$$

Assuming those are given for all $\tau = 0, 1, 2, \ldots$, those can be organized as a system of linear equations as follows

$$\begin{bmatrix} \sigma^2 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} r_y(0) & r_y(1) & r_y(2) & \ldots & r_y(M) & \ldots \\ r_y(1) & r_y(0) & & & r_y(M-1) & \ldots \\ r_y(2) & & & & & \\ & & & & & \\ \vdots & & & \ddots & & \\ & & & & & \\ r_y(M) & r_y(M-1) & & & r_y(0) & \ldots \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_M \end{bmatrix} \tag{4.39}$$

Those are known as the Yule-Walker (YW) equations. Conversely, if one knows the parameters $\{a_1, \dots, a_{m_a}\}$, the covariances $\{r_y(\tau)\}$ are given as solutions to the system

$$\begin{bmatrix} 1 & a_1 & a_2 & \dots & a_{m_a} \\ a_1 & 1+a_2 & a_3 & & 0 \\ \vdots & & \ddots & & \vdots \\ a_{m_a} & 0 & \dots & & 1 \end{bmatrix} \begin{bmatrix} r_y(0) \\ r_y(1) \\ \vdots \\ r_y(m_a) \end{bmatrix} = \begin{bmatrix} \sigma^2 \\ 0 \\ \vdots \\ 0 \end{bmatrix}. \tag{4.40}$$

Next, assume that the timeseries follows an ARMA(m_a, m_c) model given as

$$y_t - a_1 y_{t-1} - \dots - a_{m_a} y_{t-m_a} = e_t + c_1 e_{t-1} + \dots + c_{m_c} e_{t-m_c}, \tag{4.41}$$

and where $\{e_t\}$ is a white zero mean process satisfying the conditions (4.35). Then, again multiplying both sides of eq. (4.41) with $y_{t-\tau}$ and taking the expectation gives the relations

$$r_y(\tau) - a_1 r_y(\tau-1) - \dots - a_{m_a} r_y \tau - m_a = r_{ey}(\tau) + c_1 r_{ey}(\tau-1) + \dots + r_{ey}(\tau - m_c). \tag{4.42}$$

The cross-correlations $r_{ey}(\tau)$ are found as follows. Multiply both sides of eq.(4.41) with $e_{t-\tau}$ and take expectations, then we have that

$$r_{ey}(\tau) - a_1 r_{ey}(\tau-1) - \dots - a_{m_a} r_y \tau - m_a = \sigma^2 \left(c_0 \delta_\tau + \dots + c_{m_c} \delta_{\tau - m_c}\right), \tag{4.43}$$

where

$$r_{ey}(\tau) = \mathbb{E}[e_{t+\tau} y_t] = \lim_{n \to} \frac{1}{n} \sum_{t=1}^{n-\tau} e_{t+\tau} y_t. \tag{4.44}$$

As y_t is a linear combination of $\{e_s\}_{s \leq t}$, we have that $r_{ey}(\tau) = 0$ for $\tau > 0$. It as such follows that for all $\tau > m_c$ one has

$$r_y(\tau) - a_1 r_y(\tau - 1) - \dots - a_{m_a} r_y(\tau - m_a) = 0. \tag{4.45}$$

Note that those equations involve only the AR parameters of the ARMA process.

4.5.2 Spectral Factorization

In order to get an initial estimate of the MA parameters the following technique is often used. Rather than setting up the YW equations, one extracts the MA part from the covariance structure using the following Lemma.

Lemma 5 (Spectral Factorization) *Let $\phi : \mathbb{C} \to \mathbb{C}$ be a spectrum that can be written for all $z \in \mathbb{C}$ as*

$$\phi(z) = \frac{\sum_{k=-m_\beta}^{m_\beta} \beta_k z^k}{\sum_{k=-m_\alpha}^{m_\alpha} \alpha_k z^k}, \tag{4.46}$$

for $\beta_{-m_\beta}, \dots, \beta_{m_\beta}, \alpha_{-m_\alpha}, \dots, \alpha_{m_\alpha} \in \mathbb{R}$. Then there exists two functions

$$\begin{cases} A(z) = 1 + a_1 z + \dots + a_{m_\alpha} z^{m_\alpha} \\ C(z) = 1 + c_1 z + \dots + c_{m_\beta} z^{m_\beta}, \end{cases} \tag{4.47}$$

with $a_1, \dots, a_{m_\alpha}, c_1, \dots, c_{m_\beta} \in \mathbb{C}$, and a constant $\sigma > 0$ such that

1. *$A(z)$ has all zeros inside the unit circle.*
2. *$C(z)$ has all zeros inside or on the unit circle.*
3. *one has for all $z \in \mathbb{C}$ that*
$$\phi(z) = \sigma^2 \frac{C(z)}{A(z)} \frac{C^*(z^{-*})}{A^*(z^{-*})}, \tag{4.48}$$
where $A^*(z) = 1 + a_1^* z + \cdots + a_{m_\alpha}^*$ and $C^*(z^{-*}) = 1 + c_1^* z + \cdots + c_{m_\beta}^*$ and with z^* denoting the conjugate of $z \in \mathbb{C}$.

The proof of this result hinges on complex algebra, see e.g. [5].

Example 31 (MA Process) *Suppose one has given a MA(m) process such that*

$$y_t = e_t + c_1 e_{t-1} + \cdots + c_m e_{t-m}, \tag{4.49}$$

and where $\{e_t\}_t$ is zero mean, white noise with standard deviation σ. Then the covariances of this process are given as
$$r_y(\tau) = \begin{cases} \sigma^2 c_{|\tau|}^2 & |\tau| \leq m \\ 0 & \textit{elsewhere.} \end{cases} \tag{4.50}$$

The spectral density is then given for any $\omega \in]-\pi, \pi]$ as

$$\phi_y(\omega) = \frac{1}{2\pi} \sum_\tau r_y(\tau) e^{-i\omega\tau} = \frac{1}{2\pi} \left(r_y(0) + \sum_{k=1}^m c_k^2 (e^{-i\omega k} + e^{i\omega k}) \right)$$
$$= \frac{1}{2\pi} \left(r_y(0) + \sum_{k=1}^m c_k^2 \cos(\omega k) \right). \tag{4.51}$$

Chapter 5

Stochastic Setup

Niels Bohr, 1986 - as reply to a visitor to his home in Tisvilde who asked him if he really believed a horseshoe above his door brought him luck: "Of course not ... but I am told it works even if you don't believe in it."

The framework of stochastic models is often useful for implementing the following two philosophies:

(Analysis): The primary use of a stochastic framework is to assume that the experiments involved in a certain estimation task follow a proper stochastic rule set. In this way one can abstract away much of the technical irregularities while making life much easier for the analysis of the techniques. The price one has to pay in general for this convenience is that the results 'only' hold 'almost surely', that is, there is an extremely small chance that results go bogus. (Computer scientists like to use the phrase 'with overwhelming probability').

(Constructive): Recent work has shown that the device of randomization is useful in the design of algorithms. It turns out that this way one can push the boundaries of feasible computation tasks much further theoretically (w.r.t. computational complexity) as well as practically (w.r.t. large-scale computation tasks).

The predominant setup in the analysis of estimation, identification or filtering techniques is that where the involved signals are considered (partially) stochastic. Intuitively this means that the signals itself can be unspecified (to a certain degree), but that the *mechanism generating the signals* is fixed. In practice, stochastic properties manifest themselves as follows: when performing a *stochastic* experiment twice under exactly the same conditions, results could possibly differ. If performing the same experiment twice and results would always be equal, we say that the experiment were *deterministic*.

While I assume the reader experienced already an introductory class in probability theory or statistics, we will spend some effort in reviewing the basics once more. Not only for the sake of expressing results unambiguously, but also in order to pinpoint the power and limitations of surveyed results.

5.1 Getting the Basics Right

5.1.1 Events, Random variables and Derived Concepts

The following definitions establish a proper setup, as introduced by N. Kolmogorov in the 30s. Abstracting from applications, probability theory studies an experiment with a number of possible outcomes $\{\omega\}$. The totality of such outcomes is the sample space $\Omega = \{\omega\}$. An *event* - say A is a subset of the sample space. A *probability measure* P is a function from an event to a number between 0 and 1, or $P : \{\Omega\} \to [0, 1]$, with properties:

1. $0 \leq P(A) \leq 1$
2. $P(\Omega) = 1$
3. Let $\{A_i\}_i$ be any (countable many) set of disjunct events, then $\sum_i P(A_i) = P(\cup_i A_i)$.

Not all possible subsets of Ω need to be events, but the universe of events must form a sigma-field: 'if A is an event, so is $\Omega\backslash A$' and 'the union of any countable number of events must be an event', and 'Ω is an event'. Let's give some examples.

Example 32
- *Sample ω = images on web. A corresponding sample space Ω contains all images present on the web. An event A is e.g. 'all the images in Ω which are black and white' (informally, an image $\omega \in \Omega$ is black-and-white iff $\omega \in A$.)*
- *Sample ω = speech signals. A corresponding sample space Ω is the collection of all possible speech signals. An event A is e.g. the subset of speech signals only containing background noise. (informally, a speech signal ω contains only background noise iff $\omega \in A$.)*
- *Sample ω = weather in Uppsala. A corresponding sample space Ω is the collection of all possible weather regimes in Uppsala. An event A here is e.g. those cases where the weather is called sunny. (informally, a weather regime ω is called sunny iff $\omega \in A$.)*
- *Sample ω = external force on a petrochemical plant. A corresponding sample space Ω is the collection of all possible external forces which could act on the studied plant. An event A here is e.g. the collections of all those external forces which may drive the plant to an unstable working. (informally, an external force ω results in unstable working iff $\omega \in A$.)*

There are a number of derived concepts which we merely summarize:

(Joint): Let $A, B \subset \Omega$ be two events, then the joint probaility is defined as

$$P(A, B) \triangleq P(A \cup B). \tag{5.1}$$

(Independence): Let $A, B \subset \Omega$ be two events, then they are called mutually independent if

$$P(A, B) \triangleq P(A)P(B). \tag{5.2}$$

(Conditional): Let $A, B \subset \Omega$ be two events where $B \neq \{\}$, then the conditional probability is defined as

$$P(A|B) \triangleq \frac{P(A, B)}{P(B)}. \tag{5.3}$$

(Bayes): Let $A, B \subset \Omega$ be two events, then Bayes' law says that

$$P(A|B)P(B) = P(B|A)P(A) = P(A, B). \tag{5.4}$$

Often, we are interested in quantities associated with the outcome of an experiment. Such quantity is denoted as a *random variable*. Formally, a random variable is a function defined for any possible $\omega \in \Omega$. If the random variable is evaluated at the sample ω which actually occurred (the observation), we refer to it as a *realization* of this random variable. This quantity is what we intend with a *value* of a random variable. Following the convention in statistical literature, we denote a random variable as a capital letter. This notational convention makes it easier to discriminate between random variables and deterministic quantities (denoted using lower case letter). This motivates the use of the following notational convention:

$$P(X = x) \triangleq P\left(\{\omega | X(\omega) = x\}\right). \tag{5.5}$$

where $\{\omega | X(\omega) = x\}$ is the set of all samples ω which has a random value $X(\omega)$ equal to x. We have as before that $P : \{\Omega\} \to [0, 1]$, and as such $P\left(\{\omega | X(\omega) = x\}\right)$ gives a number between 0 and 1. Likewise, $P(X > 0)$ means that $P\left(\{\omega | X(\omega) > 0\}\right)$ etc. If X denotes a random variable defined over the outcome space Ω, then $X(\omega)$ denotes a realization measured when ω is sampled from Ω. Sometimes, X can only take a finite number of values, and X is as such called discrete. If not so, X is called a continuous random variable.

Example 33 *The following example illustrates ideas using a simple urn model.*

1. *Consider an urn containing $m = 10$ balls, one ball labeled '2', three balls labeled '1', and 6 of them labeled '0'. The set of all 10 balls is called the 'sampling space' Ω.*

2. *Randomness samples a ball in Ω denoted as ω. This sampling is essentially uniform, any sample comes up equally probable.*

3. *'The subset of balls with label 0' or informally 'A ball with label '0' is drawn', is an event.*

4. *Then the label of this 'random' ball - denoted as the function Z - is a random variable. The actual value $Z(\omega)$ is called a realization of this random variable.*

5. Before *the actual sampling, one could expect a value Z of $\frac{1}{10}(6*0+3*1+1*2) = 0.5$ denoted as $\mathbb{E}[Z] = 0.5$.*

6. *If repeating the experiment $n \to \infty$ times independently, one would end up with the ball labeled '2' in a fraction of $\frac{1}{10}$ of the times. This is captured by the law of large numbers.*

At this elementary level, we make already important conceptual steps:

- The sample space describes the physical reality.

- A random variable is a *mapping* of a sample to its corresponding label.

- 'Randomness' picks any sample with equal probability, while the probability of the corresponding labels is governed by the frequency of the samples with identical labels. This means that the law of probability corresponding to Z is implied by the definition of the random variable, not in the way randomness were implemented!

- Expectations are evaluated *before* the actual experiment is carried out. When doing the calculations when knowledge exists on which ω actually occurred in reality (the observation), the notion of probability is contaminated! In general, a statisticians job is finished right before the actual experiment is implemented (except for the consultancy part).

5.1.2 Continuous Random Variables

However, the above setup does not entirely characterize the intuitive concepts that we were after: a stochastic setup is adopted in order to characterize the mechanism generating the data. This probability function P is however not suited to explain the likelihood of a single sample, but focusses on sets and subsets of events. This subtle difference leads easily to a paradox, as seen in the following example. Consider an event-space such that an infinite number of events may occur. For example, consider the events of all possible 'weathers' in Uppsala: an infinite number of variations can occur, and assume (for the sake of the argument) that any 'weather' is equally probably to occur at an instance. Lets represent the weather which actually occurred as ω. Then $P(\omega) = 0$ necessarily, and the probability of this event equals zero. So it seems that this precise sample (the observation) was not possible to occur after all! This paradox arises as working with infinite sample spaces is not as straightforward as in the discrete case, and a proper notion of 'the probability of a single event' needs an additional apparatus as shown in the following subsection.

In case the sample space Ω contains an (uncountable) infinite number of elements, the above framework needs to be extended slightly in order to deal properly with measurability issues. Let us first look towards the case where a random value X defined over such a sampling space takes values in $\mathbb{R}$.

Definition 14 (CDF and PDF) *The laws of probability associated to a continuous, univariate random variable go as follows:*

(CDF): *The* Cumulative Distribution Function $F : \mathbb{R} \to [0, 1]$ *(CDF) of a univariate random variable* $X : \Omega \to \mathbb{R}$ *is defined as*

$$F(x) \triangleq P(X \leq x) \triangleq P\left(\{\omega | X(\omega) \leq x\}\right). \tag{5.6}$$

Consequently, one has that $F(-\infty) = 0$, $F(\infty) = 1$ *and the function* F *is monotonically increasing. An example is given in Fig. (5.1.a)*

(PDF): *The* Probability Density Function $f : \mathbb{R} \to \mathbb{R}_+$ *(PDF) of a univariate random variable* $X : \Omega \to \mathbb{R}$ *with a differential CDF* F *is defined as*

$$f(x) \triangleq \frac{\partial P(X \geq x)}{\partial x} = \frac{\partial F(x)}{\partial x}. \tag{5.7}$$

Those definitions are not mere academical, but clarify for example that a density function does not equal a probability law. Both notions lead also to different tools to estimate the probability laws underlying data.

(HIST): Given a sample of n samples taking values in $\mathbb{R}$, or $\{y_i\}_{i=1}^n \subset \mathbb{R}$, the histogram counts the frequency (normalized number) of samples occurring in a given interval (bin) of $\mathbb{R}$. For example, if we have 5 samples $\{1, 2, 3, 4, 5\}$, and two intervals (bins) $(-\infty, 3]$ and $(3, \infty)$, then the histogram would say $(3/5, 2/5)$. This is then an estimate of the PDF. A graphical example is given in Fig. (5.2).a of a histogram with 20 bins, and using a sample of $n = 100$. The bins are usually chosen to make the picture look 'pleasing' (ad hoc).

(ECDF): Given a sample of n samples taking values in $\mathbb{R}$, or $\{y_i\}_{i=1}^n \subset \mathbb{R}$, then the *Empirical Cumulative Distribution Function* (ECDF) is a function $\hat{F}_n : \mathbb{R} \rightarrow [0, 1]$ which is defined as

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(y_i \leq x), \tag{5.8}$$

where $I(z)$ equals one if z holds true, and zero otherwise. Note that in order to set up this function, one does not need to make choices as the location or size of the bins. This estimator is far more efficient than the histogram, albeit the latter is more often used as it is visually more appealing. A graphical example is given in Fig. (5.2).b of the ECDF using a sample of $n = 100$.

5.1.3 Normal or Gaussian Distribution

Of special (practical as well as theoretical) interest is the Gaussian or Normal distribution with mean μ and standard deviation $\sigma > 0$. Those quantities are also referred to as the first two *moments* of the distribution. The PDF is given for any x as

$$f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right). \tag{5.9}$$

The quantity σ^2 is also known as the *variance* and characterizes the *spread* of the PDF (see Fig. (5.1).a) This specific distribution is of practical as well as theoretical interest for many reasons, perhaps the most important ones being:

(CLT): (the Central Limit Theorem): This classical result states that the average of a large number n of random variables arising from independently samples tends to a normal distribution with standard deviation $O(\sqrt{n})$. This theorem has a long history, but is now often connected to J.W. Lindenberg.

(Closed): The Gaussian distribution is remarkably stable, meaning that a convolution of two Gaussians is still Gaussian. Often, when performing calculations with Gaussian distributions one can easily derive that the resulting distribution is Gaussian as well. Since the Gaussian is characterized by their first two moments only, one consequently needs only to calculate with those and sidestep working with the functional form for the rest.

(Convenience): A third reason one has for using the Gaussian distribution is its convenience. For example, from a practical point of view many related tools are available in statistical software environments. From a more pen-and-pencil perspective it is plain that it is more easy to work with the two first moments than to work with the full functional form of a distribution.

The first reason also implies that the Gaussian distribution will often turn up as a limit distribution of an estimator.

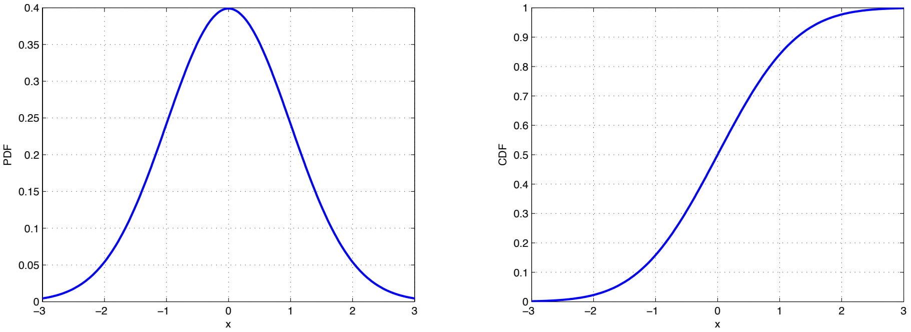


Figure 5.1: (a) PDF of the normal distribution with mean 0 and unit variance. (b) CDF of the normal distribution with mean 0 and unit variance.

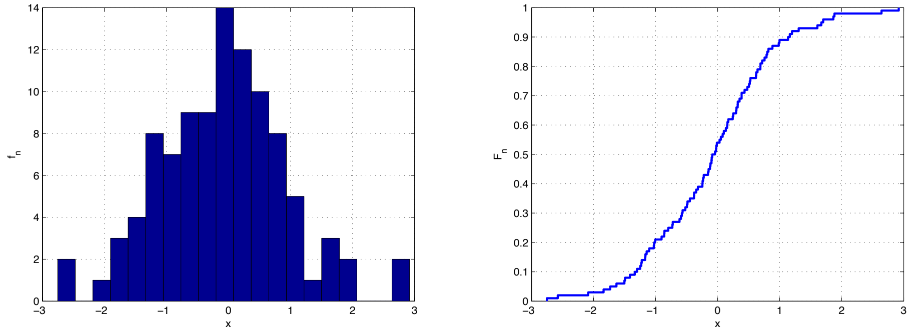


Figure 5.2: Illustration of difference of CDF versus PDF based on a sample of $n = 100$ standard Gaussian distributed values. The histogram - displaying the relative frequency of samples falling within each bin - is the better-known estimate of the pdf. The empirical CDF - defined for each $x \in \mathbb{R}$ as the relative frequency of samples smaller than x - is however much more accurate and fool-proof, but is perhaps less intuitive.

Example 34 *The following examples are instructive. Assume Z is a random variable taking values in $\mathbb{R}^d$, following a Gaussian distribution with the PDF as given in (5.9) for given parameters $\mu \in \mathbb{R}^d$ and covariance $\Sigma \in \mathbb{R}^{d\times d}$. Then*

$$\mathbb{E}[Z] = \mu, \tag{5.10}$$

and

$$\mathbb{E}\left[(Z-\mu)(Z-\mu)^T\right] = \Sigma. \tag{5.11}$$

and

$$Z \sim \mathcal{N}(\mu, I_d) \Leftrightarrow Z - \mu \sim \mathcal{N}(\mu, I_d). \tag{5.12}$$

Let $z \in \mathbb{R}^d$ be a realization of the random variable Z, then

$$\mathbb{E}[z] = z, \tag{5.13}$$

and

$$\mathbb{E}[z^T Z] = z^T \mu. \tag{5.14}$$

Hence

$$\mathbb{E}\left[(Z-\mu)(z-\mu)^T\right] = 0_d. \tag{5.15}$$

5.1.4 Random Vectors

A random vector is an array of random variables. In general, those random variables are related, and the consequent probability rules governing the sampling of the random vector summarizes both the individual laws as the dependence structure inbetween the different elements. This then leads to the notion of a joint probability distribution functions. Again, we make a difference between the joint Cumulative Distribution Function (joint CDF) and the joint Probability Density Function (joint PDF). Those are also referred to as multivariate distribution functions.

The canonical example is the multivariate Gaussian distribution. The Multivariate Gaussian PDF in d dimensions with mean vector $\mu \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$ is given for any $\mathbf{x} \in \mathbb{R}^d$ as

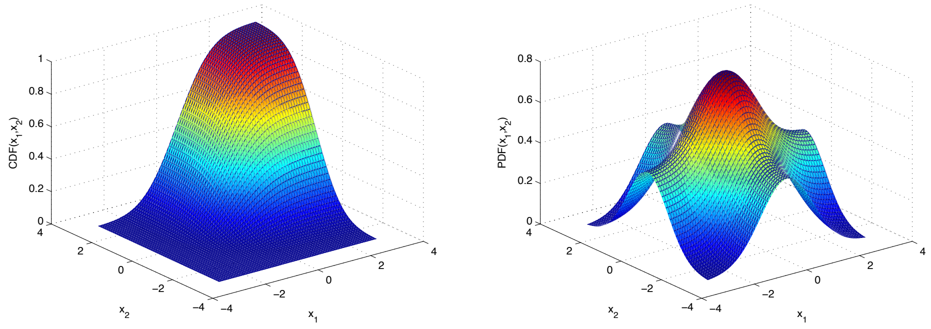
$$f(\mathbf{x};\mu,\Sigma) = \frac{1}{(2\pi)^{\frac{k}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\right), \tag{5.16}$$

where $|\Sigma|$ denotes the determinant of the matrix Σ, and we assume that Σ has a unique inverse (or the determinant does not equal 0). Figure (5.3) gives an example of the CDF and the PDF of a Multi-Variate Normal (MVN) distribution with mean $\mu = (0,0)^T$ and $\Sigma = I_2$.

5.1.5 Stochastic Processes

In the context of this course we stick to the following definition of a stochastic process.

Definition 15 (Stochastic Process) *A stochastic process Z is a sequence of random variables $Z = \{Z_1, Z_2, \ldots, Z_n\}$ where each Z_t takes values in $\mathbb{R}$. It is entirely defined by its joint probability distribution. A sequence of values $\{z_1, \ldots, z_n\}$ is a* realization *of this process if it is assumed to be sampled from mentioned stochastic process.*



(a) (b)

Figure 5.3: Example of a Multivariate Normal Distribution of two independent random variables. (a) the CDF, and (b) the PDF.

Formally, we consider again an experiment with sample space Ω. Now, a stochastic process is a mapping from a sample ω into a path, i.e. a possibly infinite sequence of numbers. The mathematical description of a path is as a function mapping time instances into its corresponding element in the array of numbers. For example let $z = (z_1, z_2, \dots)$ denote such an array, then there $z(t) = z_t$ for each $t = 1, 2, \dots$. This indicates that there is no formal difference between a function and an indexed array, either concept is a mere notational convention. Since in the context of this course we will primarily be interested in discrete stochastic processes where t could take a finite or countably infinite number of values, we will stick to the indexing notation.

While this looks like a very general definition, it excludes quite some cases which are of interest in different situations. Firstly, we restrict attention to finite sequences of random variables, where the index t ('time') runs from 1 to n. Alternatives are found when the index t can take on continuous values ('Continuous stochastic processes'), or even more complex objects belonging to a well-defined group ('Empirical processes').

The subtlety of such processes goes as follows. A stochastic process is a mapping from an event ω to a corresponding time-series, denoted as a realization of this process. The expected value of a stochastic process is the average of all time-series associated to all possible events. That is, the expected value of a stochastic process is a deterministic timeseries! Let this timeseries be denoted as $m = (\dots, m_0, m_1, m_2, \dots)$. In general, one is interested of a value of one location of this timeseries, say m_t Similarly, one can come up with a definition of the covariance associated to a stochastic process, and the covariance evaluated for certain instances. Often, one makes a simplifying assumption on this series by assuming stationarity:

Definition 16 (Stationary Process) *A stochastic process $\{Z_t\}_t$ is said to be (wide-sense) stationary in case the first two moments do not vary over time, or*

$$\begin{cases} \mathbb{E}[Z_t] = \dots \mathbb{E}[Z_t] = \dots = \mathbb{E}[Z_n] & = m \\ \mathbb{E}[(Z_t - m_t)(Z_{t-\tau} - m_{t-\tau})] & = \mathbb{E}[(Z_{t'} - m_{t'})(Z_{t'-\tau} - m_{t'-\tau})] = r(\tau), \end{cases} \tag{5.17}$$

for all t, t', where one has $|m| < C$ and $|r(\tau)| \leq c$ for some finite constants C, c.

This implies that the covariance structure of a stochastic process has a simple form: namely, that all covariances associated to two different locations are equal. This structural assumption makes stochastic processes behave very similar as the LTIs as studied before (why?). In the context of system identification, one is often working assuming a slightly weaker condition on the involved stochastic processes:

Definition 17 (Quasi-Stationary Process) *A stochastic process $\{Z_t\}_t$ is said to be quasi-stationary in case one has*

$$\begin{cases} \mathbb{E}[Z_t] = m_t \\ \mathbb{E}[Z_t Z_s] = r(t,s) \\ \lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^n r(t,t-\tau) = r(\tau), \end{cases} \tag{5.18}$$

where for all t one has $|m_t| < C$ and $|r(\tau)| \leq c$ for some finite constants C, c.

That is, we allow the mean of the signal to vary over time, but assume the average covariance be independent over time. The reason that this definition is quite useful is that systems will typically be expressed as stochastic process Y satisfying for all $t = 1, \ldots, n$ that

$$\mathbb{E}[Y_t] = h_t(u_1, \ldots, u_t), \tag{5.19}$$

where h_t is a filter, and $\{u_1, \ldots, u_n\}$ are deterministic. That means that the mean is almost never time-invariant.

An important problem is that in practice we are only given a single realization of a stochastic process. This observation seems to imply that there is nothing much we as a statistician can do. Surely, we must work with expectations of stochastic quantities for which we have only one sample from. And we know that a average of only one sample gives a very poor estimate of the expectation of this sample. Luckily, there is however a way to go ahead. We can shift a bit further in the stochastic process, and uses the so collected samples to build up a proper estimate. If such estimate would indeed converge to the expected value, one says that the process under study is *ergodic*:

Definition 18 (Ergodic Process) *A stochastic process $\{Z_t\}_t$ is said to be ergodic if for any $\tau = 0, 1, \ldots$ one has*

$$\begin{cases} \lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^n Z_t = \mathbb{E}[Z] \\ \lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^n Z_t Z_{t-\tau} = \mathbb{E}[Z_t Z_{t-\tau}]. \end{cases} \tag{5.20}$$

This notion turns out to be quite fundamental in the analysis of stochastic processes, but in practice it is often (assumed to be) satisfied.

Practically, perhaps the most useful stochastic process is the following.

Definition 19 (Zero Mean White Gaussian Noise) *A stochastic process $Z = \{Z_1, \ldots, Z_n\}$ is called a zero mean white Gaussian noise in case*

(Zero Mean): For each t one has $\mathbb{E}[Z_t] = 0$.

(Gaussian): Each subset of elements is jointly Gaussian distributed with zero mean.

(White): The different elements are uncorrelated.

slightly weaker is

Definition 20 (Zero Mean White Noise) *A stochastic process $Z = \{Z_1, \ldots, Z_n\}$ is called a zero mean white noise in case (i) For each t, $\mathbb{E}[Z_t] = 0$, (ii) the variances of the elements are bounded (i.e. the first two moments exists), and (iii) the different elements are uncorrelated.*

The naming 'white' is historically connected to the related Brownian motion, having a non-vanishing correlation matrix. A popular motivation is that such 'white' noise signal has no 'coloring' due to the fact that all frequencies in its spectrum are equally present.

5.1.6 Interpretations of Probabilities

While notions of probability, random variables and derived concepts were formulated rock-solid (i.e. axiomatic) by A.N. Kolmogorov in the 1930s, there is still ample discussion of what those quantities stand for. This discussion is not only to be fought by philosophers of science, but ones' position here has far-reaching practical impact as well. Rather than surveying the different schools of thought on this matter, let us give the following example by Chernoff suggesting that one should not be guided only by formulas, definitions and formal derivations only in this discussion: statistic is in first instance a practical tool conceived in order to assist decision making in reality. Be critical of its use!

> 'The metallurgist told his friend the statistician how he planned to test the effect of heat on the strength of a metal bar by sawing the bar into six pieces. The first two would go into the hot oven, the next two into the medium oven and the last two into the cool oven. The statistician, horrified, explained how he should randomize in order to avoid the effect of a possible gradient of strength in the metal bar. The method of randomization was applied, and it turned out that the randomized experiment called for putting the first two into the hot oven, the next two into the medium oven and the last two into the cool oven. "Obviously, we can't do that," said the metallurgist. "On the contrary, you have to do that," said the statistician."'

5.2 Statistical Inference

Given a statistical setup ('statistical system') associated to an experiment, perhaps encoded as a number of CDFs or PDFs, one can give solutions to many derived problems. For example one can quantify 'what value to expect next', 'how often does a significance test succeed in its purpose', 'when is an observation not 'typical' under this statistical model', and so on. Statistical inference then studies the question how a statistical system can be identified from associated random values. Often such random variables denote the observations which were gathered while performing an experiment of the studied system. We acknowledge at this point that a statistical system is often an highly abstracted description of the actual experiment, and one rather talks about a 'statistical model underlying the observations', however ambiguous that may sound in the context of this book.

5.2.1 In All Likelihood

Definition 21 (Likelihood Function) *Consider a random value, random vector of stochastic process Z_n which takes values in $\mathbb{Z}$, and with associated cdf F and pdf f (assuming it exists). Consider a family of functions $\{f_\theta : \mathbb{Z} \to \mathbb{R}_+\}_\theta$ indexed by $\theta \in \Theta$. The hope is that this family* *contains an element f_{θ_*} which is in some sense similar to the unknown f. Then the strictly positive likelihood function $L_n : \Theta \to \mathbb{R}_0^+$ is defined as*

$$L_n(\theta) = f_\theta(Z_n). \tag{5.21}$$

The log-Likelihood *of θ on a sample Z_n is defined as $\ell_n(\theta) \triangleq \log L_n(\theta)$*

Note the similarities as well as dissimilarities of the Likelihood function and the pdf $f(Z_n)$ evaluated in the observations. In the special case that there exist a $\theta_* \in \Theta$ such that $f_{\theta_*} = f$, one has obviously that $f(Z_n) = L_n(\theta_*)$.

Definition 22 (The Maximum Likelihood Estimator) *Assume the values $z \in \mathbb{Z}$ observed during an experiment are assumed to follow a random variable Z taking value in $\mathbb{Z}$, obeying a PDF function f which is only known up to some parameters θ. Then the Likelihood function $L_n(\theta)$ can be constructed. A Maximum Likelihood (ML) estimator $\hat\theta$ of θ is defined as*

$$\hat\theta \in \operatorname*{argmax}_\theta L_n(z). \tag{5.22}$$

A prototypical example goes as follows:

Example 35 (Average as an Estimator) *Given n i.i.d. samples from a random variable Z obeying a Gaussian distribution with fixed but unknown mean μ, and a given variance σ^2, or*

$$f(z) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(z-\mu)^2}{2\sigma^2}\right). \tag{5.23}$$

Then the ML estimator for Then given a sample $\{Z_1, \dots, Z_n\}$ of length n, each one being an independent copy of the Gaussian distribution of (5.23). Then the ML estimate of μ is given as

$$\hat\mu = \operatorname*{argmax}_\mu \ell_n(\mu) = \log \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(Z_i-\mu)^2}{2\sigma^2}\right). \tag{5.24}$$

Simplifying the expression and neglecting fixed terms gives the equivalent problem

$$\hat\mu = \operatorname*{argmax}_\mu \sum_{i=1}^n -(Z_i - \mu)^2. \tag{5.25}$$

which equals the familiar LS estimator, and the closed form formula is given as

$$\bar Z_n = \frac{1}{n} \sum_{i=1}^n Z_i. \tag{5.26}$$

Note that this equivalence does not hold any longer if σ is unknown too!

This reasoning can easily be generalized to the case where deterministic explanatory vectors $\{\mathbf{x}_i\}_i$ ('inputs') are available as well. At first, let a statistical model be assumed as follows.

$$Y = \mathbf{x}^T \theta_0 + e, \tag{5.27}$$

where $\mathbf{x} \in \mathbb{R}^d$ is a deterministic, fixed vector, $\theta_0 \in \mathbb{R}^d$ is a fixed vector which happened to be unknown. The last term e is a random variable which takes values into $\mathbb{R}$ following certain rules of probabilities. Specifically, we have that it follows a PDF given as $f_e(\cdot; \mu, \sigma)$ with mean $\mu \in \mathbb{R}$ and standard deviation $\sigma \in \mathbb{R}$, defined $\forall z \in \mathbb{R}$ as

$$f(z; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(z-\mu)^2}{2\sigma^2}\right). \tag{5.28}$$

We assume that e follows a PDF $f(\cdot, 0, \sigma)$. This model implies that also Y is a random variable following a PDF with mean $f(\cdot; \mathbf{x}^T\theta, \sigma)$. A last important assumption which is often given is that the samples we observe from this model are independently sampled. That is, the n samples $\{Y_i\}_{i=1}^n$ satisfy the model

$$Y_i = \mathbf{x}_i^T \theta_0 + e_i, \tag{5.29}$$

where $\{e_i\}_{i=1}^n$ are independent, identically distributed (i.i.d.), that is each sample e_i does not contain information about a sample e_j with $i \neq j$, except for their shared PDF function.

Definition 23 (I.I.D.) *A set of random variables $\{e_1, \ldots, e_n\}$ which each take values in $\mathbb{R}$, contains independent random variables iff for all $i \neq j = 1, \ldots, n$ as*

$$\mathbb{E}[e_i e_j] = \mathbb{E}[e_i]\mathbb{E}[e_j]. \tag{5.30}$$

Those random variables are identically distributed iff they share the same probability function, or if e_i has PDF f_i one has

$$f_i(z) = f_j(z), \tag{5.31}$$

for all $i, j = 1, \ldots, n$ and z ranging over the domain $\mathbb{R}$. If both conditions are satisfied, then the set $\{e_1, \ldots, e_n\}$ is denoted as independently and identically distributed, or abbreviated as i.i.d.

This assumption plays a paramount role in most statistical inference techniques. However, it is exactly on those assumptions that time-series analysis, and estimation for dynamical models will deviate. That is, in such context often past errors e_t will say something about the next term e_{t+1}. This cases will be investigated in some details in later chapters.

Now we can combine the different elements. The corresponding Likelihood function of the model of eq. (5.27), the assumed form of the errors as in(5.28), as well as the i.i.d. assumption results in the following Likelihood function expressed in terms of the parameter vector θ:

$$L_n(\theta) = f(Y_1, \ldots, Y_n) = \prod_{i=1} f(Y_i - \mathbf{x}_i^T\theta; 0, \sigma). \tag{5.32}$$

Note again, that this function equals the PDF of the n samples in case $\theta = \theta_0$. Now the Maximum Likelihood Estimate is given as

$$\hat{\theta} = \operatorname*{argmax}_{\theta} L_n(\theta), \tag{5.33}$$

Working out the right-hand side gives

$$\hat{\theta} = \operatorname*{argmax}_{\theta} \prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(Y_i - \mathbf{x}_i^T\theta)^2}{2\sigma}\right) \propto -\sum_{i=1}^n \left(Y_i - \mathbf{x}_i^T\theta\right)^2. \tag{5.34}$$

In this special case, it is seen that the ML estimator is found by solving the least squares problem $\hat{\theta} = \operatorname{argmin}_\theta \sum_{i=1}^n (Y_i - \mathbf{x}_i^T\theta)^2$. That is, in case $\sigma > 0$ is fixed. In case σ needs to be estimated as well, the Likelihood function becomes more intricate.

5.2.2 Power and Limitations of ML

The ML estimator has a number of exquisite properties as discussed in some detail in the abundant literature of statistical inference. Perhaps the most direct property is that the ML method is efficient, that is it performs as well as any estimator under the given assumptions.

Definition 24 (Unbiased) *Given any estimator $\theta_n = g(Y_1, \ldots, Y_n)$ which approximates a parameter (vector) θ_0 based on a sample $Y_1, \ldots, Y_n$. Then this estimator is unbiased iff*

$$\mathbb{E}[\theta_n] = \mathbb{E}[g(Y_1, \ldots, Y_n)] = \theta_0. \tag{5.35}$$

Theorem 3 (EMSE) *Given any estimator $\theta_n = g(Y_1, \ldots, Y_n)$ of θ_0 , then the performance of this estimator can be expressed as the Expected Mean Square Error (EMSE)*

$$V(g) = \mathbb{E}\left\|g(Y_1, \ldots, Y_n) - \theta\right\|_2^2 = \mathbb{E}\left\|\theta_n - \theta_0)^2\right\|_2^2. \tag{5.36}$$

Theorem 4 (Covariance of Estimate) *Consider a class of PDFs $\{f_\theta : \theta \in \mathbb{R}^d\}$ where θ_0 is the (unknown) one underlying the data observations, that is f_{θ_0} is the PDF underlying the sample $Y_1, \ldots, Y_n$. Given any estimator $\theta_n = g(Y_1, \ldots, Y_n)$ of a parameter vector $\theta_0 \in \mathbb{R}^d$, then the covariance of this estimator $\mathbf{R}(g) \in \mathbb{R}^d \times d$ can be expressed as*

$$\mathbf{R}(g) = \mathbb{E}\left[(g(Y_1, \ldots, Y_n) - \theta)^2\right] = \mathbb{E}\left[(\theta_n - \theta_0)(g(Y_1, \ldots, Y_n) - \theta)^2\right] = \mathbb{E}\left[(\theta_n - \theta_0)^T\right]. \tag{5.37}$$

Theorem 5 (Cramér-Rao Lowerbound) *Given any estimator $\theta_n = g(Y_1, \ldots, Y_n)$ of θ_0 which is unbiased, then*

$$\mathbf{R}(g) \succeq \mathbf{I}_{\theta_0}^{-1}, \tag{5.38}$$

where the so-called Fisher information matrix $\mathbf{I}_{\theta_0}$ is defined as

$$\begin{aligned} \mathbf{I}_{\theta_0} &= \mathbb{E}\left[\frac{d \log f_\theta(Y_1, \ldots, Y_n)}{d\theta} \frac{d^T \log f_\theta(Y_1, \ldots, Y_n)}{d\theta}\Big|_{\theta=\theta_0}\right] \\ &= -\mathbb{E}\left[\frac{d^2 \log f_\theta(Y_1, \ldots, Y_n)}{d\theta^2}\Big|_{\theta=\theta_0}\right]. \end{aligned} \tag{5.39}$$

The general proof can e.g. be found in Ljung's book on System Identification, Section 7.4 and Appendix 7.A. The crucial steps are however present in the following simplified form.

Lemma 6 (Cramér-Rao, simplified) *Consider the case where we have a class of PDFs with a single parameter, say $\{f_\theta : \theta \in \mathbb{R}\}$, such that there is a $\theta_0 \in \mathbb{R}$ such that f_{θ_0} underlies the sample $Y_1, \ldots, Y_n$. Let $\theta_n = g(Y_1, \ldots, Y_n)$ be an unbiased estimator of θ_0, then*

$$\mathbb{E}\left[(\theta_n - \theta_0)^2\right] \geq \frac{1}{m_{\theta_0}}. \tag{5.40}$$

where

$$m_{\theta_0} = \mathbb{E}\left[\frac{df_\theta}{d\theta}\Big|_{\theta=\theta_0}\right]^2. \tag{5.41}$$

5.3 Least Squares Revisited

Let us now turn attention once more to the least squares estimator, and derive some statistical properties on how this works. The analysis is mostly asymptotic, that is properties are derived as if we would have that $n \to \infty$. This is in practice not the case obviously, but those results give nevertheless a good indication of how the estimators behave. We will work under the assumptions that n observations $\{Y_i\}_{i=1}^n$ follow the model

$$Y_i = \mathbf{x}_i^T \theta_0 + D_i, \tag{5.42}$$

where $\theta_0 \in \mathbb{R}^d$ is the *true* parameter which is fixed and deterministic, but which happens to be unknown to us. Here $\{D_1, \dots, D_n\}$ are i.i.d. and hence is uncorrelated, all have zero mean $\mathbb{E}[D_i] = 0$, but have a fully unspecified PDF except for some regularity conditions. Still the LS estimator has very good properties, although it does not correspond necessarily to a ML estimator.

Note at this point the conceptual difference of the deterministic model

$$y_i = \mathbf{x}_i^T \theta + \epsilon_i, \tag{5.43}$$

where $\{\epsilon_1, \dots, \epsilon_n\}$ are (deterministic) residuals, depending (implicitly) on the choice of θ. For this model, there is no such thing as a true parameter. Moreover, there is no stochastic component, such that e.g. $\mathbb{E}[\epsilon_i] = \epsilon_i$. Note the important differences between the 'true' noise $\{D_i\}_i$ under model (5.42), and the residuals $\{\epsilon_i\}_i$. They only equal each other in the special case that the model (5.42) is assumed to underly the observations $\{y_i\}_i$ (that is if $\{y_i\}_i$ are samples from $\{Y_i\}_i$), and $\theta = \theta_0$ (that is, we have estimated the true parameter *exactly*). Often one makes this assumption that $\{y_i\}_i$ are samples from $\{Y_i\}_i$, but one has merely that $\theta \approx \theta_0$ and the residual terms do not obey the stochastic properties of the noise!

Example 36 (Average, Ct'd) *Consider again the model $Y_i = \theta_0 + D_i$ where $\theta_0 \in \mathbb{R}$ is fixed but unknown, and $\{D_i\}_i$ are i.i.d. random variables with zero mean and standard deviation σ. Then the LS estimator θ_n of θ_0 is solves the optimization problem*

$$V_n(\theta_n) = \min_\theta \sum_{i=1}^n (Y_i - \theta)^2, \tag{5.44}$$

for which the solution is given as $\theta_n = \frac{1}{n}\sum_{i=1}^n Y_i$. How well does θ_n estimate θ_0?

$$\mathbb{E}[\theta_0 - \theta_n]^2 = \mathbb{E}\left[\theta_0 - \frac{1}{n}\sum_{i=1}^n Y_i\right]^2 = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n (\theta_0 - Y_i)\right]^2 = \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}[D_i^2] = \frac{\sigma^2}{n}. \tag{5.45}$$

Now we answer the question whether the minimal value $V_n(\theta_n)$ says something about the standard *deviation σ. Therefore, we elaborate the objective for the optimal θ_n, which gives*

$$\begin{aligned}
V_n(\theta_n) &= \mathbb{E}\sum_{i=1}^n \left(Y_i - \frac{1}{n}\sum_{i=1}^n Y_i\right)^2 \\
&= \mathbb{E}\sum_{i=1}^n \left((Y_i-\theta_0) - (\frac{1}{n}\sum_{j=1}^n Y_j - \theta_0)\right)^2 \\
&= \mathbb{E}\sum_{i=1}^n \left((Y_i-\theta_0)^2 - 2(Y_i-\theta_0)(\frac{1}{n}\sum_{j=1}^n Y_j - \theta_0) + (\frac{1}{n}\sum_{j=1}^n Y_i - \theta_0)^2\right) \\
&= \mathbb{E}\sum_{i=1}^n \left(D_i^2 - \frac{2}{n}\sum_{j=1}^n D_i D_j + \frac{1}{n}\sum_{j=1}^n D_i^2\right) \\
&= \sum_{i=1}^n \mathbb{E}[D_i^2] - \mathbb{E}\frac{1}{n}\sum_{i=1}^n D_i^2 \\
&= (n-1)\,\sigma^2, \qquad (5.46)
\end{aligned}$$

since $\sum_{i=1}^n (Y_i - \theta_n) = 0$ by the property of least squares.

Let us now study the covariance and the expected minimal value of the OLS estimate.

Lemma 7 (Statistical Properties of OLS) *Assume the data follows a model $Y_i = \mathbf{x}_i^T\theta + D_i$ with $\{D_1,\dots,D_n\}$ uncorrelated random variables with mean zero and standard deviation $\sigma > 0$, and $\theta, \mathbf{x}_1, \dots, \mathbf{x}_n$ are deterministic vectors in $\mathbb{R}^d$. Let the matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$ enumerate those such that $\mathbf{X}_i = \mathbf{x}_i^T$ for all $i = 1,\dots,n$, and assume that $\mathbf{X}$ has full rank such that the inverse $(\mathbf{X}^T\mathbf{X})^{-1}$ is defined uniquely. Let θ_n be the LS estimate (as in Chapter 2) solving*

$$V_n(\theta) = \min_\theta \frac{1}{2}\sum_{i=1}^n (Y_i - \mathbf{x}_i^T\theta)^2, \qquad (5.47)$$

then

- *The estimate θ_n is unbiased, that is $\mathbb{E}[\theta_n] = \theta_0$.*
- *The covariance of the estimate is given as*

$$\mathbb{E}\left[(\theta_0 - \theta_n)(\theta_0 - \theta_n)^T\right] = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}. \qquad (5.48)$$

- *The estimate $V_n(\theta_n)$ implies an unbiased estimate of σ as*

$$\sigma^2 = \frac{2}{n-d}\mathbb{E}[V_n(\theta_n)]. \qquad (5.49)$$

Proof: At first, we have the normal equations characterizing θ_n as

$$\theta_n = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Y = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}\theta_0 + D). \qquad (5.50)$$

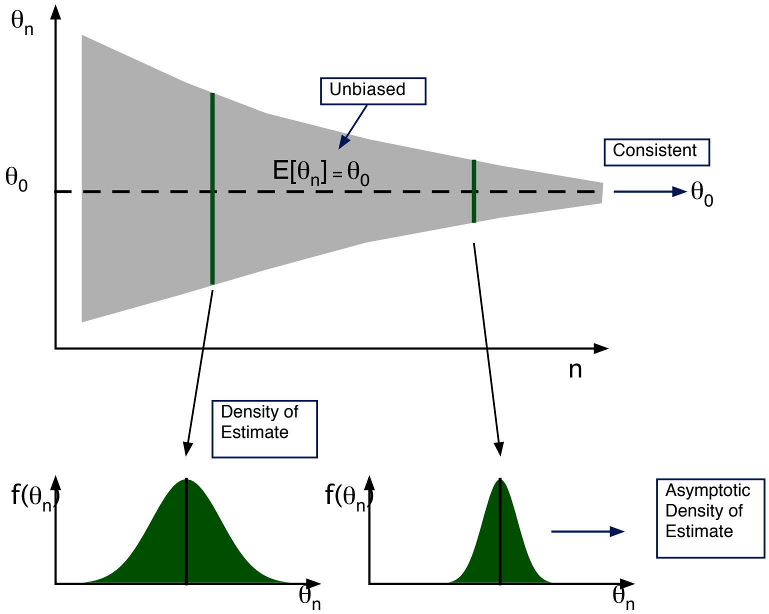


Figure 5.4: Schematical illustration of an *unbiased estimator* θ_n of θ_0. Here $n = 1, 2, \ldots$ denotes the size of the samples $\{Y_1, \ldots, Y_n\}$ on which the estimator $\theta_n = g(Y_1, \ldots, Y_n)$ is based. The estimator is called unbiased if one has for any n that $\mathbb{E}[\theta_n] = \theta_0$. The grey area denoted the possible estimates θ_n for different samples $\{Y_1, \ldots, Y_n\}$. The cross-section of this area for a given n equals the sample distribution, denoted as the 2 bell-shaped curves at the bottom.

where $Y = (Y_1, \ldots, Y_n)^T$ and $D = (D_1, \ldots, D_n)^T$ are two random vectors taking values in $\mathbb{R}^n$. Then

$$\theta_n = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}\theta_0 + D) = (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{X})\theta_0 + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T D = \theta_0 + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T D. \quad (5.51)$$

Taking the expectation of both sides gives

$$\mathbb{E}[\theta_n] = \mathbb{E}\left[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}\theta_0 + D)\right] = \theta_0 + \mathbb{E}\left[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T D\right] = \theta_0 + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbb{E}\left[D\right] = \theta_0, \quad (5.52)$$

since the vectors $\theta_0, \{\mathbf{x}_i\}$ are deterministic, and hence $\mathbb{E}[\theta_0] = \theta_0, \mathbb{E}[\mathbf{X}] = \mathbf{X}$. This proves unbiasedness of the estimator. Note that the assumption of the vectors $\theta_0, \mathbf{x}_1, \ldots, \mathbf{x}_n$ being deterministic is crucial.

Secondly, the covariance expression can be derived as follows. Here the crucial insight is that we have by assumption of zero mean i.i.d. noise (or white noise) that $\mathbb{E}[DD^T] = \sigma^2 I_n$ where $I_n = \operatorname{diag}(1, \ldots, 1) \in \mathbb{R}^{n\times n}$. Then we have from eq. (5.51) that

$$\begin{aligned}
\mathbb{E}\left[(\theta_0 - \theta_n)(\theta_0 - \theta_n)^T\right] &= \mathbb{E}\left[((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T D)(D^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-T})\right] \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbb{E}[DD^T]\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-T} \\
&= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{X})(\mathbf{X}^T\mathbf{X})^{-T} \\
&= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}. \quad (5.53)
\end{aligned}$$

Thirdly, the minimal value of the minimization problem is given as

$$\begin{aligned}
V_n(\theta_n) &= \frac{1}{2}(Y - \mathbf{X}\theta_n)^T(Y - \mathbf{X}\theta_n) \\
&= \frac{1}{2}\left(Y^TY - 2Y^T\mathbf{X}\theta_n + \theta_n^T\mathbf{X}^T\mathbf{X}\theta_n\right) \\
&= \frac{1}{2}\left(Y^TY - 2Y^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^TY + Y^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{X})(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^TY\right) \\
&= \frac{1}{2}\left(Y^TY - Y^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^TY\right). \quad (5.54)
\end{aligned}$$

Hence

$$\begin{aligned}
2V_n(\theta_n) &= Y^T\left(I_n - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right)Y \\
&= (\mathbf{X}\theta_0 + D)^T\left(I_n - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right)(\mathbf{X}\theta_0 + D) \\
&= D^T\left(I_n - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right)D. \quad (5.55)
\end{aligned}$$

Then, using the properties of the trace operator $\operatorname{tr}(\mathbf{x}^T\mathbf{A}\mathbf{x}) = \operatorname{tr}(\mathbf{x}\mathbf{x}^T\mathbf{A})$ and $\operatorname{tr}(\mathbf{A} + \mathbf{B}) = \operatorname{tr}(\mathbf{A}) +$

$\mathrm{tr}(\mathbf{B})$ for $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n\times n}$ and $\mathbf{x} \in \mathbb{R}^n$ gives

$$\begin{aligned}2\mathbb{E}[V_n(\theta_n)] &= \mathbb{E}\,\mathrm{tr}\left(D^T\left(I_n - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right)D\right)\\ &= \mathbb{E}\,\mathrm{tr}\left(D^TD\left(I_n - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right)\right)\\ &= \mathrm{tr}\left(\mathbb{E}[D^TD]\left(I_n - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right)\right)\\ &= \sigma^2\,\mathrm{tr}(I_n - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)\\ &= \sigma^2\left(\mathrm{tr}(I_n) - \mathrm{tr}(\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)\right)\\ &= \sigma^2\left(\mathrm{tr}(I_n) - \mathrm{tr}((\mathbf{X}^T\mathbf{X})(\mathbf{X}^T\mathbf{X})^{-1})\right)\\ &= \sigma^2\left(\mathrm{tr}(I_n) - \mathrm{tr}(I_d)\right) = \sigma^2(n-d). \end{aligned} \tag{5.56}$$

□

This result is slightly generalized as follows.

Theorem 6 (Gauss-Markov Theorem) *Given a model with deterministic values $\{\mathbf{x}_1, \dots, \mathbf{x}_n\} \subset \mathbb{R}^d$ with fixed but unknown $\theta_0 \in \mathbb{R}^d$ such that*

$$Y_i = \mathbf{x}_i^T\theta_0 + D_i, \tag{5.57}$$

where $\{D_1, \dots, D_n\}$ are uncorrelated, all have zero mean $\mathbb{E}[D_i] = 0$ and have (finite) equal variances, i.e. $\mathbb{E}[D_i] = \dots = \mathbb{E}[D_n]$ (i.e. $\{D_i\}_i$ is heteroskedastic). Suppose that we have an estimator $\theta_n = g(Y_1, \dots, Y_n)$. Then its performance can be measured as the variance

$$V(\theta_n) = \mathbb{E}\left[(\theta_0 - \theta_n)^2\right]. \tag{5.58}$$

Then the estimator achieving the minimal possible variance $V(\theta_n)$ such that it is unbiased $\mathbb{E}[\theta_n] = \theta_0$ is given as

$$\theta_n = \operatorname*{argmin}_{\theta} \frac{1}{2}\sum_{i=1}^{n}(Y_i - \mathbf{x}_i^T\theta)^2. \tag{5.59}$$

In turn we also have that the least squares estimate needs modification if the coloring of the noise is known to equal a matrix $\mathbf{R}$. The reasoning goes as follows. Assume again that the linear model

$$Y_i = \mathbf{x}_i^T\theta_0 + D_i, \tag{5.60}$$

with $\mathbf{x}_1, \dots, \mathbf{x}_n, \theta \in \mathbb{R}^d$, but where $\{D_1, \dots, D_n\}$ are zero mean random variables with covariance $\mathbf{R}$ such that $\mathbf{R}_{ij} = \mathbb{E}[D_iD_j]$ for all $i, j = 1, \dots, n$. Then the estimator θ_n of θ_0 with minimal expected error is given as

$$\theta_n = \operatorname*{argmin}_{\theta} \frac{1}{2}(Y - \mathbf{X}\theta)^T\mathbf{R}^{-1}(Y - \mathbf{X}\theta). \tag{5.61}$$

This estimator is known as the Best Linear Unbiased Estimate (BLUE). The following simple example illustrates this point:

Example 37 (Heteroskedastic Noise) *Consider again the model*

$$Y_i = \theta_0 + D_i, \tag{5.62}$$

where $\theta_0 \in \mathbb{R}$ and $\{D_i\}_i$ are uncorrelated (white) zero mean stochastic variables, with variances $\mathbb{E}[D_i^2] = \sigma_i^2 > 0$ which are different for all samples, i.e. for all $i = 1, \dots, n$. Then the BLUE estimator becomes

$$\theta_n = \operatorname{argmin} \sum_{i=1}^{n} (Y - 1_n\theta_0)^T \mathbf{M} (Y - 1_n\theta_0), \tag{5.63}$$

where

$$\mathbf{M} = \begin{bmatrix} \sigma_1^{-2} & & 0 \\ & \ddots & \\ 0 & & \sigma_n^{-2} \end{bmatrix}. \tag{5.64}$$

The solution is hence given as

$$1_n^T \mathbf{R} 1_n \theta_n = 1_n^T \mathbf{R} Y, \tag{5.65}$$

where $Y = (Y_1, \dots, Y_n)^T$ takes elements in $\mathbb{R}^n$. Equivalently,

$$\theta_n = \frac{1}{\sum_{i=1}^n \sigma_i^2} \sum_{i=1}^{n} \frac{Y_i}{\sigma_i^2}. \tag{5.66}$$

Note that the influence of a sample Y_i in the total sum is small in case it is inaccurate, or σ_i is large, and vice versa.

Lets now give an example where the inputs are stochastic as well, or

Example 38 (Stochastic Inputs) *Assume the observations $\{Y_i\}_i$ are modeled using the random vectors $\{X_i\}_i$ taking values in $\mathbb{R}^d$, $\theta_0 \in \mathbb{R}^d$ is deterministic but unknown*

$$Y_i = X_i^T \theta_0 + D_i, \tag{5.67}$$

where $\{D_i\}_i$ are zero mean i.i.d. and are assumed to be independent from $\{X_i\}$. *This assumption is crucial as we will see later. Then the above derivations still hold more or less. Consider the LS estimate θ_n. It is an unbiased estimate of θ_0 as could be seen by reproducing the above proof. Let $e = (D_1, \dots, D_n)^T$, then*

$$\mathbb{E}[\theta_n] = \mathbb{E}\left[(X^TX)^{-1}X^T(X\theta_0 + e)\right] = \theta_0 + \mathbb{E}\left[(X^TX)^{-1}X^Te\right] = \theta_0 + \mathbb{E}[(X^TX)^{-1}X^T]\mathbb{E}\left[e\right] = \theta_0, \tag{5.68}$$

where X is the random matrix taking elements in $\mathbb{R}^{n\times d}$ such that $\mathbf{e}_i^T X = X_i$ for all $i = 1, \dots, n$. Here we need the technical condition that $\mathbb{E}[(X^TX)^{-1}]$ exists, or that X^TX is almost surely full rank. This equation implies asymptotic unbiasedness of the estimator θ_n. Similarly, one can proof that the covariance of θ_n is given as

$$\mathbb{E}\left[(\theta_0 - \theta_n)(\theta_0 - \theta_n)^T\right] = \sigma^2 \mathbb{E}\left[(X^TX)^{-1}\right]. \tag{5.69}$$

Note that $\mathbb{E}[(X^TX)^{-1}] \neq (\mathbb{E}[X^TX])^{-1}$ exactly, although such relation holds approximatively in case $\frac{1}{n}X^TX \approx \mathbb{E}[X^TX]$. Finally, the minimal value $V_n(\theta_n)$ satisfies

$$\sigma^2 = \frac{2}{n-d}\mathbb{E}[V_n(\theta_n)]. \tag{5.70}$$

The key property which causes those complications is the fact that $\mathbb{E}[X_tD_t] \neq \mathbb{E}[X_t]\mathbb{E}[D_t]$. This condition was trivially satisfied if $\mathbf{x}_t$ were deterministic, leading to the many optimality principles of least squares estimates as stated in the Gauss-Markov theorem.

5.4 Instrumental Variables

Example 39 (Dependent Noise) *Consider again the following model using the definitions as given in the previous example:*

$$Y_t = X_t^T \theta_0 + D_t, \tag{5.71}$$

and D_t is a random variable with bounded variance and zero mean, then the least squares estimate θ_n is given by the solution of

$$\theta_n = \left(\frac{1}{n} \sum_{t=1}^{n} X_t X_t^T \right)^{-1} \left(\frac{1}{n} \sum_{t=1}^{n} X_t Y_t \right). \tag{5.72}$$

In case $n \to \infty$, one has by definition that

$$\theta_n = \left(\mathbb{E}\left[X_t X_t^T \right] \right)^{-1} \mathbb{E}\left[X_t Y_t \right]. \tag{5.73}$$

Assuming that $\mathbb{E}[X_t X_t^T]$ exists and is invertible, one can write equivalently that

$$\begin{aligned} \theta_0 - \theta_n &= \theta_0 - \left(\mathbb{E}\left[X_t X_t^T \right] \right)^{-1} \mathbb{E}\left[X_t Y_t \right] \\ &= \left(\mathbb{E}\left[X_t X_t^T \right] \right)^{-1} \mathbb{E}\left[X_t X_t^T \right] \theta_0 - \left(\mathbb{E}\left[X_t X_t^T \right] \right)^{-1} \mathbb{E}\left[X_t (X_t^T \theta_0 + D_t) \right] \\ &= \left(\mathbb{E}\left[X_t X_t^T \right] \right)^{-1} \mathbb{E}\left[X_t D_t \right]. \end{aligned} \tag{5.74}$$

And the estimate θ_n is only (asymptotically) unbiased if $\mathbb{E}\left[X_t D_t \right] = 0_d$.

This reasoning implies that we need different parameter estimation procedures in case the noise is dependent on the inputs. Such condition is often referred to as the 'colored noise' case. One way to construct such an estimator, but retaining the convenience of the LS estimator and corresponding normal equations goes as follows.

We place ourselves again in a proper stochastic framework, where the system is assumed to be

$$Y_i = X_i^T \theta_0 + D_i, \tag{5.75}$$

where $X_1, \ldots, X_n, \theta_0 \in \mathbb{R}^d$ are random vectors, and $\{D_i\}$ is zero mean stochastic noise. As in the example this noise can have a substantial coloring, and an ordinary least squares estimator wont give consistent estimates of θ_0 in general. Now let us suppose that we have the random vectors $\{Z_t\}_t$ taking values in $\mathbb{R}^d$ such that

$$\mathbb{E}\left[Z_t D_t \right] = 0_d. \tag{5.76}$$

That is, the instruments are orthogonal to the noise. Then the IV estimator θ_n is given as the solution of $\theta \in \mathbb{R}^d$ to the following system of linear equations

$$\sum_{t=1}^{n} Z_t (Y_t - X_t^T \theta) = 0_d, \tag{5.77}$$

where expectation is replaced by a sample average. That means that we estimate the parameters by imposing the sample form of the assumed independence: that is the estimated model necessarily matches the assumed moments of the involved stochastic quantities. Note that this expression looks similar to the normal equations. If $\sum_{t=1}^n (Z_t X_t^T)$ were invertible, then the solution is unique and can be written as

$$\theta_n = \left(\sum_{t=1}^n Z_t X_t^T\right)^{-1} \left(\sum_{t=1}^n Z_t Y_t^T\right). \tag{5.78}$$

So the objective for us id to design instruments, such that

- The instruments are orthogonal to the noise, or $\mathbb{E}\left[Z_t D_t\right] = 0_d$.
- The matrix $\mathbb{E}[Z_t X_t^T]$ were of full rank, such that also with high probability $\sum_{t=1}^n (Z_t X_t^T)$ has a unique inverse.

Example 40 *A common choice in the context of dynamical systems for such instruments goes as follows. Assume that the random vectors X_t consists of delayed elements of the output $Y_{t-\tau}$ of the system which cause the troublesome correlation between D_t and X_t. This is for example typically the case in an ARMAX model. Then a natural choice for the instruments would be to take delayed entries of the input $\{U_t\}_t$ of the system*

$$Z_t = (U_{t-1}, \ldots, U_{t-d}), \tag{5.79}$$

which takes values in $\mathbb{R}^d$. This is a good choice if the inputs were assumed to be independent of the (colored) noise.

Chapter 6

Prediction Error Methods

> ”How far can we go by optimizing the predictive performance of an estimated model?”

This chapter studies the parameter estimation technique called the Prediction Error Method (PEM). The idea is that rather than a plain least squares approach, or a statistical maximum likelihood approach there is a third important principle in use for estimating the parameters of a dynamic model based on recorded observations. This technique considers the accuracy of the predictions computed for the observations, rather than the model mismatch are the likelihood of the corresponding statistical model. This technique is perhaps the most tightly connected to systems theory as it explicitly exploits the dynamical structure of the studied system. Those three design principles are represented schematically in Fig. (6.1). In a number of cases the three design decision leads to the same estimators as will be discussed in some detail.

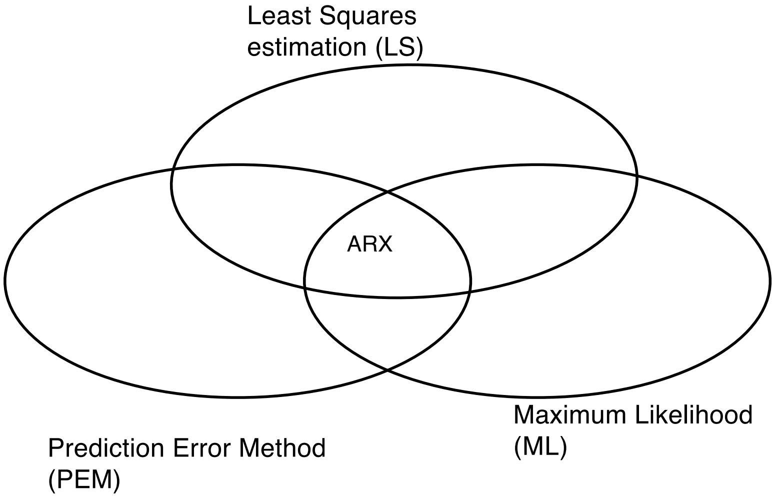


Figure 6.1: Schematic Illustration of the different approaches which one could take for estimation of parameters.

6.1 Identification of an ARX model

Let's start the discussion with a study of a convenient model. The result is not too difficult to obtain, but the consecutive steps in the analysis will come back over and over in later sections. Consider a system relating two signals $\{u_t\}_{t=-\infty}^{\infty}$ and $\{y_t\}_{t=-\infty}^{\infty}$ which is modeled as

$$A(q^{-1})y_t = B(q^{-1})u_t + e_t,\ \forall t = \dots, 0, 1, , 2, \dots, \tag{6.1}$$

where for given $n_a, n_b > 0$ one has $A(q^{-1}) = 1 + a_1 q^{-1} + \dots + a_{n_a} q^{-n_a}$ and $B(q^{-1}) = b_1 q^{-1} + \dots + b_{n_b} q^{-n_b}$, with fixed but unknown coefficients $\{a_1, \dots, a_{n_a}, b_1, \dots, b_{n_b}\}$. Here the residuals $\{e_t\}_t$ are small in some sense, but unknown otherwise. This system can be written equivalently as

$$y_t = \varphi_t^T \theta + e_t,\ \forall t = \dots, 0, 1, , 2, \dots, \tag{6.2}$$

where

$$\begin{cases} \varphi_t^T = (-y_{t-1}, \dots, -y_{t-n_a}, u_{t-1}, \dots, u_{t-n_b})^T \in \mathbb{R}^{n_a+n_b},\ \forall t \\ \theta = (a_1, \dots, a_{n_a}, b_1, \dots, b_{n_b})^T \in \mathbb{R}^{n_a+n_b}. \end{cases} \tag{6.3}$$

The model is linear in the parameters, hence it is already known how to estimate the parameter vector θ from given samples $\{(\varphi_t, y_t)\}_{t=1}^n$ induced by the signals $\{u_t\}_t$ and $\{y_t\}_t$. Note that if the signals are only recorded at time instances $t = 1, 2, \dots, n$, one can only construct the samples $\{(\varphi_t, y_t)\}_{t=1+\max(n_a,n_b)}^n$. - for notational convenience we shall assume further that the signals are observed fully such that $\{(\varphi_t, y_t)\}_{t=1}^n$ can constructed. The Least Squares (LS) estimation problem is

$$\min_{\hat\theta=(a_1,\dots,a_{n_a},b_1,\dots,b_{n_b})} \sum_{t=1}^n (y_t + a_1 y_{t-1} + \dots + a_{n_a} y_{t-n_a} - b_1 u_{t-1} - \dots - b_{n_b} u_{t-n_b})^2 = \sum_{t=1}^n (\varphi_t \hat\theta - y_t)^2, \tag{6.4}$$

and the estimate $\hat\theta$ is given as the solution to

$$\left(\frac{1}{n}\sum_{t=1}^n \varphi_t \varphi_t^T\right)\hat\theta = \left(\frac{1}{n}\sum_{t=1}^n y_t \varphi_t\right), \tag{6.5}$$

which are known as the normal equations associated to problem (6.4). If the matrix

$$\Phi = \left(\frac{1}{n}\sum_{t=1}^n \varphi_t \varphi_t^T\right), \tag{6.6}$$

is of full rank the estimate is unique and is given as

$$\hat\theta = \Phi^{-1}\left(\frac{1}{n}\sum_{t=1}^n y_t \varphi_t\right). \tag{6.7}$$

Such approach is also related as an 'equation error method' since the errors we minimize derive directly from $\{e_t\}_t$ which occur as equation errors in (6.1).

The normal equations can readily be solved with the numerical tools described in Chapter 1. For the statistical properties it is of crucial importance which setup is assumed. We will work with the assumption that $\{e_t\}_t$ are modeled as random variables, and hence so are $\{y_t\}_t$ and $\{\varphi_t\}_{t=1}^n$. This is an important difference with a classical analysis of a LS approach as given in Section ... as there one assumes φ_t is deterministic. The reason that this difference is important is that when taking expectations various quantities, it is no longer possible to treat Φ nor Φ^{-1} as a constant matrix.

The common statistical assumptions used to model and analyze this problem go as follows. Formally, let the signals $\{U_t\}_t$ and $\{Y_t\}_t$ be stationary stochastic processes related as

$$Y_t = \varphi_t^T \theta_0 + V_t, \ \forall t = \dots, 0, 1, , 2, \dots, \tag{6.8}$$

where $\theta_0 \in \mathbb{R}^{n_a+n_b}$ is the fixed but unknown 'true' parameter vector, the vector $\varphi_t = (-Y_{t-1}, \dots, Y_{t-n_a}, U_{t-1}, \dots, U_{t-n_b})^T$ which takes values in $\mathbb{R}^{n_a+n_b}$, and where we assume that $\{V_t\}_t$ is a stationary stochastic process independent of the input signal $\{U_t\}_t$. If an estimate $\hat{\theta}$ is 'good', it should be in some sense 'close' to θ_0. Lets examine then how good the LS estimator is. From the normal equations one gets

$$\begin{aligned}\hat{\theta} - \theta_0 &= \left(\frac{1}{n}\sum_{t=1}^n \varphi_t \varphi_t^T\right)^{-1} \left(\frac{1}{n}\sum_{t=1}^n \varphi_t Y_t\right) - \left(\frac{1}{n}\sum_{t=1}^n \varphi_t \varphi_t^T\right)^{-1} \left(\frac{1}{n}\sum_{t=1}^n \varphi_t \varphi_t^T\right)\theta_0 \\ &= \left(\frac{1}{n}\sum_{t=1}^n \varphi_t \varphi_t^T\right)^{-1} \left(\frac{1}{n}\sum_{t=1}^n V_t \varphi_t\right). \end{aligned} \tag{6.9}$$

Under weak conditions, the normalized sums tend to their expected values when n tends to infinity. Hence $\hat{\theta} \to \theta_0$, or $\hat{\theta}$ is consistent if

$$\begin{cases} \mathbb{E}\left[\varphi_t \varphi_t^T\right] \text{ is nonsinguar} \\ \mathbb{E}[\varphi_t V_t] = 0. \end{cases} \tag{6.10}$$

The first condition ('nonsingular') is often satisfied, but there are a few important exceptions:

- The inputs $\{U_t\}$ is not sufficiently rich: it is not PE of order n_b.
- The data is noise-free (i.e. $V_t = 0$ for all t), and the model orders are chosen too high: this implies that $A_0(q^{-1})$ and $B_0(q^{-1})$ associated with θ_0 have common factors (are not coprime).
- The input signal $\{U_t\}_t$ is generated by a linear low-order feedback law from the output $\{Y_t\}_t$.

Unlike the 'nonsingular' condition, the requirement $\mathbb{E}[\varphi_t V_t] = 0$ is in general *not* satisfied. An important exception is when $\{V_t\}_t$ is white noise, i.e. is a sequence of uncorrelated random variables. In such case, $\{V_t\}_t$ will be uncorrelated with all past data, and in particular V_t will be uncorrelated with φ_t, implying the condition.

The LS estimation technique is certainly simple to use. In case those requirements are not at all satisfied, we need modifications to the LS estimate to make it 'work', i.e. make the estimate consistent or at least not too biased. We will study two such modifications.

- Minimization of the prediction error for 'more detailed' model structures. This idea leads to the class of Prediction Error Methods (PEM) dealt with in this chapter.
- Modification of the normal equations associated to the LS estimator. This idea leads to the class of Instrumental Variables dealt with in Chapter ... .

6.2 Optimal Prediction and PEM

A model obtained by identification can be used in many ways depending on the purpose of modeling. In may applications the aim of the model is prediction in some way. It therefore makes sense to determine the model such that the prediction error would be minimal. Let us consider the SISO case at first. We denote the model prediction error here as

$$\epsilon_t(\theta) = y_t - f_{t|t-1}(\theta),\ \forall t = 1, 2, \dots, \tag{6.11}$$

where θ represents the parameters of the current model, and $f_{t|t-1}(\theta)$ represents the prediction of the outcome y_t using all past information and the model determined by θ. In case of an ARX model as described in the previous chapter, we have obviously that

$$f_{t|t-1}(\theta) = \varphi_t^T \theta. \tag{6.12}$$

In the context of PEM methods one is in general interested in more general models. Suppose a general LTI describes the signals $\{u_t\}_t$ and $\{y_t\}_t$ as

$$y_t = G(q^{-1}, \theta)u_t + H(q^{-1}, \theta)V_t,\ \forall t = \dots, 0, 1, \dots, \tag{6.13}$$

where we assume that $\{V_t\}_t$ is a stochastic process with $\mathbb{E}[V_s V_t^T] = \sigma^2 \delta_{s,t}$ with $\delta_{s,t} = 1$ if $s = t$, and zero otherwise. For notational convenience, assume that $G(0; \theta) = 0$, i.e. that the model has at least one pure delay from input to output. Then, the optimal predictor can be written as

$$f_{t|t-1}(\theta) = L_1(q^{-1}, \theta)y_t + L_2(q^{-1}, \theta)u_t,\ \forall t = \dots, 0, 1, \dots, \tag{6.14}$$

which is a function of the past data only if $L_1(0, \theta) = L_2(0, \theta) = 0$. Suppose we have for our model (H, G) corresponding mappings (L_1, L_2). Now, a PEM method will estimate the parameter vector θ by optimizing the prediction performance, i.e.

$$\hat{\theta} = \operatorname*{argmin}_{\theta} \frac{1}{n} \sum_{t=1}^{n} \ell(\epsilon_t(\theta)), \tag{6.15}$$

where $\ell : \mathbb{R} \to \mathbb{R}$ is a loss-function. E.g. $\ell(e) = e^2$.

Now we study the question how to go from an LTI model (G, H) to the corresponding predictors $(L_1, L2)$. Again let us introduce ideas using a series of elementary examples.

Example 41 (White Noise) *Assume a realization $\{e_1, e_2, \dots, e_t\}$ of zero mean white (uncorrelated) noise. Given the values of $(e_1, e_2, \dots, e_{t-1})$, the best estimate of e_t in L_2 sense is then $\hat{e}_t = 0$, That is*

$$\hat{e}_t = \operatorname*{argmin}_{\sum_{\tau=1}^{t+1} h_\tau q^{-\tau}} \mathbb{E}\left(e_t - \sum_{\tau=1}^{t+1} h_\tau e_{t-\tau}\right)^2 = \operatorname*{argmin}_{\sum_{\tau=1}^{t+1} h_\tau q^{-\tau}} \mathbb{E}[e_t^2] + \sum_{\tau=1}^{t-1} h_\tau \mathbb{E}[e_{t-\tau}], \tag{6.16}$$

and the minimum is clearly achieved when $h_1 = \dots = h_\tau = 0$.

Example 42 (FIR(d)) *Given a deterministic sequence $\{u_t\}_{t=1}^n$, and given a realization $\{y_t\}_t$ of a process $\{Y_t\}_{t=1}^n$ which satisfies a FIR system, or*

$$Y_t = b_1 u_{t-1} + \dots + b_d u_{t-d} + D_t, \tag{6.17}$$

where $\{D_1, \dots, D_n\}$ is a zero mean white noise sequence with bounded variance. Then the optimal prediction at instance $t+1$ is clearly

$$\hat{y}_{t+1} = b_1 u_t + \cdots + b_d u_{t-d+1}, \tag{6.18}$$

for any $t = d, \dots, n-1$.

Example 43 (AR(d)) *Given a realization $\{y_t\}_t$ of a process $\{Y_t\}_{t=1}^n$ which satisfies a AR(d) system, or*

$$Y_t + a_1 Y_{t-1} + \dots a_d Y_{t-d} = D_t, \tag{6.19}$$

where $\{D_1, \dots, D_n\}$ is a zero mean white noise sequence with bounded variance. Then the optimal prediction at instance $t+1$ is clearly

$$\hat{y}_{t+1} = a_1 y_t + \cdots + a_d y_{t-d+1}, \tag{6.20}$$

for any $t = d, \dots, n-1$.

Example 44 (MA(d)) *Given a realisation $\{y_t\}_t$ of a process $\{Y_t\}_{t=1}^n$ which satisfies a MA(d) system, or*

$$Y_t = D_t + c_1 D_{t-1} + \dots c_d D_{t-d}, \tag{6.21}$$

where $\{D_1, \dots, D_n\}$ is a zero mean white noise sequence with bounded variance. Equivalently,

$$Y_t = C(q^{-1}) D_t, \tag{6.22}$$

for any $t = d, \dots, n-1$. Thus

$$\begin{cases} D_t = C^{-1}(q^{-1}) Y_t \\ Y_t = (C(q^{-1}) - 1) D_t + D_t, \end{cases} \tag{6.23}$$

where the second equality separates nicely the contribution of the past noise $D_{t-1}, D_{t-2}, \dots$ on which we have some knowledge, and the present term D_t which is entirely unknown to us. This is a consequence of the fact that C is a monomial, i.e. the zeroth order term equals 1. Then it is not too difficult to combine both equations in (6.23) and then we find the corresponding optimal predictor as

$$\hat{Y}_t = \left(C^{-1}(q^{-1}) - 1\right) Y_t. \tag{6.24}$$

Those elementary reasonings lead to the optimal predictors corresponding to more complex models, as e.g.

Example 45 (ARMAX(1,1,1) model) *Consider the stochastic signals $\{U_t\}_t$ and $\{Y_t\}_t$ both taking values in $\mathbb{R}$ which follow a fixed but unknown system*

$$Y_t + aY_{t-1} = bU_{t-1} + V_t + cV_{t-1}, \ \forall t = \dots, 0, \dots, \tag{6.25}$$

where $\{V_t\}_t$ is zero mean white noise with $\mathbb{E}[V_t V_s] = \delta_{t,s}\lambda^2$. The parameter vector is $\theta = (a, b, c)^T \in \mathbb{R}^3$. Assume V_t is independent of U_s for all $s < t$, and hence the model allows for feedback from $\{Y_t\}_t$ to $\{U_t\}_t$. The output at time t satisfies

$$y_t = (-aY_{t-1} + bU_{t-1} + cV_{t-1}) + V_t, \ \forall t = \dots, 0, \dots, \tag{6.26}$$

and the two terms on the right hand side (r.h.s.) are independent by assumption. Now let $y_t^* \in \mathbb{R}$ be any number serving as a prediction, then one has for t that

$$\mathbb{E}[Y_t - y_t^*]^2 = \mathbb{E}[-aY_{t-1} + bU_{t-1} + cV_{t-1}]^2 + \mathbb{E}[V_t]^2 \geq \lambda^2, \tag{6.27}$$

giving as such a lower-bound to the prediction error variance. An optimal predictor $\{f_{t|t-1}(\theta)\}_t$ is one which achieves this lower-bound. This is the case for

$$f_{t|t-1}(\theta) = -aY_{t-1} + bU_{t-1} + cV_{t-1}. \tag{6.28}$$

The problem is of course that this predictor cannot be used as it stands as the term V_{t-1} is not measurable. However, it V_{t-1} may be reconstructed from past data as the residual in the previous iteration, and as such

$$\begin{aligned} f_{t|t-1}(\theta) &= -aY_{t-1} + bU_{t-1} + cV_{t-1} \\ &= -aY_{t-1} + bU_{t-1} + c\left(Y_{t-1} + aY_{t-2} - bU_{t-2} - cV_{t-2}\right) \\ &= -aY_{t-1} + bU_{t-1} + c\left(Y_{t-1} + aY_{t-2} - bU_{t-2}\right) - c^2\left(Y_{t-2} + aY_{t-3} - bU_{t-3} - cV_{t-3}\right) \\ &= \ldots \\ &= \sum_{i=1}^{t-1}(c-a)(-c)^{i-1}Y_{t-i} - a(-c)^{t-1}Y_0 + b\sum_{i=1}^{t-1}(-c)^{i-1}U_{t-i} - (-c)^t V_0. \end{aligned} \tag{6.29}$$

Under assumption that $|c| < 1$ the last term can be neglected for large t as it will have an exponentially decaying transient effect. Then we get a computable predictor. However we reorder terms to get a more practical expression as

$$f_{t|t-1}(\theta) = f(t-1|t-2, \theta) + (c-a)Y_{t-1} + bU_t, \tag{6.30}$$

which gives a simple recursion for computing the optimal prediction corresponding to past observations and the model parameter vector θ. We can compute the corresponding prediction error $\epsilon_t(\theta) = Y_t - f_{t|t-1}(\theta)$ similarly as

$$\epsilon_t(\theta) + c\epsilon_{t-1}(\theta) = Y_t + cY_{t-1} - \left((c-a)Y_{t-1} + bU_{t-1}\right) = Y_t + aY_{t-1} - bU_{t-1}, \tag{6.31}$$

for any $t = 2, \ldots, n$. This recursion needs an initial value $\epsilon_t(\theta)$ which is in general unknown and often set to 0. Observe that we need the statistical framework only for a definition of what an optimal predictor means exactly as in (6.27).

The above analysis can be stated more compactly using the polynomials,

Example 46 (An ARMAX(1,1,1), bis) *Consider $\{u_t\}_t$ and $\{y_t\}_t$ obeying the system*

$$(1 + aq^{-1})y_t = (bq^{-1})u_t + (1 + cq^{-1})e_t, \tag{6.32}$$

$$\begin{cases} f(t|t-1,\theta) & = \mathbf{H}^{-1}(q^{-1},\theta)\mathbf{G}(q^{-1},\theta)U_t + \left(1-\mathbf{H}^{-1}(q^{-1},\theta)\right)Y_t \\ \epsilon_t(\theta) = V_t & = \mathbf{H}^{-1}(q^{-1},\theta)(Y_t - \mathbf{G}(q^{-1},\theta)U_t). \end{cases} \tag{6.37}$$

Figure 6.2: The optimal expected least squares predictor for a general LTI.

for all t. Then

$$\begin{aligned} y_t &= \frac{(bq^{-1})}{(1+aq^{-1})}u_t + \frac{(1+cq^{-1})}{(1+aq^{-1})}e_t \\ &= \frac{(bq^{-1})}{(1+aq^{-1})}u_t + \frac{(c-a)q^{-1}}{(1+aq^{-1})}e_t + \frac{(1+aq^{-1})}{(1+aq^{-1})}e_t \\ &= \frac{(bq^{-1})}{(1+aq^{-1})}u_t + \frac{(c-a)q^{-1}}{(1+aq^{-1})}\left(\frac{(1+aq^{-1})y_t - (bq^{-1})u_t}{(1+cq^{-1})}\right) + e_t \\ &= \left(\frac{(bq^{-1})}{(1+aq^{-1})} - \frac{(c-a)q^{-1}}{(1+aq^{-1})}\frac{(bq^{-1})}{(1+cq^{-1})}\right)u_t + \frac{(c-a)q^{-1}}{(1+aq^{-1})}\frac{(1+aq^{-1})}{(1+cq^{-1})}y_t + e_t \\ &= \frac{(bq^{-1})}{(1+cq^{-1})}u_t + \frac{(c-a)q^{-1}}{(1+cq^{-1})}y_t + e_t, \end{aligned} \tag{6.33}$$

and again because of the noise terms e_t cannot be predicted from the past or the model parameters θ, the best any predictor can do is

$$f_{t|t-1}(\theta) = \frac{(bq^{-1})}{(1+cq^{-1})}u_t + \frac{(c-a)q^{-1}}{(1+cq^{-1})}y_t, \tag{6.34}$$

yielding the result. When working with filters in this way it is assumed that data are available from the infinite past. Since this wouldn't be the case in practical situations, one has to take into account transient effects before implementing thus predictors.

In general the derivation goes as follows. Assume the data $(y_1, y_2, \dots)$ and $(u_1, u_2, \dots)$ follows an LTI model where

$$y_{t+1} = H(q^{-1};\theta_0)u_{t+1} + G(q^{-1};\theta_0)e_{t+1}. \tag{6.35}$$

where

$$\begin{cases} H(q^{-1};\theta_0) = 1 + h_1q^{-1} + \dots + h_{m_h}q^{-m_h} \\ G(q^{-1};\theta_0) = 1 + g_1q^{-1} + \dots + g_{m_g}q^{-m_g}, \end{cases} \tag{6.36}$$

where $m_h \geq 1$ and $m_g \geq 1$ denote the orders of both monic polynomials, and $\theta_0 = (h_1, \dots, h_{m_h}, g_1, \dots, g_{m_g}) \in \mathbb{R}^{m_g+m_h-2}$. Then we face the question what value of

6.3 Statistical Analysis of PEM methods

The statistical analysis of PEM estimates starts off similar as in the least squares case. Assume that the observed signals satisfy a stochastic signal, or that

$$Y_t = G(q^{-1},\theta_0)U_t + G(q^{-1},\theta_0)D_t. \tag{6.38}$$

- The observed signals $\{u_t\}_{t=1}^n \subset \mathbb{R}$ and $\{y_t\}_{t=1}^n \subset \mathbb{R}$ are assumed to be samples from quasi-stationary stochastic processes $\{U_t\}_t$ and $\{Y_t\}_t$.
- The noise $\{D_t\}$ is assumed to be a stationary process with zero mean.
- The input is with high probability Persistently Exciting (PE) of sufficient order. that is $\mathbb{E}[V_t V_t^T] \succ 0$ where $V_t = (U_{t-1}, \dots, U_{t-d})$, and hence $\sum_{t=1}^n \mathbf{u}_t \mathbf{u}_t^T \succ 0$ where $\mathbf{u}_t = (u_{t-1}, \dots, u_{t-d}) \in \mathbb{R}^d$.
- The filters $G(q^{-1}, \theta)$ and $H(q^{-1}, \theta)$ are smooth (differentiable) functions of the parameters.

Then a PEM approach would be to solve for θ

$$V_n^* = \min_\theta V_n(\theta) = \frac{1}{2} \sum_{t=1} n \left(Y_t - (\mathbf{H}^{-1}(q^{-1}, \theta)\mathbf{G}(q^{-1}, \theta)U_t + \left(1 - \mathbf{H}^{-1}(q^{-1}, \theta)\right) Y_t)\right)^2. \tag{6.39}$$

This approach is in general different from a LS estimate. We also need the following assumption, namely that

- The Hessian $V_n''(\theta)$ is non-singular at least for the parameters θ close to the true parameters θ_0. This implies that no different parameters can solve the PEM objective asymptotically, and is thus in a sense closely related to Persistency of Excitation (PE).

The proof that the PEM would result in accurate estimates in that case is quite involved, but the main reasoning is summarized in Fig. (6.3). This result is then found strong enough also to quantify the variance of the estimates if n tends to infinity. Specifically we have that

$$\sqrt{n}(\theta_n - \theta_0) \sim \mathcal{N}(0, \mathbf{P}), \tag{6.40}$$

where

$$\mathbf{P} = \mathbb{E}[D_t^2]\mathbb{E}\left[\varphi_t(\theta_0)\varphi_t(\theta_0)^T\right]^{-1}, \tag{6.41}$$

and where

$$\varphi_t(\theta_0) = \frac{d\epsilon_t(\theta)}{d\theta}\Big|_{\theta=\theta_0}. \tag{6.42}$$

That is, the estimates are asymptotically unbiased and have asymptotic variance which is given by the Fisher information matrix based on the gradients of the prediction errors evaluated at the true parameters.

6.4 Computational Aspects

The practical difference of a PEM approach to a LS approach is that the solution is not given in closed form as the normal equations before did. In general, one has to resort to numerical optimization tools to solve the optimization problem. While good software implementations exists that can do this task very accurate, it is useful to write out some common approaches for getting a feeling how to interpret results from such a software.

The prototypical approach goes as follows. Let us abstract the problem as the following optimization problem over a vector $\theta \in \mathbb{R}^d$ as

$$\theta^* = \operatorname*{argmin}_\theta J(\theta), \tag{6.43}$$

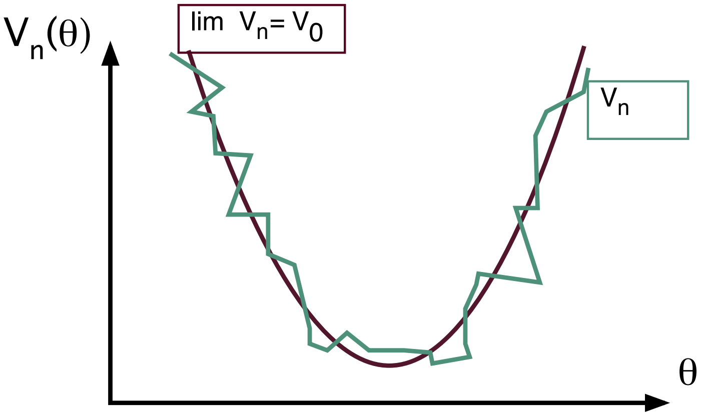


Figure 6.3: Schematic example of the PEM cost function. Here θ denotes the parameter vector to be minimized over. In case only a finite number of samples n is available, the PEM objective is a noisy version of the asymptotical loss $V_0(\theta) = \lim_{n\to\infty} V_n(\theta)$. Two results are stated then: (i) the true parameters θ_0 are the minimizer of the asymptotic objective function, and (ii) the asymptotic objective function $V_0\theta)$ differs not too much from the sample objective function $V_n(\theta)$ for *any* ('uniform) θ. Hence the minimizer θ_n to V_n is not too different from the true parameters.

where $J : \mathbb{R}^d \to \mathbb{R}$ is a proper cost function (i.e. a minimal value exists). We have an iterative regime, and in each iteration the previous estimate is refined slightly. Formally, we generate a sequence of vectors from an initial estimate $\theta^{(0)}$, obeying the recursion

$$\theta^{(k+1)} = \theta^{(k)} + \gamma \mathbf{b}(J, \theta^{(k)}), \tag{6.44}$$

where $\mathbf{b}(J, \theta^{(k)}) \in \mathbb{R}^d$ is a correction ('step') which refines the estimator. The algorithm then hopefully converges, in the sense that $\theta^{(k)} \to \theta^*$ when k increases. See Fig. (6.4.a) for an cost function in 2D, and Fig. (6.4.b) for an iterative algorithm at work in 1D. Now different algorithms specialize further using different quantities.

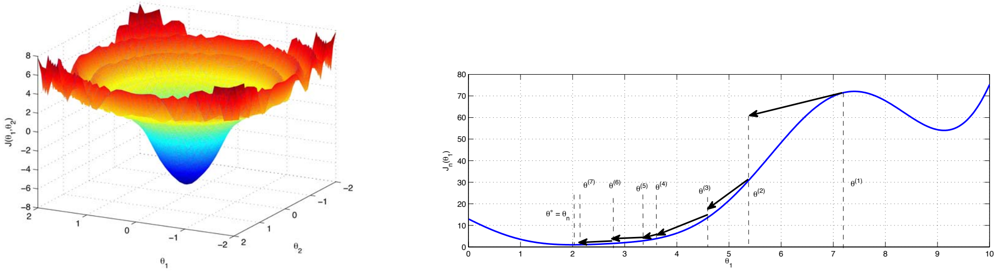


Figure 6.4: An example of an iterative optimization routine of J over a parameter θ.

The prototypical algorithm goes as follows. Here, the correction factor is determined by using a quadratic approximation of the cost function J at the current estimate $\theta^{(k)}$. The algorithm follows the recursion

$$\theta^{(k+1)} = \theta^{(k)} - \alpha_k \left(V_n''(\theta^{(k)}) \right)^{-1} V_n'(\theta^{(k)}), \tag{6.45}$$

where

- α_k is the step size, typically taken as a positive decreasing function of k.
- $V_n'(\theta_n) \in \mathbb{R}^d$ denotes the gradient of the cost function J at $\theta^{(k)}$.
- $V_n''(\theta_n) \in \mathbb{R}^{d\times d}$ denotes the Hessian matrix of the cost function J at $\theta^{(k)}$.

This algorithm is referred to as Newton-Raphson.

In the optimal point θ^* for the PEM problem one has a simplified approximative expression for the cost function $J(\theta^*)$ given as

$$V_n''(\theta^*) \approx \frac{2}{n} \sum_{i=1}^{n} \psi_t^T(\theta^*)\mathbf{H}\psi_t(\theta^*), \tag{6.46}$$

where $\mathbf{H}$ is a given matrix, and $\psi_t(\theta^*)$ equals the (first order) influence of the tth sample on the loss function of the PEM objective. Using this approximation in an iterative optimization gives the Gauss-Newton recursive algorithm given as

$$\theta^{(k+1)} = \theta^{(k)} - \alpha_k \left(\sum_{i=1}^{n} \psi_t^T(\theta^{(k)})\mathbf{H}\psi_t(\theta^{(k)}) \right)^{-1} \left(\sum_{i=1}^{n} \psi_t^T(\theta^*)\mathbf{H}\epsilon_t(\theta^{(k)}) \right), \tag{6.47}$$

where here $\epsilon_t(\theta^{(k)})$ denotes the prediction error on y_t using the past samples and the model with parameters $\theta^{(k)}$. When n is quite large both algorithms (6.48) and (6.45) behave quite similarly. But in general, the Newton-Raphson converges with quadratic speed $1/n^2$. The Gauss-Newton approach converges 'only' (super-) linear, but has the additional advantage that each iteration of the algorithm can be computed and stored much more efficiently.

If computational issues are even more important in the case at hand one may resort to a steepest descent algorithm, implementing the recursion

$$\theta^{(k+1)} = \theta^{(k)} - \alpha_k V_n'(\theta^{(k)}), \tag{6.48}$$

where $\theta^{(0)} \in \mathbb{R}^d$ is an appropriate initial estimate. Such algorithm is referred to as a steepest descent or gradient descent algorithm.

There are three important caveats when using such an approaches.

- The numerical software might be stuck in local minima, most likely giving parameter estimates which are not useful. This depends largely on the shape of the loss function. If V_n were (almost) a positive quadratic function, then there are no such local 'false' minima a numerical optimization routine could be stuck in. A simple way to circumvent this problem is to let the optimizer run based on a number of different starting points: if the optima are not mostly not equal, the problem of local minima is severely present. On the other hand, if the optimizers were equal for most of them, it is not too large a stretch to assume that the global minimizer were found successfully.

- In a number of cases the loss function V_n might not be differentiable at certain values for θ, or lead to large values, preventing the solver to converge properly. This is typically the case when the underlying dynamics are almost unstable, and slight deviations of the parameters might lead to unbounded predictions.

- The uncertainties computed by the software based on the discussion in the previous section is often not valid for finite n. Specifically, the derivation there assumes that θ_n is close enough to θ_0 in order to admit a quadratic expansion of the loss between either. This is clearly not valid θ_n were only a local minimizer. Hence, the estimated variances in software routines are valid conditioned on the fact that the optimizer worked well.

Chapter 7

Model Selection and Model Validation

”How can we be sure that an estimated model serves its future purpose well?”

Suppose I were a teacher, and you as a student had to work out a case-study of an identification experiment. The question now is how to verify wether you attempt is fruitful? Or are your efforts more fruitful than the ones of your colleagues? When is an approach not acceptable? The question of how to come up with a preference amongst different models for a given system is in practice more important than the actual method used for estimating one model: that is, even if your toolbox contains a method which gives a useless result on the task at hand, proper model selection will reveal this weak approach and prefer other tools. The aims of model selection are as follows

- Given a model class, parameter estimation techniques give you the best model in this class. Model selection on the other hand describes how to arrive at this model class at the first place. This amounts to the decision of (i) what sort of model structure suffices for our needs (e.g. ARX, BJ, State Space), and (ii) what model orders we would need (e.g. an ARX(1,1) or a ARX(100,10)).

- Which (stochastic) assumptions are reasonable to drive the analysis? Are the conditions valid under which the employed parameter estimation techniques ’work’ in the studied case?

- Is the model we have identified in some sense close to the real system? Or perhaps more realistically, is the model we have identified sufficient for our needs? We will refer to this aim as *model validation*.

Of course those objectives will be entangled in practice, and be closely related to the parameter estimation task at hand. It is as such no surprise that the same themes as explored in earlier chapters will pop up in a slightly different form. In fact, more recent investigations argue for a closer integration of parameter estimation and model selection problems at once, a theme which we will explore in the Chapter on nonlinear modeling.

The central theme of model selection and model validation will be to avoid the effect of ’overfitting’. This effect is understood as follows

Definition 25 (Overfitting) *If we have a large set of models in the model class with respect to the number of data, it might well be possible that the estimated model performs well on the data used to tune the parameters to, but that this model performs arbitrary bad in new cases.*

The following is a prototypical example

Example 47 (Fitting white noise) *Let $\{e_t\}_t$ be zero mean white noise with variance σ^2. Consider the system*

$$y_t = e_t, \ \forall t = -\infty, \dots, \infty, \tag{7.1}$$

and suppose we observe corresponding to y_t an input φ_t which is unrelated. Consider the case where the estimated model contains ('remembers') all mappings from observed inputs to corresponding outputs $\{\varphi_t \to y_t\}$. Then the estimated error on the set used for building up the model will be zero (i.e. it can be reconstructed exactly). The error on new data will be σ^2 in case $\ell(e) = e^2$.

The tasks of model selection and model validation is characterized by different trade-offs one has to make. A trade-off will in general arise as one has to pay a price for obtaining more accurate or complex models. Such trade-offs come into the form of variance of the estimates, complexity of the algorithms to be used, or even approaches which lead necessarily to unsuccessful estimates.

> "Essentially, all models are wrong, but some are useful", G.Box, 1987

This is a mantra that every person who deals with (finite numbers of) observed data implements in one way or another.

- *Bias-Variance Trade-off:* In general one is faced with a problem of recovering knowledge from a finite set of observations, referred to as an 'inverse problem'. If the model class is 'large' and contains the 'true' system, the bias of a technique might be zero, but the actual deviation of the estimated parameters from the true one might be large ('large variance'). On the other hand, if the model class is small, it might be easy to find an accurate estimate of the best candidate in this model class ('low variance'), but this one might be far off the 'true' system ('large bias). This intuition follows the bias-variance decomposition Lemma given as

  Lemma 8 (Bias-Variance Decomposition) *Let θ_n be an estimate of $\theta_0 \in \mathbb{R}^d$ using a random sample of size n, then*

  $$\mathbb{E}\|\theta_0 - \theta_n\|_2^2 = \mathbb{E}\|\theta_0 - \mathbb{E}[\theta_n]\|_2^2 + \mathbb{E}\|\mathbb{E}[\theta_n] - \theta_n\|_2^2, \tag{7.2}$$

  where $\mathbb{E}\|\theta_0 - \mathbb{E}[\theta_n]\|_2$ is often referred to as the bias, *and $\mathbb{E}\|\mathbb{E}[\theta_n] - \theta_n\|_2^2$ as the variance associated to the estimator θ_n.*

  This result follows directly by working out the squares as

  $$\mathbb{E}\|\theta_0-\theta_n\|_2^2 = \mathbb{E}\|(\theta_0-\mathbb{E}[\theta_n])+(\mathbb{E}[\theta_n]-\theta_n)\|_2^2 = \mathbb{E}\|(\theta_0-\mathbb{E}[\theta_n])\|_2^2+\mathbb{E}\|(\mathbb{E}[\theta_n]-\theta_n)\|_2^2+2\mathbb{E}[(\theta_0-\mathbb{E}[\theta_n])^T(\mathbb{E}[\theta_n]-\theta_n)], \tag{7.3}$$

  and since $\mathbb{E}[(\theta_0-\mathbb{E}[\theta_n])^T(\mathbb{E}[\theta_n]-\theta_n)]$ equals $(\theta_0-\mathbb{E}[\theta_n])^t\mathbb{E}[\mathbb{E}[\theta_n]-\theta_n] = ((\theta_0-\mathbb{E}[\theta_n]))^T 0_d = 0$, since θ_0 and $\mathbb{E}[\theta_n]$ are deterministic quantities.

- *Algorithmic Issues:* In practice, when the data can only point to a specific model in a model class with large uncertainty, the parameter estimation problem will often experience algorithmic or numeric problems. Of course, suboptimal implementations could give problems even if the problem at hand is not too difficult. Specifically, in case one has to use heuristics, one might want to take precautions against getting stuck in 'local optima', or algorithmic 'instabilities'.

The theory of algorithmic, stochastic or learning complexity studies the theoretical link between either, and what a 'large' or 'small' model class versus a 'large' number of observations mean. In our case it is sufficient to focus on the concept of Persistency of Excitation (PE), that is, a model class is not too large w.r.t. the data if the data is PE of sufficient order. This notion is in turn closely related to the condition number of the sample covariance matrix, which will directly affect numeric and algorithmic properties of methods to be used for estimation.

7.1 Model Validation

Let us first study the 'absolute' question: 'is a certain model sufficient for our needs?' Specifically, we will score a given model with a measure how well it serves its purpose. This section will survey some common choices for such scoring functions.

7.1.1 Cross-validation

A most direct, but till today a mostly unrivaled choice is to assess a given model on how well it performs on 'fresh' data. Remember that the error on the data used for parameter estimation might be spoiled by overfitting effects, that is, the model might perform well on the specific data to which the estimated model is tuned but can perform very badly in new situations. It is common to refer to the performance of the model on the data used for parameter estimation as the *training performance*. Then, the training performance is often a biased estimate of the actual performance of the model. A more accurate estimate of the performance of the model can be based on data which is in a sense independent of the data used before.

The protocol goes as follows

1. Set up a first experiment and collect signals $\{u_t\}_{t=1}^{n}$ and $\{y_t\}_{t=1}^{n}$ for $n > 0$;

2. Estimate parameters $\hat{\theta}_n$ (model) based on those signals;

3. Set up a new experiment and collect signals $\{u_t^v\}_{t=1}^{n^v}$ and $\{y_t^v\}_{t=1}^{n^v}$ for $n^v > 0$;

4. Let $\ell : \mathbb{R} \to \mathbb{R}$ as before a loss function. Score the model as

$$V_{n^v}^v(\hat{\theta}_n) = \frac{1}{n^v} \sum_{t=1}^{n^v} \ell\left(y_t^v - f_{\hat{\theta}_n,t}\right), \tag{7.4}$$

where $f_{\hat{\theta}_n,t}$ is the shorthand notation for the predictor based on the past signals $\{y_s^v\}_{s<t}$ and $\{u_s^v\}_{s\leq t}$, and the parameters $\hat{\theta}_n$.

The crucial bit is that inbetween the two experiments, the studied system is left long enough so that the second experiment is relatively 'independent' from what happened during the first one. This issue becomes more important if we have only one set of signals to perform estimation and validation on. A first approach would be to divide the signals in two non-overlapping, consecutive blocks of length (usually) 2/3 and 1/3. The first one is then used for parameter estimation ('training'), the second black is used for model validation. If the blocks are small with respect to the model order (time constants), transient effects between the training block and validation block might affect model validation. It is then up to the user to find intelligent approaches to avoid such effects, e.g. by using an other split training-validation.

7.1.2 Information Criteria

In case cross-validation procedures become too cumbersome, or lead to unsatisfactory results e.g. as not enough data is available, one may resort to the use of an Information Criterion (IC) instead. An information criterion in general tries to correct analytically for the overfitting effect in the training performance error. They come in various flavors and are often relying on statistical assumptions on the data. The general form of such an IC goes as follows

$$w_n = V_n(\theta_n)\left(1 + \beta(n, d)\right), \tag{7.5}$$

where $\beta : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^+$ is a function of the number of samples n and the number of free parameters d. In general β should decrease with growing n, and increase with larger d. Moreover β should tend to zero if $n \to \infty$. An alternative general form is

$$\tilde{w}_n = n \log V_n(\theta_n) + \gamma(n, d), \tag{7.6}$$

where $\gamma : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ penalizes model structures with larger complexity. The choice of $\gamma(n, d) = 2d$ gives Akaike's IC (AIC):

$$\underset{n}{\mathrm{AIC}}(d) = n \ln V_n(\theta_n) + 2d, \tag{7.7}$$

where θ_n is the LS/ML/PEM parameter estimate. It is not too difficult to see that the form w_n and $\tilde{w}_n$ are approximatively equivalent for increasing n.

Several different choices for γ have appeared in the literature, examples of which include the following:

(FPE): The *Finite Prediction Error* (FPE) is given as

$$\mathrm{FPE}(d) = V_n(\theta_n)\frac{1 + d/n}{1 - d/n}, \tag{7.8}$$

which gives an approximation on the prediction error on future data. This criterium can be related to AIC's and the significance test for testing the adequacy of different model structures.

(BIC): The *Bayesian Information Criterion* (BIC) is given as

$$\mathrm{BIC}(d) = -2 \ln V_n(\theta_n) + d \ln(n), \tag{7.9}$$

where θ_n is the ML parameter estimate, and the correction term is motivated from a Bayesian perspective.

In general, an individual criterion can only be expected to give consistent model orders under strong assumptions. In practice, one typically choses a model which minimizes approximatively all criteria simultaneously.

7.1.3 Testing

In case a stochastic setup were adopted, a model comes with a number of stochastic assumptions. If those assumptions hold, the theory behind those methods ensures (often) that the estimates are good (i.e. efficient, optimal in some metric, or leading to good approximations). What is left for the practitioner is that we have to verify the assumptions for the task at hand. This can be approached using statistical significance testing. If such a test gave evidence that the assumptions one adopted do not hold in practice, it is only reasonable to go back to the drawing table or the library. The basic ideas underlying significance testing go as follows.

> 'Given a statistical model, the resulting observations will follow a derived statistical law. If the actual observations are not following this law, the assumed model cannot be valid.'

This inverse reasoning has become the basis of the scientific method. Statistics being stochastic is not about deterministic laws, but will describe how the observations would be like most probable. As such it is not possible to refute a statistical model completely using only a sample, but merely to accumulate evidence that it is not valid. Conversely, if no such evidence for the contrary is found in the data, it is valid practice to go ahead as if the assumptions were valid. Note that this does not mean that the assumptions are truly valid!

The translation of this reasoning which is often used goes as follows

$$Z_1, \ldots, Z_n \sim \mathcal{H}_0 \Rightarrow T(Z_1, \ldots, Z_n) \sim \mathcal{D}_n(\beta, \mathcal{H}_0), \tag{7.10}$$

where

n: The number of samples, often assumed to be large (or $n \to \infty$)

Z_i: An observation modeled as a random variable

$\sim$: means here '*is distributed as* '

$\Rightarrow$: Implies

$\mathcal{H}_0$: or the *null distribution*

$T(\ldots)$: or the *statistic*, which is in turn a random variable.

$\mathcal{D}_n(\beta, \mathcal{H}_0)$: or the *limit distribution* of the statistic under the null distribution $\mathcal{H}_0$. Here β denotes a parameter of this distribution.

Typically, we will have asymptotic null distributions (or 'limit distributions'), characterized by a PDF f. That is, assuming that n tends to infinity, $\mathcal{D}_n(\beta, \mathcal{H}_0)$ tends to a probability law with PDF $f_{\beta,\mathcal{H}_0}$. Shortly, we write that

$$T(Z_1, \ldots, Z_n) \to f_{\beta,\mathcal{H}_0}. \tag{7.11}$$

Now, given a realization $z_1, \ldots, z_n$ of a random variable $Z'_1, \ldots, Z'_n$, the statistic $t_n = T(z_1, \ldots, z_n)$ can be computed for this actual data. A statistical hypothesis test checks wether this value t_n is likely to occur in the theoretical null distribution $\mathcal{D}_n(\beta, \mathcal{H}_0)$. That is, if the value t_n were rather unlikely to occur under that model, one must conclude that the statistical model which were assumed to underly $Z_1, \ldots, Z_n$ are also not too likely. In such a way one can build up *evidence* for the assumptions **not** to be valid. Each test comes with its associated $\mathcal{H}_0$ to test, and with a corresponding test statistic. The derivation of the corresponding limit distributions is often available in reference books and implemented in standard software packages.

A number of classical tests are enumerated:

F-test: The one-sample z-test checks wether a univariate sample $\{y_i\}_{i=1}^n$ originating from a normal distribution with given variance has mean value zero or not. The null-hypothesis is that the sample is sampled i.i.d. from zero mean Gaussian distribution with variance σ^2. The test statistic is computed as

$$T_n\left(\{y_i\}_{i=1}^n\right) = \frac{\sum_{i=1}^n y_i}{\sigma\sqrt{n}}, \tag{7.12}$$

and when the null-distribution were valid it would be distributed as a standard normal distribution, i.e. $T_n \to \mathcal{N}(0,1)$. In other words, if $Y_1, \ldots, Y_n$ are i.i.d. samples from f_0, then $f_T \to \mathcal{N}(0,1)$ when n tends to be large (in practice $n > 30$ is already sufficient for the asymptotic results to kick in!). Based on this limit distribution one can reject the null-hypothesis with a large probability if the test-statistic computed on the observed sample would have a large absolute value.

χ^2-test: Given is a set of n i.i.d. samples $\{y_i\}_{i=1}^n$ following a normal distribution. The standard χ^2-test checks wether this normal distribution has a pre-specified standard deviation σ_0. The test statistic is given as

$$T_n\left(\{y_i\}_{t=1}^n\right) = \frac{(n-1)s_n^2}{\sigma_0^2}, \tag{7.13}$$

where the sample variance is computed as $\frac{1}{n}\sum_{i=1}^n (y_i - m_n)^2$, and the sample mean is given as $\frac{1}{n}\sum_{i=1}^n y_i$. Then the limit distribution of this statistic under the null-distribution is known to follow a χ^2-distribution with $n-1$ degrees of freedom, the PDF and CDF of this distribution is computed in any standard numerical software package.

Example 48 (Lady Tasting Tea) *(Wikipedia) - The following example is summarized from Fisher, and is known as the Lady tasting tea example. Fisher thoroughly explained his method in a proposed experiment to test a Lady's claimed ability to determine the means of tea preparation by taste. The article is less than 10 pages in length and is notable for its simplicity and completeness regarding terminology, calculations and design of the experiment. The example is loosely based on an event in Fisher's life. The Lady proved him wrong.*

1. *The null hypothesis was that the Lady had no such ability.*
2. *The test statistic was a simple count of the number of successes in 8 trials.*
3. *The distribution associated with the null hypothesis was the binomial distribution familiar from coin flipping experiments.*
4. *The critical region was the single case of 8 successes in 8 trials based on a conventional probability criterion (< 5%).*
5. *Fisher asserted that no alternative hypothesis was (ever) required.*

If and only if the 8 trials produced 8 successes was Fisher willing to reject the null hypothesis effectively acknowledging the Lady's ability with > 98% confidence (but without quantifying her ability). Fisher later discussed the benefits of more trials and repeated tests.

7.1.4 Testing LTI Models

In the context of testing the results of a model of a linear system, the following tests are often used. The following setup is typically considered. Given timeseries $\{u_t\}_t$ and $\{Y_t\}_t$ as well as (estimated) parameters $\hat\theta$ of a model structure $\mathcal{M}$. Then, one can compute the corresponding optimal predictions $\{\hat y\}_t$ and the prediction errors (or residuals) $\{\hat\epsilon_t\}_t$ corresponding to this estimate $\hat\theta$. Now, a common statistical model assumes that $\epsilon_t = Y_t - \hat y_t(\theta_0)$ (the innovations) is zero mean, white stochastic noise. Then, if $\theta_0 \approx \hat\theta$, also the estimated innovations $\{\hat\epsilon_t\}_t$ would be similar to zero mean Gaussian noise. A statistical test could then be used to collect evidence for the $\{\hat\epsilon_t\}_t$ not too be a white noise sequence, hence implying that the parameter estimate is not adequate. Two typical tests which check the whiteness of a sequence $\{\epsilon_t\}_{t=1}^n$ go as follows.

- (Portmanteau):
$$T_n(\{\epsilon_i\}) = \frac{n}{\hat r_0^T} \sum_{\tau=1^m} \hat r_\tau^2, \tag{7.14}$$
where the sample auto-covariances are computed as $\hat r_\tau = \frac{1}{n}\sum_{i=1}^{n-\tau} \epsilon_i \epsilon_{i+\tau}$. This test statistic follows a χ^2 distribution with m degrees of freedom, if $\{\epsilon_t\}$ where indeed samples from a zero mean, white noise sequence. So, if the test-statistic computed using the estimated innovations $T_n\left(\{\hat\epsilon_t\}\right)$ were really large, evidence is collected for rejecting the null-hypothesis - the estimate $\hat\theta$ were close to the true value θ_0. This reasoning can be quantified exactly using the above expressions.

- (Normal): A simple test for checking where a auto-covariance at a lag $\tau > 0$ is zero based on a sample of size n is given by the statistic
$$T_n(\{\epsilon_i\}) = \sqrt{n}\frac{\hat r_\tau^2}{\hat r_0^2}, \tag{7.15}$$
with a distribution under the null-hypothesis (i.e. $r_\tau = 0$) which tends to a normal distribution with unit variance and zero mean when $n \to \infty$.

- (Cross-Correlation Test:) Now let us shift gears. Assume that the input timeseries $\{U_t\}_t$ is stochastic as well. If the model were estimated adequate, no dynamics are left in the residuals. Hence, it makes sense to test wether there are cross-correlations left between input signals and residuals. This can be done using the statistic
$$T_n(\{\epsilon_i\}) = \sqrt{n}\hat{\mathbf{r}}^T(\hat r_0^2 \hat R_u)^{-1}\hat{\mathbf{r}}, \tag{7.16}$$
where for given m and τ' one has the sample quantities
$$\begin{cases} \hat r = \left(\hat r^{u,\epsilon}_{\tau'+1}, \dots, \hat r^{u,\epsilon}_{\tau'+m}\right)^T \\ \hat r^{u,\epsilon}_\tau = \frac{1}{n}\sum_{t=1-\min(0,\tau)}^{n-\max(\tau,0)} U_t \epsilon_{t+\tau} \\ \hat R_u = \frac{1}{n}\sum_{t=m+1}^{n} U'_t {U'_t}^T \\ U'_t = \left(U_{t-1}, \dots, U_{t-m}\right)^T . \end{cases} \tag{7.17}$$
This test statistic has a distribution under the null-hypthesis (i.e. $r^{\epsilon u} = 0$) which tends to a χ^2 distribution with m degrees of freedom when $n \to \infty$.

- (Sign Test) Rather than looking at second moments, it was argued to look at different properties of the residuals. For example one could calculate the number of flips of signs in consequent values. This lead to the statistic

$$T_n(\{\epsilon_i\}) = \frac{1}{\sqrt{n/2}}\left(\sum_{t=1}^{n-1} I(\epsilon_t \epsilon_{t+1} < 0) - \sqrt{n/2}\right), \tag{7.18}$$

  with $I(z)$ equal to one if z is true, and zero otherwise. This statistic has a (sample) distribution under the null-hypthesis (i.e. $\{\epsilon_i\}$ were zero mean white) which tends to a normal distribution with unit variance and zero mean when $n \to \infty$.

Evidently, not all test are equally muscular. When applying a test, there i always a chance the the null-hypothesis were rejected even when it were actually true, or vice versa. The former risk - the so-called type 1 risk) of false positives is captured by the threshold α used in the test. In general, when decreasing this risk factor, one necessarily increases the risk of a false negative. However, this risk is much more difficult to characterize, and requires a proper characterization of the alternative hypothesis.

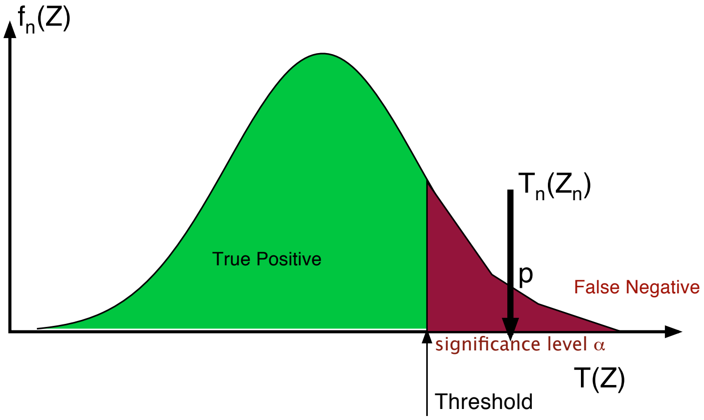


Figure 7.1: Schematic illustration of a statistical hypothesis test. A test is associated with a statistic T_n computed on a sample of size n. One is then after accepting or rejecting a null-hypothesis $\mathbf{H}_0$ underlying possibly the sample. If the sample follows indeed $\mathbf{H}_0$, one can derive theoretically the corresponding null-distribution of T_n. If the statistic computed on the sample is rather atypical under this null-distribution, evidence is found that $\mathbf{H}_0$ is not valid. If the statistic computed on the sample is likely under this null-distribution, no evidence is found to reject $\mathbf{H}_0$. The exact threshold where to draw a distinction between either to conclusions is regulated by a significance level $0 < \alpha < 1$.

7.2 Model Class Selection

Let us study the 'relative' question: 'which of two models is most useful for our needs?' The approach will be to score both models with a given model validation criterion, and to prefer the one with the best score.

The classical way to compare two candidate models are the so-called 'goodness-of-fit hypothesis tests'. Perhaps the most common one is the likelihood ratio test. Here we place ourselves again in a proper stochastic framework. Let $\hat{\theta}_1$ be the maximum likelihood estimator of the parameter θ_0 in the model structure M_1, and let $\hat{\theta}_2$ be the maximum likelihood estimator of the parameter θ'_0 in the model structure M_2. We can evaluate the likelihood function $L_{Z_n}(\hat{\theta}_1)$ in the sample Z_n under model M_1 with parameter $\hat{\theta}_1$, as well as the likelihood function $L'_{Z_n}(\hat{\theta}_2)$ in the sample Z_n under model M_2 with parameter $\hat{\theta}_2$. Then we can compute the test statistic

$$T_n(Z_n) = \frac{L_{Z_n}(\hat{\theta}_1)}{L'_{Z_n}(\hat{\theta}_1)}. \tag{7.19}$$

Let $\mathcal{M}_1$ and $\mathcal{M}_2$ be two different model structures, such that $\mathcal{M}_1 \subset \mathcal{M}_2$. That is, they are hierarchically structured. For example, both are ARX model structures but the orders of $\mathcal{M}_1$ are lower than the orders of $\mathcal{M}_2$. A related approach based on the loss functions is given makes use of the following test statistic

$$T_n(Z_n) = \frac{V_n^1 - V_n^2}{V_n^2}, \tag{7.20}$$

where V_n^1 is the minimal squared loss of the corresponding LSE in $\mathcal{M}_1$, and V_n^2 is the minimal squared loss of the corresponding LSE in $\mathcal{M}_2$. Lets consider the null-hypothesis $\mathbf{H}_0$ that the model structure $\mathcal{M}_1$ were describing the observations adequately enough. Then it is not too difficult to derive that the sample distribution under the null-hypothesis tends to a χ^2 distribution with degrees of freedom $|\theta_2|_0 - |\theta_1|_0$ (i.e. the difference in number of parameters of either model). This test is closely related to the F-test introduced earlier.

A more pragmatic approach is to use the validation criteria as follows. Consider a collection of estimated models in different model structures. For example, fix the model sort, and constructs estimates of the parameters for various model orders. Then you can score each candidate using an appropriate validation criterion. The model structure (order) leading to the best score is then obviously preferred. This approach is typically taken for model design parameters with no direct physical interpretation.

7.2.1 A Priori Considerations

Model selection is but a formalization of the experience of model building. One could easily imagine that a person which is experienced in the field does not need to implement one of the above methods to guide his design decisions. But in order to assist such an expert, a fe tricks of the trade are indispensable, a few of which are enumerated here.

- Plot the data in different ways.
- Look out for trends or seasonal effects in the data.
- Is an LTI model sufficient for our needs?
- Look at the spectra of the data.
- Try to capture and to explain the noise process. Where does the noise come from in your case study? Consequently, what is

- Try a naively simple model, and try to figure out how you see that this is not sufficient for you. Again, if you manage to formalize the exact goal you're working to, you're halfway the modeling process.

Chapter 8

Recursive Identification

> "Given a current estimated model and a new observation, how should we update this model in order to take this new piece of information into account?"

In many cases it is beneficial to have a model of the system available online while the system is in operation. The model should then be based on the observations up till the current time. A naive way to go ahead is to use all observations up to t to compute an estimate $\hat{\theta}_t$ of the system parameters. In recursive identification methods, the parameter estimates are computed recursively over time: suppose we have an estimate $\hat{\theta}_{t-1}$ at iteration $t-1$, then recursive identification aims to compute a new estimate $\hat{\theta}_t$ by a 'simple modification' of $\hat{\theta}_{t-1}$ when a new observation becomes available at iteration t. The counterpart to online methods are the so-called offline or batch methods in which all the observations are used simultaneously to estimate the model.

Recursive methods have the following general features:

- They are a central part in adaptive systems where the next action is based on the latest estimated parameters. Typical examples are found in adaptive control or adaptive filtering applications.

- Memory and computational requirements at any timestep has to be modest. Specifically, one often requires that both are independent to the length of the history at any timestep.

- They are often applied to real-time settings, where the 'true' underlying parameters are changing over time (i.e. tracking applications).

- They are often used for fault detection systems. Here one wants to detect when the observed signals or the underlying system differs significantly from what one would associate from a normal operation modus.

In general, the techniques go with the same statistical properties as their counterparts in 'batch' setting. For example, the RLS gives consistent estimates under the conditions as discussed in Section 5.3. That is, the discussion on the recursive estimators is often concerned with computational issues.

8.1 Recursive Least Squares

Let us start this section with perhaps the simplest application possible, nevertheless introducing ideas.

Example 49 (RLS for Estimating a Constant) *Given the following system*

$$y_t = \theta_0 + e_t,\ \forall t = 1, 2, \ldots. \tag{8.1}$$

In chapter 2, example 1 we derive how the least squares estimate of θ_0 *using the first* t *observations is given as the arithmetic (sample) mean, i.e.*

$$\hat{\theta}_t = \frac{1}{t}\sum_{i=1}^{t} y_i. \tag{8.2}$$

Now it is not too difficult to rewrite this in a recursive form.

$$\hat{\theta}_t = \frac{1}{t}\left(\sum_{i=1}^{t-1} y_i + y_t\right) = \frac{1}{t}\left((t-1)\hat{\theta}_{t-1} + y_t\right) = \hat{\theta}_{t-1} + \frac{1}{t}\left(y_t - \hat{\theta}_{t-1}\right). \tag{8.3}$$

This result is quite appealing: the new estimate $\hat{\theta}_t$ *equals the previous estimate* $\hat{\theta}_{t-1}$ *plus a small correction term. The correction term is proportional to the deviation of the prediction* $\hat{\theta}_{t-1}$ *and the observation* y_t*. Moreover the correction term is weighted by the term* $\frac{1}{t}$*, which implies that the magnitude of the correction will decrease in time. Instead the estimate* $\hat{\theta}_{t-1}$ *will become more reliable. The variance of* $\hat{\theta}_t$ *becomes*

$$P_t = \frac{1}{t}, \tag{8.4}$$

which can in turn be computed recursively as

$$P_t = \frac{1}{P_{t-1}^{-1} + 1} = \frac{P_{t-1}}{1 + P_{t-1}}. \tag{8.5}$$

In order to generalize the result, we need the following well-known matrix properties.

Lemma 9 (Matrix Inversion Lemma) *Let* $\mathbf{Z} \in \mathbb{R}^{d\times d}$ *be a positive definite matrix with unique inverse* $\mathbf{Z}^{-1}$*, and let* $\mathbf{z} \in \mathbb{R}^d$ *be any vector, then*

$$\mathbf{Z}_+^{-1} = (\mathbf{Z} + \mathbf{z}\mathbf{z}^T)^{-1} = \mathbf{Z}^{-1} - \frac{\mathbf{Z}^{-1}\mathbf{z}\mathbf{z}^T\mathbf{Z}^{-T}}{1 + \mathbf{z}^T\mathbf{Z}^{-1}\mathbf{z}}, \tag{8.6}$$

where $\mathbf{Z}_+^{-1} = \mathbf{Z} + \mathbf{z}\mathbf{z}^T$*.*

In words, the inverse of a matrix with a rank-one update can be written in closed form using the inverse of the matrix and a small correction. *Proof:* The proof is instrumental.

$$\begin{aligned}
(\mathbf{Z} + \mathbf{z}\mathbf{z}^T)\left(\mathbf{Z}^{-1} - \frac{\mathbf{Z}^{-1}\mathbf{z}\mathbf{z}^T\mathbf{Z}^{-1}}{1 + \mathbf{z}^T\mathbf{Z}^{-1}\mathbf{z}}\right) &= I_d - \mathbf{Z}\left(\frac{\mathbf{Z}^{-1}\mathbf{z}\mathbf{z}^T\mathbf{Z}^{-1}}{1 + \mathbf{z}^T\mathbf{Z}^{-1}\mathbf{z}}\right) + \mathbf{z}\mathbf{z}^T\mathbf{Z}^{-1} - (\mathbf{z}\mathbf{z}^T)\left(\frac{\mathbf{Z}^{-1}\mathbf{z}\mathbf{z}^T\mathbf{Z}^{-1}}{1 + \mathbf{z}^T\mathbf{Z}^{-1}\mathbf{z}}\right) \\
&= I_d - \left(\frac{\mathbf{z}\mathbf{z}^T\mathbf{Z}^{-1}}{1 + \mathbf{z}^T\mathbf{Z}^{-1}\mathbf{z}}\right) + \left(\frac{\mathbf{z}\mathbf{z}^T\mathbf{Z}^{-1}}{1 + \mathbf{z}^T\mathbf{Z}^{-1}\mathbf{z}}\right)(1 + \mathbf{z}\mathbf{Z}^{-1}\mathbf{z}) - \left(\frac{\mathbf{z}\mathbf{z}^T\mathbf{Z}^{-1}\mathbf{z}\mathbf{z}^T\mathbf{Z}^{-1}}{1 + \mathbf{z}^T\mathbf{Z}^{-1}\mathbf{z}}\right) \\
&= I_d + \left(\frac{\mathbf{z}\mathbf{z}^T\mathbf{Z}^{-1}}{1 + \mathbf{z}^T\mathbf{Z}^{-1}\mathbf{z}}\right)(\mathbf{z}\mathbf{z}^T\mathbf{Z}^{-1}) - \left(\frac{\mathbf{z}\mathbf{z}^T\mathbf{Z}^{-1}\mathbf{z}\mathbf{z}^T\mathbf{Z}^{-1}}{1 + \mathbf{z}^T\mathbf{Z}^{-1}\mathbf{z}}\right).
\end{aligned}$$

Now, note that $(\mathbf{z}^T\mathbf{Z}^{-1}\mathbf{z})$ is a scalar, and thus

$$= I_d + \left(\frac{\mathbf{z}\mathbf{z}^T\mathbf{Z}^{-1}}{1+\mathbf{z}^T\mathbf{Z}^{-1}\mathbf{z}}\right)(\mathbf{z}^T\mathbf{Z}^{-1}\mathbf{z}) - \left(\frac{(\mathbf{z}^T\mathbf{Z}^{-1}\mathbf{z})\mathbf{z}\mathbf{z}^T\mathbf{Z}^{-1}}{1+\mathbf{z}^T\mathbf{Z}^{-1}\mathbf{z}}\right) = I_d, \tag{8.7}$$

as desired.

□

The previous example serves as a blueprint of the Recursive Least Squares (RLS) algorithm, which we now will develop in full. Given a model for the observations $\{(\mathbf{x}_t, y_t)\}_t \subset \mathbb{R}^{d\times 1}$ given as

$$y_t = \theta_0^T \mathbf{x}_t + e_t,\ \forall t = 1, 2, \ldots, \tag{8.8}$$

where $\theta_0 \in \mathbb{R}^d$ and the terms $\{e_t\}_t$ are the corresponding residuals. Then chapter 2 learns us that the LS solution based on the observations $x_i : i = 1, \ldots, t$ will be given as the solution to the normal equations

$$\left(\sum_{i=1}^t \mathbf{x}_i\mathbf{x}_i^T\right)\hat{\theta}_t = \left(\sum_{i=1}^t y_i\mathbf{x}_i\right). \tag{8.9}$$

Assume for now that the solution $\hat{\theta}_t$ is unique, i.e. the matrix $R_t = \left(\sum_{i=1}^t \mathbf{x}_i\mathbf{x}_i^T\right)$ can be inverted. Since trivially one has

$$\mathbf{R}_{t-1} = \mathbf{R}_t - \mathbf{x}_t\mathbf{x}_t^T, \tag{8.10}$$

if follows that

$$\begin{aligned} \hat{\theta}_t &= \mathbf{R}_t^{-1}\left(\sum_{i=1}^{t-1} y_i\mathbf{x}_i + y_t\mathbf{x}_t\right) \\ &= \mathbf{R}_t^{-1}\left(\mathbf{R}_{t-1}\hat{\theta}_{t-1} + y_t\mathbf{x}_t\right) \\ &= \hat{\theta}_{t-1} + \mathbf{R}_t^{-1}\left(y_t\mathbf{x}_t - (\mathbf{x}_t\mathbf{x}_t^T)\hat{\theta}_{t-1}\right) \\ &= \hat{\theta}_{t-1} + \mathbf{R}_t^{-1}\mathbf{x}_t\left(y_t - \hat{\theta}_{t-1}^T\mathbf{x}_t\right), \end{aligned} \tag{8.11}$$

and in summary

$$\begin{cases} \epsilon_t = (y_t - \mathbf{x}_t^T\hat{\theta}_{t-1}) \\ \mathbf{K}_t = \mathbf{R}_t^{-1}\mathbf{x}_t \\ \hat{\theta}_t = \hat{\theta}_{t-1} + \mathbf{K}_t\epsilon_t. \end{cases} \tag{8.12}$$

Here the term ϵ_t will be interpreted as the prediction error: it is the difference between the observed sample y_t and the predicted value $\mathbf{x}_t^T\hat{\theta}_{t-1}$. If ϵ_t is 'small', the estimate $\hat{\theta}_{t-1}$ is good and should not be modified much. The matrix $\mathbf{K}_t$ is interpreted as the weighting or 'gain' matrix characterizing how much each element of the parameter vector $\hat{\theta}_{t-1}$ should be modified by ϵ_t.

The RLS algorithm is completed by circumventing the matrix inversion of $\mathbf{R}_t$ in each timestep. Hereto, we can use the matrix inversion Lemma.

$$\mathbf{R}_t^{-1} = \mathbf{R}_{t-1}^{-1} - \frac{\mathbf{R}_{t-1}^{-1}\mathbf{x}_t\mathbf{x}_t^T\mathbf{R}_{t-1}^{-1}}{1+\mathbf{x}_t^T\mathbf{R}_{t-1}^{-1}\mathbf{x}_t}. \tag{8.13}$$

Note that as such we substitute the matrix inversion by a simple scalar division.

$$\mathbf{K}_t = \mathbf{R}_t^{-1}\mathbf{x}_t = \mathbf{R}_{t-1}^{-1}\mathbf{x}_t - \frac{\mathbf{R}_{t-1}^{-1}\mathbf{x}_t(\mathbf{x}_t^T\mathbf{R}_{t-1}^{-1}\mathbf{x}_t)}{1+\mathbf{x}_t^T\mathbf{R}_{t-1}^{-1}\mathbf{x}_t} = \left(\frac{1}{1+\mathbf{x}_t^T\mathbf{R}_{t-1}^{-1}\mathbf{x}_t}\right)\mathbf{R}_{t-1}^{-1}\mathbf{x}_t. \tag{8.14}$$

Given initial values $\mathbf{R}_0^{-1}$ and $\hat\theta_0$, the final RLS algorithm can as such be written as

$$\begin{cases} \epsilon_t = (y_t - \mathbf{x}_t^T\hat\theta_{t-1}) \\ \mathbf{P}_t = \mathbf{P}_{t-1} - \frac{\mathbf{P}_{t-1}\mathbf{x}_t\mathbf{x}_t^T\mathbf{P}_{t-1}}{1+\mathbf{x}_t^T\mathbf{P}_{t-1}\mathbf{x}_t} \\ \mathbf{K}_t = \mathbf{P}_t\mathbf{x}_t = \left(\frac{1}{1+\mathbf{x}_t^T\mathbf{P}_{t-1}\mathbf{x}_t}\right)\mathbf{P}_{t-1}\mathbf{x}_t \\ \hat\theta_t = \hat\theta_{t-1} + \mathbf{K}_t\epsilon_t, \end{cases} \tag{8.15}$$

where we use $\mathbf{P}_t = \mathbf{R}_t^{-1}$ for any t. For efficiency reasons, one can We will come back to the important issue on how to choose the initial values $\mathbf{P}_0$ and $\hat\theta_0$ in Subsection 8.1.2.

8.1.1 Real-time Identification

This subsection presents some ideas which are useful in case the RLS algorithm is applied for tracking time-varying parameters. This is for example the case when the 'true' parameter vector θ_0 varies over time, and are as such denoted as $\{\theta_{0,t}\}_t$. This setting is referred to as the *real-time identification* setting. There are two common approaches to modify the RLS algorithm to handle such case: (i) use of a forgetting factor (this subsection); (ii) use of a Kalman filter as a parameter estimator (next subsection).

The 'forgetting factor' approach starts from a slightly modified loss function

$$V_t(\theta) = \sum_{s=1}^{t} \lambda^{t-s}(y_t - \theta^T\mathbf{x}_t)^2. \tag{8.16}$$

The Squared Loss function lying at the basis of RLS is recovered when $\lambda = 1$. If λ is set to some value slightly smaller than 1 (say $\lambda = 0.99$ or $\lambda = 0.95$), one has that for increasing t past observations are discounted. The smaller λ got, the quicker information obtained from previous data will be forgotten, and hence the name. It is now straightforward to re-derive the RLS based on (8.16), and the modified RLS becomes:

$$\begin{cases} \epsilon_t = (y_t - \mathbf{x}_t^T\hat\theta_{t-1}) \\ \mathbf{P}_t = \frac{1}{\lambda}\left(\mathbf{P}_{t-1} - \frac{\mathbf{P}_{t-1}\mathbf{x}_t\mathbf{x}_t^T\mathbf{P}_{t-1}}{\lambda+\mathbf{x}_t^T\mathbf{P}_{t-1}\mathbf{x}_t}\right) \\ \mathbf{K}_t = \mathbf{P}_t\mathbf{x}_t = \left(\frac{1}{\lambda+\mathbf{x}_t^T\mathbf{P}_{t-1}\mathbf{x}_t}\right)\mathbf{P}_{t-1}\mathbf{x}_t \\ \hat\theta_t = \hat\theta_{t-1} + \mathbf{K}_t\epsilon_t. \end{cases} \tag{8.17}$$

Example 50 (Estimator Windup) *Often, some periods of the identification experiment exhibit poor excitation. This causes problems for the identification algorithms. Consider the situation where* $\varphi_t = 0$ *in the RLS algorithm, then*

$$\begin{cases} \hat\theta_t = \hat\theta_{t-1} \\ \mathbf{P}_t = \frac{1}{\lambda}\mathbf{P}_{t-1}, \end{cases} \tag{8.18}$$

- *Notice that $\hat{\theta}$ remains constant during this period,*
- *... and $\mathbf{P}$ increases exponentially with time when $\lambda < 1$.*

When the system is excited again ($\varphi_t \neq 0$), the estimation gain $\mathbf{K}$ will be very large, and there will be an abrupt change in the estimate, despite the fact that the system has not changed. This effect is referred to as 'estimator windup'.

Since the study of Kalman filters will come back in some detail in later chapters, we treat the Kalman filter interpretation as merely an example here.

Example 51 (RLS as a Kalman Filter) *A stochastic state-space system takes the form*

$$\begin{cases} X_{t+1} = \mathbf{F}_t X_t + V_t \\ Y_t = \mathbf{H}_t X_t + W_t \end{cases} \quad \forall t = 1, 2, 3, \dots, \tag{8.19}$$

where

- *$\{X_t \in \mathbb{R}^n\}_t$ denote the stochastic states,*
- *$\{Y_t \in \mathbb{R}^m\}_t$ denote the observed outcomes.*
- *$\{V_t \in \mathbb{R}^n\}_t$ denote the process noise.*
- *$\{W_t \in \mathbb{R}^m\}_t$ denote the observation noise.*
- *$\{\mathbf{F}_t \in \mathbb{R}^{n\times n}\}_t$ are called the* system matrices
- $\{\mathbf{H}_t \in \mathbb{R}^{m\times n}\}_t$

Now it is easily seen that the problem of time-invariant RLS estimation can be written as

$$\begin{cases} \theta_{t+1} = \theta_t \\ Y_t = \mathbf{x}_t^T \theta_t + E_t \end{cases} \quad \forall t = 1, 2, 3, \dots, \tag{8.20}$$

where $\theta = \theta_1 = \cdots = \theta_t = \dots$ is the unknown state one wants to estimate based on observations $\{Y_t\}_t$. Hence one can phrase the problem as a filtering problem, where the Kalman filter provides the optimal solution to under appropriate assumptions, eventually reducing to (8.15). The benefit of this is that one can extend the model straightforwardly by including unknown process noise terms $\{V_t\}_t$, modeling the drifting true values as a random walk - approaching effectively the real-time identification problem. Suppose $\{V_1, \dots, V_t, \dots\}$ are sampled independently from a Gaussian with mean zero and covariance $\mathbf{V} \in \mathbb{R}^{n\times n}$, then the Kalman filter would become

$$\begin{cases} \epsilon_t = (y_t - \mathbf{x}_t^T \hat{\theta}_{t-1}) \\ \mathbf{P}_t = \left(\mathbf{P}_{t-1} - \frac{\mathbf{P}_{t-1}\mathbf{x}_t\mathbf{x}_t^T\mathbf{P}_{t-1}}{1+\mathbf{x}_t^T\mathbf{P}_{t-1}\mathbf{x}_t}\right) + \mathbf{V} \\ \mathbf{K}_t = \mathbf{P}_t \mathbf{x}_t \\ \hat{\theta}_t = \hat{\theta}_{t-1} + \mathbf{K}_t \epsilon_t. \end{cases} \tag{8.21}$$

Observe that both in case (8.17) as in (8.21) the basic RLS algorithm is modified such that $\mathbf{P}_t$ will no longer tend to zero. In this way $\mathbf{K}_t$ also is prevented from decreasing to zero. The parameter estimate will therefore change continuously.

8.1.2 Choosing Initial Values

The choice of the initial values is paramount in real life application of RLS schemes. Close inspection of the meaning of $\mathbf{P}_t$ helps us here. In the Kalman filter interpretation of RLS $\mathbf{P}_t$ plays the role of the covariance matrix of the estimate $\hat{\theta}_t$, as such suggesting that in case one is not at all certain of a certain choice of $\hat{\theta}_0$, one should take a large $\mathbf{P}_0$; if one is fairly confident in a certain choice of $\hat{\theta}_0$, $\mathbf{P}_0$ should be taken small. If $\mathbf{P}_0$ is small, so will $\{\mathbf{K}_t\}_{t>0}$ and the estimate $\{\hat{\theta}_t\}_t$ will not change too much from $\hat{\theta}_0$. If $\mathbf{P}_0$ would be large, $\hat{\theta}_t$ will quickly jump away from $\hat{\theta}_0$. Without apriori knowledge, it is common practice to take the following initial values

$$\begin{cases} \hat{\theta} = 0_d \\ \mathbf{P}_0 = \rho I_d, \end{cases} \tag{8.22}$$

with $I_d = \mathrm{diag}(1, \ldots, 1) \in \mathbb{R}^{d \times d}$ the identity matrix, and $\rho > 0$ a 'large number'.

The effect on the choice of the initial values (or the 'transient behavior) can be derived algebraically. Consider the basic RLS algorithm (8.15). Then

$$\mathbf{R}_t = \mathbf{R}_0 + \sum_{s=1}^{t} \mathbf{x}_t \mathbf{x}_t. \tag{8.23}$$

Now set

$$\mathbf{z}_t = \mathbf{R}_t \hat{\theta}_t. \tag{8.24}$$

Then

$$\mathbf{z}_t = \mathbf{R}_t \hat{\theta}_{t-1} + \mathbf{x}_t \epsilon_t = \left(\mathbf{R}_{t-1} + \mathbf{x}_t \mathbf{x}_t^T\right) \hat{\theta}_{t-1} + \mathbf{x}_t \left(y_t - \hat{\theta}_{t-1}^T \mathbf{x}_t\right) = \mathbf{z}_{t-1} + \mathbf{x}_t y_t = \mathbf{z}_0 + \sum_{s=1}^{t} \mathbf{x}_s y_s. \tag{8.25}$$

And hence

$$\hat{\theta}_t = \mathbf{P}_t \mathbf{z}_t = \left(\mathbf{R}_0 + \sum_{s=1}^{t} \mathbf{x}_s \mathbf{x}_s^T\right)^{-1} \left(\mathbf{R}_0 \hat{\theta}_0 + \sum_{s=1}^{t} \mathbf{x}_s y_s\right). \tag{8.26}$$

So, if $\mathbf{R}_0$ is small (i.e. $\mathbf{P}_0$ is large), then $\hat{\theta}_t$ is close to the offline estimate

$$\theta_t^* = \operatorname*{argmin}_{\theta} \sum_{s=1}^{t} (y_s - \theta^T \mathbf{x}_t)^2, \tag{8.27}$$

as seen by comparison of (8.26) with the normal equations associated to (8.27)

$$\theta_t^* = \left(\sum_{s=1}^{t} \mathbf{x}_s \mathbf{x}_s^T\right)^{-1} \left(\sum_{s=1}^{t} \mathbf{x}_s y_s\right). \tag{8.28}$$

The methods discussed in the above subsections are appropriate to systems that are known to change slowly over time. In such cases λ is chosen close to 1, or $\mathbf{V}$ is chosen as a small non-negative positive definite matrix. If the system exhibits more likely from time to time some abrupt changes of the parameters, techniques based on fault detection might be more suitable.

8.1.3 An ODE Analysis

Simulation no doubt gives useful insight. However, it is also clear that it does not permit generally valid conclusions to be drawn, and therefore it is only a complement to theory. The scope of a theoretical derivation would in particular be to study whether the parameter estimates $\hat{\theta}_t$ *converge* as t tends to infinity. If so, to what limit? And if possible also to establish the limiting distribution of $\hat{\theta}_t$.

A successful approach considers the sequence $\{\hat{\theta}_t\}_{t=0,1,2}$ as approximating a continuous vector valued function $\{\theta : \mathbb{R}_+ \to \mathbb{R}^d\}$. This continuos function evaluated at a time instant $\tau > 0$ is denoted as $\theta(\tau)$, and the whole sequence is described as an Ordinary Differential Equation. Such approach typically adopts a stochastic setting where $\{Y_t\}_t$is a stochastic process, and $\{X_t\}_t$ is a vector valued stochastic process, and both have bounded first- and second moments. Recall that the minimal MMSE $\theta_* \in \mathbb{R}^d$ is then given as the solution to

$$\mathbb{E}\left[X_t X_t^T\right]\theta_* = \mathbb{E}\left[X_t Y_t\right]. \tag{8.29}$$

Define again $\mathbf{R} = \mathbb{E}\left[X_t X_t^T\right]$, and suppose this one is invertible. Define the functional r (recall that θ here is a function) as:

$$r(\theta) = \mathbb{E}\left[X_t(Y_t - X_t^T\theta)\right]. \tag{8.30}$$

Now, consider the following ODE

$$\frac{\partial\theta(\tau)}{\partial\tau} = \mathbf{R}^{-1} r(\theta(\tau)). \tag{8.31}$$

If this ODE is solved numerically by an Euler method on discretization steps $\tau_1, \tau_2, \ldots$ one gets

$$\theta_{\tau_k} \approx \theta_{\tau_{k-1}} + (\tau_k - \tau_{k-1})\mathbf{R}^{-1}\mathbb{E}[Y_t - X_t^T\theta_{\tau_{k-1}}]. \tag{8.32}$$

Note the similarity between (8.32) and the algorithm (8.15), suggesting that the solutions to the deterministic ODE will be close in some sense. Specifically, consider the following recursion described by the algorithm

$$\hat{\theta}_t = \hat{\theta}_{t-1} + \gamma_t \mathbf{R}^{-1} X_t (Y_t - X_t^T\hat{\theta}_t). \tag{8.33}$$

with $\hat{\theta}_0$ given. Then the paths described by this discrete recursion will be similar to the solutions $\{\theta_{\tau_k}\}_k$ using the timescale $(\tau_k - \tau_{k-1}) = \gamma_t$. The above statements can be made quite precise as was done in [**?**]. The study of this ODE gives us new insight in the RLS algorithm, including:

1. The trajectories which solve the ODE are the expected paths of the algorithm.
2. Assume that there is a positive function $V(\theta, \mathbf{R})$ such that along along the solutions of the ODE we have that $\frac{\partial}{\partial\tau}V(\theta(\tau), \mathbf{R}) \leq 0$. Then as $\tau \to \infty$, $\theta(\tau)$ either tend to the set

   $$D_c = \left\{\theta_* \;\middle|\; \frac{\partial}{\partial\tau}V(\theta(\tau), \mathbf{R}) = 0\right\}, \tag{8.34}$$

   or to the boundary of the set of feasible solutions. In other words, $\theta(\tau)$ for $\tau \to \infty$ go to the stable stationary points of the ODE. Equivalently, $\hat{\theta}_t$ converge locally to a solution in D_c.

8.2 Other Algorithms

Since the problem of recursive identification, adaptive filtering or online estimation is so ubiquitous, it comes as no surprise that many different approaches exist. This section reviews three common variations.

8.2.1 Recursive Instrumental Variables

Recall that instrumental variable techniques come into the picture when the noise is known to be strongly colored, and a plain LSE is not consistent. An instrumental variable estimator uses random instruments $\{Z_t\}$ which are known to be independent to the noise of the system. Then we look for the parameters which match this property using the sample correlations instead. Formally, consider the statistical system

$$Y_t = \mathbf{x}_t^T \theta_0 + D_t, \tag{8.35}$$

where $\{D_t\}$ is colored noise, and $\mathbf{x}_t \in \mathbb{R}^d$, with deterministic but unknown vector θ_0. Suppose we have d-dimensional instruments Z_t such that

$$\mathbb{E}\left[Z_t D_t\right] = 0_d. \tag{8.36}$$

That is, the instruments are orthogonal to the noise. Then the (batch) IV estimator θ_n is given as the solution of $\theta \in \mathbb{R}^d$ to

$$\sum_{t=1}^n Z_t(Y_t - \mathbf{x}_t^T\theta) = 0_d, \tag{8.37}$$

which look similar to the normal equations. If $\sum_{t=1}^n (Z_t\mathbf{x}_t^T)$ were invertible, then the solution is unique and can be written as

$$\theta_n = \left(\sum_{t=1}^n Z_t\mathbf{x}_t^T\right)^{-1}\left(\sum_{t=1}^n Z_tY_t^T\right), \tag{8.38}$$

Now we can use the techniques used for RLS to construct a recursive method to estimate θ_t when the data comes in. It is a simple example to derive the algorithm, which is given as

$$\begin{cases} \epsilon_t = (y_t - \mathbf{x}_t^T\hat{\theta}_{t-1}) \\ \mathbf{P}_t = \mathbf{P}_{t-1} - \frac{\mathbf{P}_{t-1}Z_t\mathbf{x}_t^T\mathbf{P}_{t-1}}{1+\mathbf{x}_t^T\mathbf{P}_{t-1}Z_t} \\ \mathbf{K}_t = \mathbf{P}_tZ_t \\ \hat{\theta}_t = \hat{\theta}_{t-1} + \mathbf{K}_t\epsilon_t. \end{cases} \tag{8.39}$$

The discussion on the behavior of RLS w.r.t. initial variables and forgetting factor remains valid.

8.2.2 Recursive Prediction Error Method

Recall that a PEM method bases inference on maximizing performance of the best predictor corresponding to a model. Also this technique is straightforwardly to phrase in a recursive form.

$$\theta_t = \operatorname*{argmin}_\theta V_t(\theta) = \sum_{k=1}^t \lambda^{t-k}\epsilon_k(\theta), \tag{8.40}$$

where $0 < \lambda \leq 1$ is typically chosen as $0.99, 0.95, 0.9$. As before $\epsilon_t(\theta)$ denotes the prediction errors of corresponding to model parameters θ, that is $\epsilon_t(\theta) = y_t - \hat{y}_t(\theta)$ where $\hat{y}_t(\theta)$ is the optimal predictor at the tth instance. Now, unlike the previous algorithms, no closed form solution in exists in general, and one resorts to numerical optimization tools. But there is an opportunity here: it is not too difficult to integrate -say- a Gauss-Newton step in the optimizer with the online protocol.

To see how this goes, consider again the second order Taylor decomposition of the loss function. Lets assume we have a fairly good estimate $\hat{\theta}_{t-1}$ at the previous instance

$$V_t(\theta) = V_t(\hat{\theta}_{t-1}) + V'(\hat{\theta}_{t-1})^T(\theta - \hat{\theta}_{t-1}) + \frac{1}{2}(\theta - \hat{\theta}_{t-1})^T V_t''(\hat{\theta}_{t-1})(\theta - \hat{\theta}_{t-1}). \tag{8.41}$$

Now, the challenge is to compute gradient V_t' and Hessian V_t'' recursively. Details can be found in the SI book, but are necessarily tied to the adapted model and are often approximative in nature.

8.2.3 Recursive Pseudo-linear Least Squares

The following example expresses an ARMAX as a pseudo-linear model as follows.

Example 52 (ARMAX) *Given an ARMAX system*

$$A(q^{-1})y_t = B(q^{-1})u_t + C(q^{-1})e_t, \tag{8.42}$$

of orders n_a, n_b, n_c. Then this system can 'almost' be written as a LIP model as follows

$$y_t = \varphi_t^T \theta_0 + e_t, \tag{8.43}$$

where

$$\begin{cases} \varphi_t = (-y_{t-1}, \dots, -y_{t-n_a}, u_{t-1}, \dots, u_{t-n_b}, \hat{e}_{t-1}, \dots, \hat{e}_{t-n_c})^T \\ \theta_0 = (a_1, \dots, a_{t-n_a}, b_1, \dots, b_{t-n_b}, c_1, \dots, c_{t-n_c}), \end{cases} \tag{8.44}$$

where $\hat{e}_t$ is the prediction error computed based on the model parameters $\hat{\theta}_{t-1}$. The rationale is that in case $\theta_{t-1} \approx \theta_t$, $\hat{e}_t$ is a good proxy to the prediction errors e_t based on the parameters θ_t. Then the Recursive Partial Least Squares algorithm implements a RLS strategy based on this 'linearized' model.

Indeed one can prove that the resulting estimates do converge if the system is obeys some regularity conditions. Specifically, if the system is almost unstable the recursive estimates are often unstable (and diverging) as well. In practice, the resulting algorithm needs monitoring of the resulting estimates in order to detect such divergent behavior.

8.2.4 Stochastic Approximation

The class of stochastic approximation techniques take a quite different perspective on the recursive identification problem. Here the parameter estimate $\hat{\theta}_{t-1}$ obtained previously is modified such that it is better suited for explaining the new sample related to (φ_t, y_t). Formally, a new estimate $\hat{\theta}_t$ is obtained from the given $\hat{\theta}_{t-1}$ and the sample (φ_t, y_t) by solving for a given $\gamma > 0$ the optimization problem

$$\hat{\theta}_t = \operatorname*{argmin}_{\theta} J_t(\theta) = (\theta - \hat{\theta}_{t-1})^T(\theta - \hat{\theta}_{t-1}) + \gamma \left(\varphi_t^T \theta - y_t\right)^2. \tag{8.45}$$

The optimal result is then given directly as

$$\hat{\theta}_t = \hat{\theta}_{t-1} - \gamma \left(\varphi_t^T \theta - y_t \right) \varphi_t, \tag{8.46}$$

obtained by equating the derivative of $J_t(\theta)$ to zero. The algorithm is then completed by specification of the initial estimate $\hat{\theta}_0$. This recursion gives then what is called the Least Mean Squares (LMS) algorithm. This is the building stone of many implementations of adaptive filtering. The naming convention 'stochastic approximation' is motivated as follows. The correction at instance t is based on the gradient of a single point (φ_t, y_t), and is a very 'noisy' estimate of the overall gradient. A variation of this algorithm is given by the recursion

$$\hat{\theta}_t = \hat{\theta}_{t-1} - \frac{\gamma}{\|\varphi_t\|_2 + \epsilon} \left(\varphi_t^T \theta - y_t \right), \tag{8.47}$$

with $\epsilon > 0$ small, and where $\hat{\theta}_0$ is given. This recursion is the basis of the Normalized LMS algorithm. The rationale is that here each sample modifies the present estimate proportional how close the estimate is to the working point 0_d.

8.3 Model Selection

As in the batch setting, it is paramount to be able to qualify and quantify how well our recursive algorithms succeeds in its task. But the conceptual and practical ways to do turn out to be entirely different. As it stands there is no comprehensive theoretical for this question, but some insight is gained in the following example.

Example 53 (Predicting Random noise) *As seen, a lot of fancy mathematics can be brought in to form complex recursive schemes, but at the end of the day the methods implemented need 'merely' may good predictions. It helps to reason about this objective by considering the prediction of random white noise: by construction this is impossible to do better than $\hat{y}_t = 0$ (why?). A method trying to fit a complex model to such data will necessarily do worse than this simple predictor, and the example is often used as a validity check of a new method.*

Except for the traditional considerations of bias and variance of a model, and the statistical uncertainty associated with estimating parameters, other issues include the following:

- Initialization of the parameters. If the initial guess of the parameters is not adequate, the recursive algorithm might take much samples before correcting this (transient effect).

- Forgetting Factor. The choice of a forgetting factor makes a trade-off between flexibility and accuracy.

- Window. If the window used for estimating then one must decide on how many samples are used for estimating at a certain instance t.

- Stability of the estimate. If the algorithm at hand is not well-tuned to the task at hand, it may display diverging estimates. This is clearly undesirable, and some algorithms go with guarantees that no such unstable behavior can occur.

- Gain. A typical parameter which needs t be tuned concerns the size of the update made at a new sample. If the gain is too low, a resulting algorithm will not converge fastly. If the gain is too large, one may risk unstable behavior.

In order to check wether a recursive identification is well-tuned for a certain application, it is instrumental to monitor closely the online behavior of the method, and to make appropriate graphical illustrations of the method.

Part II

Advanced Topics