Parameter estimation and model structures

Luc Pronzato
Laboratoire I3S
CNRS/Université de Nice–Sophia Antipolis
Les Algorithmes, 2000 route des Lucioles
BP 121, 06903 Sophia-Antipolis Cedex, France
e-mail: pronzato@i3s.unice.fr

Foreword

These notes, largely based on the books [36, 38], result from many years of collaboration with several co-authors, especially Éric Walter (Research Director at CNRS, Gif-sur-Yvette, France) and Andrej Pázman (Professor at Comenius University, Bratislava). They intend to give only a short overview of the subject, decomposed into the following steps:

1. Models and their structural properties
2. Estimators
3. Optimisation of estimation criteria
4. Experimental design
5. (In)validation & testing

Points 3 and 5 will be almost skipped and point 4 will receive much less attention than it deserves. This means that most of the developments will concern point 1, with particular attention to the issues of *identifiability* and *distinguishability* of model structures, and point 2, where the problem caused by the presence of *local optimas* of the estimation criterion is explained in details. Examples prevail over mathematical developments (and rigor) throughout the notes, but references are given where precise results can be found.

Vocabulary & notations

We call *system* and denote $\mathcal{S}$ the physical "reality" on which observations are collected; we act on the system through some *inputs* $u(t)$ and collect information through the observations $y(t)$ of its *outputs*. For the sake of simplicity, mainly the case of only one observed output will be considered ($y(t)$ is then a scalar). Because $\mathcal{S}$ is a real system, it is subject to random disturbances (perturbations), that we call *errors* and denote $\varepsilon(t)$. They will be considered as random variables. For the moment one may consider t as time, and think of a dynamical system. Later on t will be replaced by some other, more general, *explanatory variables*, or *experimental conditions*, x (not to be confounded with the state variable in a state-space representation).

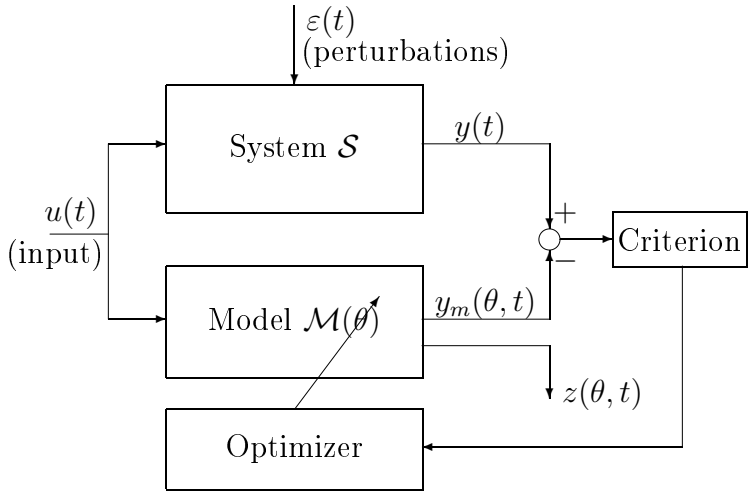

Figure 1: Flow of information in identification

The *model* $\mathcal{M}(\theta)$ corresponds to our theoretical vision of $\mathcal{S}$ (some mathematical equations), which depends on θ, a vector of (unknown) *parameters*, $\theta \in \mathbb{R}^p$. Only parametric models will be considered. The model receives the same inputs $u(t)$ as the system and responds by $y_m(\theta,t)$. Hopefully $y_m(\theta,t)$ will look like the observations $y(t)$ collected on $\mathcal{S}$. For that, we need to choose a proper set of equations (the model structure) and find the good tuning of the parameters. This is the purpose of *system identification*. Note the difference between the model structure (the equations) and the model of parameters θ. In some cases the prediction of quantities that cannot be observed on $\mathcal{S}$ but can be predicted by $\mathcal{M}(\theta)$ is the main objective of construction of the model. Such quantities are denoted $z(\theta,t)$.

The flow of information in system identification can be summarized as in Figure 1. The criterion defines the estimator and corresponds to a sort of distance between the quantities $y(t)$ that are observed on the system and those $y_m(\theta,t)$ that are predicted by the model of parameters θ. The role of the optimizer is to find the best value of θ in terms of criterion value. The presence of perturbations $\varepsilon(t)$ makes the observations $y(t)$ random (observed quantities can be considered as realizations of random variables). The value $\hat{\theta}$ that optimizes the criterion, the *estimate* of θ, is obtained from the observations, and is also random. We can thus consider the mean, variance, etc., of $\hat{\theta}$, and possibly try to minimize the variance to make the estimation precise, this will be one of the objectives considered in Section 4.

1 Model structures

The choice of a model structure is called *characterization*. It is critical, and intuition often plays a key role; it is of special importance since it defines the class within which the best model will be looked for. The choice may in particular depend on

- the objective of modelling: analysis of physical phenomena, prediction of unobservable quantities $z(\theta,t)$, test, diagnosis, teaching, prediction (short or long term, e.g. for control), regulation, signal processing (compression, filtering, echo cancellation...), simulation...
- the conditions under which the model will be used: operating range (large/small signals)

- the amount of information available, etc.

1.1 Knowledge-based or reproduction of some behavior?

Phenomenological models (knowledge-based) are based on physical equations. We give a simple example.

Example 1 *Consider a chemical reaction described as*

$$A \underset{\theta_2}{\overset{\theta_1}{\rightleftarrows}} B \xrightarrow{\theta_3} C\,.$$

The assumption of first-order kinetics leads to the equations

$$\frac{d[A]}{dt} = -\theta_1[A] + \theta_2[B]\,,\ \frac{d[B]}{dt} = \theta_1[A] - (\theta_2+\theta_3)[B]\,,\ \frac{d[C]}{dt} = \theta_3[B]\,,$$

where $[A]$, $[B]$, $[C]$, θ_1, θ_2, θ_3 *have a physical meaning. The differential equations themselves are imposed by the prior knowledge on (and assumptions about) the underlying phenomenon.*

The simulation of a phenomenological model is sometimes complicated, but often accurate and provides good predictions.

On the other hand, *behavioral models* only aim at reproducing some observed behavior. Polynomials, splines, neural networks, support vector machines, etc., are possible candidates. Their simulation and the estimation of their parameters are generally simple, but they should be considered *more like interpolators than predictors* (in the sense that predictions outside the range of observed outputs are not reliable).

1.2 Linear and nonlinear models

It is of paramount importance to distinguish clearly between two types of nonlinearities.

We say that a model structure is *linear with respect to its inputs* (LI) when

$$y_m(\theta, t, \alpha_1 u_1 + \alpha_2 u_2) = \alpha_1 y_m(\theta, t, u_1) + \alpha_2 y_m(\theta, t, u_2)\,,$$

which corresponds to a linear model in the sense of *control theory*.

We say that a model structure is *linear with respect to its parameters* (LP) when

$$y_m(\alpha_1\theta_1 + \alpha_2\theta_2, t, u) = \alpha_1 y_m(\theta_1, t, u) + \alpha_2 y_m(\theta_2, t, u)\,,$$

which corresponds to a linear model in the sense of *statistics*.

LI structures usually have a limited domain of validity, but their study is facilitated by the existence of classical results from linear control theory. Parameter estimation, and the characterization of its precision, is easy for LP structures (in some cases, the estimator is given by an explicit expression, as for Least-Squares, see (2) in Section 2.1). On the other hand, non-LP (NLP) structures require the use of nonlinear statistics for studying the properties of estimators. Most often knowledge based model structures, for which parameters have a physical meaning, are NLP.

1.3 Continuous or discrete time

The evolution of a dynamical system can be described by a set of differential or difference (i.e. recurrence) equations, which respectively corresponds to a continuous or discrete-time model. Note that a continuous-time model can be discretized to produce a discrete-time model, e.g. in order to facilitate its simulation. On the other hand, some discrete-time models may have no continuous-time counterpart. Discrete-time models impose constraints on the measurement times (they should be multiple of the sampling period), whereas the measurement times can be chosen individually for continuous-time models.

1.4 Choice of complexity

Complex structures have more degrees of freedom (they have more parameters) and therefore more capability to reproduce complex behaviors: the model responses after estimation of the parameters will be closer to the observations for a complex structure than for a simple one. However, the additional degrees of freedom of a complex structure should not be employed to reproduce the perturbations! The choice of a model structure of appropriate complexity (a problem especially for behavioral models) is thus a difficult issue. We shall see in Section 2.4.2 how to make a compromise between complexity and performance (robustness) through the definition of a suitable criterion.

We can already give the recommendation (see also Section 5) to spare some validation data (which will not be used for parameter estimation), and try (at the very end) to predict them by the model that has been constructed. This test often permits to reject overcomplicated structures that have produced a good fitting of the particular data used for parameter estimation, but lack of robustness and are not able to suitably reproduce data associated with different realizations of the perturbations.

1.5 Regression models

Regression models will often be used in the rest of the notes. They correspond to the situation where the perturbations that corrupt the observations can be considered as independent with zero mean (we even often assume that they are i.i.d. —independently and identically distributed). The term *output-error model* is sometimes used.

For instance, let $y_m(\theta, t)$ be the solution of a differential equation, it may correspond to $[A](t)$ as in Example 1. Suppose that the observation number k is taken at time t_k, so that

$$y(t_k) = y_m(\bar{\theta}, t_k) + \varepsilon_k$$

with (ε_k) a sequence of independent random variables, with zero mean. Here $\bar{\theta}$ denotes the unknown "true" value of θ. This means that we make the *very strong assumption* that the system corresponds to a model structure of the type we have chosen, with some particular value $\bar{\theta}$ for θ, and an additive noise corrupting the observations.

More generally, in a regression model we shall denote

$$y(x_k) = \eta(\bar{\theta}, x_k) + \varepsilon_k \tag{1}$$

with $\eta(\theta, x)$ the model response, equal to the mean of $y(x)$ for given x and θ,

$$\eta(\theta, x) = \mathtt{E}_{x,\theta}\{y\}$$

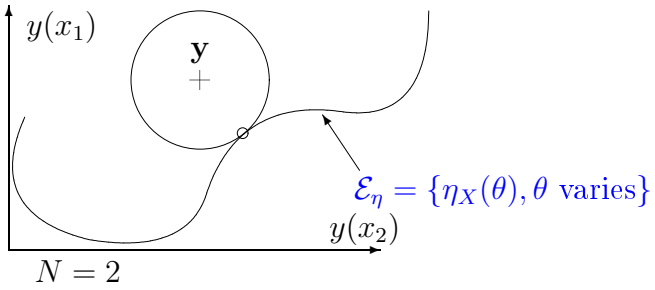

Figure 2: Expectation surface $\mathcal{E}_\eta$ and observations $\mathbf{y}$

and x the *design variable*, which corresponds to the *experimental conditions*, such as time, temperature, pressure, etc. We shall need vector notations to collect N quantities, with N the number of observations (scalar), and we write

$X = X_1^N = (x_1, \ldots, x_N)$	the design
$\mathbf{y} = [y(x_1), \ldots, y(x_N)]^\top$	the (column) vector of observations
$\eta_X(\theta) = [\eta(\theta, x_1), \ldots, \eta(\theta, x_N)]^\top$	the (column) vector of model responses

1.6 Identifiability

Once a model structure has been chosen that depends on parameters θ, one may naturally wonder if **(A)** one has a chance to estimate θ uniquely. However, this question is too vague to receive a precise answer. We shall see that for regression models a related issue concerns **(B)** the influence of a variation of θ on $\eta_X(\theta)$. These questions are related since if a unique θ can be associated with any given $\eta_X(\theta)$, then the answer to the first question is positive: θ will be estimated uniquely (almost surely...). This is due to the following, see [25].

Theorem 1 *In a regression model, if the p.d.f. of ε is absolutely continuous w.r.t. Lebesgue measure, then there exists a unique vector $\eta_X(\theta)$ closest to* $\mathbf{y}$ with probability one.

Here, closest is in the sense of euclidian distance, which corresponds to Least-Squares (LS) estimation, see Section 2.1. The LS estimator $\hat{\theta}_{LS}^N$ corresponds to the value of θ such that the distance from $\mathbf{y}$ to $\eta_X(\theta)$ is minimum. The set of vectors $\eta_X(\theta)$ when θ varies in its admissible set $\Theta \subset \mathbb{R}^p$ defines the *expectation surface* $\mathcal{E}_\eta$. It is indeed a p-dimensional surface in $\mathbb{R}^N$, and in the regression model (1) $\eta_X(\bar{\theta})$ is the expectation of $\mathbf{y}$. Figure 2 illustrates the situation when $p = 1$ and $N = 2$. The small circle indicates the position of $\eta_X(\hat{\theta}_{LS}^N)$, the point of $\mathcal{E}_\eta$ closest to $\mathbf{y}$.

If a unique θ can be associated with any $\eta_X(\theta)$, clearly the estimator $\hat{\theta}_{LS}^N$ is unique. The analysis is quite simple for LP structures. Indeed, in that case $\eta_X(\theta) = \mathbf{R}\theta + \mathbf{c}$ for some matrix $\mathbf{R}$ and vector $\mathbf{c}$, and $\eta_X(\theta_1) = \eta_X(\theta_2)$ is equivalent to $\mathbf{R}\theta_1 = \mathbf{R}\theta_2$. The solution is thus unique if and only if $\mathbf{R}$ has full rank, which is easy to test.

For a NLP structure, the problem is more difficult, the answer depending on the design X. We thus consider an idealized framework where

- the system $\mathcal{S}$ is replaced by $\mathcal{M}(\bar{\theta})$, for some unknown $\bar{\theta}$;
- there are no perturbations;

- we can collect as many data as we wish, and choose the design X (in particular the input $u(t)$) as we want.

Let $\mathtt{M}(\theta)$ denote all possible behaviors for a model with parameters θ (for a regression model, this denotes all possible vectors of responses $\eta_X(\theta)$ when X varies and N is arbitrarily large).

We then consider the following question **(C)**: suppose that $\mathtt{M}(\bar{\theta}) \equiv \mathtt{M}(\theta)$ for some θ (which means that $\mathcal{M}(\theta)$ and $\mathcal{M}(\bar{\theta})$ have the same behavior), does it imply $\theta = \bar{\theta}$? The answer is, in general, of the following type.

Definition 1

- $[\theta]_i$ *is* structurally globally identifiable (s.g.i.) *if for almost any* $\bar{\theta}$, $\mathtt{M}(\hat{\theta}) \equiv \mathtt{M}(\bar{\theta}) \Rightarrow [\hat{\theta}]_{\mathtt{i}} = [\bar{\theta}]_{\mathtt{i}}$.
  If each $[\theta]_i$ *is s.g.i.,* $i = 1, \ldots, p$, $\mathcal{M}(\cdot)$ *is s.g.i.*

- $[\theta]_i$ *is* structurally locally identifiable (s.l.i.) *if for almost any* $\bar{\theta}$, $\exists$ *some neighborhood* $\mathcal{V}(\bar{\theta})$ *such that* $\hat{\theta} \in \mathcal{V}(\bar{\theta})$ *and* $\mathtt{M}(\hat{\theta}) \equiv \mathtt{M}(\bar{\theta}) \Rightarrow [\hat{\theta}]_{\mathtt{i}} = [\bar{\theta}]_{\mathtt{i}}$.
  If each $[\theta]_i$ *is s.l.i.,* $i = 1, \ldots, p$, $\mathcal{M}(\cdot)$ *is s.l.i.*

- $[\theta]_i$ *is* structurally unidentifiable (s.u.i.) *if for almost any* $\bar{\theta}$ *and any neighborhood* $\mathcal{V}(\bar{\theta})$, $\exists \hat{\theta} \in \mathcal{V}(\bar{\theta})$ *such that* $[\hat{\theta}]_i \neq [\bar{\theta}]_i$ *and* $\mathtt{M}(\hat{\theta}) \equiv \mathtt{M}(\bar{\theta})$.
  $[\theta]_i$ *is s.u.i.* $\Rightarrow \mathcal{M}(\cdot)$ *is s.u.i.*

Note that s.g.i. implies s.l.i. and that one component $[\theta]_i$ of θ may be s.g.i. and another one $[\theta]_j$, $j \neq i$, s.u.i.

There exist various methods to test model structures, LI or not, for identifiability, see e.g. the surveys [35, 37]. Here we only indicate a simple one for *stationary LI* structures: in that case the transfer function (matrix) characterizes all the input/output behavior $\mathtt{M}(\theta)$ and the analysis is as follows.

1. Write the transfer function $H(\theta, s)$ (or matrix $\mathbf{H}(\theta, s)$), such that $\hat{y}_m(\theta, s) = H(\theta, s)\hat{u}(s)$ with $\hat{y}_m(\theta, s)$ and $\hat{u}(s)$ the Laplace transforms of $y_m(\theta, t)$ and $u(t)$ respectively;
2. put it under a canonical form;
3. solve (w.r.t. $\hat{\theta}$): $H(\bar{\theta}, s) = H(\hat{\theta}, s)$ (or $\mathbf{H}(\bar{\theta}, s) = \mathbf{H}(\hat{\theta}, s)$).
   - If there is a unique solution $[\hat{\theta}]_i = [\bar{\theta}]_i$, $[\hat{\theta}]_i$ is s.g.i.
   - If there is a finite number of solutions for $[\hat{\theta}]_i$, $[\hat{\theta}]_i$ is s.l.i.
   - If there are several solutions for $[\hat{\theta}]_i$ in any neighborhood of $\bar{\theta}$, $[\hat{\theta}]_i$ (and thus $\mathcal{M}(\cdot)$) is s.u.i.

The importance of using a canonical form can be seen from the following simple example.

Example 2 *Suppose that*

$$H(\theta, s) = \frac{\theta_1}{\theta_2 + \theta_3 s}\,.$$

Then, $H(\hat{\theta}, s) = H(\bar{\theta}, s)$ *does not imply* $\{\hat{\theta}_1 = \bar{\theta}_1, \hat{\theta}_2 = \bar{\theta}_2, \hat{\theta}_3 = \bar{\theta}_3\}$. *A canonical form can be obtained for instance by simplifying numerators and denominators so that the highest degree coefficient in* s *equals 1.*

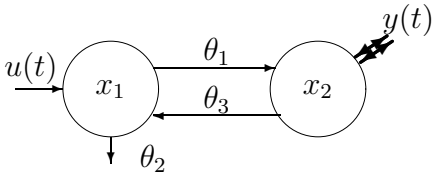

Figure 3: Compartmental model in Example 3

Example 3 *Figure 3 presents a compartmental model, where x_1 and x_2 may either denote the quantities of two different products (like in a chemical reaction) or quantities of the same product in two different places (like in reservoirs). The parameters θ_i, $i = 1, \ldots, 3$, are inverse of time constants, and the model is described by the following state-space equations*

$$\begin{aligned} \frac{\partial x_1(t)}{dt} &= -(\theta_1 + \theta_2)x_1(t) + \theta_3 x_2(t) + u(t)\,, \quad x_1(0) = 0\,, \\ \frac{\partial x_2(t)}{dt} &= \theta_1 x_1(t) - \theta_3 x_2(t)\,, \quad x_2(0) = 0\,. \end{aligned}$$

The double arrow that connects $y(t)$ to the x_2 compartment in Figure 3 means that x_2 is observed, that is, the observation equation is

$$y_m(\theta, t) = x_2(t)\,.$$

The transfer function (in canonical form) of this structure is easily computed as

$$H(\theta, s) = \frac{\theta_1}{s^2 + s(\theta_1 + \theta_2 + \theta_3) + \theta_2\theta_3}\,.$$

We can then test the structure for identifiability and

$$\begin{aligned} [\mathtt{M}(\hat{\theta}) \equiv \mathtt{M}(\bar{\theta})] &\Leftrightarrow [H(\hat{\theta}, s) = H(\bar{\theta}, s) \forall s]\,, \\ &\Leftrightarrow \begin{cases} \hat{\theta}_1 = \bar{\theta}_1\,, \\ \hat{\theta}_1 + \hat{\theta}_2 + \hat{\theta}_3 = \bar{\theta}_1 + \bar{\theta}_2 + \bar{\theta}_3\,, \\ \hat{\theta}_2\hat{\theta}_3 = \bar{\theta}_2\bar{\theta}_3\,. \end{cases} \end{aligned}$$

Therefore, two solutions exist for $\hat{\theta}$:

$$\begin{aligned} \hat{\theta} &= (\bar{\theta}_1, \bar{\theta}_2, \bar{\theta}_3)\,, \\ \hat{\theta}' &= (\bar{\theta}_1, \bar{\theta}_3, \bar{\theta}_2)\,, \end{aligned}$$

which means that θ_1 is s.g.i. and θ_2 and θ_3 are s.l.i. (and $\mathcal{M}(\cdot)$ is s.l.i.).

Remark 1

1. In a regression model an estimator typically minimizes a distance between $\mathbf{y}$ (observed) and $\eta_X(\theta)$. In Example 3 if we use an optimisation algorithm to find the estimator, we shall find say $\hat{\theta}$, but from the identifiability analysis above we know that $\hat{\theta}' \neq \hat{\theta}$ gives exactly the same distance to $\mathbf{y}$.

2. In Example 3 if $\bar{\theta}_1 = 0$, then θ_2 and θ_3 are not identifiable. However, this corresponds to an atypical value for $\bar{\theta}_1$, hence the words "almost any $\bar{\theta}$" in the definition of identifiability, which

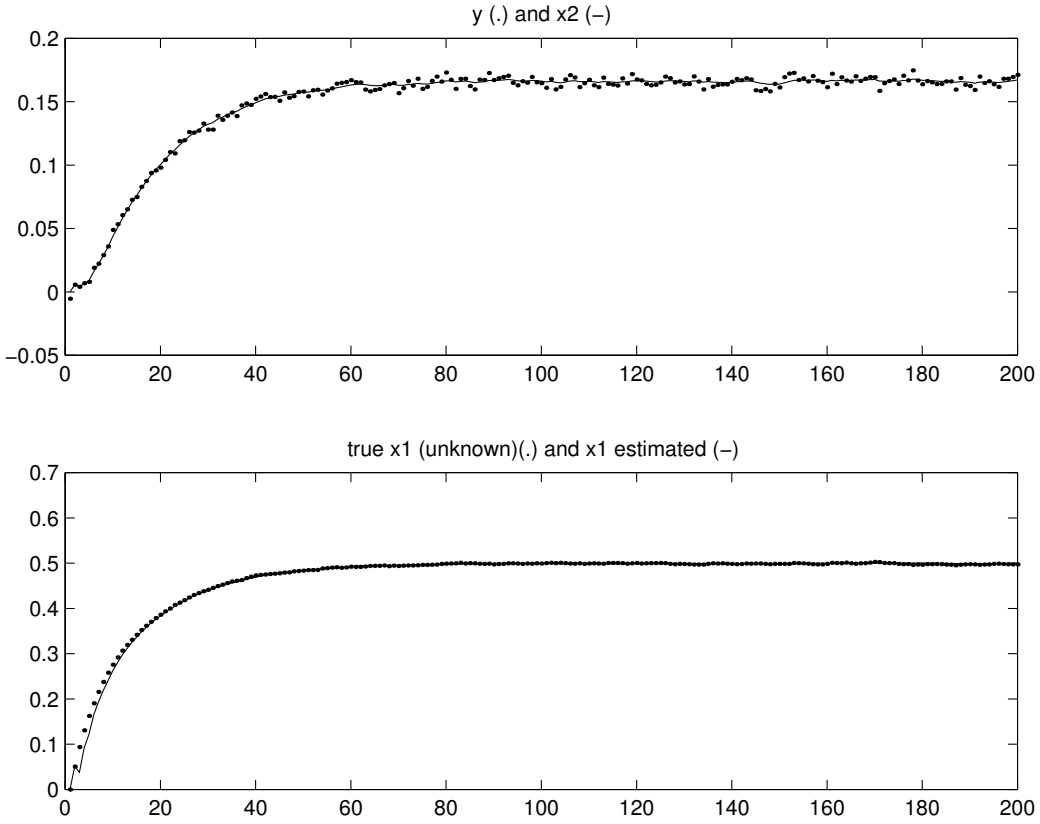

Figure 4: Estimated and true states in Example 3 when $\hat{\theta} = \bar{\theta}$

corresponds to a structural property. *(We shall see in Section 1.7 that there exist situations where no structural conclusion can be drawn.)*

3. Identifiability is an important notion both for parameter and state *estimations: in Example 3, $x_1(t)$ is not observed, assume that it is reconstructed (e.g., by Kalman filtering). Then, for any $u(t)$ there exist two solutions (trajectories) for $x_1(t)$, and we can* never *know which one is correct (but we know that we cannot know!). Figure 4 presents a realization of the true and estimated states $x_1(t)$ and $x_2(t)$ when there is no misspecification of $\hat{\theta}$, that is, $\hat{\theta} = \bar{\theta}$. Figure 5 corresponds to $\hat{\theta} = (\bar{\theta}_1, \bar{\theta}_3, \bar{\theta}_2)$.*

4. Whatever the method that is used to test a model structure, at some point the problem amounts to solving a system of (often polynomial) equations. Computer algebra softwares (MAPLE, Mathematica, Reduce, etc.) are then especially useful.

5. NLI structures are more difficult to test than LI ones and require particular techniques. At the same time, they are generally "more identifiable" (and it is non trivial to exhibit an example of a nonacademic NLI structure that is not s.g.i., even if such examples exist).

1.7 Distinguishability

The question now concerns the model structure itself, namely, **(A)** do we have a chance to determine the right structure for $\mathcal{S}$? Again, this is too vague to receive a precise answer, and we shall consider an idealized framework where

- the data are generated by $\bar{\mathcal{M}}(\bar{\theta})$ for some unknown $\bar{\theta}$;

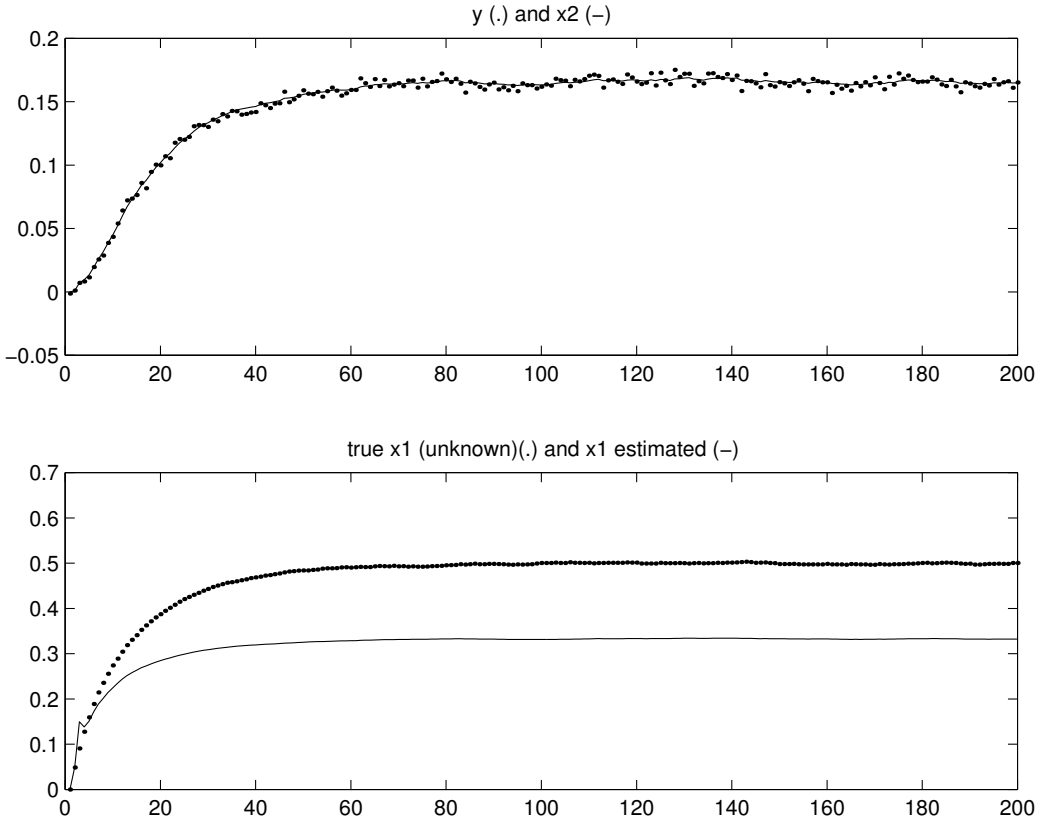

Figure 5: Estimated and true states in Example 3 when $\hat{\theta} = (\bar{\theta}_1, \bar{\theta}_3, \bar{\theta}_2)$

- there are no perturbations;
- we can collect as many data as we wish, and choose the design X (in particular the input $u(t)$) as we want.

We thus replace the system $\mathcal{S}$ by a model structure $\bar{\mathcal{M}}(\bar{\theta})$, with $\bar{\theta}$ unknown. We propose a structure $\hat{\mathcal{M}}(\cdot)$ different from $\bar{\mathcal{M}}(\cdot)$, now the question is **(B)** does $\hat{\theta}$ exist, such that $\hat{\mathtt{M}}(\hat{\theta}) \equiv \bar{\mathtt{M}}(\bar{\theta})$? The answer may be as follows.

Definition 2

- $\hat{\mathcal{M}}(\cdot)$ *is* structurally distinguishable (s.d.) *from* $\bar{\mathcal{M}}(\cdot)$ *if for almost any* $\bar{\theta}$*, there exists no* $\hat{\theta}$ *such that* $\hat{\mathtt{M}}(\hat{\theta}) \equiv \bar{\mathtt{M}}(\bar{\theta})$*.*
- *If* $\hat{\mathcal{M}}(\cdot)$ *is s.d. from* $\bar{\mathcal{M}}(\cdot)$ *and* $\bar{\mathcal{M}}(\cdot)$ *is s.d. from* $\hat{\mathcal{M}}(\cdot)$*, then* $\hat{\mathcal{M}}(\cdot)$ *and* $\bar{\mathcal{M}}(\cdot)$ *are said s.d.*

Note that the definition of s.d. is not symmetrical. The techniques used for studying identifiability can be used to test structures for distinguishability (note, however, that we hope that a unique solution exists when testing for identifiability, whereas we hope that there are no solutions when testing for distinguishability).

Example 4 *Consider the two compartmental models presented in Figure 6. Using their transfer functions to characterize their behaviors, similarly to Example 3, we obtain that the two structures are s.g.i., but they are not s.d.*

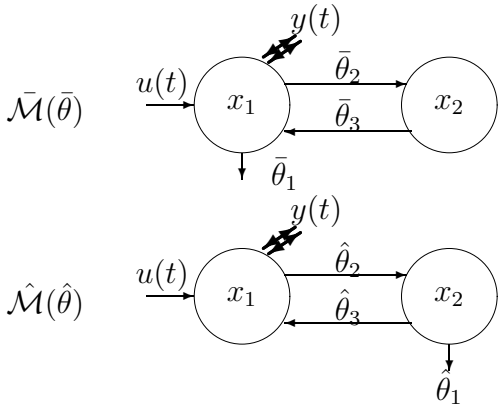

Figure 6: Compartmental models tested for distinguishability in Example 4

Remark 2

1. Example 4 shows that s.g.i. of model structures does not imply their s.d. Conversely, there exist structures that are s.d. but not s.g.i., see [34].

2. Consider the model structures associated with the following transfer functions

$$\bar{H}(\bar{\theta}, s) = \frac{1}{s^2 + \bar{\theta}_1 s + \bar{\theta}_2}, \quad \hat{H}(\hat{\theta}, s) = \frac{1}{(s + \hat{\theta}_1)(s + \hat{\theta}_2)}, \quad \bar{\theta}, \hat{\theta} \in I\!R^2 .$$

Then $\hat{\mathcal{M}}(\cdot)$ is not distinguishable from $\bar{\mathcal{M}}(\cdot)$ if $\bar{H}$ has two real poles, but is distinguishable otherwise. Since none of these cases can be considered as atypical, no structural conclusion can be drawn in that case.

1.8 An example

We conclude this section by drawing attention to the fact that identifiability (and distinguishability) issues are often non trivial, even for LI structures, as illustrated by the next example.

Example 5 *It corresponds again to a compartmental model, as shown in Figure 7. There is one input $u(t)$, that corresponds to the oral administration of a drug D. Four different outputs can be observed:*

- *the concentration of D in the blood, $y_1 = \theta_6 D_S$;*
- *the concentration of metabolite M in the blood, $y_2 = \theta_7 M_S$;*
- *the urinary execration of D, $y_3 = \theta_5 D_S$;*
- *the urinary execration of M, $y_4 = \theta_4 M_S$.*

There are seven parameters to estimate and fifteen possible input/ouput configurations. The results of the identifiability study for each configuration are reported in Table 5 which indicates the parameters that are s.g.i. and those that are s.l.i. (with the number of solutions).

Note that the model structure is never s.g.i., although there are only seven parameters for four different outputs that are observed. When the structure is s.l.i., we can calculate all the values of θ that give the same behavior. Notice that observing y_1, y_2 and y_4 is qualitatively

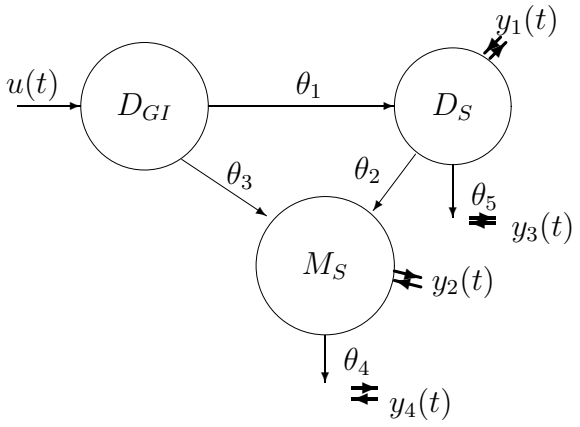

Figure 7: Compartmental model tested for identifiability in Example 5

observed outputs	structure	s.g.i. parameters	s.g.i. parameters
1	s.n.i.		
2	s.n.i.		θ_4 (3 sol.)
3	s.n.i.		
4	s.n.i.		$\theta_1, \theta_2, \theta_3, \theta_4, \theta_5$ (6 sol.)
1 & 2	s.n.i.	θ_4	
1 & 3	s.n.i.		
1 & 4	s.n.i.	θ_2, θ_4	$\theta_1, \theta_3, \theta_5, \theta_6$ (2 sol.)
2 & 3	s.n.i.	θ_4	$\theta_1, \theta_2, \theta_3, \theta_5, \theta_7$ (2 sol.)
2 & 4	s.n.i.		$\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_7$ (6 sol.)
3 & 4	s.n.i.	θ_2, θ_4	$\theta_1, \theta_3, \theta_5$ (2 sol.)
1, 2 & 3	s.l.i.	θ_4	$\theta_1, \theta_2, \theta_3, \theta_5, \theta_6, \theta_7$ (2 sol.)
1, 2 & 4	s.l.i.	$\theta_2, \theta_4, \theta_7$	$\theta_1, \theta_3, \theta_5, \theta_6$ (2 sol.)
1, 3 & 4	s.n.i.	θ_2, θ_4	$\theta_1, \theta_3, \theta_5, \theta_6$ (2 sol.)
2, 3 & 4	s.n.i.	$\theta_2, \theta_4, \theta_7$	$\theta_1, \theta_3, \theta_5$ (2 sol.)
1, 2, 3 & 4	s.l.i.	$\theta_2, \theta_4, \theta_7$	$\theta_1, \theta_3, \theta_5, \theta_6$ (2 sol.)

Table 1: Results of identifiability tests in Example 5

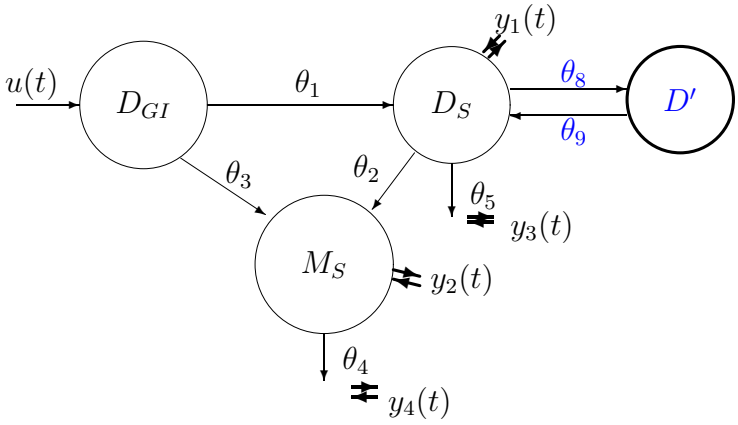

Figure 8: This structure is s.g.i.!

equivalent to observing y_1, y_2, y_3 *and* y_4. *(However, it is not equivalent* quantitatively*, in terms of precision of the estimation.)*

The situation will appear still less trivial when it will be noted that a more complex structure can be s.g.i.! Indeed, consider the structure defined by Figure 8. There are two additional parameters θ_8 *and* θ_9 *with respect to the structure of Figure 7, we do not observe any additional output, but* the structure is now s.g.i. when y_1, y_2, y_3 and y_4 are observed!

2 Estimators

An estimator $\hat{\theta}^N$ minimizes a *"distance"* between the response of $\mathcal{M}(\theta)$ and the observations collected on $\mathcal{S}$. For a regression model, the estimation criterion thus measures the distance between the expectation surface $\mathcal{E}_\eta$ and $\mathbf{y}$, see Figure 2. The most intuitive estimator corresponds to the euclidian distance, which is the situation presented in the figure.

2.1 Least squares

For N observations $\mathbf{y}$, the (ordinary) LS estimator $\hat{\theta}^N_{LS}$ minimizes

$$J_N(\theta) = \frac{1}{N}\|\mathbf{y} - \eta_X(\theta)\|^2 = \frac{1}{N}\sum_{k=1}^{N}[y(x_k) - \eta(\theta, x_k)]^2$$

with respect to $\theta \in \Theta \subset I\!R^p$. Here Θ is some feasible parameter set, usually a compact subset of $I\!R^p$. The method can be traced back to Gauss and Legendre. A straightforward extension corresponds to *weighted LS*, where the observation number k is weighted by some $w_k \geq 0$, $\hat{\theta}^N_{WLS}$ thus minimizes

$$J_N(\theta) = \frac{1}{N}\sum_{k=1}^{N} w_k[y(x_k) - \eta(\theta, x_k)]^2 .$$

The weight w_k may depend on x_k and we shall write $w_k = w(x_k)$.

When the model structure is LP, $\hat{\theta}^N_{WLS}$ can be calculated explicitly. Indeed, we have $\eta(\theta, x) = \mathbf{r}^\top(x)\theta$ for any x and $\eta_X(\theta)$ can be written as $\eta_X(\theta) = \mathbf{R}_X\theta$. Therefore, $\hat{\theta}^N_{WLS}$ minimizes $(\mathbf{y} - \mathbf{R}_X\theta)^\top\mathbf{W}(\mathbf{y} - \mathbf{R}_X\theta)$ with $\mathbf{W} = \mathrm{diag}(w_1, \ldots, w_N)$. This gives

$$\hat{\theta}^N_{WLS} = (\mathbf{R}_X^\top\mathbf{W}\mathbf{R}_X)^{-1}\mathbf{R}_X^\top\mathbf{W}\mathbf{y} \tag{2}$$

provided that $\mathbf{R}_X^\top\mathbf{W}\mathbf{R}_X$ has full rank (a condition for identifiability, see Section 1.6).

2.1.1 Data recursive LS

When the observations $y(x_k)$ are collected one after the other, *on-line estimation* is possible and the data recursive WLS estimator is obtained by the following recurrence equations

$$\begin{aligned}
\mathbf{P}_{k+1} &= \mathbf{P}_k - \frac{\mathbf{P}_k \mathbf{r}(x_{k+1}) \mathbf{r}^\top(x_{k+1}) \mathbf{P}_k}{w^{-1}(x_{k+1}) + \mathbf{r}^\top(x_{k+1}) \mathbf{P}_k \mathbf{r}(x_{k+1})}\,, \\
\hat{\theta}^{k+1}_{WLS} &= \hat{\theta}^k_{WLS} + \frac{\mathbf{P}_k \mathbf{r}(x_{k+1})}{w^{-1}(x_{k+1}) + \mathbf{r}^\top(x_{k+1}) \mathbf{P}_k \mathbf{r}(x_{k+1})} \\
&\quad \times [y(x_{k+1}) - \mathbf{r}^\top(x_{k+1}) \hat{\theta}^k_{WLS}]\,.
\end{aligned}$$

Let k_0 be the first integer such that $\mathbf{r}(x_1), \ldots, \mathbf{r}(x_{k_0})$ span $\mathbb{R}^p$. The recursion can be initialized at $k = k_0$ by

$$\begin{aligned}
\mathbf{P}_{k_0} &= \left[\sum_{i=1}^{k_0} w(x_i) \mathbf{r}(x_i) \mathbf{r}^\top(x_i) \right]^{-1}\,, \\
\hat{\theta}^{k_0}_{WLS} &= \mathbf{P}_{k_0} \sum_{i=1}^{k_0} \mathbf{r}(x_i) w(x_i) y(x_i)\,.
\end{aligned}$$

With this initialisation the estimator exactly coincides with expression (2). If N is large enough, one can simply initialize $\mathbf{P}_0$ at $C\mathbf{I}_p$, with $\mathbf{I}_p$ the p-dimensional identity matrix and C a large positive constant, and $\hat{\theta}^0_{WLS}$ at $\mathbf{0}$, the p-dimensional null vector. Indeed, if $(1/N) \sum_{i=1}^N w(x_i) \mathbf{r}(x_i) \mathbf{r}^\top(x_i)$ tends to a non-singular matrix as $N \to \infty$, the influence of this initialisation will asymptotically vanish.

2.1.2 Repetitions of observations

Suppose that only m values of x_i and w_i are different, with n_i observations $y_j(x_i)$, $j = 1, \ldots, n_i$, collected for the same x_i, $\sum_{i=1}^m n_i = N$.

One can then easily check that the WLS estimator is not modified when the $y_j(x_i)$'s, with $j = 1, \ldots, n_i$, are replaced by their mean $\bar{y}(x_i) = (1/n_i) \sum_{j=1}^{n_i} y_j(x_i)$, that is, $\hat{\theta}^N_{WLS}$ minimizes

$$J'_N = \frac{1}{N} \sum_{i=1}^m n_i w_i [\bar{y}(x_i) - \eta(\theta, x_i)]^2\,. \tag{3}$$

Indeed, we can write

$$\begin{aligned}
J_N(\theta) &= \frac{1}{N} \sum_{i=1}^m \sum_{j=1}^{n_i} w_i [y_j(x_i) - \eta(\theta, x_i)]^2 \\
&= \frac{1}{N} \sum_{i=1}^m \sum_{j=1}^{n_i} w_i [y_j(x_i) - \bar{y}(x_i) + \bar{y}(x_i) - \eta(\theta, x_i)]^2 \\
&= \frac{1}{N} \sum_{i=1}^m n_i w_i [\bar{y}(x_i) - \eta(\theta, x_i)]^2 + \frac{1}{N} \sum_{i=1}^m w_i \sum_{j=1}^{n_i} [y_j(x_i) - \bar{y}(x_i)]^2
\end{aligned}$$

where the second term does not depend on θ.

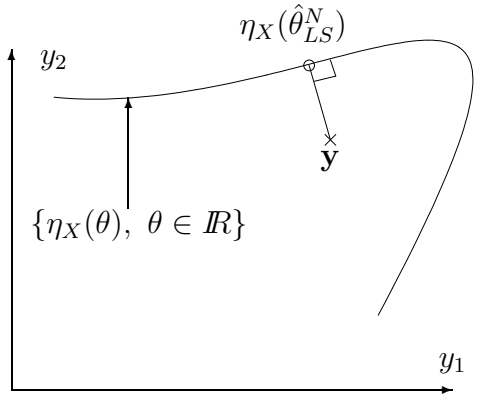

Figure 9: $\eta_X(\hat{\theta}^N_{LS})$ is the orthogonal projection of $\mathbf{y}$ on $\mathcal{E}_\eta$

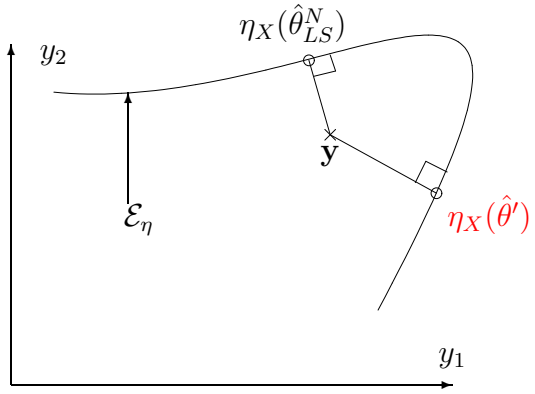

Figure 10: Local minima may always exist when $\mathcal{S}_\eta$ is curved

Suppose that the system corresponds to a model with parameters $\bar{\theta}$ corrupted by additive independent perturbations. The repetition of observations at x_i has then the same effect as if $\bar{y}(x_i)$ was moved closer to $\eta(\bar{\theta}, x_i)$ as the n_i's increase. In the space of arithmetic means of observations $\bar{y}(x_i)$, the expectation surface $\mathcal{E}_\eta$ is invariant by repetitions, and $\mathbf{y}$ thus moves closer to $\eta_X(\bar{\theta})$, a point on $\mathcal{E}_\eta$. In terms of criterion value, it corresponds to (3) tending to zero as the n_i's increase.

2.1.3 Local minima

We only consider the case of the LS estimator $\hat{\theta}^N_{LS}$, but the developments are valid for WLS estimation too (with a modification of the metric of the space).

The vector $\eta_X(\hat{\theta}^N_{LS})$ corresponds to the orthogonal projection of $\mathbf{y}$ on the expectation surface $\mathcal{E}_\eta = \{\eta_X(\theta),\ \theta \in I\!R^p\}$, see Figure 9. It should be clear from Figure 10 that local minima may always exist when the surface $\mathcal{S}_\eta$ is curved. When local minima exist, it is difficult to know if the solution to the minimisation of the estimation criterion $J_N(\theta)$ given by the optimizer corresponds to the global minimum or a local minimum, which can be far from $\hat{\theta}^N_{LS}$.

The existence of local minima depends on the distance of $\mathbf{y}$ to $\mathcal{E}_\eta$, on the curvature of $\mathcal{E}_\eta$ and on the size of the admissible parameter space Θ, see for instance [5, 6, 7, 8]. The following example (see [9]) illustrates the difficulties.

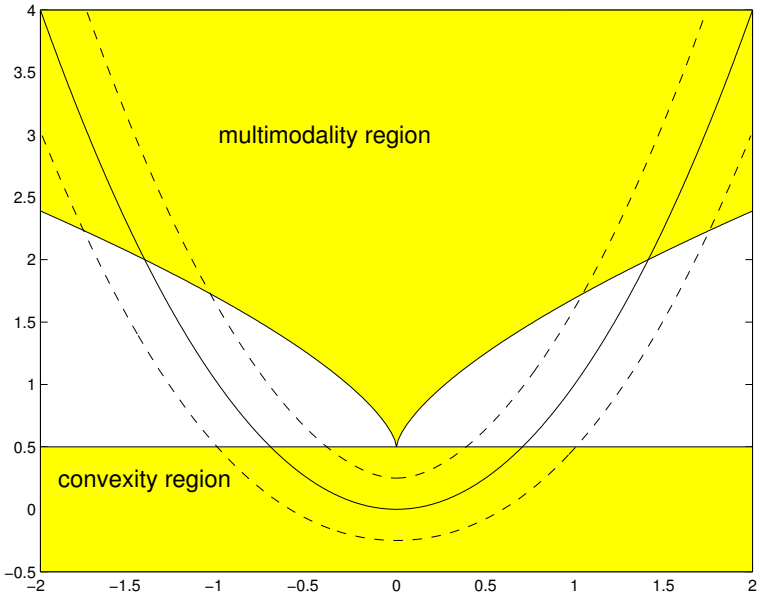

Figure 11: Local minima in Example 6 depending on the location of $\mathbf{y}$

Example 6 *We take* $\eta(\theta, x) = \theta\{x\}_1 + \theta^2\{x\}_2$, $\theta \in I\!R$, *and perform two observations, at* $x_1 = (1, 0)$ *and* $x_2 = (0, 1)$. *Three regions are delimited in Figure 11: the full-line parabola corresponds to* $\mathcal{E}_\eta$, *when* $\mathbf{y}$ *belongs to the multimodality region on the top (the region where* $J_N(\theta)$ *has two minima), a local minimum always exists; when* $\mathbf{y}$ *is in the convexity region on the bottom, the criterion* $J_N(\theta)$ *is a convex function of* θ, *which can only have one (global) minimum; the intermediate region corresponds to observations such that* $J_N(\theta)$ *has a unique minimum but is not convex. Also, one can show that when* $\mathbf{y}$ *lies between the two dashed curves, if two minima* $\hat{\theta}$ *and* $\hat{\theta}'$ *exist they are necessarily distant from each other by more than 1, that is,* $\|\hat{\theta} - \hat{\theta}'\| > 1$.

Such a precise analysis is possible only in very simple cases. Also, the distance of $\mathbf{y}$ *to* $\mathcal{E}_\eta$ *and the curvature of* $\mathcal{E}_\eta$ *do not give all the information, as shown by the following modification of the example.*

We modify $\eta(\theta, x)$ *when* $x < 0$ *and take now* $\eta(\theta, x) = [\theta\{x\}_1 + \theta^2\{x\}_2]\mathtt{I}_{I\!R^+}(\theta) + [\sin(\theta)\{x\}_1 + 2[1 - \cos(\theta)]\{x\}_2]\mathtt{I}_{I\!R^-}(\theta)$, *where* $\theta \in [-5.5, \infty)$ *and where* $\mathtt{I}_{\mathcal{A}}$ *denotes the indicator of the set* $\mathcal{A}$: $\mathtt{I}_{\mathcal{A}}(x) = 1$ *if* $x \in \mathcal{A}$ *and equals 0 otherwise. The two observations are still performed at* $x_1 = (1, 0)$ *and* $x_2 = (0, 1)$. *Figure 12 presents the new expectation surface* $\mathcal{S}_\eta$: *local minima may now exist whatever the value of* $\mathbf{y}$.

When optimizing $J_N(\theta)$ with some numerical algorithm it is important to know that there is no other better minimum than the estimated value which is obtained.

There is a trivial situation where the minimum $\eta_X(\hat{\theta})$ is unique: it is when $\mathcal{E}_\eta$ is *flat*: $\eta_X(\hat{\theta})$ is then the projection of $\mathbf{y}$ on a p-dimensional hyperplane. We can try to obtain this situation by playing with the design.

Example 6 *(continued)*

Repeat two observations *at* $x_2 = x_1 = (1, 0)$; $\mathcal{E}_\eta$ *then becomes*

$$\mathcal{E}_\eta = \{(-1, -1)^\top + \alpha(1, 1)^\top\,,\ \alpha > 0\}\,,$$

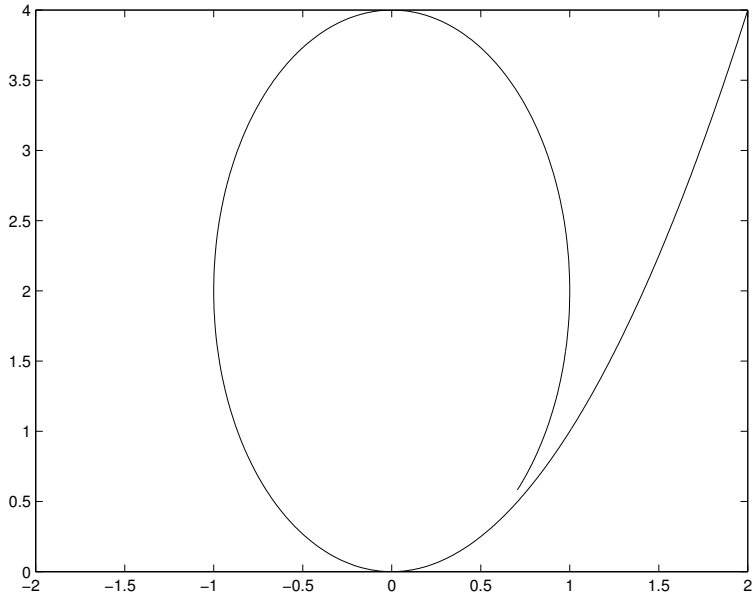

Figure 12: $\mathcal{E}_\eta$ for a modification of $\eta(\theta, x)$ in Example 6, two observations at $x_1 = (1, 0)$ and $x_2 = (0, 1)$

that is, a straight line. To any given $\mathbf{y}$ *corresponds a unique* $\hat{\eta}_X$*, its orthogonal projection on the line.*

However, we omitted one difficulty in previous example: the intrinsic curvature of the model is zero for almost any[1] θ but it is infinite for some particular values of θ, and this induces a loss of global identifiability. Everything happens as if we had folded $\mathcal{S}_\eta$ over itself.

Example 6 *(continued) Take* x_2 *not exactly equal to* x_1 *but close to* x_1*, e.g.,* $x_2 = (1, 0.1)$*. Figure 13 shows* $\mathcal{S}_\eta$*: it is folded and close to the straight line obtained when* $x_2 = x_1$*, the curvature is small for most values of* θ*, but large where the folding occurs.*

When $x_2 = x_1 = (1, 0)$*,* $\mathcal{S}_\eta$ *is completely folded and three different* θ *may give the same* $\eta_X(\theta)$*. Suppose that it is the case,* $\mathbf{y}$ *being projected on* $\hat{\eta} = \eta_X(\hat{\theta}_1) = \eta_X(\hat{\theta}_2) = \eta_X(\hat{\theta}_3)$ *for three different values* $\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3$*. We now perform a third observation, at* $x_3 \neq x_2 = x_1$*. The design* X' *is given by* $X' = (x_1, x_2 = x_1, x_3 \neq x_1)$ *with* $x_1 = (1, 0)$*,* $x_3 = (0, 1)$*. We can replace the first two observations by their arithmetic mean* $\bar{y}(x_1)$*, see Section 2.1.2. Consider the two dimensional space formed by* $\bar{y}(x_1)$ *and* $y(x_3)$*. In this space,* $\mathcal{S}_\eta$ *has the same form as in Figure 12 and the three estimates* $\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3$ *give three points* A, B, C *on* $\mathcal{E}_\eta$*, see Figure 14;* $\mathbf{y}$ *should then be close to* A*,* B *or* C*. One can then initialize the optimizer (the search algorithm) at* $\hat{\theta}_1$*, then* $\hat{\theta}_2$*, then* $\hat{\theta}_3$*, compare the estimators obtained in terms of* J_N *and decide which one to retain. The decision may be difficult when* $\mathbf{y}$ *is close to* A *or* B*, but we shall see in Section 2.1.4 how repetitions of the design* X' *can help.*

Following the steps presented in Example 6, we formulate the next recommendations to face the presence of local minima.

1. Start with a design X consisting of repetitions of observations (X should contain p different experimental conditions when $\dim(\theta) = p$). $\mathcal{E}_\eta$ is then flat and one given $\mathbf{y}$ results

[1] When the *intrinsic curvature* of the model is zero for any θ, the model is said *intrinsically linear*.

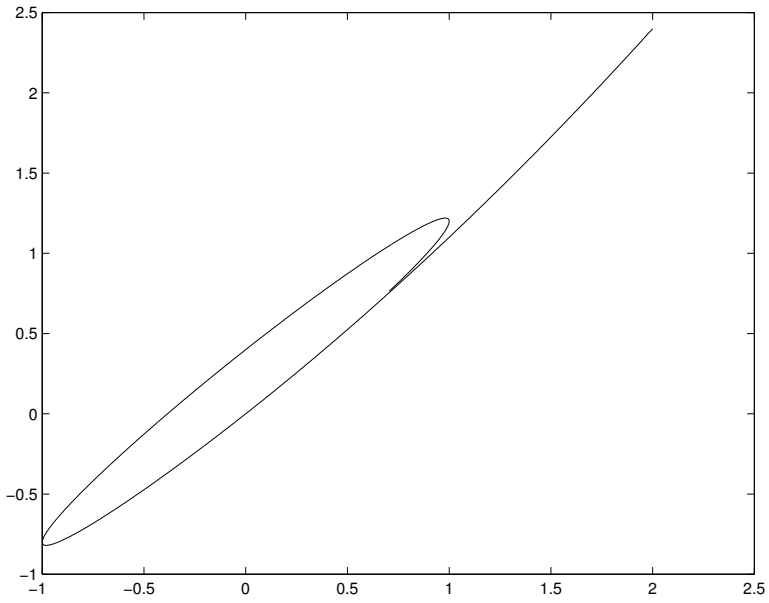

Figure 13: $\mathcal{E}_\eta$ for the same model as in Figure 12 but with $x_2 = (1, 0.1)$

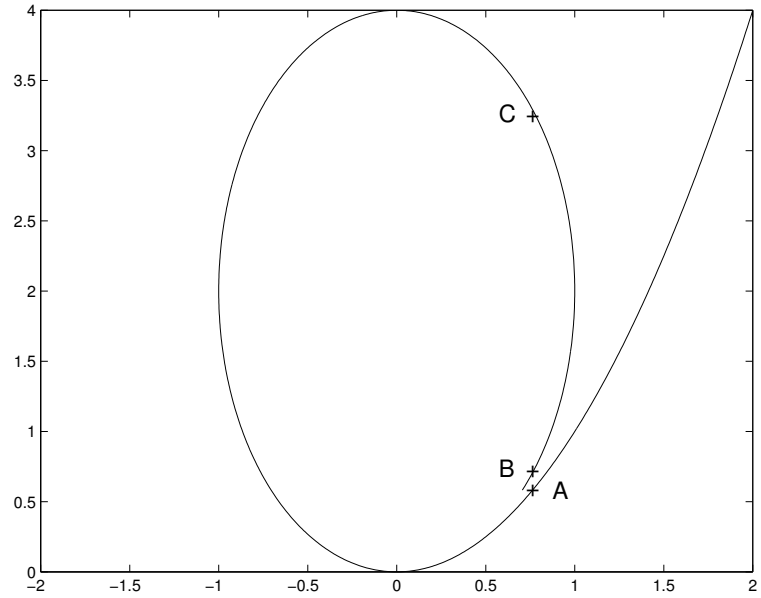

Figure 14: $\mathcal{E}_\eta$ for three observations at $x_1 = x_2 = (1, 0)$ and $x_3 = (0, 1)$ in Example 6

in one unique projection $\hat{\eta} = \eta_X(\hat{\theta})$. However, several $\hat{\theta}_i$'s may exist that give the same $\hat{\eta}$. One must thus solve the equations $\eta_X(\hat{\theta}) = \hat{\eta}$ w.r.t. $\hat{\theta}$ and *obtain all solutions* $\hat{\theta}_i$.

2. Use previous solutions $\hat{\theta}_i$ to initialize the optimizer with a design $X' \supset X$ (containing more points than X) such that there is global identifiability for X': $\eta_{X'}(\theta) = \eta_{X'}(\theta') \Rightarrow \theta = \theta'$. (Such a design exists if the structure is s.g.i.)

3. Even if X' ensures global identifiability, values of θ that are far away may correspond to values of $\eta_X(\theta)$ that are almost equal, see A, B in Figure 14. When $\mathbf{y}$ falls in such an area different estimates may exist that are associated with values of $J_N(\theta)$ almost similar. Deciding which estimated value to use is then difficult. However, repeating observations with X' has the same effect as moving $\mathbf{y}$ closer to $\mathcal{E}_\eta$, see Section 2.1.2, which helps to decide. This will receive a more formal interpretation in Section 2.1.4.

We conclude this section on local minima by a numerical example.

Example 7 *The regression model is* $\eta(\theta, \mathbf{x}) = \theta_1 x_1 + \theta_2 x_2 + \theta_1^3(1 - x_1) + \theta_2^2(1 - x_2)$. *When three observations are collected at* $\mathbf{x}^1 = (1,\ 0)^\top$, $\mathbf{x}^2 = (1,\ 1)^\top$ *and* $\mathbf{x}^3 = (0,\ 1)^\top$ *the surface* $\mathcal{E}_\eta$ *is curved, and local minima may therefore exist. We follow the steps recommended above, and start with a design* X *with repetitions:* $\mathbf{x}^1 = (1,\ 0)^\top$, $\mathbf{x}^2 = \mathbf{x}^3 = (1,\ 1)^\top$. *The associated vector of observations is* $\mathbf{y} = (5, -12, -8)^\top$. *Figure 15 presents the level sets for the estimation criterion* $J_N(\theta)$ *(distance from* $\mathbf{y}$ *to* $\eta_X(\theta)$*). The folding of* $\mathcal{E}_\eta$ *makes the model only locally identifiable:* $\hat{\theta}^1 = (\hat{\theta}_1, \hat{\theta}_2)^\top$ *and* $\hat{\theta}^2 = (\hat{\theta}_1 + 2\hat{\theta}_2 - 1, 1 - \hat{\theta}_2)^\top$ *give the same* $\eta_X(\theta)$. *For any* $\mathbf{y}$, *the projection on* $\mathcal{E}_\eta$ *is unique, the optimizer will yield one of the two points* $\hat{\theta}^1$ *or* $\hat{\theta}^2$, *from which the other is easily obtained.*

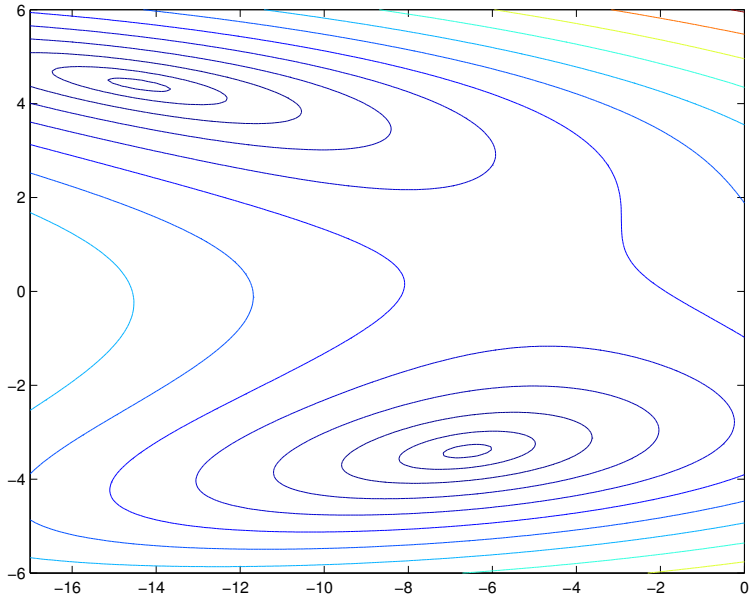

Figure 15: Level sets for $J_N(\theta)$ for three observations at $\mathbf{x}^1 = (1,\ 0)^\top$, $\mathbf{x}^2 = \mathbf{x}^3 = (1,\ 1)^\top$, $\mathbf{y} = (5, -12, -8)^\top$, Example 7

We complement X *by a fourth observation at* $\mathbf{x}^4 = (0,\ 1)^\top$, *that is,* X' *consists of* $\mathbf{x}^1 = (1,\ 0)^\top$, $\mathbf{x}^2 = \mathbf{x}^3 = (1,\ 1)^\top$ *and* $\mathbf{x}^4 = (0,\ 1)^\top$. *The vector of observations becomes* $\mathbf{y} =$

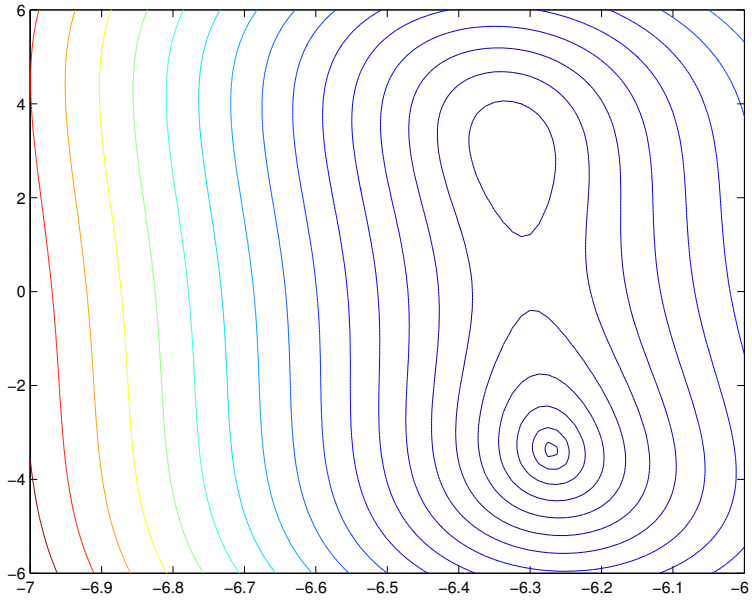

Figure 16: Level sets for $J_N(\theta)$ for four observations at $\mathbf{x}^1 = (1,\ 0)^\top$, $\mathbf{x}^2 = \mathbf{x}^3 = (1,\ 1)^\top$ and $\mathbf{x}^4 = (0,\ 1)^\top$, $\mathbf{y} = (5, -12, -8, -250)^\top$, Example 7

$(5, -12, -8, -250)^\top$. *Figure 16 presents the new configuration of the level sets for* $J_N(\theta)$. *When the search is successively initialized at each of the minima in Figure 15, the global optimum will easily be located.*

2.1.4 Asymptotic properties

In order to study the asymptotic properties of an estimator we need to specify how the experimental variables (design points) x_k's are generated. In particular, the proofs are "easy" in two cases[2]:

A) the x_k's form a sequence of i.i.d. random variables (vectors) with probability measure ξ (*random design*);

B) the sequence (x_k) accumulates on a finite number K of points x^i receiving weights $\xi(x^i) > 0$ with $\sum_{i=1}^K \xi(x^i) = 1$ (convergence to a *discrete design measure*).

We suppose A or B satisfied in what follows. Note in particular that B contains the case where a given design X is repeated. We write a.s. for *almost surely*.

Consider an estimator $\hat{\theta}^N$ that minimizes $J_N(\theta)$. Then the asymptotic properties of $\hat{\theta}^N$ are related to those of $J_N(\theta)$ and

- if $J_N(\theta) \overset{\text{a.s.}}{\to} J(\theta)$ (almost surely) uniformly in θ (when $J_N(\theta)$ can be written as a sum of stochastically independent terms, this corresponds to a *uniform Strong Law of Large Numbers*) and if $J_N(\theta)$ is continuous in θ for any N and $J(\theta)$ has a unique minimum at $\theta = \bar{\theta}$, then

$$\hat{\theta}^N \overset{\text{a.s.}}{\to} \bar{\theta}\,,\ N \to \infty\,;$$

[2]The standard reference for the asymptotic properties of the LS estimator in a general situation is [17]. One can also refer to [1] for rigorous developments. In both cases the proofs are more complicated than in the situations A and B below.

- the local behavior of $J(\theta)$ around $\bar{\theta}$ (its derivatives) and the uniform almost sure convergence of the derivatives of $J_N(\theta)$ give the asymptotic normal distribution of $\hat{\theta}^N$ around $\bar{\theta}$. We shall denote $\mathcal{N}(\mathbf{m}, \mathbf{C})$ the normal distribution with mean $\mathbf{m}$ and variance-covariance matrix $\mathbf{C}$.

Consider (W)LS estimation in nonlinear regression. We suppose that

$$y(x_k) = \eta(\bar{\theta}, x_k) + \varepsilon_k\,,\ \bar{\theta} \in \Theta\,,\ x_k \in \mathcal{X}\,,\ k = 1, 2, \ldots \tag{4}$$

where (ε_k) is a sequence of independent random variables with $\mathtt{E}_x(\varepsilon_k) = 0$ and $\mathtt{E}_x(\varepsilon_k^2) = \sigma^2(x)$ with $0 < a < \sigma^2(x) < b < \infty$ for any $x \in \mathcal{X}$.

We shall use the following technical assumptions:

$\mathbf{H}_\Theta$: Θ is a compact subset of $\mathbb{R}^p$ such that $\Theta \subset \overline{\mathrm{int}(\Theta)}$, the closure of the interior $\mathrm{int}(\Theta)$ of Θ.

$\mathbf{H1}_\eta$: $\eta(\theta, x)$ is bounded on $\mathcal{X} \times \Theta$ and $\eta(\theta, x)$ is continuous in $\theta \in \Theta\ \forall x \in \mathcal{X}$.

$\mathbf{H2}_\eta$: $\bar{\theta} \in \mathrm{int}(\Theta)$ and $\forall x \in \mathcal{X}$, $\eta(\theta, x)$ is two times continuously differentiable with respect to $\theta \in \mathrm{int}(\Theta)$, these first two derivatives are bounded on $\mathcal{X} \times \mathrm{int}(\Theta)$.

The WLS estimator $\hat{\theta}^N_{WLS}$ minimizes

$$J_N(\theta) = \frac{1}{N} \sum_{k=1}^{N} w(x_k)[y(x_k) - \eta(\theta, x_k)]^2\,,$$

with $w(x)$ bounded on $\mathcal{X}$. One can show that it satisfies the following:

- if $\mathbf{H}_\Theta$ and $\mathbf{H1}_\eta$ are satisfied together with the estimability condition[3]

$$\int_{\mathcal{X}} w(x)[\eta(\theta, x) - \eta(\theta', x)]^2 \xi(dx) = 0 \Leftrightarrow \theta' = \theta\,,$$

then $\hat{\theta}^N_{WLS} \overset{\text{a.s.}}{\to} \bar{\theta}$, $N \to \infty$.

- If, moreover, $\mathbf{H2}_\eta$ is satisfied and the matrix

$$\mathbf{M}_1(\xi, \bar{\theta}) = \int_{\mathcal{X}} w(x) \frac{\partial \eta(\theta, x)}{\partial \theta}_{|\bar{\theta}} \frac{\partial \eta(\theta, x)}{\partial \theta^\top}_{|\bar{\theta}} \xi(dx)$$

has full rank, then

$$\sqrt{N}(\hat{\theta}^N_{WLS} - \bar{\theta}) \overset{\text{d}}{\to} z \sim \mathcal{N}(0, \mathbf{C}(w, \xi, \bar{\theta}))\,,\ N \to \infty\,,$$

where

$$\mathbf{C}(w, \xi, \theta) = \mathbf{M}_1^{-1}(\xi, \theta)\mathbf{M}_2(\xi, \theta)\mathbf{M}_1^{-1}(\xi, \theta)$$

with

$$\mathbf{M}_2(\xi, \theta) = \int_{\mathcal{X}} w^2(x)\sigma^2(x) \frac{\partial \eta(\theta, x)}{\partial \theta} \frac{\partial \eta(\theta, x)}{\partial \theta^\top} \xi(dx)\,.$$

[3]Note that for an asymptotically discrete design, which corresponds to situation B, an integral written $\int_{\mathcal{X}} f(x)\,\xi(dx)$ corresponds to the discrete sum $\sum_{i=1}^{K} f(x^i)\xi(x^i)$.

One can show that $\mathbf{C}(w,\xi,\bar{\theta}) - \mathbf{M}^{-1}(\xi,\bar{\theta})$ is non-negative definite for any $w(x)$, where

$$\mathbf{M}(\xi,\bar{\theta}) = \int_{\mathcal{X}} \sigma^{-2}(x) \frac{\partial \eta(\theta,x)}{\partial \theta}_{|\bar{\theta}} \frac{\partial \eta(\theta,x)}{\partial \theta^\top}_{|\bar{\theta}} \xi(dx)$$

and $\mathbf{C}(w,\xi,\bar{\theta}) = \mathbf{M}^{-1}(\xi,\bar{\theta})$ for $w(x) = c\,\sigma^{-2}(x)$ with c a positive constant. Weighting by the inverse of the variance of the errors is thus optimum among WLS estimators in terms of asymptotic variance of the estimator.

Consider now the case of n *repetitions of a design* $X = (x^1, \ldots, x^m)$, with $N = m\,n$ the total number of observations. The criterion $J'_N(\theta)$ obtained by replacing the observations by their mean, see (3), satisfies

$$J'_N(\theta) \overset{\text{a.s.}}{\to} \frac{1}{m} \sum_{i=1}^{m} w(x^i)[\eta(\theta,x^i) - \eta(\bar{\theta},x^i)]^2\,,\ n \to \infty$$

and the convergence is uniform in θ under $\mathbf{H}_\Theta$ and $\mathbf{H1}_\eta$. Repetitions may thus help to solve ambiguity problems such as that encountered in Example 6. Consider again Figure 14, if $\mathbf{y}$ is close to A or B there exist two distant estimates $\hat{\theta}_1$ and $\hat{\theta}_2$ associated with almost similar criterion values $J_N(\hat{\theta}_1) \simeq J_N(\hat{\theta}_2)$. By repetitions of the design, J'_N will tend to zero for one the the $\hat{\theta}_i$'s only, thus indicating that the other estimate corresponds to a local optimum.

2.1.5 Errors with parameterized variance

In some situations the assumption of constant variance σ^2 for the errors ε_k in the model (4) is not satisfied. When the variance is non stationary, it is reasonable to suppose that it depends on x. Assume that

$$\mathtt{E}_{x_k}(\varepsilon_k^2) = \beta\lambda(\bar{\theta},x_k)\,,\ \beta > 0\,.$$

We shall use the following technical assumptions.

H1$_\lambda$: $\lambda(\bar{\theta},x)$ is bounded on $\mathcal{X}$, $\lambda^{-1}(\theta,x)$ is bounded on $\mathcal{X}\times\Theta$ and $\lambda(\theta,x)$ is continuous on Θ $\forall x \in \mathcal{X}$.

H2$_\lambda$: $\forall x \in \mathcal{X}$, $\lambda(\theta,x)$ is two times continuously differentiable with respect to $\theta \in \mathrm{int}(\Theta)$, these first two derivatives are bounded on $\mathcal{X}\times\mathrm{int}(\Theta)$.

The ordinary LS estimator is still (strongly) consistent and asymptotically normal in this context of non stationary variance under the assumptions used in the previous section. However, we have seen that among WLS estimators the variance was minimum when the weights were given by $\lambda^{-1}(\bar{\theta},x_k)$. The problem is that they cannot be used since $\bar{\theta}$ is unknown...

It would be tempting to use weights that depend on θ, that is, to minimize

$$\frac{1}{N}\sum_{k=1}^{N} \frac{[y(x_k) - \eta(\theta,x_k)]^2}{\lambda(\theta,x_k)}\,.$$

However, this method *must be rejected* since one can easily show that the corresponding estimator is *not consistent*.

A simple method consists in using two steps of LS estimation:

1. use ordinary LS ($w(x) \equiv 1$), which gives the estimator $\hat{\theta}^N_{LS}$;

2. use WLS with weights $w(x_k) = \lambda^{-1}(\hat{\theta}^N_{LS}, x_k)$; denote by $\hat{\theta}^N_{TSLS}$ the (Two-Stage LS) estimator obtained at this second stage.

One can then show that when the experimental design satisfies the conditions A or B of Section 2.1.4

- $\mathbf{H}_\Theta$, $\mathbf{H1}_\eta$ and $\mathbf{H1}_\lambda$ together with the estimability condition

$$\int_{\mathcal{X}} \lambda^{-1}(\bar{\theta}, x)[\eta(\theta, x) - \eta(\theta' x)]^2 \xi(dx) = 0 \Leftrightarrow \theta' = \theta\,,$$

  imply $\hat{\theta}^N_{TSLS} \overset{\text{a.s.}}{\to} \bar{\theta}$, $N \to \infty$.

- If, moreover, $\mathbf{H2}_\eta$ and $\mathbf{H2}_\lambda$ are satisfied and the matrix

$$\mathbf{M}(\xi, \bar{\theta}) = \int_{\mathcal{X}} \lambda^{-1}(\bar{\theta}, x) \frac{\partial \eta(\theta, x)}{\partial \theta}_{|\bar{\theta}} \frac{\partial \eta(\theta, x)}{\partial \theta^\top}_{|\bar{\theta}} \xi(dx)$$

  has full rank, then

$$\sqrt{N}(\hat{\theta}^N_{TSLS} - \bar{\theta}) \overset{\text{d}}{\to} z \sim \mathcal{N}(0, \beta \mathbf{M}^{-1}(\xi, \bar{\theta}))\,,\ N \to \infty\,.$$

Note that under suitable conditions $\hat{\theta}^N_{TSLS}$ thus has minimum variance among WLS estimators.

The same idea can be repeated for more than two steps, with $\hat{\theta}^N_k$ minimizing

$$J_{k,N}(\theta) = \frac{1}{N} \sum_{k=1}^{N} \lambda^{-1}(\hat{\theta}^N_{k-1}, x_k)[y(x_k) - \eta(\theta, x_k)]^2\,,\ k \geq 2$$

and $\hat{\theta}^N_1 = \hat{\theta}^N_{LS}$. For a fixed number N of observations, one can use similar steps until convergence of the estimator, that is $\hat{\theta}^N_k \simeq \hat{\theta}^N_{k-1}$ (one can show that for N large enough convergence will occur with probability one). The corresponding estimator is called *iteratively re-weighted LS*. It has the same asymptotic properties as the TSLS estimator presented here, although their performances may differ for finite N.

One can refer to [18, 10] for an alternative method, inspired from maximum likelihood estimation, see Section 2.4, and to [4] for a comparison of this method with the TSLS approach presented here (which is more robust to misspecification of the variance function).

2.2 Other distances, other estimators

Taking other distances than the quadratic (Euclidian) gives other estimators than LS. For instance, one can use absolute values (L_1-norm), which results in the criterion

$$J_N(\theta) = \frac{1}{N} \|\mathbf{y} - \eta_X(\theta)\|_1 = \frac{1}{N} \sum_{i=1}^{N} |y(x_k) - \eta(\theta, x_k)|$$

for a regression model. It is important to note that $J_N(\theta)$ is not differentiable everywhere, which may cause difficulties for its optimisation. We shall see in Section 2.3 (Huber's M-estimator) how $J_N(\theta)$ can be modified to avoid this difficulty.

The optimum of $J_N(\theta)$ is not always unique, not even for s.g.i. models structures, as illustrated by the following example.

Example 8 *Take* $\eta(\theta, x_1) = \theta_1 + \theta_2^2$, $\eta(\theta, x_2) = \eta(\theta, x_3) = \theta_1 + \theta_2$ *and suppose that* $\mathbf{y} = (5, 2, 4)$. *Then* $J_N(\theta) = 2$ *for any* θ *such that* $\theta_1 + \theta_2^2 = 5$ *and* $2 \leq \theta_1 + \theta_2 \leq 4$.

Also, the asymptotic properties are more difficult to derive than for LS estimation due to the lack of differentiability, see for instance Chap. 5 of [32].

There is, however, one situation where L_1-estimation is easy to use: when the model structure is LP. Suppose that $\eta(\theta, x) = \mathbf{r}^\top(x)\theta$. Then, the minimisation of $J_N(\theta)$ is equivalent to the minimisation of $\sum_{k=1}^N \alpha_k$ under the constraints

$$-\alpha_k \leq y(x_k) - \mathbf{r}^\top(x_k)\theta \leq \alpha_k\,,\ k = 1, \ldots, N\,.$$

There are $N + p$ variables ($\alpha = (\alpha_1, \ldots, \alpha_N)^\top$ and θ), $2N$ constraints linear in α and θ, and the objective $\sum_{k=1}^N \alpha_k$ is linear too. This is thus a linear programming problem and many tools exist to solve it.

Despite all these difficulties L_1-estimation has an advantage over LS estimation. It concerns *robustness* with respect to outliers ("bad data"), which we illustrate by an example.

Example 9 *Consider a regression model with observations given by*

$$y(x_i) = \bar{\theta}_1 \exp(-\bar{\theta}_2 x_i) + \epsilon_i\,,$$

where the unknown value[4] *of* $\bar{\theta}$ *is* $(1, 2)^\top$, *the errors* ϵ_i *are i.i.d. normal* $\mathcal{N}(0, \sigma^2)$ *with* $\sigma = 5\ 10^{-3}$. *The design corresponds to 200 points* x_i *equally spaced in* $[0, 1]$.

Suppose that a failure of the sensor used to collect observations occurs from $i = 11$ *to* $i = 30$ *and from* $i = 50$ *to* $i = 64$; *the corresponding values of* $y(x_i)$ *are then equal to zero. The observations correspond to the dots in Figure 17, with* i *on the horizontal axis. The predicted response* $\eta(\hat{\theta}_{LS}^N, x_i)$ *for the LS estimator is indicated by the dashed line in the same figure. The outliers corresponding to the failure of the sensor have a very strong influence on* $\hat{\theta}_{LS}^N$ *and the response is strongly attracted by the observations set at zero. This is due to the fact that the LS criterion uses squared errors, so that large errors have a very strong effect. The predicted response for* L_1*-estimation is indicated by the full line, the criterion now uses absolute values of errors, large errors have less influence and the curve is less attracted towards the* x *axis.*

2.3 M-estimation

LS and L_1-estimation are particular cases of M-estimation where the criterion to be minimized is given by

$$J_N(\theta) = \frac{1}{N} \sum_{i=1}^N \rho[y(x_i) - \eta(\theta, x_i)]$$

with ρ a function which is minimum at zero.

For instance, the Huber estimator satisfies

$$\rho(e) = \begin{cases} e^2/2 & \text{if } |e| < \delta \\ \delta|e| - \delta^2/2 & \text{otherwise,} \end{cases}$$

with δ some positive threshold. A plot of $\rho(e)$ is given in Figure 18 (full line), together with a plot of the absolute value of its first derivative (dashed line).

[4] The value is unknown for the estimator, we use this value for simulating data, which will permit to compare the estimated response with the true one.

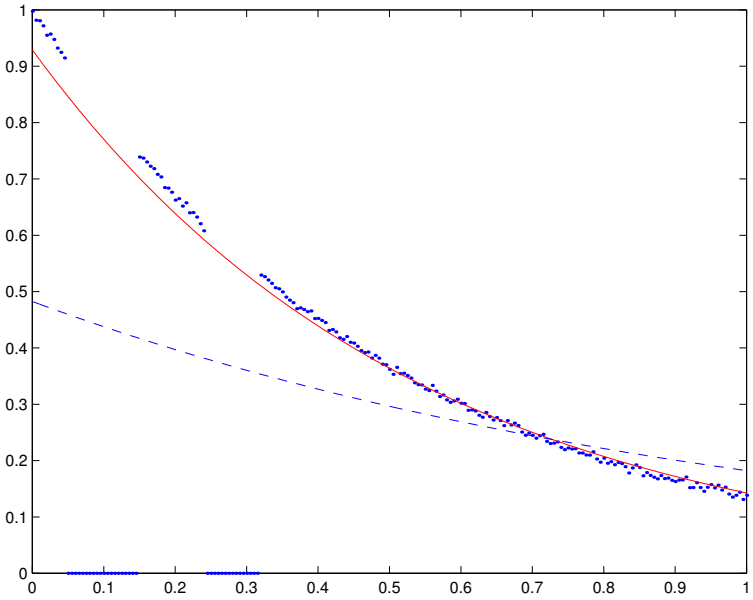

Figure 17: Observations (dots), model response for LS estimation (dashed line), model response for L_1-estimation (full line), Example 9

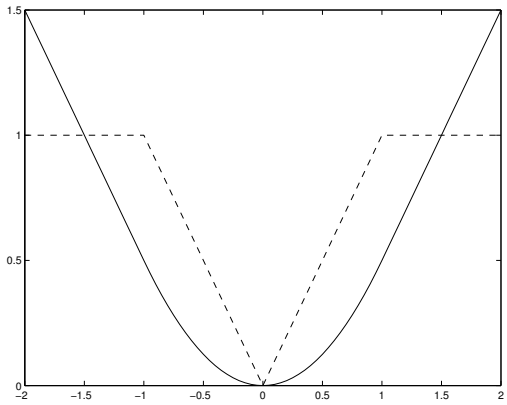

Figure 18: Huber's function ρ (full line) and absolute value of its derivative ρ' (dashed line)

Tukey's estimator is given by

$$\rho(e) = \begin{cases} [e^2 - e^4/\delta^2 + e^6/(3\delta^4)]/2 & \text{if } |e| < \delta \\ \delta^2/6 & \text{otherwise.} \end{cases}$$

A plot of $\rho(e)$ is given in Figure 19 (full line), together with a plot of the absolute value of its first derivative (dashed line).

The behavior of the derivative ρ' explains why Huber's estimator is called *non-redescending* whereas Tukey's is called *redescending*. Notice that Huber's estimator makes a smooth compromise between LS estimation (for small errors, $|e| < \delta$) and L_1-estimation (for large errors). It thus preserves the robustness property of L_1-estimation and at the same time is differentiable everywhere, so that standard optimisation algorithm can be used for its minimisation.

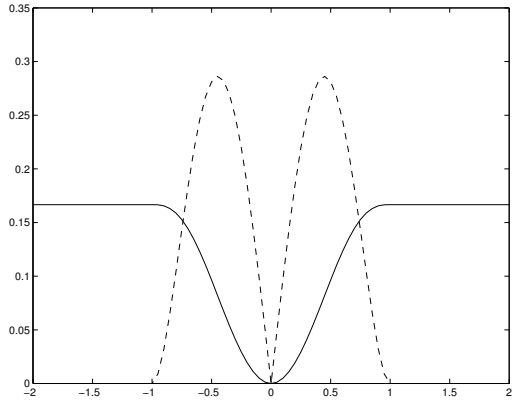

Figure 19: Tukey's function ρ (full line) and absolute value of its derivative ρ' (dashed line)

2.4 Maximum likelihood

2.4.1 The ML estimator

The likelihood of $\mathbf{y}$ corresponds to the conditional probability density of the observations $\mathbf{y}$ given the model parameters θ, which we shall denote $\pi(\mathbf{y}|\theta)$. When we estimate parameters, $\mathbf{y}$ is given (observations) and the maximum likelihood (ML) estimator $\hat{\theta}^N_{ML}$ maximizes $\pi(\mathbf{y}|\theta)$.

Consider a regression model with N observations corrupted by i.i.d. errors with probability density function (p.d.f.) $\varphi(\cdot)$. Then,

$$\pi(\mathbf{y}|\theta) = \prod_{k=1}^{N} \pi[y(x_i)|\theta] = \prod_{k=1}^{N} \varphi[y(x_k) - \eta(\theta, x_k)] \,.$$

Since the function logarithm is strictly increasing, we can equivalently minimize

$$-\log \pi(\mathbf{y}|\theta) = \sum_{k=1}^{N} -\log \varphi[y(x_k) - \eta(\theta, x_k)] \,.$$

The ML estimator is therefore a M-estimator for $\rho(e) = -\log \varphi(e)$: the estimation method is adapted to the distribution of the perturbations. Different p.d.f. φ yield different estimators: LS for Gaussian distributions, L_1-estimation for Laplace distributions, etc.

More generally, $\varphi(\cdot)$ may depend on x, with φ_{x_k} the p.d.f. of the error ε_k that corrupts the observation $y(x_k)$, and $\hat{\theta}^N_{ML}$ then minimizes

$$\sum_{k=1}^{N} -\log \varphi_{x_k}[y(x_k) - \eta(\theta, x_k)] \,.$$

One can show that under suitable assumptions, similar to those used in Section 2.1.4, $\sqrt{N}(\hat{\theta}^N_{ML} - \bar{\theta})$ is asymptotically normal with *minimum variance*, given by the inverse of the *Fisher information matrix*:

$$\sqrt{N}(\hat{\theta}^N_{ML} - \bar{\theta}) \xrightarrow{\mathrm{d}} z \sim \mathcal{N}(0, \mathbf{M}_F^{-1}(\xi, \bar{\theta})) \,, \; N \to \infty \,,$$

with

$$\mathbf{M}_F(\xi, \bar{\theta}) = \int_{\mathcal{X}} I_\varphi(x) \, \frac{\partial \eta(\theta, x)}{\partial \theta}_{|\bar{\theta}} \frac{\partial \eta(\theta, x)}{\partial \theta^\top}_{|\bar{\theta}} \xi(dx)$$

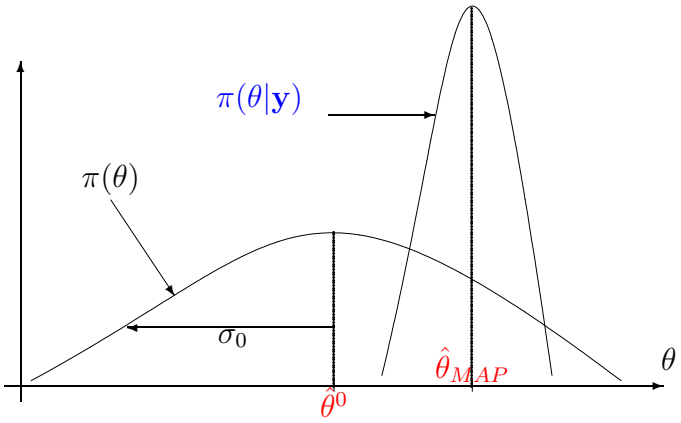

Figure 20: Bayesian estimation

where $I_\varphi(x)$ is the Fisher information for location,

$$I_\varphi(x) = \int_{-\infty}^{\infty} \left[\frac{\varphi_x'(e)}{\varphi_x(e)}\right]^2 \varphi_x(e)\, de\,.$$

2.4.2 Model structure selection

Consider the situation where several structures $\mathcal{M}_i$ compete to describe the behavior of a system $\mathcal{S}$. For instance, they may correspond to behavioral models with increasing complexity, that is, increasing number of parameters, with p_i parameters for $\mathcal{M}_i$.

We naturally wish to find the best structure and estimate its parameters. The *Akaike Information Criterion* gives probably the most famous method. It is given by

$$j_{AIC}(\theta, i) = \frac{1}{N}\left[-\log \pi(\mathbf{y}|\theta) + p_i\right]$$

to be minimized with respect to θ and i. When the model structure is fixed, minimizing $j_{AIC}(\theta, i)$ with respect to θ corresponds to maximum likelihood estimation. Let $\hat{\theta}_i$ be the estimated value of θ for $\mathcal{M}_i$, the idea is to choose $\mathcal{M}_{\hat{i}}$ with $j_{AIC}(\hat{\theta}_i, i)$ minimum. Note that complex structures are penalized due to the term p_i.

Other selection methods exist that rely on similar ideas of penalizing complex structures to avoid over-parameterization (unnecessary complexity), see, e.g., [33]. They differ by their definition of penalization (the way p_i enters the criterion).

2.5 Bayesian estimation

In Bayesian estimation the parameters θ are considered as random variables, with a known (or rather guessed) p.d.f. $\pi(\theta)$ (the *prior distribution*). After N observations $\mathbf{y}$ we can then construct the *posterior distribution* of the parameters $\pi(\theta|\mathbf{y})$. It will be more concentrated than the prior due to the information added by the observations. Figure 20 shows the situation.

Applying Bayes rule we obtain,

$$\pi(\theta|\mathbf{y}) = \frac{\pi(\theta, \mathbf{y})}{\pi(\mathbf{y})} = \frac{\pi(\mathbf{y}|\theta)\pi(\theta)}{\pi(\mathbf{y})}$$

The maximum *a posteriori* estimator $\hat{\theta}^N_{MAP}$ maximizes $\pi(\theta|\mathbf{y})$, that is, maximizes

$$\underbrace{\log \pi(\mathbf{y}|\theta)}_{\text{log-likelihood}} + \underbrace{\log \pi(\theta)}_{\text{log prior}} .$$

Both the information on the nature of the perturbations (log-likelihood) and the information on the model parameters (log prior) are thus taken into account.

Under suitable assumptions, $\hat{\theta}^N_{MAP}$ has the same asymptotic properties as the ML estimator.

2.6 Prediction-error methods

So far we only detailed the case of regression models. There exist situations where the observations y_k are not independent. The idea of prediction-error methods is to construct variables $e(x_k, \theta)$, the prediction errors, that form a sequence of *independent* random variables when $\theta = \bar{\theta}$ (under the assumption that the data are generated by the model $\mathcal{M}(\bar{\theta})$ corrupted by some perturbations). For a regression model, they correspond to the output errors, the differences between the observations and the model responses.

Consider for instance the Box & Jenkins model, frequently used in time series. One has

$$y = F(\theta, q)u + G(\theta, q)\epsilon \,,$$

where q^{-1} corresponds to the delay operator, that is, $q^{-i}x(k) = x(k-i)$ for any variable x, and F and G are rational functions in q^{-1}, e.g., $F(\theta, q) = (\theta_1 q^{-1} + \theta_2 q^{-3})/(q^{-1} + \theta_3 q^{-2})$. The prediction errors are then given by

$$e(k, \theta) = G^{-1}(\theta, q)[y(k) - F(\theta, q)u(k)]$$

and are obtained by filtering the sequences of inputs $u(k)$ and outputs $y(k)$.

The estimation criteria presented in the previous sections must then be applied to the prediction errors. One may refer especially to [23] for a detailed presentation of identification for such dynamical systems, see also [14, 3].

3 Optimisation

The problem consists in minimizing $J_N(\theta)$ with respect to $\theta \in \Theta$. The method to be used depends on several factors among which:

1. the definition of the admissible parameter set which may involve some constraints to be taken into account;
2. the fact that $J_N(\theta)$ may or may not be differentiable with respect to θ;
3. the possible presence of local minima;
4. the dimension p of θ;
5. the online collection of the data which may call for the online estimation of the model parameters.

Optimisation is by itself a very broad topic, and we shall only briefly comment the five points above.

1. Parameter estimation most often does not involve constraints. Indeed, suppose that a model structure requires some function $f(\theta)$ to be negative for the model to make sense. Imposing the constraint during the optimisation of $J_N(\theta)$ may then produce two types of solutions. Either $f(\hat{\theta}^N) < 0$, in which case the model makes sense, but ignoring the constraint would have produced the same estimate $\hat{\theta}^N$, or the constraint is active and $f(\hat{\theta}^N) = 0$. In that case the model is just on the edge, and it is advisable to check its validity.

   The typical case where constraints have to be introduced is when the computation of the model response involves some simulations that can be carried out only for some admissible parameter values. For instance, a differential equation may become unstable depending on the value of θ. The role of the constraints is then to avoid some "bad regions" *during the search* of the estimator carried out by the optimizer.

   Optimisation in the presence of constraints is more difficult than without, and constraints should be removed as far as possible. Sometimes this can be done by a reparameterization of the model (e.g., replace θ_1 that should be positive by $\theta_1'^2$ which is always positive).

2. It is important that $J_N(\theta)$ is differentiable. Optimisation methods for non-differentiable criteria exist, but are less standard and usually slower than methods for differentiable problems. Also, standard methods that do not use derivatives *do not necessarily work when the criterion is not differentiable*: for instance, the Nelder-Mead simplex algorithm [24] or the Powell method [27] should not be used for non-differentiable criteria.

   It is therefore advisable to modify a non-differentiable criterion to make it differentiable, compare for instance Huber's M-estimator of Section 2.3 with the L_1-estimator of Section 2.2.

3. The presence of local minima may require the use of global optimisation algorithms. Most of them (genetic algorithms, simulated annealing, etc.) do not guarantee that they have reached the global minimum. Interval methods (see e.g. [16]) provide guaranteed results but can be slow when $\dim(\theta)$ is large. Hence one should try to remove local minima as far as possible. Some indications on how to proceed have been given in Section 2.1.3.

4. When $J_N(\theta)$ is differentiable, there are no constraints on θ and local solutions are acceptable, one still has to choose the method which is most adapted (i.e., that gives the solution in shortest time). This is not mandatory if the problem has only to be solved one time, but it becomes an issue if similar problems have to be solved routinely. Then, if the problem corresponds to LS estimation, the Gauss-Newton algorithm is recommended. If not, quasi-Newton methods [12] (also called variable metric) can be used for $\dim(\theta)$ not too large (say less than 20, because matrices are manipulated), and conjugate gradients algorithms [13] for large to very large dimensions (in image processing applications for instance). Different implementations exist, giving different names for algorithms, for instance Davidon-Fletcher-Powell for quasi-Newton and Polak-Ribière for conjugate gradients.

5. There exist recursive methods that update the estimator $\hat{\theta}$ after the arrival of each new observation. So, if $\hat{\theta}^N$ is the estimator after $(x_1, y_1), \ldots, (x_N, y_N)$ are known, after the

new data (x_{N+1}, y_{N+1}) the estimator becomes $\hat{\theta}^{N+1} = \hat{\theta}^N + f(\hat{\theta}^N, y_{N+1}, x_{N+1})$. In the case of LS estimation this is similar to a linearisation of the problem, followed by the application of one step of recursive LS, see Section 2.1.1. Such recursive methods also exist for maximum likelihood estimation. Under suitable assumptions the asymptotic behavior of recursive methods is the same as for off-line, non recursive, ones (but the proofs of these asymptotic properties are more difficult).

A general recommendation is *not to try to implement one's own algorithm and use existing subroutines from software libraries.* Also, comparing different methods on the same problem often proves useful.

Efficient optimizers require the computation of the first-order derivatives of the criterion with respect to the components $[\theta]_i$ of θ, $i = 1, \ldots, p$. For regression models, the estimation criterion $J_N(\theta)$ depends on θ through the model responses $\eta(\theta, x_k)$, so that the derivatives of $J_N(\theta)$ can be obtained by the computation of the sensitivity functions[5] $\partial\eta(\theta, x_k)/\partial[\theta]_i$, $i = 1, \ldots, p$. When $\eta(\theta, x)$ is obtained by simulating a differential (respectively recurrence) equation of order q, $\partial\eta(\theta, x)/\partial[\theta]_i$ is also obtained by simulating a differential (respectively recurrence) equation of order q. Simplifications are possible for LI structures with known initial conditions, so that an equation of order $2q$ only has to be simulated whatever p (and not of order $p\,q$), see chap. 4 of [36, 38]. An adjoint state method is also presented there for models given by recurrence equations, which permits to compute the derivatives of $J_N(\theta)$ with respect to the $[\theta]_i$'s in two simulations only: one goes forward, that is, in direct time, as the recurrence that computes the model response, and calculates $J_N(\theta)$; the other goes backward and calculates the derivatives. Since a computer code can be considered as a recurrence equation (consider the line number k as being executed at time k), it comes as no surprise that the adjoint state method can be extended for computing the derivatives of *any function* $J(\theta)$. This corresponds to the dual (or adjoint) code technique used in automatic differentiation, see, e.g., [15].

A final recommendation is to *use one of the methods just mentioned to compute exact derivatives*, rather than using approximation by finite differences, which may be not only very approximate by also much slower.

4 Experimental design

Again, this is quite a broad topic which we shall only briefly touch, mainly through examples (hopefully motivating).

4.1 Parameter estimation

Experimental design for parameter estimation aims at providing a small dispersion for the parameter estimates.

Example 10 *We have to determine the weights of eight objets with a spring balance, as shown in Figure 21. The objects have weights m_i, $i = 1, \ldots, 8$, the errors ϵ_i are i.i.d. $\mathcal{N}(0, \sigma^2)$.*

We shall use two methods.

Method 1 *We weigh each objet successively. This gives*

$$y(i) = m_i + \epsilon_i\,,\ \ i = 1, \ldots, 8\,,$$

[5] When the prediction-error method is used, sensitivities of the prediction error have to be used.

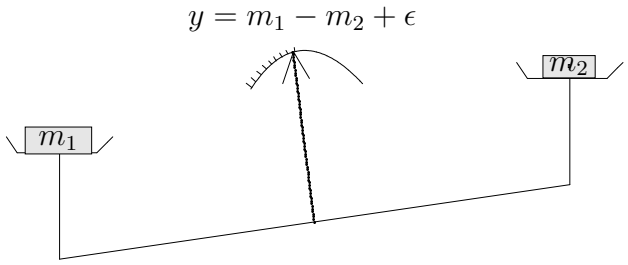

Figure 21: The spring balance used in Example 10

and the estimated weight of objet i is simply $\hat{m}_i = y(i)$, which is distributed $\sim \mathcal{N}(m_i, \sigma^2)$.

To gain in precision, we repeat the experiment 8 times and average the results: it gives new estimates $\hat{\hat{m}}_i \sim \mathcal{N}(m_i, \sigma^2/8)$. This method uses 64 observations.

Method 2 *We use 8 different configurations, with for each of them some objects on the left and some on the right:*

$$\begin{aligned}
y(1) &= m_1 + m_2 + m_3 + m_4 + m_5 + m_6 + m_7 + m_8 + \epsilon_1\,,\\
y(2) &= m_1 + m_2 + m_3 - m_4 - m_5 - m_6 - m_7 + m_8 + \epsilon_2\,,\\
y(3) &= m_1 - m_2 - m_3 + m_4 + m_5 - m_6 - m_7 + m_8 + \epsilon_3\,,\\
y(4) &= m_1 - m_2 - m_3 - m_4 - m_5 + m_6 + m_7 + m_8 + \epsilon_4\,,\\
y(5) &= -m_1 + m_2 - m_3 + m_4 - m_5 + m_6 - m_7 + m_8 + \epsilon_5\,,\\
y(6) &= -m_1 + m_2 - m_3 - m_4 + m_5 - m_6 + m_7 + m_8 + \epsilon_6\,,\\
y(7) &= -m_1 - m_2 + m_3 + m_4 - m_5 - m_6 + m_7 + m_8 + \epsilon_7\,,\\
y(8) &= -m_1 - m_2 + m_3 - m_4 + m_5 + m_6 - m_7 + m_8 + \epsilon_8\,.
\end{aligned}$$

Since there are 8 observations for 8 parameters, we obtain the estimates by solving previous equations when setting the ϵ_i's to zero,

$$\hat{m}_1 = \frac{y(1) + y(2) + y(3) + y(4) - y(5) - y(6) - y(7) - y(8)}{8}\,,\ \textit{etc.}$$

Therefore,

$$\hat{m}_1 = m_1 + \frac{\epsilon_1 + \epsilon_2 + \epsilon_3 + \epsilon_4 - \epsilon_5 - \epsilon_6 - \epsilon_7 - \epsilon_8}{8}\,,\ \textit{etc.}$$

Since the ϵ_i' are independent, we obtain $\hat{m}_i \sim \mathcal{N}(m_i, \sigma^2/8)$ but with 8 observations only!

In Example 10, the construction of a "good" design corresponds to a combinatorial problem. When the design variables x_k are real numbers, optimum design for parameter estimation is (classically) obtained by optimizing a scalar function of the (asymptotic) covariance matrix of the estimator.

For instance, for WLS estimation in a regression model with weighting by the inverse of the variance $\sigma^2(x)$ of the errors, the asymptotic covariance matrix of $\hat{\theta}^N_{WLS}$ is the inverse of

$$\mathbf{M}(\xi, \bar{\theta}) = \int_{\mathcal{X}} \sigma^{-2}(x)\, \frac{\partial \eta(\theta, x)}{\partial \theta}_{|\bar{\theta}} \frac{\partial \eta(\theta, x)}{\partial \theta^\top}_{|\bar{\theta}} \xi(dx)\,,$$

see Section 2.1.4. Other matrices are obtained for other estimators, see e.g. Sections 2.1.5, 2.4. Note the role of the sensitivity functions, already useful to compute the derivative of the estimation criterion, see Section 3.

One should thus choose ξ that makes $\mathbf{M}^{-1}(\xi, \bar{\theta})$ as small as possible. Note that $\mathbf{M}$ depends on $\bar{\theta}$ which is unknown. A classical approach called *local design* consists in using a prior guess $\hat{\theta}$, hopefully not too far from $\bar{\theta}$. We then work with $\mathbf{M}^{-1}(\xi, \hat{\theta})$. For instance, one can maximize $\log\det \mathbf{M}(\xi, \hat{\theta})$, which is called D-optimum design. Classical references are [11, 31].

The optimal ξ^* (specific algorithms are indicated in the above references) takes the form of a discrete distribution on the experimental domain $\mathcal{X}$, with k support points $x^1, \ldots, x^k$ receiving weights $w_1, \ldots, w_k$, and naturally $\sum_{i=1}^k w_i = 1$ (it can be considered as the ideal distribution of experimental effort that one should use). One can show that $k \leq 1 + p(p+1)/2$ is always enough, with p the number of parameters. Quite often for D-optimal experiments $k = p$. When one plans to collect N observations, one should then try to distribute them as closely to ξ^* as possible. That is, n_i observations should be made for $x = x^i$ where n_i/N approximates the optimal weight w_i.

The same methodology can be used to design optimal inputs for parameter estimation in dynamical systems, see [14, 39]. For LI structures, one can either choose the input sequence $u(k)$ (for a discrete-time system) or the shape of the spectrum (power spectral density) of u (for a discrete or continuous-time system). In the latter case, the optimum corresponds to a discrete spectrum (sinusoids) with a few components only.

We already mentioned the dependence of the optimum design on a prior guess $\hat{\theta}$ (this is always the case for NLP structures). Three approaches can be used to go beyond, see e.g. chap. 6 of [36, 38].

1. One can use the expected value of the design criterion, e.g., $\log\det \mathbf{M}(\xi, \hat{\theta})$, with respect to $\hat{\theta}$ distributed with some density π and maximize $\int \log\det \mathbf{M}(\xi, \hat{\theta})\pi(\hat{\theta})d\hat{\theta}$.

2. One can use a maximin approach, and consider the worst value of the design criterion with respect to θ in some set $\hat{\Theta}$. We then maximize $\min_{\hat{\theta}\in\hat{\Theta}} \log\det \mathbf{M}(\xi, \hat{\theta})$.

3. One can use a sequential design approach, where each design phase is followed by an observation and an estimation stage:

$$\cdots \text{ design } x_k \rightarrow \text{ observe } y(x_k) \rightarrow \text{ estimate } \hat{\theta}^k \rightarrow \text{ design } x_{k+1} \cdots$$

Remark 3

1. *It is important in sequential design that at each estimation stage the estimator $\hat{\theta}^k$ uses* all the data available $y(x_1), \ldots, y(x_k)$.

   The asymptotic properties of the estimator $\hat{\theta}^k$, $k \rightarrow \infty$, are not the same for a sequential design as when the x_k's are chosen independently of the observations (as random or deterministic constants). This remark applies even for LS, even in linear models, see e.g. [21, 22, 20].

2. *The design of the experiment for NLP models can also be based on non-asymptotic properties of estimators. However, this requires rather complicated developments, see, e.g., [26, 29].*

3. *In some situations the estimation of parameters is only an intermediate step towards a more specific objective, for instance the optimisation of the system. The position of the optimum then depends on the parameters to be estimated, which can be taken into account*

in the definition of a design criterion especially adapted to this final objective, see [11, 31]. When the design is sequential, this is strongly related to adaptive control *in control theory, see, e.g., [28, 30].*

4.2 Model discrimination

This is the quantitative counterpart to distinguishability considered in Section 1.7. We only present one example taken from [2]. The method used is described in the same reference.

Example 11 ([2]) *The system corresponds to a chemical reaction $A \to B$; there are two explanatory variables:* $\mathbf{x} = (\text{time } t,\ \text{temperature } T)$. *We wish to know whether the reaction is of 1st, 2nd, 3rd or 4th order.*

Therefore, there are four competing model structures:

$$\begin{aligned}
\eta^{(1)}(\theta_1, \mathbf{x}) &= \exp[-\theta_{11} t \exp(-\theta_{12}/T)]\,, \\
\eta^{(2)}(\theta_2, \mathbf{x}) &= \frac{1}{1 + \theta_{21} t \exp(-\theta_{22}/T)}\,, \\
\eta^{(3)}(\theta_3, \mathbf{x}) &= \frac{1}{[1 + 2\theta_{31} t \exp(-\theta_{32}/T)]^{1/2}}\,, \\
\eta^{(4)}(\theta_4, \mathbf{x}) &= \frac{1}{[1 + 3\theta_{41} t \exp(-\theta_{42}/T)]^{1/3}}\,,
\end{aligned}$$

each one depending on two parameters. One can check that they are distinguishable.

We perform some simulations and generate observations with the second structure, that is,

$$y(\mathbf{x}_j) = \eta^{(2)}(\bar{\theta}_2, \mathbf{x}_j) + \epsilon_j$$

with $\bar{\theta}_2 = (400, 5000)^\top$ *the true value of the parameters in model 2. The* $(\epsilon_j)_j$*'s are i.i.d.* $\mathcal{N}(0, \sigma^2)$ *with* $\sigma = 0.05$.

The admissible experimental domain $\mathcal{X}$ *is given by:* $0 \le t \le 150$, $450 \le T \le 600$, *see Figure 23.*

The design approach we consider is sequential. After the observation of $y(\mathbf{x}_j)$, $j = 1, \ldots, k$, *we*

- *estimate* $\hat{\theta}_i^k$ *by LS for each structure;*
- *estimate the probability* $\pi_i(k)$ *that the model* i *is correct,* $i = 1, \ldots, 4$ *(with the procedure described in [2]).*

The design process is initialized by assuming equal probabilities for all models, $\pi_i(0) = 1/4$, $i = 1, \ldots, 4$, *and choosing the first four design points* $\mathbf{x}_1, \ldots, \mathbf{x}_4$ *reasonably spread in* $\mathcal{X}$. *Figure 22 presents the evolution of the four probabilities* $\pi_i(k)$ *as functions of* k. *We quickly detect that models 1 and 4 are not correct. The four initial points tend to indicate that model 3 has a higher probability than model 2 of being correct. However, the careful choice of the successive points made by the procedure yields a correct decision after a reasonably small number of observations.*

The sequence of design points generated by the procedure is presented in Figure 23. The first four points have been chosen without prior knowledge. The next points tend to accumulate on a small number of different conditions, considered as the most informative for discriminating between the competing structures.

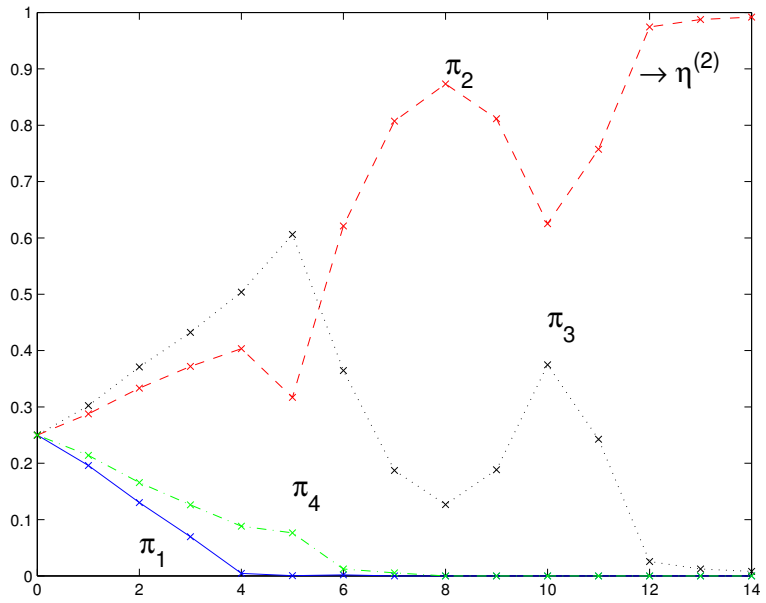

Figure 22: Evolution of the 4 probabilities $\pi_i(k)$ in Example 11

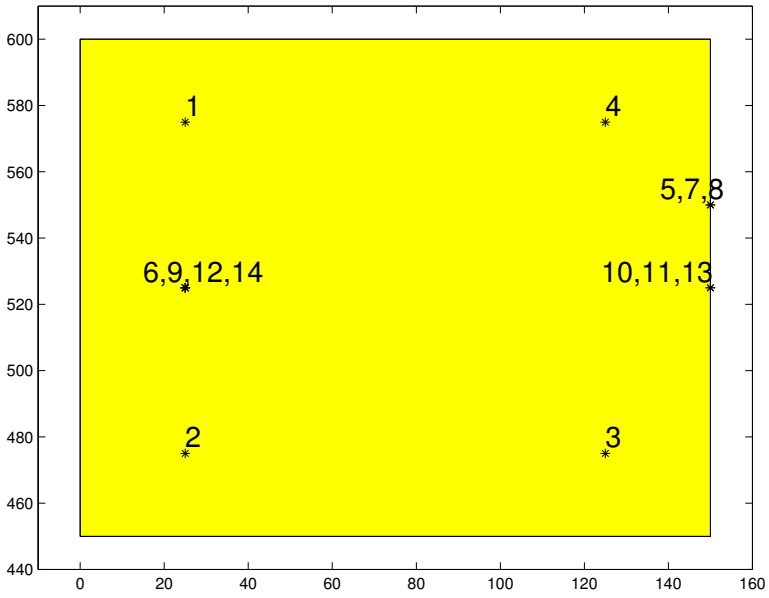

Figure 23: Sequence of design points $\mathbf{x}_k$ in Example 11

5 (In)validation and testing

A first approach for testing the model consists in testing the residuals, $y(x_k) - \eta(\hat{\theta}^N, x_k)$ in a regression model, or more generally the prediction errors $e(k, \hat{\theta}^N)$. They should correspond to a sequence of independent random variables, and independence is a statistical property that can be tested. Also, different assumptions that have been used, such as stationarity or normality of the distribution of the perturbations, can also be tested on the residuals, see, e.g., [19]. Quite often, it is already instructive to simply plot the residuals as a function of k.

A second approach is based on validation data $(x'_j, y(x'_j))$, $j = 1, \ldots, M$, not used for the construction of the model, and consists in comparing the $y(x'_j)$'s with their predictions $\eta(\hat{\theta}^N, x'_j)$ by the model constructed from the estimation data $(x_j, y(x_j))$, $j = 1, \ldots, N$.

When the results are not satisfying, the model (and maybe the assumptions) must be changed. E.g., one may have to modify the estimator because the initial assumptions on the distribution of the perturbations proved to be wrong, see Section 2.4, or one may have to abandon the stationarity assumption, and assume that the variance depends on x, see Section 2.1.5. Sometimes the collection of more data will prove necessary to choose between rival models. Experimental design should then be considered, see Section 4.2.

References

[1] H. Bierens. *Topics in Advanced Econometrics.* Cambridge University Press, Cambridge, 1994.

[2] G. Box and W. Hill. Discrimination among mechanistic models. *Technometrics*, 9(1):57–71, 1967.

[3] P. Caines. *Linear Stochastic Systems.* Wiley, New York, 1988.

[4] R. Carroll and D. Ruppert. A comparison between maximum likelihood and generalized least squares in a heteroscedastic linear model. *Journal of the American Statistical Association*, 77(380):878–882, 1982.

[5] G. Chavent. Local stability of the output least square parameter estimation technique. *Matematicada Applicada e Computacional*, 2(1):3–22, 1983.

[6] G. Chavent. Identifiability of parameters in the output least square formulation. In E. Walter, editor, *Identifiability of Parametric Models*, chapter 6, pages 67–74. Pergamon, Oxford, 1987.

[7] G. Chavent. A new sufficient condition for the wellposedness of non-linear least-square problems arising in identification and control. In A. Bensoussan and J. Lions, editors, *Analysis and Optimization of Systems*, pages 452–463. Springer, 1990.

[8] G. Chavent. New size $\times$ curvature conditions for strict quasiconvexity of sets. *SIAM Journal on Control and Optimization*, 29(6):1348–1372, 1991.

[9] E. Demidenko. Is this the least squares estimate? *Biometrika*, 87(2):437–452, 2000.

[10] D. Downing, V. Fedorov, and S. Leonov. Extracting information from the variance function: optimal design. In A. Atkinson, P. Hackl, and W. Müller, editors, *mODa6 – Advances in Model–Oriented Design and Analysis*, pages 45–52. Physica Verlag, Heidelberg, 2001.

[11] V. Fedorov. *Theory of Optimal Experiments.* Academic Press, New York, 1972.

[12] R. Fletcher and M. Powell. A rapidly convergent descent method for minimization. *Computer Journal*, (june):163–168, 1963.

[13] R. Fletcher and M. Reeves. Function minimization by conjugate gradients. *Computer Journal*, (july):149–154, 1964.

[14] G. Goodwin and R. Payne. *Dynamic System Identification: Experiment Design and Data Analysis.* Academic Press, New York, 1977.

[15] A. Griewank and G. Corliss, editors. *Automatic Differentiation of Algorithms: Theory, Implementation and Application.* SIAM, Philadelphia, 1991.

[16] E. Hansen. *Global Optimization Using Interval Analysis.* Marcel Dekker, New York, 1992.

[17] R. Jennrich. Asymptotic properties of nonlinear least squares estimation. *Annals of Math. Stat.*, 40:633–643, 1969.

[18] J. Jobson and W. Fuller. Least squares estimation when the covariance matrix and parameter vector are functionally related. *Journal of the American Statistical Association*, 75(369):176–181, 1980.

[19] G. Kanji. *100 Statistical Tests.* Sage Pub., London, 1993.

[20] T. Lai. Asymptotic properties of nonlinear least squares estimates in stochastic regression models. *Annals of Statistics*, 22(4):1917–1930, 1994.

[21] T. Lai and C. Wei. Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *Annals of Statistics*, 10(1):154–166, 1982.

[22] T. Lai and C. Wei. Asymptotically efficient self-tuning regulators. *SIAM J. Control and Optimization*, 25(2):466–481, 1987.

[23] L. Ljung. *System Identification, Theory for the User.* Prentice-Hall, Englewood Cliffs, 1987.

[24] J. Nelder and R. Mead. A simplex method for function minimization. *Computer Journal*, 7:308–313, 1965.

[25] A. Pázman. *Nonlinear Statistical Models.* Kluwer, Dordrecht, 1993.

[26] A. Pázman and L. Pronzato. Nonlinear experimental design based on the distribution of estimators. *Journal of Statistical Planning and Inference*, 33:385–402, 1992.

[27] M. Powell. *Nonlinear Optimization.* Academic Press, New York, 1981.

[28] L. Pronzato. Adaptive optimisation and D-optimum experimental design. *Annals of Statistics*, 28(6):1743–1761, 2000.

[29] L. Pronzato and A. Pázman. Second-order approximation of the entropy in nonlinear least-squares estimation. *Kybernetika*, 30(2):187–198, 1994. *Erratum* 32(1):104, 1996.

[30] L. Pronzato and E. Thierry. Sequential experimental design and response optimisation. *Statistical Methods and Applications*, 11(3):277–292, 2003.

[31] S. Silvey. *Optimal Design*. Chapman & Hall, London, 1980.

[32] A. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, Cambridge, 1998.

[33] S. Veres. *Structure Selection for Stochastic Dynamic Systems: the Information Criterion Approach*. Gordon and Breach, New York, 1991.

[34] E. Walter, Y. Lecourtier, and J. Happel. On the structural output distinguishability of parametric models, and its relation with structural identifiability. *IEEE Transactions on Automatic Control*, 29:56–57, 1984.

[35] E. Walter and L. Pronzato. Identifiabilités et non linéarités. In A. Fossard and D. Normand-Cyrot, editors, *Systèmes non linéaires*, pages 113–146. Masson, Paris, 1993.

[36] E. Walter and L. Pronzato. *Identification de Modèles Paramétriques à Partir de Données Expérimentales*. Masson, Paris, 1994. 371 pages.

[37] E. Walter and L. Pronzato. Identifiabilities and nonlinearities. In A. Fossard and D. Normand-Cyrot, editors, *Nonlinear Systems. Modeling and Estimation*, chapter 3, pages 111–143. Chapman & Hall, London, 1995.

[38] E. Walter and L. Pronzato. *Identification of Parametric Models from Experimental Data*. Springer, Heidelberg, 1997.

[39] M. Zarrop. *Optimal Experiment Design for Dynamic System Identification*. Springer, Heidelberg, 1979.