From robust estimation to hybrid system identification

Laurent Bako

ÉCOLE CENTRALE LYON

From robust estimation to hybrid system identification

Mémoire

présenté publiquement le 02 décembre 2016

pour l'obtention de

du diplôme d'Habilitation à Diriger des Recherches de l'Ecole Centrale de Lyon

par

Laurent Bako

Composition du jury:

Rapporteurs :	Jamal Daafouz	Professeur des Universités Université de Lorraine, CRAN
	Andrea Garulli	Professeur des Universités Universita' degli Studi di Siena, DIISM
	Roland Tòth	Professeur assistant Eindhoven University of Technology, DEE
Examinateurs :	Xavier Bombois	Directeur de Recherche au CNRS Ecole Centrale de Lyon, Ampère
	Pascal Dufour	Maître de Conférences HDR Université Claude Bernard Lyon 1, LAGEP
	Gérard Scorletti	Professeur des Universités Ecole Centrale de Lyon, Ampère

UNIVERSITÉ DE LYON

Version provisoire du 19 décembre 2016

Ampère

Remerciements

Pour commencer, je remercie chaleureusement Jamal Daafouz (Université de Lorraine), Andrea Garulli (University of Siena) et Roland Tòth (Eindhoven University of Technology) qui m'ont fait l'honneur d'être les rapporteurs de ce mémoire. J'exprime aussi ma sincère reconnaissance à mes collègues de Lyon, Xavier Bombois, Pascal Dufour et Gérard Scorletti pour avoir accepté d'être mes examinateurs.

Les réflexions présentées dans ce mémoire s'inscrivent sur le plan thématique, dans la continuité de mes travaux de thèse réalisés sous la direction de Stéphane Lecoeuche et Guillaume Mercère sur l'identification de systèmes hybrides. Cependant, une importante rupture méthodologique concerne l'introduction de l'optimisation parcimonieuse et sa maturation progressive au fil des années et de mes collaborations avec mes collègues de la communauté. Je remercie tous mes collaborateurs et co-auteurs avec qui j'ai eu la chance de travailler pendant ces dix années de recherche. J'ai une pensée pleine de reconnaissance pour mes anciens collègues du département Informatique et Automatique de Mines Douai. En particulier, je remercie mes anciens étudiants Khaled Boukharouba et Dulin Chen pour leurs contributions à l'avancement de la réflexion sur ce sujet.

Je remercie enfin mes collègues du département EEA de l'Ecole Centrale de Lyon et ceux du laboratoire Ampère qui m'ont aidé et encouragé dans mon projet de HDR. Je remercie plus particulièrement Bruno Allard, Laurent Krähenbühl, Christian Vollaire pour leurs conseils et leur aide administrative précieuse dans la préparation de mon inscription à l'HDR. Merci enfin à Edith Bergeroux qui a été le pilier de l'organisation du déplacement des membres de mon jury.

Abstract

This report elaborates on my research activities with a particular focus put on the period that ranges from my graduation until now (2009-2016). Only a brief overview of scientific results obtained on this period is presented. The research topic discussed is mainly concerned with hybrid system identification from input-output measurements. Hybrid systems form a class of dynamic systems where discrete and continuous dynamics interact. The global behavior results from switching among a finite number of subsystems. A fundamental challenge associated with the identification of such systems is that the available data points are not labelled beforehand in the sense that one does not know a priori which data point is generated by which subsystem. Ideally, one would like to partition the data points into a finite number of groups each of which is relevant to a single subsystem. However this is typically a nonconvex procedure which does not admit any numerically efficient solution.

We propose a robust identification approach whose principle is to fit appropriately the entire mixed dataset to a single equation. A common thread of our results is the concept of sparse optimization with its associated convex relaxations. This common idea is presented in the third chapter as a solution to the robust regression problem. It is later applied to the identification of switched linear systems and piecewise affine systems.

Contents

List of abbreviations

SISO	Single Input Single Output
MISO	Multiple Input Single Output
MIMO	Multiple Input Multiple Output
PWA	PieceWise Affine
PWA-DI	PieceWise Affine Differential (Difference) Inclusion
ARX	Auto-Regressive Exogenous
SARX	Switched Auto-Regressive Exogenous
SLS	Switched Linear Systems
RBF	Radial Basis Function
PV	parameter vector

Notations

$\mathbb{R}$	set of real numbers
$\mathbb{Z}$	set of integers
$\mathbb{R}^n$	set of n-dimensional real vectors
$\mathbb{R}^{m\times n}$	set of $m \times n$ real matrices
$\mathcal{N}(c, Q)$	Gaussian distribution with center c and covariance matrix Q
N	number of data points
X	regressor matrix
Y	output matrix
x_t, y_t	regressor and output vectors at time t (resp.)
$\xi(X)$	self-decomposability amplitude of matrix X
$\nu_n(X)$	n-genericity index of matrix X
$\|\cdot\|$	some norm
$\|\cdot\|_p$	vector or matrix p-norm
$\arg\min$	minimizing argument
$\mathrm{im}(\cdot)$	range space (of a matrix)
$\mathrm{int}(\cdot)$	interior (of a set)

Chapter 1

Introduction

Contents

1.1 General context

In many fields of system engineering such as machine learning, process control, signal processing or finance, one is frequently confronted with the necessity of finding the underlying relationship that relates a set of signals of interest. Such a relationship is commonly called a *system* at least in system theory. In informal terms, a system can be viewed as a modeling abstraction of a causality relationship between some phenomena which can be quantitatively described by signals. A model of the system is then a mathematical object (or a set of mathematical objects) which formally (and quantitatively) describes the interaction of the signals brought into play by the system. Hence a model attempts to capture as well as possible the (true, physical) hidden relationship. In some simple specific applications, there are well-established physical laws from which the control engineer can readily obtain a useful model. For example, the control of a DC-motor requires the knowledge of the mathematical model that relates the input voltage to its angular position. For such a simple system, a model is obtainable by combining the laws of electricity and Newton's laws of dynamics. Most of the time however such physical laws are either not clearly known or are too complex or involve unknown parameters. In these situations, an alternative approach for constructing a model is by processing experimental measurements of the signals of interest. The process of building a model from experimental data is referred to as *system identification* or experimental/data-driven modeling.

The primary interest in a model resides in its capacity to predict the values of some signals based on the knowledge of some others. This makes it a valuable tool for analysis, simulation, control design, filtering, fault detection, etc. Perhaps an application which illustrates somewhat strikingly the benefit of models is that of computer simulation. Indeed if one can find a reliable mathematical model of a certain physical process under study, then it is possible to substitute a numerical experiment on a computer for a possibly long and costly physical experiment. Therefore the availability of a mathematical model allows for a great deal of flexibility in the study of

true industrial processes by making it possible to experiment a rich set of operating conditions some of which would not have been possible on the true system.

The standard procedure in system identification consists in many steps: (1) run an experiment on the real system of interest and measure input-output data samples; (2) setup a model structure that is, a family of relevant (generally parametrized) candidate models; (3) estimate a model, that is, determine an instance of the model structure that explains best the data using, for example, optimization methods; (4) check validity or the representativity of the candidate model with respect to the usage for which it is intended. Depending on whether or not the validation step is conclusive it might be necessary to repeat all the above steps until a model which achieves the desired performance is obtained. The approaches to system identification are essentially classified with respect to the class of models which are treated: linear models [37, 59], nonlinear smooth models [58, 35, 50, 61], Linear Parameter Varying (LPV) models [25, 62], PieceWise Affine (PWA) models. A secondary classification is based on the nature of the estimation algorithms: batch or adaptive. The batch estimator operates on a finite collection of data samples and produces a single estimate along with an uncertainty set. In contrast, the adaptive estimator processes sequentially (and possibly in real-time) the data samples and produces a parameter estimate at any time ; it acts as a dynamic system whose input is the sequence of input-output pairs of the to-be-identified system.

1.2 Research interests and contributions

Our research focuses essentially on system identification. Given input-output measurements collected from a real system, the system identification problem aims at constructing a mathematical model which describes as accurately as possible the behavior of the system. Our research covers various classes of dynamic and static systems but a strong focus is put on the class of hybrid dynamical systems. These are systems in which continuous dynamics interact with discrete-event dynamics. Loosely speaking, hybrid systems can be thought of as systems whose global dynamics are generated by different operating regimes together with internally or externally controlled switchings, jumps or transitions among these regimes. The discrete dynamics are typically induced by the presence of logic devices, switching circuits, valves, computer programs, ... The class of hybrid systems is virtually universal in the sense that almost any system can be formally represented as a hybrid system either naturally or somewhat artificially through some modeling abstraction. Typical examples of hybrid systems are: chemical processes, electrical networks, air traffic modeling/management, biological systems, etc.

The identification of hybrid systems is a relatively recent research topic in the control community. It was pioneered in the beginning of the 2000s by the works [28, 66, 11, 36, 52]. See also the tutorial paper [45] and references therein. Other researchers including myself joined later the research effort in striving to devise efficient algorithms for estimating hybrid systems from empirical observations. The sum of these efforts has incrementally produced now a wealth of results based on various mathematical and learning concepts. A quite large number of approaches have been developed: algebraic-geometric, set-membership based, bayesian, clustering-based, convex optimization based, ... For an overview of the existing methods we refer the interested reader to the recent survey [29]. However a fair assessment of the current state-of-art reveals that there are still numerous challenges ahead. We will get back to some of these in Section 5.2.

Hybrid system identification

From a mathematical point of view, a hybrid system is informally defined by two types of interacting components:

- a finite number of subsystems described by ordinary differential (or difference) equations
- an event-generator generating the switching signal that controls the activation of the different subsystems over time.

Therefore the general problem of hybrid system identification consists in determining only from a collection of input-output measurements, a model of each individual subsystem and a model of the switching law whenever such a law exists. This is a very challenging problem because the switching signal is not observed. As a result, by just looking at the global input-output data we do not know a priori which subsystem has been activated at which time. With respect to the nature of the switching mechanism and the structure of the models used to describe the subsystems, different classes of models can be considered for identification: switched input-output models, switched state-space models, jump markov models, piecewise affine models, hinging hyperplanes, ...
Our contributions cover most of these models. We have developed a variety of algorithms for the estimation of models for hybrid systems. A general idea of our approach is to formulate the hybrid system identification problem as a *sparse optimization* problem which, in some sense, can be viewed as a robust estimation scheme. By sparse optimization, we refer here to an optimization problem which aims at optimizing the number of nonzero (or zero) entries in a vector or a matrix. Note in passing that this class of problems include matrix rank minimization as a special case since this latter is equivalent to minimizing the number of nonzero entries in the vector formed with the singular values.
We will consider in this report two classes of hybrid systems:

- switched linear systems with subsystems described by ARX models (SARX). To tackle the identification of this class of systems, we develop a sparse optimization approach. Implementing this scheme directly however comes with a huge price in complexity. We therefore resort to a more affordable convex relaxation. An analysis of the equivalence between the original sparse optimization and the relaxed version is proposed.
- piecewise affine systems with subsystems described by ARX models (PWARX). These can be viewed as particular switched systems where the regressor domain is partitioned into a finite number of polyhedral regions with each region associated to one subsystem. The switching signal is then internally controlled by the regressor being member of one region or another. The identification problem aims at identifying the parameters of the submodels and the boundary hyperplanes of the validity regions. The report will present two approaches: the first is a nonsmooth convex optimization formulation inspired by our general framework of sparse optimization; the second is a recursive scheme which performs alternately clustering and identification for the simultaneous estimation of both the parameter vectors and the associated regions.

Robust estimation

One of our approach to hybrid system identification is by employing robust estimation tools to identify the submodels one after another from the mixed data set. For example, for the case of switched systems, if one concentrates on the estimation of a single submodel, then the data pertaining to the other submodels can be regarded as outliers to be detected and corrected. Therefore to carry out properly the estimation task for hybrid systems we need to design a robust identifier which would be insensitive to multiple gross errors. This view has shifted our attention to the development and analysis of a class of estimators that may be robust against a relatively large number of outliers. We propose a class of robust estimators which contains the well-known least deviation (LAD) estimator as a special member. In particular, we consider the problem of identifying a linear model from measurements which are corrupted by two types of noise: a dense noise sequence and a sparse noise sequence. While the dense noise is generally assumed to be of moderate amplitude and zero-mean, the sparse noise shows up only intermittently in time but when it does, it can take on arbirarily large values. This is a fundamental problem in many estimation-related applications such as fault detection, state estimation in lossy networks, hybrid system identification, etc. In its most natural formulation, the problem is computationally hard to handle because it exhibits some intrinsic combinatorial features. Therefore, obtaining an effective solution necessitates relaxations that are both solvable at a reasonable cost and effective in the sense that they can return the true parameter vector under specific circumstances.

- We introduce a new, quantitative and computable measure of the richness properties of the regression data called self-decomposability amplitude. Based on this number an underestimate of the bound on the number of correctable outliers is obtained. It is shown that when the measurements are affected by only a sparse noise sequence, exact recovery of the true parameter vector is possible provided the number of outliers is no larger than a bound depending on the richness of the regression data.

- When sparse and dense noises are simultaneously active within the data, (computable) parametric error bounds are derived in function of the amplitude of the dense noise and the number of outliers. These bounds give rise to an uncertainty set containing the true parameter matrix.

Some other research topics

We have also conducted research on various other topics which are closely related to that of estimating hybrid systems. These works will not be described in details here but we do provide a summary.

- **Subspace clustering:** This is the problem of identifying a finite number of subspaces from mixed observations that lie in the union of those subspaces. Indeed the intrinsic challenge of this problem is basically of the same nature as that of switched system identification. The data points being in the union of the subspaces, it is not known which data point originates from which subspace. Consequently, our robust identifier can be slightly adapted for dealing with this scenario [2]. More precisely, we show that the clustering problem is amenable to a sparse optimization problem. Considering a candidate subspace and the distances of the data points to that subspace, the foundation of the proposed approach lies in the maximization of the number of zero distances. This can be relaxed into a

convex optimization (a second order cone programming). Efficiency of the relaxation can be significantly increased by solving a sequence of reweighted convex optimization problems. The problem of subspace clustering has numerous applications in Machine Learning and Computer Vision (e.g., face clustering under varying illumination, temporal video segmentation).

- **State observer design for switched systems:** While our research focuses mainly on parameter estimation, we have also explored the problem of state estimation for switched linear systems. For a dynamic system, the state usually refers to a vector of signals that encodes at each time instant, from a modeling perspective, the full information about the past of that system. There are many practical engineering situations in which an accurate estimate of the state is desirable. For example, this can help get around the necessity of instrumenting the system with possibly expensive state sensors. Another application of state estimation is in fault detection. In effect, comparing a model-based estimate of some function of the state (e.g., the output) to its measured version can bring out model inconsistencies thereby enabling the detection of changes in the system whose nominal behavior is described by that model. Also, in state feedback control systems a complete knowledge of the state is required. Our contribution on this topic consists in the development of a (continuous) state observer for discrete-time linear switched systems under the assumptions that neither the continuous state nor the switching signal are known. A specificity of the proposed observer is that, in contrast to the state of the art, it does not require an explicit prior estimation of the discrete state. The key idea of the method consists in minimizing a nonsmooth weighted cost function which is formed from the matrices of all the subsystems regardless of when each of them is active [9].

- **Time-optimal control of linear systems:** We show that this problem can be solved via nonsmooth optimization. The minimum-time (or time-optimal) control problem consists in finding a control policy that will drive a given dynamic system from a given initial state to a given target state (or a set of states) as quickly as possible. This is a well-known challenging problem in optimal control theory for which closed-form solutions exist only for a few systems of small dimensions. We have proposed a very generic solution to the minimum-time problem for arbitrary discrete-time linear systems. This is a numerical solution based on sparse optimization, that is, the minimization of the number of nonzero elements in the state sequence over a fixed control horizon. We consider both single input and multiple inputs systems. An important observation is that, contrary to the continuous-time case, the minimum-time control for discrete-time systems is not necessarily entirely bang-bang, see [7, 22].

- **Realization and identifiability of switched systems:** In collaboration with M. Petreczky (CR, CRIStAL) I have been involved in the development of a piece of research concerning the realization of discrete-time switched systems. System identification techniques typically return input-output models. This is more particularly so when switched systems are concerned. However control methods are based most of the time on state-space models. A question of great interest then is whether state-space realizations exist naturally for input-output descriptions of switched systems and if not, characterize the conditions of existence and develop algorithms for computing such realizations. Our work focuses on this analysis. A fundamental result is as follows: an input-output map admits a finite dimensional realization if and only if it has a *generalized convolution representation* and the

rank of the Hankel matrix formed with its *Markov parameters* (which are interpretable as generalized impulse responses) is finite, see [49] for further details. If a state-space realization exists for a given input-output map, then there are infinitely many such realizations. Is of particular interest the subset of minimal realizations, that is, those realizations which have the smallest state-dimension. It can indeed be shown that such minimal realizations are isomorphic. We provide rank conditions for characterizing system-theoretic concepts such as observability, reachability and minimality. We also provide concrete realization algorithms for computing a minimal realization from the Hankel matrix formed with the Markov parameters and for model reduction.

The realization theory of switched linear systems lays down the background for studying identifiability of those systems [48, 47]. Identifiability of parametrized model structures is a central question in the theory of system identification. This is the qualitative, formal and yet fundamental question of whether attempting to infer a given parametrized model from noise-free input-output data is a well-posed problem. More precisely, this is related to the injectivity of the parameterization map which maps a parameter space to a set of dynamic models (here, the switched models). The answer to this question has a number of implications for the design of informative experiments, the development of parameter estimation algorithms, the analysis of identification methods and the significance of the estimated models. In fact, determining whether the model structure is identifiable is an essential step in the theoretical analysis of identification algorithms.

1.3 Outline of this report

The next chapter (Chapter 2) presents a somewhat general robust estimation framework for models with linear-dependency on parameters. In Chapter 3, we formulate the problem of switched linear system identification. Recalling that the inherent challenge posed by this problem is the fact that the data are not partitioned per subsystem, we propose a sparse optimization approach. As such however we are still facing a NP-hard complexity. Hence for implementation purpose, a convex relaxation scheme is adopted which turns out to lie in the framework developed in the first chapter.
Chapter 4 deals with the identification of the class of piecewise affine systems. These can be viewed as models of nonlinear systems and as such, they are of major interest from the modeling perspective. Finally Chapter 5 draws a picture of the research line we are planning to pursue in the future.

Chapter 2

Robust regression

Contents

This chapter is partially based on our paper [10] and on an unpublished technical report. It stands more as a generalization of the reflections reported in [10]. We present a class of optimization-based robust estimators which aim at recovering a parameter matrix from data which are subject to outliers and dense noise. We also derive some fundamental properties of those estimators in terms of the number of admissible outliers and error bounds. The obtained results are expected to be useful for analyzing later on the hybrid system identifiers.

2.1 The robust regression problem

Consider a system described by an equation of the form

$$y_t = A^o x_t + f_t + e_t \tag{2.1}$$

where $y_t \in \mathbb{R}^m$ and $x_t \in \mathbb{R}^n$ are respectively the output and the regressor vector at time t; $A^o \in \mathbb{R}^{m\times n}$ is an unknown parameter matrix; f_t and e_t are some noise terms which are unobserved.

Given a finite collection $\{x_t, y_t\}_{t=1}^N$ of measurements obeying the relation (2.1), the robust regression problem of interest here is the one of finding an estimate of the parameter matrix A^o under the assumptions that:

- $\{e_t\}$ is a dense noise sequence with bounded elements accounting for moderate model mismatches or measurement noise.
- $\{f_t\}$ is such that the majority of its elements are equal to zero while the remaining nonzero elements can be of arbitrarily large magnitude. The nonzero elements of that sequence are usually termed gross errors or outliers. They can account for possible intermittent sensor faults. We will refer to $\{f_t\}$ as the sequence of sparse noise.

For the time being, these are just informal descriptions of the characteristics of the sequences $\{f_t\}$ and $\{e_t\}$. They will be made more precise whenever necessary in the sequel for the need of stating more formal results. The question of whether the considered signals can be stochastic or not does not matter much provided, in the case stochasticity is assumed, each realization satisfies the required assumptions almost surely (i.e., with probability one). Alternatively, the analysis can be extended to the expected values of the different signals.
Let $Y \in \mathbb{R}^{m\times N}$ and $X \in \mathbb{R}^{n\times N}$ be data matrices formed respectively with N output measurements and regressor vectors. Then it follows from (2.1) that

$$Y = A^o X + E + F, \tag{2.2}$$

where $E \in \mathbb{R}^{m\times N}$ and $F \in \mathbb{R}^{m\times N}$ are unknown noise components. The matrices Y and X can be structured or not, depending on whether the system (2.1) is dynamic or not. For example, when the model (2.2) is of MIMO FIR type, Y contains a finite collection of output measurements while X is a Hankel matrix containing lagged inputs of the system. In this case Y and X take the form

$$Y = \begin{bmatrix} y_1 & y_2 & \cdots & y_N \end{bmatrix},$$

$$X = \begin{bmatrix} u_1 & u_2 & \cdots & u_N \\ u_0 & u_1 & \cdots & u_{N-1} \\ \vdots & \vdots & \cdots & \vdots \\ u_{1-n_f} & u_{2-n_f} & \cdots & u_{N-n_f} \end{bmatrix}.$$

where $\{u_t\}$ and $\{y_t\}$ stand respectively for the input and output of the system and the maximum lag n_f is called the order of the model. In the sequel, the notations of the type y_t and x_t with subindex $t \in \mathbb{I} \triangleq \{1, \ldots, N\}$ refer to the columns of Y and X respectively.

Relevant prior works. The so formulated regression problem is called a robust regression problem in connection with the fact that the error matrix F assumes columns of (possibly) arbitrarily large amplitude. It has applications in e.g., the identification of switched linear systems [1, 44], subspace clustering [2], etc. Existing approaches for solving the robust regression problem can be roughly divided into two groups: methods from the field of robust statistics [54, 39, 34] which have been developed since the early 60s and a class of more recent methods inspired by the compressed sensing paradigm [10, 19, 57, 67, 40]. The first group comprises methods such as the least absolute deviation (LAD) estimator [33], the least median of squares [53], the least trimmed squares [54], the family of M-estimators [34]. The latter group can be viewed essentially as a refreshed look at the so-called least absolute deviation method. There has been however a fundamental shift of philosophy in the analysis. While in the framework of

robust statistics, robustness of an estimator is measured in terms of the breakdown point (the asymptotic minimum proportion of points which cause an estimator to be unbounded if they were to be arbitrarily corrupted by gross errors), in the compressed-sensing-inspired category of robust methods, the analysis aims generally at characterizing properties of the data that favor exact recovery of the true parameter matrix A^o. In this latter group, the LAD estimator is regarded as a convex relaxation of a combinatorial sparse optimization problem.

Contributions. In the work presented hereafter we propose and analyze a class of optimization-based robust estimators. It is shown that the robust properties of the estimator are essentially inherited from a key property of the to-be-optimized performance function (or loss function) called column-wise summability. The proposed framework admits the LAD estimator and its usual variants as special cases. Moreover it applies to both SISO and MIMO systems. When the dense noise component E in (2.2) is identically equal to zero, we derive bounds on the number of gross errors (nonzero columns of F) that the estimator is able to accommodate while still returning the true parameter matrix A^o. The proposed bounds have the important advantage that they are numerically computable through convex optimization. When both E and F are active, exact recovery of the true parameter matrix is no longer possible. In this scenario, we derive upper bounds on the parametric estimation error in function of the amplitude of E and the number of nonzero columns of F. Again, computable but (possibly) looser versions of those bounds are obtainable.

To the best of our knowledge, only the papers [57] provides an explicit bound on the estimation error induced by the LAD estimator. However that bound does not apply to the current setting since the estimators although similar are of different natures. Indeed, the LAD estimator stands only as a special case of the current framework. Moreover the bound in [57] is not easily computable while ours is. The references [19] and [40] provide some bounds for a noise-aware version of the LAD estimator which are based respectively on the Restricted Isometry Property (RIP) and a measure based on subspace angles. Unfortunately numerical evaluation of those bounds is a process of exponential complexity, a price that is unaffordable in practice.

Notations. This is a glossary of notations applicable to this chapter and all the following ones.

$\mathbb{I} = \{1, \ldots, N\}$ is the index set of the measurements. $X = [x_1 \ \cdots \ x_N] \in \mathbb{R}^{n\times N}$ denotes the matrix formed with the available regressors $\{x_t\}_{t=1}^N$ and $Y = [y_1 \ \cdots \ y_N] \in \mathbb{R}^{m\times N}$ denotes the output matrix. If $A \in \mathbb{R}^{m\times n}$, then $\mathbb{I}^0(A) = \{t \in \mathbb{I} : y_t - Ax_t = 0\}$ and $\mathbb{I}^c(A) = \{t \in \mathbb{I} : y_t - Ax_t \neq 0\}$. In the special case where $m = 1$, the matrix A will be preferentially replaced by a row vector $\theta^\top$ with $\theta \in \mathbb{R}^n$. In this case a partition of the set of indices $\mathbb{I}$ is defined by $\mathbb{I}^-(\theta) = \{t \in \mathbb{I} : y_t - \theta^\top x_t < 0\}$, $\mathbb{I}^+(\theta) = \{t \in \mathbb{I} : y_t - \theta^\top x_t > 0\}$, $\mathbb{I}^0(\theta) = \{t \in \mathbb{I} : y_t - \theta^\top x_t = 0\}$. Hence $\mathbb{I}^c(\theta) = \mathbb{I}^-(\theta) \cup \mathbb{I}^+(\theta)$.

Cardinality of a finite set. Throughout this report, whenever $\mathcal{S}$ is a finite set, the notation $|\mathcal{S}|$ will refer to the cardinality of $\mathcal{S}$. However, for a real number x, $|x|$ will denote the absolute value of x.

Submatrices and subvectors. If $I \subset \mathbb{I}$, the notation X_I denotes a matrix in $\mathbb{R}^{n\times|I|}$ formed with the columns of X indexed by I. Likewise, with $\boldsymbol{y} = [y_1 \ \cdots \ y_N]^\top \in \mathbb{R}^N$, $\boldsymbol{y}_I$ is the vector in $\mathbb{R}^{|I|}$ formed with the entries of $\boldsymbol{y}$ indexed by I. We will use the convention that $X_I = 0 \in \mathbb{R}^n$ (resp. $\boldsymbol{y}_I = 0 \in \mathbb{R}$) when the index set I is empty.

Vector norms. $\|\cdot\|_p$, $p = 1, 2, \ldots, \infty$, denote the usual p-norms for vectors defined for any vector $z = [z_1 \ \cdots \ z_N]^\top \in \mathbb{R}^N$, by $\|z\|_p = (|z_1|^p + \cdots + |z_N|^p)^{1/p}$. Note that in the limiting case

where $p = \infty$, $\|z\|_\infty = \max_{i=1,\dots,N} |z_i|$. The ℓ_0 "norm" of z is defined to be the number of nonzero entries in z, i.e., $\|z\|_0 = |\{i : z_i \neq 0\}|$.
Matrix norms. The notation $\|\cdot\|_p$, with $p = 1, 2, \dots, \infty$, refers to the standard matrix p-norm. $\|\cdot\|_{2,\text{col}}$ and $\|\cdot\|_{2,\infty}$ are specific matrix-norms defined as follows. For a matrix $A = [a_1 \quad \cdots \quad a_N] \in \mathbb{R}^{n\times N}$ with $a_i \in \mathbb{R}^n$,

$$\|A\|_p = \sup_{x\in\mathbb{R}^N, \|x\|_p=1} \|Ax\|_p \quad \text{and} \quad \|A\|_{2,\infty} = \max_{i=1,\dots,N} \|a_i\|_2 .$$

2.2 A class of robust estimators

Let $\mathcal{D}_N$ be the set of N data points generated by system (2.1) for any possible values of the noise sequences, i.e.,

$$\mathcal{D}_N = \left\{(Y, X) \in \mathbb{R}^{m\times N} \times \mathbb{R}^{n\times N} : \exists (E, F) \in \mathcal{G}_N^e \times \mathcal{G}_N^f, \text{ (2.2) holds}\right\},$$

with $\mathcal{G}_N^e \subset \mathbb{R}^{m\times N}$ and $\mathcal{G}_N^f \subset \mathbb{R}^{m\times N}$ denoting the set of dense and sparse noise matrices respectively. The estimation problem aims at determining the unknown parameter matrix A^o given a point (Y, X) in $\mathcal{D}_N$. Of course, this quest would not make sense if the noises E and F were completely arbitrary since in this case, we would have $\mathcal{D}_N = \mathbb{R}^{m\times N} \times \mathbb{R}^{n\times N}$ hence loosing any informativity concerning the data-generating system. Therefore some minimum constraints need to be put on E and F as informally described above.
With respect to the estimation problem just stated, an estimator is a set-valued map $\Psi : \mathcal{D}_N \to \mathscr{P}(\mathbb{R}^{m\times n})$, $(Y, X) \mapsto \Psi(Y, X)$ which is defined from the data space $\mathcal{D}_N$ to the power set $\mathscr{P}(\mathbb{R}^{m\times n})$ of the parameter space. For (Y, X) generated by a system of the form (2.1), one would like to design an estimator achieving, whenever possible, the ideal property that $\Psi(Y, X) = \{A^o\}$. In default of that ideal situation, a more pragmatic goal is to search for a Ψ so that $A^o \in \Psi(Y, X)$ and $\Psi(Y, X)$ is of small size in some sense despite the troublesome effects of the unknown noise components E and F. The design of an optimal estimator requires specifying a performance index (usually called a loss function) which is to be minimized.

In this paper, we study the properties of the estimator of the parameter matrix A^o in (2.2) defined by

$$\Psi(Y, X) = \arg\min_{A\in\mathbb{R}^{m\times n}} \varphi(Y - AX) \tag{2.3}$$

where $\varphi : \mathcal{M}(\mathbb{R}) \to \mathbb{R}_{\geq 0}$ is a *convex function* defined on the set $\mathcal{M}(\mathbb{R})$ of (all) real matrices. It is assumed that φ has the following properties:

P1. For all $A, B \in \mathcal{M}(\mathbb{R})$ of compatible dimensions,

$$\varphi([A \quad B]) = \varphi(A) + \varphi(B) \tag{2.4}$$

with $[A \quad B]$ denoting the matrix formed by concatenating column-wise A and B.

P2. There exists a matrix norm $\ell : \mathcal{M}(\mathbb{R}) \to \mathbb{R}_{\geq 0}$ such that for all $A, B \in \mathcal{M}(\mathbb{R})$, conformable for addition,

$$\varphi(A) \leq \varphi(A - B) + \ell(B) \tag{2.5}$$

P3. There exists a constant real number $\varepsilon \geq 0$ such that for all $A \in \mathcal{M}(\mathbb{R})$ with n rows and

N columns,

$$\ell(A) - |I_\varepsilon^c(A)|\,\varepsilon \leq \varphi(A) \leq \ell(A) \tag{2.6}$$

where

$$I_\varepsilon^c(A) = \{i \in \{1, \ldots, N\} : \ell(a_i) > \varepsilon\}$$

and $|I_\varepsilon^c(A)|$ is the cardinality of $I_\varepsilon^c(A)$ and $a_i \in \mathbb{R}^n$ is the ith column of the (n, N)-matrix A.

The property (2.4) will be called column-wise summability. Since φ is a function defined over the space of real matrices of any dimensions, it is also defined for n-dimensional vectors of real numbers. Hence according to property (2.4), if $A = \begin{bmatrix} a_1 & \cdots & a_N \end{bmatrix}$ with column vectors $a_i \in \mathbb{R}^n$, then

$$\varphi(A) = \sum_{i=1}^{N} \varphi(a_i).$$

The so-defined function φ is not necessarily a norm. For any $\varepsilon^o \geq 0$ and any vector norm ℓ^o, it can be verified that the function φ defined by

$$\varphi(A) = \sum_{i=1}^{N} \max(0, \ell^o(a_i) - \varepsilon^o) \tag{2.7}$$

is positive and convex and satisfies properties (2.4)-(2.6) but it is not a norm for $\varepsilon^o > 0$ since in this case, $\varphi(A) = 0$ does not imply that $A = 0$. But if $\varepsilon^o = 0$ in (2.7), then $\varphi = \ell$ by (2.6) so that φ corresponds to the matrix norm defined by $\varphi(A) = \sum_{i=1}^{N} \ell^o(a_i)$. We note in this latter case that (2.6) is trivial while (2.5) reduces to the triangle inequality.

We will show in the sequel that the estimator Ψ in (2.3) enjoys some impressive robustness properties with respect to the sparse noise matrix F. The term sparse is used here to mean that a relatively large proportion of the column vectors of F are equal to zero. And saying that Ψ is robust with respect to F means that $\Psi(Y, X)$ does not depend on (or is insensitive to) the magnitudes of the nonzero columns of F under the sparsity condition. Therefore those few columns which are nonzero can have arbitrarily large magnitude. As will be shown in the sequel, the robustness properties of Ψ are inherited from the properties P1-P3 of the objective function φ. In the special case where φ is a norm, the properties P2-P3 are automatically satisfied so that P1 becomes the only key property required. As to the convexity of φ, it is intended just for computational reasons as it eases the solving of the optimization problem in (2.3).

2.3 Properties of the estimators

2.3.1 Exact recoverability

We first study the conditions under which the true parameter matrix A^o in (2.2) can be exactly recovered. Theorem 2.1 and Theorem 2.2 stated next show that if the number of nonzero columns in the matrix $V \triangleq E + F$ is less than a certain threshold, then $\Psi(Y, X) = \{A^o\}$.

Theorem 2.1 (A necessary and sufficient condition)**.** *Let φ be a function satisfying (2.4)-(2.6) with $\varepsilon = 0$ and Ψ be defined as in (2.3). Let d be an integer and assume that* $\operatorname{rank}(X) = n$. *For any $A \in \mathbb{R}^{m\times n}$ and $Y \in \mathbb{R}^{m\times N}$, let $\mathbb{I}^c(A, Y) = \{t \in \mathbb{I} : y_t - Ax_t \neq 0\}$. Then the following statements are equivalent.*

(i)

$$\forall A \in \mathbb{R}^{m\times n}, \forall Y \in \mathbb{R}^{m\times N},\ |\mathbb{I}^c(A,Y)| \le d \quad \Rightarrow \quad \Psi(Y,X) = \{A\} \tag{2.8}$$

(ii)

$$\max_{\substack{I^c\subset\mathbb{I}:\\ |I^c|=d}} \max_{\substack{\Lambda\in\mathbb{R}^{m\times n}\\ \Lambda\neq 0}} \left[\frac{\varphi(\Lambda X_{I^c})}{\varphi(\Lambda X)}\right] < \frac{1}{2} \tag{2.9}$$

Here and in the following, the notation $\mathbb{I} \triangleq \{1,\dots,N\}$ *is used to denote the index set for the columns of the data matrices.*

Proof. We first note that the rank assumption on X is intended to insure that (2.9) is well-defined since then, with φ being a norm, $\varphi(\Lambda X) \neq 0$ whenever $\Lambda \neq 0$.
(i) $\Rightarrow$ (ii): Assume that (i) holds.
Consider an arbitrary subset I^c of $\mathbb{I}$ such that $|I^c| = d$. Let Λ be any matrix in $\mathbb{R}^{m\times n}$ satisfying $\Lambda \neq 0$. Finally, consider a matrix $Y \in \mathbb{R}^{m\times N}$ defined by $Y_{I^c} = 0$ and $Y_{I^0} = \Lambda X_{I^0}$ where $I^0 = \mathbb{I} \setminus I^c$. Then $\mathbb{I}^c(\Lambda, Y) \subset I^c$ and so $|\mathbb{I}^c(\Lambda, Y)| \le d$. Hence by (i) $\{\Lambda\} = \arg\min_H \varphi(Y - HX)$ which means that $\varphi(Y - \Lambda X) < \varphi(Y - HX)$ for any $H \in \mathbb{R}^{m\times n}$, $H \neq \Lambda$. In particular, by taking $H = 0$ we get $\varphi(Y - \Lambda X) < \varphi(Y)$. It follows from the property (2.4) that

$$\varphi(Y_{I^c} - \Lambda X_{I^c}) + \varphi(Y_{I^0} - \Lambda X_{I^0}) < \varphi(Y_{I^c}) + \varphi(Y_{I^0}).$$

Using now the relations $Y_{I^c} = 0$ and $Y_{I^0} = \Lambda X_{I^0}$ yields $\varphi(\Lambda X_{I^c}) < \varphi(\Lambda X_{I^0})$ or, equivalently, $\varphi(\Lambda X_{I^c}) < 1/2\varphi(\Lambda X)$. Eq. (2.9) then follows from the fact that I^c and Λ are arbitrary.
(ii) $\Rightarrow$ (i): To begin with, note that if Eq. (2.9) holds for some d, then it holds also for any $d_0 \le d$. As a result, the equality $|I^c| = d$ in (2.9) can be changed to $|I^c| \le d$. Assuming (ii), let $A \in \mathbb{R}^{m\times n}$ and $Y \in \mathbb{R}^{m\times N}$ be matrices satisfying $|\mathbb{I}^c(A,Y)| \le d$. Set $I^c = \mathbb{I}^c(A,Y)$ and $I^0 = \mathbb{I} \setminus I^c$. Then for all $\Lambda \in \mathbb{R}^{m\times n}$ such that $\Lambda \neq 0$,

$$2\varphi(\Lambda X_{I^c}) < \varphi(\Lambda X) = \varphi(\Lambda X_{I^c}) + \varphi(\Lambda X_{I^0})$$

where the equality is obtained by the property (2.4) of φ. It follows that

$$\varphi(\Lambda X_{I^c}) < \varphi(Y_{I^0} - (A + \Lambda)X_{I^0}) \tag{2.10}$$

On the other hand, we know by (2.5) that

$$\varphi(Y_{I^c} - AX_{I^c}) - \varphi(Y_{I^c} - (A + \Lambda)X_{I^c}) \le \varphi(\Lambda X_{I^c})$$

Combining with the inequality (2.10) yields,

$$\varphi(Y - AX) < \varphi(Y - (A + \Lambda)X)$$

Since Λ is an arbitrary nonzero matrix, this inequality says that A is the unique minimizer of $V(H) = \varphi(Y - HX)$. □

Consider a data pair (Y, X) generated by (2.2). By letting

$$\pi^c_\varphi(X) = \max\{d : \text{Eq. (2.9) holds}\}, \tag{2.11}$$

and assuming that $\pi_\varphi^c(X) > 0$ we can see that whenever $|\mathbb{I}^c(A^o, Y)| \leq \pi_\varphi^c(X)$, A^o can be exactly recovered by computing $\Psi(Y, X)$. Of course this is likely to hold only if the dense noise component E does not exist. So in the situation where $E = 0$, the theorem says that A^o can be uniquely obtained by convex optimization provided that the number of outliers (nonzero columns of F) is less than or equal to $\pi_\varphi^c(X)$. For the condition of exact recoverability to be checkable we must be able to compute $\pi_\varphi^c(X)$. The bad news are that evaluating numerically such a number is likely to be NP-hard in most cases.
In the sequel, we investigate sufficient conditions of exact recovery which are more tractable from a numerical standpoint. For this purpose let us introduce some definitions.

Definition 2.1. *A matrix $X = [x_1 \quad \cdots \quad x_N] \in \mathbb{R}^{n\times N}$ is said to be self-decomposable if*

- *X has full row rank,* $\operatorname{rank}(X) = n$
- *For all $k \in \mathbb{I}$, $x_k \in \operatorname{im}(X_{\neq k})$ where $X_{\neq k} \triangleq X_{\mathbb{I}\setminus\{k\}}$ is the matrix obtained from X by removing its k-th column and* $\operatorname{im}(\cdot)$ *refers to range space.*

For a matrix to be self-decomposable it is enough that $X_{\neq k}$ be full row rank for any $k \in \mathbb{I}$. Achieving this condition in practice seems easy provided that the number N of measurements is large enough compared to the dimension n of X.

Definition 2.2 (self-decomposability amplitude)**.** *Let $X \in \mathbb{R}^{n\times N}$ be a self-decomposable matrix. We call* self-decomposability amplitude *of X, the number $\xi(X)$ defined by*

$$\xi(X) = \max_{k\in\mathbb{I}} \min_{\gamma_k\in\mathbb{R}^{N-1}} \Big\{ \|\gamma_k\|_\infty : x_k = X_{\neq k}\gamma_k \Big\}. \tag{2.12}$$

The so-defined $\xi(X)$ constitutes a quantitative measure of richness (or genericity) of the regressor matrix X. By richness it is meant here how much in a global sense the columns of X are linearly independent. $\xi(X)$ is expected to be small if the columns of X are somehow strongly linearly independent.

Remark 2.1. *If for some k the norm of x_k was to be considerably large in comparison to the norm of the other columns of X, then $\xi(X)$ would get large hence reducing recoverability capacity of the considered class of estimators (see also Eq. (2.9)). Such situations can be alleviated by normalizing each column of X, i.e., for example by replacing (y_k, x_k) by $(\tilde{y}_k, \tilde{x}_k) \triangleq (y_k/\|x_k\|, x_k/\|x_k\|)$ under the assumption that $x_k \neq 0$ for all $k \in \mathbb{I}$.*

With the help of the device of self-decomposability amplitude (2.12), we can state a condition for exact recovery of the parameter matrix A^o by solving the optimization problem in (2.3). A similar result was proven in [10] for the Least Absolute Deviation (LAD) estimator.

Theorem 2.2 (A sufficient condition for exact recovery)**.** *Let φ be a function satisfying (2.4)-(2.6) with $\varepsilon = 0$ and Ψ be defined as in (2.3). Assume that X is self-decomposable. Then the following statement is true:*

$$\begin{aligned} \forall A \in \mathbb{R}^{m\times n},\ \forall Y \in \mathbb{R}^{m\times N},\ |\mathbb{I}^c(A, Y)| < T\big(\xi(X)\big) \\ \Rightarrow \quad \Psi(Y, X) = \{A\}. \end{aligned} \tag{2.13}$$

where $T : \mathbb{R}_{>0} \to \mathbb{R}_{>0}$ is the function defined by $T(\alpha) = \dfrac{1}{2}(1 + \dfrac{1}{\alpha})$.

Proof. The proof is completely parallel to that of Theorem 11 in [10]. From the assumptions, each x_k, $k \in \mathbb{I}$, can be written as a linear combination of the columns of $X_{\neq k}$. Let $\gamma_k \in \mathbb{R}^{N-1}$ be any vector satisfying $x_k = X_{\neq k}\gamma_k$. It follows that for any $\Lambda \in \mathbb{R}^{m\times n}$,

$$\varphi(\Lambda x_k) = \varphi\Big(\sum_{t\in\mathbb{I}\setminus\{k\}} \gamma_{k,t}\Lambda x_t\Big)$$

with $\gamma_{k,t}$ denoting the entry of $\gamma_k \in \mathbb{R}^{N-1}$ indexed by t. Under the assumptions of the theorem, φ is a norm. So, it is positive and satisfies the triangle inequality property. As a result we can write

$$\varphi(\Lambda x_k) \le \sum_{t\neq k} |\gamma_{k,t}|\, \varphi(\Lambda x_t) \le \|\gamma_k\|_\infty \left(\varphi(\Lambda X) - \varphi(\Lambda x_k)\right)$$

Since this holds for any γ_k such that $x_k = X_{\neq k}\gamma_k$, it holds also for

$$\gamma_k^\star = \arg\min_{\gamma\in\mathbb{R}^{N-1}} \Big\{ \|\gamma\|_\infty : x_k = X_{\neq k}\gamma \Big\}.$$

Hence,

$$\varphi(\Lambda x_k) \le \xi(X)\left(\varphi(\Lambda X) - \varphi(\Lambda x_k)\right) \quad \forall k \in \mathbb{I}, \forall \Lambda \in \mathbb{R}^{m\times n}. \tag{2.14}$$

or equivalently

$$\varphi(\Lambda x_k) \le \frac{\xi(X)}{1+\xi(X)}\varphi(\Lambda X) \quad \forall k \in \mathbb{I},\ \forall \Lambda \in \mathbb{R}^{m\times n}.$$

Let I^c be any subset of $\mathbb{I}$ and pose $|I^c| = d$. Summing the previous inequality over the set I^c yields

$$\max_{\Lambda\neq 0} \frac{\varphi(\Lambda X_{I^c})}{\varphi(\Lambda X)} \le \frac{1}{2T\big(\xi(X)\big)}\,|I^c| \tag{2.15}$$

Therefore (2.9) holds if $|I^c| < T\big(\xi(X)\big)$ and the conclusion follows from Theorem 2.1. □

Remark 2.2. *The statement of Theorem 2.2 still holds true if we replace $\xi(X)$ with $\delta_\varphi(X)$ defined by*

$$\delta_\varphi(X) = \max_{k\in\mathbb{I}} \sup_{\Lambda\neq 0} \frac{\varphi(\Lambda x_k)}{\varphi(\Lambda X_{\neq k})} \tag{2.16}$$

when it is assumed that φ is a norm and $\operatorname{rank}(X_{\neq k}) = n$ for all k. Doing so will give a less conservative condition for exact recovery. However $\delta_\varphi(X)$ seems much harder to evaluate numerically than $\xi(X)$.

Remark 2.3 (A few useful properties of $\xi(X)$)**.**

- *For any nonsingular matrix $T \in \mathbb{R}^{n\times n}$, $\xi(TX) = \xi(X)$. It follows that the number $\xi(X)$ depends only on the subspace spanned by the rows of the regressor matrix X.*
- *For any self-decomposable $X \in \mathbb{R}^{n\times N}$, $\xi(X)$ is lower-bounded in the following sense*

$$\xi(X) \ge \frac{1}{N-1},$$

This follows from the more general observation that

$$\xi(X) \ge \max_{k\in\mathbb{I}} \frac{\|x_k\|}{\sum_{t\neq k}\|x_t\|}$$

for any vector norm $\|\cdot\|$. *As a result,* $T(\xi(X)$ *is upper-bounded as follows*

$$T(\xi(X)) \leq \frac{N}{2}.$$

Theorem 2.2 provides a sufficient condition for exact recovery in the situation where the function φ is a norm. Next, another condition is stated which holds in the general case.

Proposition 2.1. *Consider a triplet* $(\varphi, \ell, \varepsilon)$ *satisfying* (2.4)-(2.6). *For* $A \in \mathbb{R}^{m\times n}$ *and* $Y \in \mathbb{R}^{m\times N}$, *pose* $I^c = \mathbb{I}^c(A, Y)$, $I^0 = \mathbb{I}\backslash I^c = \{t \in \mathbb{I} : y_t - Ax_t = 0\}$ *and* $I^c_\varepsilon(\Lambda X_{I^0}) = \{t \in I^0 : \ell(\Lambda x_t) > \varepsilon\}$. *Then* $\Psi(Y, X) = \{A\}$ *if*

$$|I^c_\varepsilon(\Lambda X_{I^0})|\,\varepsilon < \ell(\Lambda X_{I^0}) - \ell(\Lambda X_{I^c}), \quad \forall \Lambda \in \mathbb{R}^{m\times n}, \Lambda \neq 0 \tag{2.17}$$

Proof. $\Psi(Y, X) = \{A\}$ is equivalent to

$$\varphi(Y - AX) < \varphi(Y - (A + \Lambda)X)$$

for any $\Lambda \in \mathbb{R}^{m\times n}$, $\Lambda \neq 0$. Using the definitions of the sets I^0 and I^c and applying property (2.4) of φ yields the equivalent relation

$$\varphi(Y_{I^c} - AX_{I^c}) - \varphi(Y_{I^c} - (A + \Lambda)X_{I^c}) < \varphi(\Lambda X_{I^0}).$$

By (2.5), we can note that $\varphi(Y_{I^c} - AX_{I^c}) - \varphi(Y_{I^c} - (A+\Lambda)X_{I^c}) \leq \ell(\Lambda X_{I^c})$. It then follows that

$$\ell(\Lambda X_{I^c}) < \varphi(\Lambda X_{I^0})$$

is a sufficient condition for $\Psi(Y, X) = \{A\}$. Finally, invoking (2.6) allows us to observe that $\ell(\Lambda X_{I^0}) - |I^c_\varepsilon(\Lambda X_{I^0})|\,\varepsilon \leq \varphi(\Lambda X_{I^0})$ which implies that $\ell(\Lambda X_{I^c}) < \ell(\Lambda X_{I^0}) - |I^c_\varepsilon(\Lambda X_{I^0})|\,\varepsilon$ is a sufficient condition for $\Psi(Y, X) = \{A\}$. We have hence proved the proposition. □

2.3.2 Uncertainty set induced by dense noise

When both E and F are nonzero in the data-generating system (2.2), $\Psi(Y, X)$ is likely to be a non-singleton subset of $\mathscr{P}(\mathbb{R}^{m\times n})$ especially if we consider all possible realizations of the unknown components E and F. In this case the desirable properties of the estimator are in default of better (i) that it contains A^o and (ii) that its size with respect to some metric is as small as possible. In this section we are interested in estimating the size of $\Psi(Y, X)$ when both dense noise E and sparse noise F are active in the data-generating system (2.2).

A notion of estimator gain. Similarly to the concept of system gain in control [68], one could define the gain of an estimator, that is, a quantitative measure of the sensitivity of the estimator with respect to the perturbations affecting the measurements. Consider a data pair (Y, X) generated by a system of the form (2.2) with A^o being the parameter matrix sought for. Let us fix the sparse noise matrix F or view it somehow as part of the data-generating system. This consideration proceeds from the fact that Ψ can be insensitive to F (when acting alone) under for example the condition derived in Theorem 2.2. Let E be bounded in the sense that $\ell(E)$ is finite with ℓ being the norm appearing in (2.6). Then we can define a gain of the estimator with respect to the dense noise component E. More specifically, an (ℓ, q)-gain of the

estimator Ψ with respect to the dense noise E may be defined by

$$g_{\ell,q}(Y,X) = \sup_{\substack{A^\star \in \Psi(Y,X) \\ 0<\ell(E)<\infty \\ F \text{ sparse}}} \frac{\|A^\star - A^o\|_q}{\ell(E)}. \tag{2.18}$$

Here $\|\cdot\|_q$ denotes matrix q-norm. The so-defined number $g_{\ell,q}(Y,X)$ provides an upper bound on the distance from the set $\Psi(Y,X)$ to A^o in function of the amount of dense noise. The following theorem and its corollaries show that if the number of nonzero columns in F is no larger than a certain threshold, then $g_{\ell,q}(Y,X)$ exists and is finite.

Theorem 2.3. *Let (Y,X) be the data generated by system (2.2) subject to the noise components E and F. Consider a triplet $(\varphi,\ell,\varepsilon)$ satisfying (2.4)-(2.6). Let $S^0 \subset \mathbb{I}$ be a set such that $F_{S^0} = 0$ and let $S^c = \mathbb{I} \setminus S^0$. Assume that the matrix X and the partition (S^0,S^c) are such that there exists $\alpha > 0$ such that*

$$\ell(\Lambda X_{S^0}) - \ell(\Lambda X_{S^c}) \geq \alpha \|\Lambda\|_q \ \forall \Lambda \in \mathbb{R}^{m\times n}, \tag{2.19}$$

with $\|\cdot\|_q$ denoting some matrix q-norm.
Then for any $A^\star \in \Psi(Y,X)$, it holds that

$$\|A^\star - A^o\|_q \leq \frac{1}{\gamma_{\ell,q}(X,S^c)}\left[2\ell(E_{S^0}) + |I^c_\varepsilon|\,\varepsilon\right] \tag{2.20}$$

with[1] $I^c_\varepsilon = I^c_\varepsilon(Y_{S^0} - A^\star X_{S^0}) = \{t \in S^0 : \ell(y_t - A^\star x_t) > \varepsilon\}$ *and*

$$\gamma_{\ell,q}(X,S^c) = \inf_{\substack{\Lambda\in\mathbb{R}^{m\times n} \\ \Lambda\neq 0}} \frac{\ell(\Lambda X_{S^0}) - \ell(\Lambda X_{S^c})}{\|\Lambda\|_q} \tag{2.21}$$

where $\|\cdot\|_q$ refers to matrix q-norm.

Proof. By definition of $\Psi(Y,X)$ in (2.3),

$$\varphi(Y - A^\star X) \leq \varphi(Y - AX)\ \forall A \in \mathbb{R}^{m\times n}$$

By letting $\Lambda = A - A^o$, $\Lambda^\star = A^\star - A^o$ and applying (2.2), the last inequality takes the form

$$\varphi(F + E - \Lambda^\star X) \leq \varphi(F + E - \Lambda X)\ \forall \Lambda \in \mathbb{R}^{m\times n}.$$

In particular, for $\Lambda = 0$, we get $\varphi(F + E - \Lambda^\star X) \leq \varphi(F + E)$ which, thanks to property (2.4), takes the form

$$\begin{aligned}\varphi(F_{S^c} + E_{S^c} - \Lambda^\star X_{S^c}) + \varphi(E_{S^0} - \Lambda^\star X_{S^0}) \\ \leq \varphi(F_{S^c} + E_{S^c}) + \varphi(E_{S^0}).\end{aligned}$$

Now applying property (2.5) to the first member of the left hand side and rearranging yields

$$\varphi(E_{S^0} - \Lambda^\star X_{S^0}) - \ell(\Lambda^\star X_{S^c}) \leq \varphi(E_{S^0}).$$

[1] The notation I^c_ε is used for simplicity reasons.

Using (2.6) gives

$$\ell(E_{S^0} - \Lambda^\star X_{S^0}) - |I_\varepsilon^c|\,\varepsilon - \ell(\Lambda^\star X_{S^c}) \leq \varphi(E_{S^0}) \leq \ell(E_{S^0}).$$

Here we used the fact that $I_\varepsilon^c(E_{S^0} - \Lambda^\star X_{S^0})$ is equal to the set I_ε^c defined in the statement of the theorem.
Applying the triangle inequality property of ℓ, it can be seen that $\ell(\Lambda^\star X_{S^0}) - \ell(E_{S^0}) \leq \ell(E_{S^0} - \Lambda^\star X_{S^0})$. Combining with the previous inequality yields

$$\ell(\Lambda^\star X_{S^0}) - \ell(\Lambda^\star X_{S^c}) \leq 2\ell(E_{S^0}) + |I_\varepsilon^c|\,\varepsilon.$$

Finally, it follows from the definition of $\gamma_{\ell,q}(X, S^c)$ in (2.21) that

$$\gamma_{\ell,q}(X, S^c)\,\|\Lambda^\star\|_q \leq [2\ell(E_{S^0}) + |I_\varepsilon^c|\,\varepsilon].$$

The condition (2.19) guarantees that $\gamma_{\ell,q}(X, S^c)$ is well-defined and positive. Hence the statement of the theorem is established. □

Theorem 2.3 constitutes an interesting stability result in that it provides a finite upper bound on the distance from A^o to the set $\Psi(Y, X)$ as a function of the amplitude of the dense noise matrix E. It applies to any estimator Ψ defined as in (2.3) with φ a function obeying (2.4)-(2.6). In particular, in the situation where φ is a norm (in which case ε can be taken equal to zero in (2.6)), the inequality in (2.20) simplifies to

$$\|A^\star - A^o\|_q \leq \frac{2}{\gamma_{\ell,q}(X, S^c)}\ell(E_{S^0}). \tag{2.22}$$

If φ is defined as in (2.7) and if the dense noise matrix E is such that $\ell^o(e_t) \leq \varepsilon^o$ for all $t \in \mathbb{I}$, then by taking $\varepsilon = \varepsilon^o$ the set I_ε^c defined in the statement of Theorem 2.3 corresponds to the empty set so that (2.22) holds as well in this case. In connection with the idea of gain discussed earlier, one can interpret the factor $2/\gamma_{\ell,q}(X, S^c)$ as an estimate of the gain (of the estimator) with respect to dense noise.

Lastly, it is interesting to see that when φ is a norm, if $E = 0$ then the result of Theorem 2.3 implies that $\Psi(Y, X) = \{A^o\}$ provided (2.19) is true.

2.4 Discussions on some special cases

For the purpose of illustrating the extent of the results above, let us discuss further the situation where φ reduces to a norm.

2.4.1 Scenario when the loss function is a norm

Corollary 2.1. *Let (Y, X) be the data generated by system (2.2) subject to the noise components E and F. Let S^0 and S^c be defined as in the statement of Theorem 2.3. Assume that φ is a norm i.e., it satisfies (2.4)-(2.6) with $\varepsilon = 0$.*
If X is self-decomposable and $|S^c| < T(\xi(X))$, then for any $A^\star \in \Psi(Y, X)$,

$$\|A^\star - A^o\|_q \leq \mathscr{B}_\varphi(|S^0|, X)\varphi(E_{S^0}) \tag{2.23}$$

where

$$\mathscr{B}_{\varphi}(\alpha, X) = \frac{2}{\sigma_{\varphi,q}(X)\Big[1 - \dfrac{N-\alpha}{T(\xi(X))}\Big]}, \tag{2.24}$$

$$\sigma_{\varphi,q}(X) = \inf_{\Lambda \neq 0} \frac{\varphi(\Lambda X)}{\|\Lambda\|_q} \tag{2.25}$$

Proof. The principle of the proof is to show that $\gamma_{\ell,q}(X, S^c)$ is well-defined and find a positive underestimate of it. Using the property (2.4) of φ and the fact that $\varphi = \ell$, we can write

$$\frac{\ell(\Lambda X_{S^0}) - \ell(\Lambda X_{S^c})}{\|\Lambda\|_q} = \frac{2\varphi(\Lambda X)}{\|\Lambda\|_q}\left[\frac{1}{2} - \frac{\varphi(\Lambda X_{S^c})}{\varphi(\Lambda X)}\right].$$

On the other hand we know from the proof of Theorem 2.2 (see Eq. (2.15)) that

$$\frac{\varphi(\Lambda X_{S^c})}{\varphi(\Lambda X)} \leq \frac{1}{2T(\xi(X))}\,|S^c|$$

so that

$$\left[1 - \frac{|S^c|}{T(\xi(X))}\right]\frac{\varphi(\Lambda X)}{\|\Lambda\|_q} \leq \frac{\ell(\Lambda X_{S^0}) - \ell(\Lambda X_{S^c})}{\|\Lambda\|_q}$$

Taking now the infimum on both sides of the inequality symbol over all nonzero matrices $\Lambda \in \mathbb{R}^{m\times n}$ yields

$$\sigma_{\varphi,q}(X)\left[1 - \frac{|S^c|}{T(\xi(X))}\right] \leq \gamma_{\ell,q}(X, S^c).$$

It follows from the rank condition imposed on X that $\sigma_{\varphi,q}(X) > 0$. This shows that $\gamma_{\ell,q}(X, S^c)$ is well defined and is strictly positive. Finally, since $\varphi = \ell$, invoking (2.22) gives the result. □

Two important comments can be made at this stage.

- First it is interesting to note that the bound $\mathscr{B}_{\varphi}(\alpha, X)$ is an increasing function of $\xi(X)$. Therefore it is all the smaller as $\xi(X)$ is small. That is the error bound will be small if the data matrix X is rich enough.
- Second, $\mathscr{B}_{\varphi}(\alpha, X)$ is a decreasing function of α. This means that the upper bound on the estimation error decreases when the number of gross error column in F decreases.

Beyond these observations it should be noted that a key assumption of Corollary 2.1 is that $|S^c| < T(\xi(X))$ with S^c being the index set of the nonzero columns in F. Realizing this condition requires on the one hand that the number of nonzero columns in the sparse noise matrix F be small and on the other hand that $\xi(X)$ be small[2] (which means that the data must be generic). Indeed this condition is not necessarily as strong as it might appear to be at first sight. For example, it can be relaxed as follows. Observe that the sum $E + F$ is not uniquely defined from model (2.2). Taking advantage of this, one can always absorb in E all nonzero columns of F whose magnitude does not exceed a certain level. To see this let $I = \{t \in S^c : \ell(e_t + f_t) \leq \varepsilon^o\}$ where $\varepsilon^o = \max_{t\in\mathbb{I}} \ell(e_t)$. Then we can define $\tilde{E}$ and $\tilde{F}$ such that $E + F = \tilde{E} + \tilde{F}$ and $\tilde{F}_{S^0\cup I} = 0$ that is, we set $\tilde{e}_t = f_t + e_t$ for any $t \in I$. As a consequence, E and F in Corollary 2.1 can be

[2]Recall that T is a decreasing function hence implying that $T(\xi(X))$ is large when $\xi(X)$ is small.

replaced by $\tilde{E}$ and $\tilde{F}$ respectively so that $|S|$ and $|S^c|$ are replaced by $|S|+|I|$ and $|S^c|-|I|$. The condition of the corollary becomes $|S^c|-|I|<T(\xi(X))$ which is potentially easier to fulfill.

Remark 2.4 (sum of p-norms)**.** *Evaluating numerically the bound $\mathscr{B}_\varphi(\alpha,X)$ might prove to be a hard problem due to the potential difficulty in computing the term $\sigma_{\varphi,q}(X)$ in (2.25). A particular case of interest is when φ consists of a sum of p-norms of the column vectors, i.e. when it is defined by $\varphi(A)=\sum_{i=1}^N \|a_i\|_p$. In this case if we take $q=2$ in (2.23) and (2.25), it is easy to see that $\lambda_{\min}^{1/2}(XX^\top)\leq\sigma_{\varphi,2}(X)$ with $\lambda_{\min}^{1/2}$ denoting the square root of the minimum eigenvalue. Replacing $\sigma_{\varphi,2}(X)$ with $\lambda_{\min}^{1/2}(XX^\top)$ in (2.24) yields an overestimate of $\mathscr{B}_\varphi(\alpha,X)$ which is computable.*

Remark 2.5. *Corollary 2.1 still holds true if one replaces $T(\xi(X))$ with $\pi_\varphi^c(X)$ defined in (2.11). As shown in [57], the number $\pi_\varphi^c(X)$ in (2.11) is computable although at the price of a combinatorial complexity. However if the n-dimension of X is small enough the complexity of the algorithm proposed there can be affordable. Then by using our formula (2.24) and Remark 2.4 above, it is possible therefore obtain a smaller bound on the estimation error.*

2.4.2 Single output case: ℓ_1 norm

The results obtained above apply very interestingly as well to single-output systems defined by

$$y_t=(\theta^o)^\top x_t+f_t+e_t \tag{2.26}$$

where y_t, e_t, f_t are scalars and x_t and θ^o are n-dimensional vectors. By letting $Y=[y_1\quad\cdots\quad y_N]\in\mathbb{R}^{1\times N}$ and defining E and F similarly, we obtain

$$Y=(\theta^o)^\top X+F+E. \tag{2.27}$$

This last equation corresponds indeed to (2.2) where the matrix A^o reduces to the row vector $(\theta^o)^\top$. In this case, if we let $\varphi(A)=\sum_{i=1}^N\|a_i\|_2$ then for any $\theta\in\mathbb{R}^n$, the columns of (the row vector) $Y-AX$ are scalars so that

$$\varphi(Y-\theta^\top X)=\sum_{t=1}^N\|y_t-\theta^\top x_t\|_2=\sum_{t=1}^N|y_t-\theta^\top x_t|. \tag{2.28}$$

As a result, Ψ coincides in this case with the Least Absolute Deviation (LAD) estimator. The following corollary specializes the result of Theorem 2.3 to the LAD estimator.

Corollary 2.2. *Let $(Y,X)\in\mathbb{R}^{1\times N}\times\mathbb{R}^{n\times N}$ be generated by model (2.26). Let $S^c=\{t\in\mathbb{I}:f_t\neq0\}$, $S^0=\mathbb{I}\setminus S^c$. Assume that X is self-decomposable and $|S^c|<T(\xi(X))$. Then for any $\theta^\star\in\arg\min_{\theta\in\mathbb{R}^n}\|Y-\theta^\top X\|_1$,*

$$\|\theta^\star-\theta^o\|_2\leq\mathscr{B}_1(|S^0|,X)\,\|E_{S^0}\|_1$$

where

$$\mathscr{B}_1(\alpha,X)=\frac{2}{\sigma_{1,2}(X)\Big[1-\dfrac{N-\alpha}{T(\xi(X))}\Big]},$$

$$\sigma_{1,2}(X)=\inf_{\eta\neq0}\frac{\left\|X^\top\eta\right\|_1}{\|\eta\|_2}$$

Again here the bound $\mathscr{B}_1(\alpha, X)$ can be numerically overestimated by following the idea of Remark 2.4.

2.4.3 Further analysis of a special case

The elegance of the above results resides in the fact that they read nicely as a simple threshold condition on the number of nonzeros columns (outliers) in the sparse noise matrix F under which either exact recovery or estimation error boundedness is guaranteed. See for example Theorems 2.2 and 2.3. The derived threshold depends solely on the informativity properties of the regressor matrix X. In addition, the threshold is computable. Note however that this readability of the recoverability conditions is obtained at the price of introducing some pessimism.
In this section, we state necessary and sufficient conditions for exact recovery of the parameter matrix. Considering the case where the function φ in (2.3) is defined by $\varphi(A) = \sum_{i=1}^{N} \|a_i\|_2$, the underlying optimization problem is nonsmooth and convex. Applying then the subdifferential condition for optimality yields the following theorem [10].

Theorem 2.4. *Consider the estimator Ψ in (2.3) with φ being defined from a sum of 2-norms by $\varphi(A) = \sum_{i=1}^{N} \|a_i\|_2$. Then a matrix $A^\star \in \mathbb{R}^{m\times n}$ lies in $\Psi(Y, X)$ if and only if any of the following equivalent statements holds:*

T1. *There exists a sequence of vectors $\{\beta_t\}_{t\in\mathbb{I}^0(A^\star)} \subset \mathcal{B}_2(0, 1)$ such that*

$$\sum_{t\notin\mathbb{I}^0(A^\star)} v_t^\star x_t^\top + \sum_{t\in\mathbb{I}^0(A^\star)} \beta_t x_t^\top = 0, \tag{2.29}$$

where $v_t^\star = (y_t - A^\star x_t)/\|y_t - A^\star x_t\|_2$. Here, $\mathcal{B}_2(0, 1) \subset \mathbb{R}^m$ is the Euclidean unit ball of $\mathbb{R}^m$.

T2. *For any matrix $\Lambda \in \mathbb{R}^{m\times n}$,*

$$\Big|\sum_{t\notin\mathbb{I}^0(A^\star)} {v_t^\star}^\top \Lambda x_t\Big| \le \sum_{t\in\mathbb{I}^0(A^\star)} \|\Lambda x_t\|_2 \,. \tag{2.30}$$

T3. *The optimal value of the problem*

$$\min_{Z\in\mathbb{R}^{m\times p}} \|Z\|_{2,\infty} \text{ subject to } V^\star X_{\mathbb{I}^c(A^\star)}^\top = Z X_{\mathbb{I}^0(A^\star)}^\top \tag{2.31}$$

is smaller than 1. *Here, $p = |\mathbb{I}^0(A^\star)|$ and $V^\star$ being a matrix formed with the unit ℓ_2-norm vectors $v_t^\star$, for $t \in \mathbb{I}^c(A^\star)$.*

Moreover, the solution $A^\star$ is unique if and only if any of the following assertions is true.

T1'. (2.29) *holds and* $\operatorname{rank}(X_{\mathcal{T}}) = n$ *where* $\mathcal{T} = \{t \in \mathbb{I}^0(A^\star) : \|\beta_t\|_2 < 1\}$.

T2'. (2.30) *holds with strict inequality symbol for all $\Lambda \in \mathbb{R}^{m\times n}$, $\Lambda \neq 0$.*

In the special case of single output systems the ℓ_2-norm reduces to absolute value and Theorem 2.4 then takes a simpler form as recalled below (See Theorem 4 in [10]). For the sake of simplicity we will discuss more the single output case instead of the general multivariable one.

Theorem 2.5 (Solution to the ℓ_1 problem)**.** *Consider the estimator Ψ in (2.3) with $m = 1$ and φ being defined from a sum of 2-norms by $\varphi(A) = \sum_{i=1}^{N} \|a_i\|_2$. Then a vector $\theta^\star \in \mathbb{R}^n$ lies in $\Psi(Y, X)$ if and only if any of the following equivalent statements hold:*

S1. There exist some numbers $\lambda_t \in [-1,\ 1]$, $t \in \mathbb{I}^0(\theta^\star)$, such that[3]

$$\sum_{t\in\mathbb{I}^+(\theta^\star)} x_t - \sum_{t\in\mathbb{I}^-(\theta^\star)} x_t = \sum_{t\in\mathbb{I}^0(\theta^\star)} \lambda_t x_t. \tag{2.32}$$

S2. For any $\eta \in \mathbb{R}^n$,

$$\Big| \sum_{t\in\mathbb{I}^+(\theta^\star)} \eta^\top x_t - \sum_{t\in\mathbb{I}^-(\theta^\star)} \eta^\top x_t \Big| \le \sum_{t\in\mathbb{I}^0(\theta^\star)} |\eta^\top x_t|. \tag{2.33}$$

Moreover, the solution $\theta^\star$ is unique if and only if any of the following is true.

S1'. (2.32) holds and $\operatorname{rank}(X_S) = n$ where $S = \{t \in \mathbb{I}^0(\theta^\star) : |\lambda_t| < 1\}$.

S2'. (2.33) holds with strict inequality symbol for all $\eta \in \mathbb{R}^n$, $\eta \neq 0$.

A number of important comments follow from Theorem 2.5.

- One first consequence of the theorem is that θ^o can be computed exactly from a finite set of erroneous data (by solving the convex problem in (2.3)) provided it satisfies the conditions S1' or S2' of the theorem. Note that there is no explicit boundedness condition imposed on the error sequence $\{f_t\}$. Hence the nonzero errors in this sequence can have arbitrarily large magnitudes as long as the optimization problem makes sense, i.e., provided $\varphi(Y - (\theta^o)^\top X)$ remains finite.

- Second, the true parameter vector θ^o can be exactly recovered in the presence of, say, infinitely many nonzero errors f_t (see also Proposition 2.2). For example, if the regressors $\{x_t\}$ satisfy

$$\sum_{t\in\mathbb{I}^+(\theta^o)} x_t - \sum_{t\in\mathbb{I}^-(\theta^o)} x_t = 0, \tag{2.34}$$

  and $\operatorname{rank}(X_{\mathbb{I}^0(\theta^o)}) = n$, then by condition S2' θ^o is the unique element of (2.3) regardless of the number of errors affecting the data. This situation is graphically illustrated in Figure 2.1.

- Third, if $\Psi(Y, X)$ admits a member $\theta^\star \in \mathbb{R}^n$ that satisfies $y_t - x_t^\top \theta^\star \neq 0$ for all $t = 1, \ldots, N$, then $\theta^\star$ is non-unique. In effect, $\mathbb{I}^0(\theta^\star) = \emptyset$ in this case and so, $\operatorname{rank}(X_{\mathbb{I}^0(\theta^\star)}) = 0 < n$ which, by Theorem 3.2, implies non-uniqueness. Indeed this is typically the case whenever the noise $\{e_t\}$ is nonzero.

Although it is unlikely that a condition of the form (2.34) holds in general, it is very instructive to see the implication of that property. For example, a question one may ask is whether it can hold approximately, i.e., $\sum_{t\in\mathbb{I}^+(\theta^o)} x_t - \sum_{t\in\mathbb{I}^-(\theta^o)} x_t \approx 0$. Next, we discuss a special case in which the true parameter vector θ^o in (2.26) can, in principle, be obtained asymptotically in the presence of an infinite number of nonzero errors f_t's.

[3] Eq. (2.32) should be understood here with the implicit convention that any of the three terms is equal to zero whenever the corresponding index set is empty.

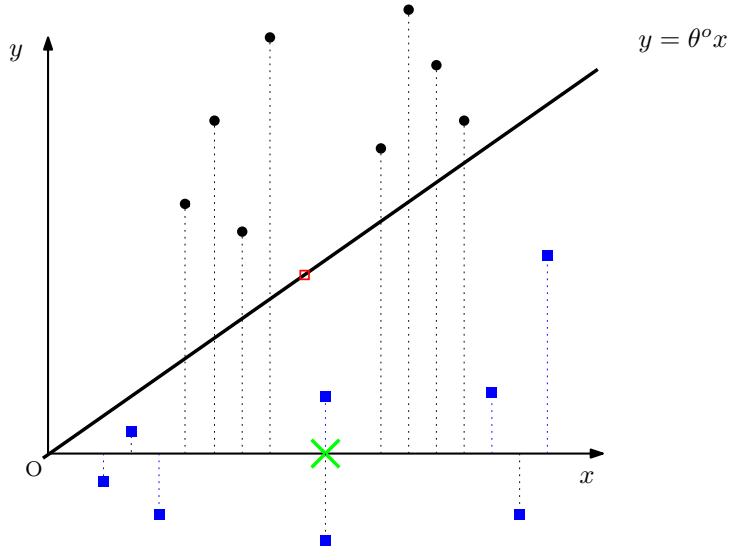


Figure 2.1: Illustration of the situation where condition (2.34) is satisfied. Here, $\mathbb{I}^+(\theta^o)$ is formed by the index set of the black points; $\mathbb{I}^-(\theta^o)$ indexes the blue squares; finally $\mathbb{I}^0(\theta^o)$ indexes the red boxes. Note that the cardinalities of $\mathbb{I}^+(\theta^o)$ and $\mathbb{I}^-(\theta^o)$ are both equal to 8. The average of the x-coordinates of the points indexed by $\mathbb{I}^+(\theta^o)$ and $\mathbb{I}^-(\theta^o)$ coincide at the point which is materialized by a green cross. The dimension of the regressor x is one and there are at least one point (distinct from the origin) that lies exactly on the line to be recovered. It follows from the above remarks that θ^o can be uniquely recovered by convex optimization from the whole data samples.

Proposition 2.2 (Infinite number of outliers)**.** *Assume that the error sequence* $\{e_t\}$ *in* (2.2) *is identically equal to zero. Assume further that the data* $\{(x_t, y_t)\}_{t=1}^N$ *are generated such that:*

- *There is a set* $I^0 \subset \mathbb{I}$ *with* $|I^0| \geq n$*, such that for any* $t \in I^0$*,* $f_t = 0$ *and* $\operatorname{rank}(X_{I^0}) = n$*,*
- *For any* $t \notin I^0$*,* f_t *is sampled from a distribution which is symmetric around zero.*
- *The regression vector sequence* $\{x_t\} \subset \mathbb{R}^n$ *is drawn from a probability distribution having a finite first moment.*

Then

$$\lim_{N\to\infty} \operatorname*{arg\,min}_{\theta\in\mathbb{R}^n} \frac{1}{N} \sum_{t=1}^{N} \left| y_t - x_t^\top \theta \right| = \{\theta^o\} \tag{2.35}$$

with probability one.

Another immediate consequence of Theorem 2.5 can be stated as follows.

Corollary 2.3 (On the special case of affine model)**.** *If the system* (2.2) *is affine in the sense that the regressor* x_t *has the form* $x_t = [\tilde{x}_t^\top \; 1]^\top$*, with* $\tilde{x}_t \in \mathbb{R}^{n-1}$*, then a necessary condition for* $\theta^\star$ *to be in* $\Psi(Y, X)$ *is that*

$$\left| \left| \mathbb{I}^+(\theta^\star) \right| - \left| \mathbb{I}^-(\theta^\star) \right| \right| \leq \left| \mathbb{I}^0(\theta^\star) \right|. \tag{2.36}$$

Here, the outer bars $|\cdot|$ *refer to absolute value while the inner ones which apply to sets refer to cardinality.*

Eq. (2.36) implies that if the measurement model is affine (for example, (2.26)) and all the f_t's have the same sign, i.e., if one of the cardinalities $|\mathbb{I}^+(\theta^o)|$ or $|\mathbb{I}^-(\theta^o)|$ is equal to zero, then (2.3) cannot contain the true θ^o whenever more than 50% of the elements of the sequence $\{f_t\}$ are nonzero.

The comments above show that exact recovery can happen in various circumstances and even in situations where the number of outliers is very large. For that, it suffices that the signs of the f_t's be appropriately distributed.

2.5 Practical implementation aspects

On the computation of $\Psi(Y, X)$. In the previous sections we have studied the properties of the estimator Ψ defined in (2.3) as induced by those of the loss function φ. An interesting question we shall discuss now is the computability of $\Psi(Y, X)$ given the data (Y, X). For this purpose recall that the convexity assumption imposed on φ aims simply at easing the numerical computation of Ψ. As long as φ is convex, the optimization problem which defines Ψ is a convex optimization problem. And these types of problems can be efficiently solved using numerous and well-documented solvers. In general, especially when dense noise is active, Ψ is likely to contain an infinite number of elements so that it is impossible to enumerate them all. A pragmatic way is to compute a single element of $\Psi(Y, X)$ and then find a set of larger size containing $\Psi(Y, X)$ using for example the bounding results of Theorem 2.3.

Enforcing recoverability by iterative re-weighting. The parameter matrix A^o from the model (2.2) can be uniquely recovered by solving the convex problem (2.3) if A^o satisfies, for example, condition (2.13) of Theorem 2.2. In case this condition is not naturally satisfied, an interesting question is how we can process the data in order to promote it. In this section we discuss an algorithmic strategy for enhancing the recoverability of A^o. Our discussion is inspired by [21]. The idea is to solve a sequence of problems of the type (2.3) with different weights computed iteratively [21, 1]. The iterative scheme can be defined for a fixed number $r_{\max}$ of iterations as follows. At iteration $r = 0, \ldots, r_{\max}$, compute

$$A^{(r)} \in \underset{A \in \mathbb{R}^{m \times n}}{\arg\min}\, \varphi\big((Y - AX)W^{(r)}\big), \tag{2.37}$$

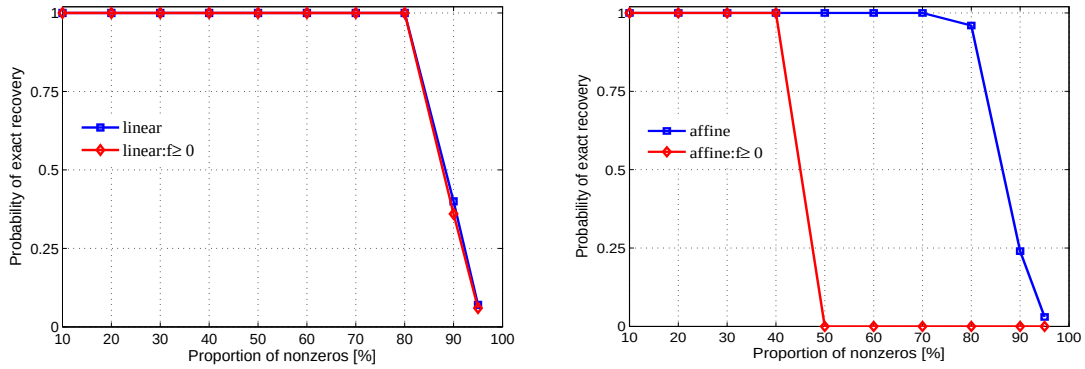
with weighting matrix $W^{(r)}$ defined by $W^{(r)} = \operatorname{diag}\left(w_1^{(r)}, \ldots, w_N^{(r)}\right)$ with the weights defined for all t, by $w_t^{(0)} = 1/N$, and

$$w_t^{(r)} = \frac{\xi_t^{(r)}}{\sum_{t=1}^N \xi_t^{(r)}}, \quad \text{if } r \geq 1,$$

where

$$\xi_t^{(r)} = \frac{1}{\varphi(y_t - A^{(r-1)}x_t) + \delta},$$

with $\delta > 0$ a small number whose role is to prevent division by zero and r is the iteration number. Since we are dealing here with a sequence of convex optimization problems, they can be numerically implemented using any convex solver. In particular the `CVX` software package [32] provides a handy environment for shaping this category of problems in a Matlab.



(a) Linear static model

(b) Affine static model

Figure 2.2: Estimates of probabilities of exact recovery when noise $\{e_t\}$ is equal to zero.

2.6 Numerical illustrations

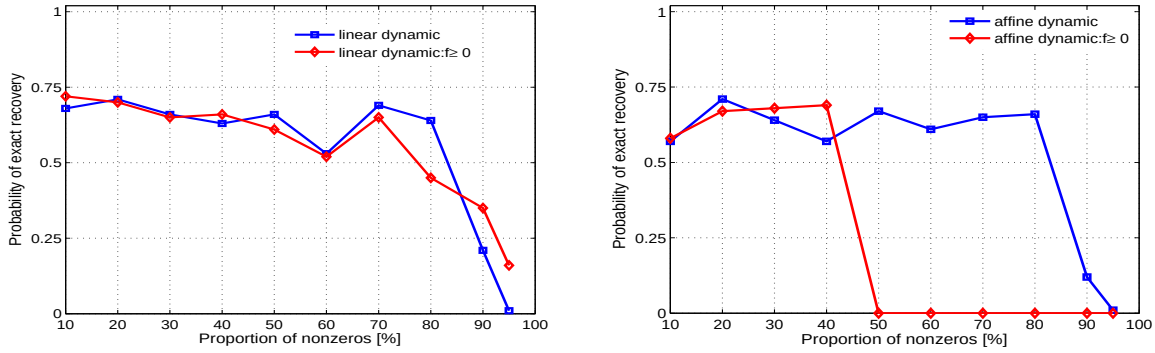
In this section we illustrate numerically the robustness property of the discussed estimators for different configurations of datasets.

2.6.1 Exact recovery

Static models subject to intermittent gross errors. In our first experiment we consider static linear and affine models of the form (2.26) with $n = 4$ and $N = 500$. The affine model refers to the case where the regressor x_t has the form $x_t = [\tilde{x}_t^\top \ 1]^\top$. The goal is to estimate the probability of exact recovery of the true parameter vector by minimizing the objective (2.28) as a function of the number of nonzero elements in the sequence $\{f_t\}$. For this purpose, the noise $\{e_t\}$ is set to zero. The nonzero elements of $\{f_t\}$ are drawn from a Gaussian distribution with mean 100 and variance 1000^2. For each level of sparsity (i.e., proportion of nonzeros), a Monte Carlo simulation of size 100 is carried out with randomly generated static/affine models and 500 data samples at each run. Repeating this for four situations (linear/affine and linear/affine with positive f_t's), we obtain the results depicted in Figure 2.2. We observe that in the linear case, the true parameter vector is the unique element of $\Psi(Y, X)$ when the output is affected by up to 80% of nonzero gross errors. This is because the data $\{x_t\}$ which were sampled from a Gaussian distribution are very generic. In the case of affine models, the performance is a little less good. If we set all f_t's to have the same sign, then, the percentage of outliers that can be corrected by the estimator Ψ cannot exceed 50%. This is consistent with (2.36).

Dynamic linear models subject to sensor intermittent faults. In the case when (2.26) represents a dynamic ARX model subject to gross errors, it can be observed (see Fig. 2.3) that the probabilities of exact recovery are much smaller than in the static case studied above. This difference is related to the richness (or genericity) of the regression vectors (columns of X) involved in each case. In the static example above, the vectors $\{x_t\}$ are freely sampled in any direction of $\mathbb{R}^n$ by following a Gaussian distribution. In the dynamic system case however, the data vectors $\{x_t\}$ constructed as

$$x_t = \begin{bmatrix} y_{t-1} & \cdots & y_{t-n_a} & u_t^\top & u_{t-1}^\top & \cdots & u_{t-n_b}^\top \end{bmatrix}^\top$$



(a) Linear dynamic model (b) Affine dynamic model

Figure 2.3: Estimates of probabilities of exact recovery when noise $\{e_t\}$ is equal to zero. Results of a Monte-Carlo simulation of size 100 with randomly generated linear ARX systems of order $n_a = n_b = 2$.

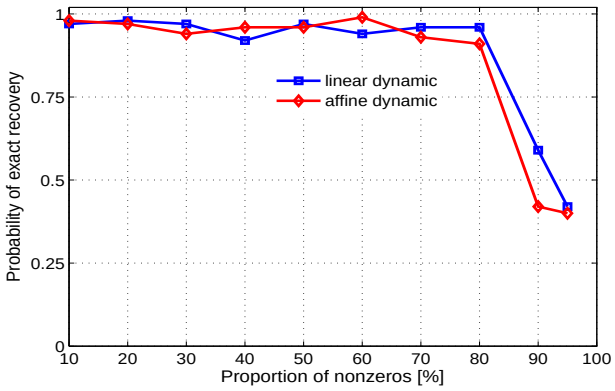
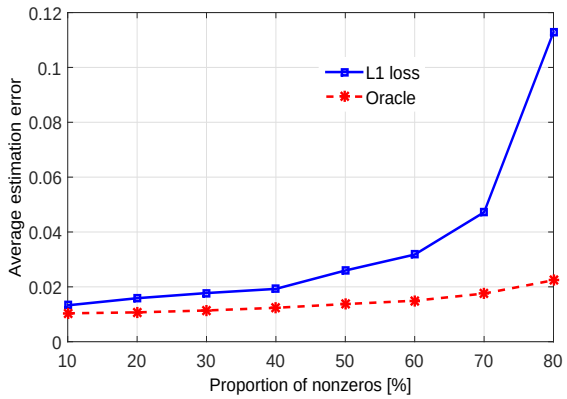


Figure 2.4: Estimates of probabilities of exact recovery by reweigthed ℓ_1 minimization when noise $\{e_t\}$ is equal to zero. Results of a Monte-Carlo simulation of size 100 with randomly generated linear ARX systems with orders $n_a = n_b = 2$.

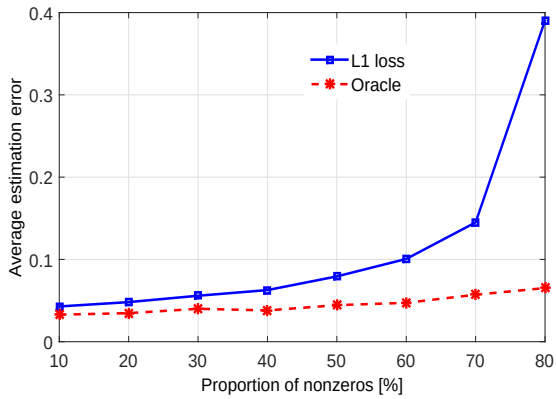
are constrained to lie on a manifold. As a result, the data matrix X generated by the dynamic system is less generic. According to the condition (2.13), there is a threshold depending on the richness of the data such that exact recovery is guaranteed whenever the number of nonzero entries in $\boldsymbol{f}$ is smaller than this threshold. So, the more generic the data contained in X are, the more outliers can be removed by the estimator. Note that the lack of sufficient genericity can be compensated (to some extent) by implementing the iterative sparsity enhancing technique (the ℓ_1 reweighted algorithm) described in Section 2.5. This leads, for only two iterations, to significantly improved results as represented in Figure 2.4.

2.6.2 Presence of both dense and sparse noise

Empirical estimation error. Consider now the case of static models of the form (2.2) in the presence of both $\{e_t\}$ and $\{f_t\}$ sampled from Gaussian distributions $\mathcal{N}(0, \sigma_e^2)$ and $\mathcal{N}(100, 1000^2)$ respectively. The variance σ_e^2 is selected so as to achieve a certain signal to noise ratio before the



(a) Static model: SNR = 20 dB



(b) Static model: SNR = 10 dB

Figure 2.5: Average relative estimation error versus sparsity level.

gross error sequence is added to the output. Again, by carrying out a Monte-Carlo simulation of size 100 with different sparsity levels and randomly generated models at each run, we obtain the average errors plotted in Figure 2.5. The performance can be assessed by comparing with an "oracle" estimate i.e., the least squares estimate one would obtain if the locations of zeros in the sequence $\{f_t\}$ were known. The results in Figure 2.5 tend to suggest that the proposed approach performs very well. For the current numerical experiment, our results are very close to the ideal estimate when the proportion of nonzeros is less than 70%. Above this proportion, the estimation error presents an important jump.

Evaluation of theoretical error bounds. Here we evaluate numerically an estimate of the gain of the estimator based on Corollary 2.1 and Remark 2.4. The estimation is carried out for the case where φ consists in the sum of 2-norms and $q = 2$. Four different cases are studied:

(a) Static data: $X \in \mathbb{R}^{2\times 200}$ is sampled from a *Gaussian distribution* $\mathcal{N}(0, I_2)$ with zero-mean and identity-covariance.

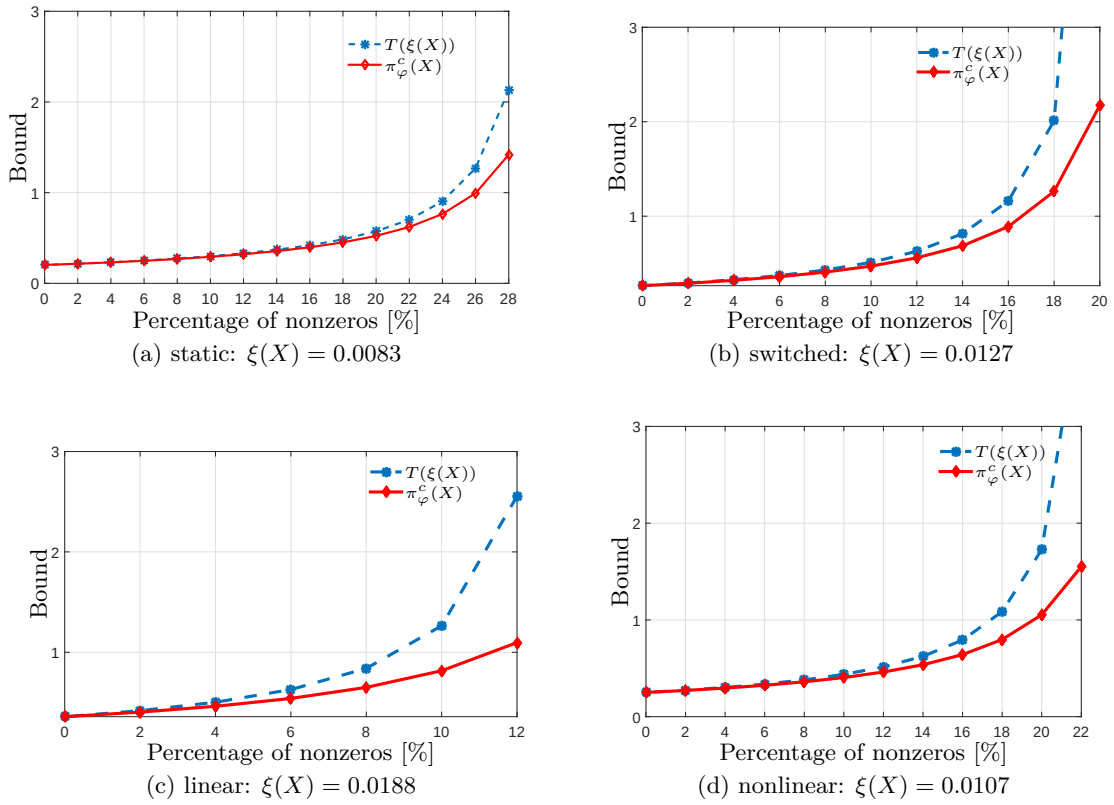
(b) Dynamic data generated by a *switched linear system*: $X \in \mathbb{R}^{2\times 200}$ is formed with the regressors (y_{t-1}, u_{t-1}) generated by a switched linear system composed of 3 subsystems of order 1. This is a switched ARX system defined by $y_t = a_{\sigma(t)}y_{t-1} + b_{\sigma(t)}u_{t-1}$ with the switching signal $\sigma(t) \in \{1, 2, 3\}$ generated from a uniform distribution and input u_t being a white noise with Gaussian distribution; $(a_1, b_1) = (-0.40, -0.15)$, $(a_2, b_2) = (1.55, -2.10)$ and $(a_3, b_3) = (1, -0.65)$.

(c) Dynamic data generated by a *linear ARX system* defined by $y_t = a_1 y_{t-1} + b_1 u_{t-1}$

(d) Dynamic data generated by a *nonlinear NARX system* defined by $y_t = (y_{t-1} + 2.5)/(1 + y_{t-1}^2) + u_{t-1}$

Following Remark 2.1, the columns of all data matrices X have been normalized to unit 2-norm before being processed.

Figure 2.6 plots the obtained estimate of the estimator gain against the proportion of correctable outliers. As remarked in Section 2.4, the gain estimate increases as the proportion of outliers gets larger. But the growth rate of the gain estimate depends on the genericity of the data



(a) static: $\xi(X) = 0.0083$

(b) switched: $\xi(X) = 0.0127$

(c) linear: $\xi(X) = 0.0188$

(d) nonlinear: $\xi(X) = 0.0107$

Figure 2.6: An overestimate of $\mathscr{B}_\varphi$ using respectively $\pi^c_\varphi(X)$ and $T(\xi(X))$ for a data matrix $X \in \mathbb{R}^{2\times 200}$: (a) static data sampled from a Gaussian distribution; (b) data generated by a switched system; (c) data generated by a linear dynamic system ; (d) data generated by a dynamic nonlinear system.

matrix X. The more generic the columns of X are, the smaller the growth rate of the estimation error is when regarded as a function of the proportion of outliers. The experiment confirms also the intuition according to which static data tend to be more generic than data generated by a dynamic system. Among the three cases of dynamic systems, the linear system appears to be the one generating the least generic data.

2.7 Conclusion

In this chapter we have discussed a somewhat general framework for designing a robust estimator. Given the training data, the estimator is defined as the minimizing set of a certain performance index applying to the data. We have shown that if the performance function possesses some key properties, then the so-defined estimator will inherit robustness properties. Considering a data set generated by a linear model subject to both sparse and dense noises, we showed that the estimator is insensitive to the sparse noise provided the number of its nonzero components is no larger than a certain (computable) threshold. Conditions are proposed for the exact recovery of the true parameter matrix when only the sparse noise is active. When both types of noises affect the measurements we propose computable bounds on the parametric estimation error which depend on the amplitude of the dense noise and the number of nonzero elements of the

sequence of sparse noise. Finally we note that the proposed robust estimation and analysis framework is sufficiently general to cover the LAD estimator and its main variants. Also, it applies to both SISO and MIMO types of systems.

Robust estimation has been studied here in a batch identification mode. The resulting estimator is defined from a finite collection of measurements. In this case, the estimator takes the form of a static operator. One can also envision studying robust estimation in an adaptive framework. In this latter case the estimator becomes a dynamic operator which produces at each time a set-valued output. See for example our papers [23, 24, 3].

A relevant open problem is how to design the excitation input of a dynamic system so as to maximize the richness of the resulting regressor (which can be measured in the sense of $\xi(X)$).

Chapter 3

Identification of switched ARX systems

Contents

3.1 Introduction

This chapter is based on [1] where we consider the problem of identifying a switched linear system from input-output data. In our setting, each subsystem is described by an ARX model. The main challenge with this identification problem is that the data are available only as a mixture of observations generated by a finite set of different interacting linear subsystems so that one does not know a priori which subsystem has generated which data.
The contribution of this work consists in the development of an identification method for switched linear systems. Data vectors generated by such systems lie in the union of a finite set of linear hyperplanes. Therefore we pose the identification of a specific submodel as the problem of extracting the hyperplane that contains the largest number of data. The corresponding submodel is hence the one that, among all submodels, achieves, over the whole dataset, the sparsest vector

of fitting errors. With this formulation, one submodel can be estimated directly without any prior clustering, by means of sparse optimization, i.e., the minimization of the number of nonzero components in an error vector. Since sparse optimization is in general non-convex, it is classical to consider instead a convex ℓ_1 relaxation of this problem. We then present sufficient conditions under which the ℓ_1 relaxation is guaranteed to recover exactly the solution of the initial sparse optimization problem. In the case when these conditions are not satisfied, we show that all the PVs can still be identified by slightly adapting an iterative reweighted ℓ_1 optimization technique proposed in [21]. In contrast to most of the existing methods for hybrid system identification, our method lends itself to a relatively easy analysis. For example, conditions for optimality even though somewhat conservative, can be derived.

3.2 Switched ARX systems

We consider a discrete-time MISO switched linear system (SLS) represented by

$$\begin{aligned} y_t = a^1_{\sigma(t)} y_{t-1} + \cdots + a^{n_a}_{\sigma(t)} y_{t-n_a} \\ + (b^1_{\sigma(t)})^\top u_{t-1} + \cdots + (b^{n_b}_{\sigma(t)})^\top u_{t-n_b} + e_t \end{aligned} \tag{3.1}$$

where $u_t \in \mathbb{R}^{n_u}$ and $y_t \in \mathbb{R}$ denote respectively the input and the output of the system. The integers n_a and n_b in (3.4) are the output and input lags (also called the orders of the system). $\{e_t\}$ models a deterministic sequence of bounded errors referring to potential model mismatch and measurement noise. We will assume that for all time t, $|e_t| \leq \varepsilon$ with ε possibly unknown. $\sigma(t) \in \mathbb{S} \triangleq \{1, \ldots, s\}$ is the discrete mode (or discrete state), that is, the index of the active subsystem at time t; for $j \in \mathbb{S}$, $a^i_j \in \mathbb{R}$ and $b^q_j \in \mathbb{R}^{n_u}$, $i = 1, \ldots, n_a$, $q = 1, \ldots, n_b$, are the parameters of the system. The model (3.2) is called a Switched Auto-Regressive eXogenous (SARX) model. For convenience, we rewrite (3.1) in the form

$$y_t = x_t^\top \theta^o_{\sigma(t)} + e_t, \tag{3.2}$$

where $\theta^o_{\sigma(t)} \in \mathbb{R}^n$, $n = n_a + n_b n_u$, is the parameter vector (PV) associated with the mode $\sigma(t)$,

$$\theta^o_{\sigma(t)} = \begin{bmatrix} a^1_{\sigma(t)} & \cdots & a^{n_a}_{\sigma(t)} & (b^1_{\sigma(t)})^\top & \cdots & (b^{n_b}_{\sigma(t)})^\top \end{bmatrix}^\top \tag{3.3}$$

and $x_t \in \mathbb{R}^n$ is the regressor at time $t \in \mathbb{Z}_+$ defined by

$$x_t = \begin{bmatrix} y_{t-1} & \cdots & y_{t-n_a} & u_{t-1}^\top & \cdots & u_{t-n_b}^\top \end{bmatrix}^\top. \tag{3.4}$$

Here we do not require that $n_a \geq n_b$. For example, if $n_a = 0$ in Eq. (3.4) then model (3.2) corresponds to a switched FIR (Finite Impulse Response) model.

We consider the problem of inferring a model of the form (3.2) from a finite collection of measurements $\{(x_t, y_t)\}_{t=1}^N$. We shall solve the identification problem without any knowledge of the switching signal $\{\sigma(t)\}$. This means that we do not know beforehand which data pair is associated with which parameter vector.

3.2.1 On the identifiability of the SARX model

As discussed in [65], the problem of inferring a switched model such as (3.2) from a *finite set measurements* is not well-posed if the structure of the model is not properly set. In effect, if the structural indices n_a and n_b are not fixed, then one can find for example a trivial switched linear model consisting of one single submodel with large orders that fits all the finite data set. Even if finite and fixed values are assigned to n_a and n_b, there are still infinitely many switched models that explain the data. For example, it can be simply verified that there is a switched linear model with $s = N$ submodels that can reproduce exactly the data.
In order to remove the identifiability issue, we will assume here that

- The orders n_a and n_b are finite, equal for all submodels and known a priori. This fixes the form of the model and thereby the dimension of the parameter space.
- Each individual ARX subsystem is minimal in the ordinary sense.[1]

With this setting for the structural indices n_a and n_b, the SARX of interest will be viewed as the one that, among all switched linear models consistent with the data, has the minimum number of submodels. Note that by the results of [47], the second assumption implies minimality of the SARX system. The interested reader is referred to the papers [47, 48] for a more complete treatment of the identifiability problem for switched linear systems in both the frameworks of state-space models and input-output models.

3.2.2 Preliminary discussions

The most direct approach to the SARX identification problem would be to solve the following optimization problem

$$\min_{\substack{\theta_1,\dots,\theta_s \\ \sigma(1),\dots,\sigma(N)}} \sum_{t=1}^{N} \left(y_t - \theta_{\sigma(t)}^\top x_t\right)^2 \tag{3.5}$$

that is, search jointly for a switching sequence $\{\sigma(t)\}_{t=1}^N$ and a set of s parameter vectors $\{\theta_i\}_{i=1}^s$ so as to minimize the average discrepancy between the measured output y_t and the output $\theta_{\sigma(t)}^\top x_t$ of the SARX model. A major obstacle however on the path to the solution to (3.5) is computational complexity. In effect, solving (3.5) might require an exhaustive search over the (discrete) set of length-N switching sequences $\{\sigma(t)\}_{t=1}^N$ whose cardinality is about s^N, a number which grows exponentially fast with respect to the number N of samples. Note that an equivalent formulation of the problem can be the following

$$\min_{\theta_1,\dots,\theta_s} \Big[\sum_{t=1}^{N} \min_{i=1,\dots,s} (y_t - \theta_i^\top x_t)^2\Big] \tag{3.6}$$

where the discrete decision variables have been removed. Unfortunately, this is a nonconvex problem.

A geometrical interpretation. From a geometrical perspective, the switched system identification problem formulated above is equivalent to that of subspace clustering, i.e., the problem of estimating subspaces from unlabeled data that lie in the union of those subspaces (an illustration

[1] i.e, the numerator and the denominator polynomials of the associated transfer function are coprime.

is depicted in Figure 3.1). In effect, if we neglect the noise and introduce the notations,

$$\bar{\theta}_i = [1 \quad \theta_i^\top]^\top \text{ and } \bar{x}_t = [y_t \quad -x_t^\top]^\top. \tag{3.7}$$

then for any time instant t, there is $i \in \{1, \ldots, s\}$ such that $y_t - \theta_i^\top x_t = \bar{x}_t^\top \bar{\theta}_i = 0$. Hence the data record $\{\bar{x}_t\}_{t=1}^N$ lie in the union of s linear hyperplanes whose normal directions are given by the parameter vectors $\bar{\theta}_i$, $i = 1, \ldots, s$. Estimating these normal vectors may require to group data lying in each hyperplane and then proceed with standard linear identification techniques for each group. Instead of doing so, we will extract the parameter vectors θ_i one after another, starting directly from the entire data set. In a sense, our method can be thought of as a robust identification approach. In fact, the method can only identify one submodel at a time and so, when identifying one submodel, data from other submodels are roughly treated as outliers or gross errors to be corrected.

3.3 The sparse optimization approach

3.3.1 The rationale of the idea

For the sake of clarity, assume for now that the noise sequence $\{e_t\}$ is identically null. Let $\theta \in \mathbb{R}^n$ denote a candidate parameter vector. Given the data $\{(x_t, y_t)\}_{t=1}^N$ generated by the system (3.2), we form the error vector

$$\boldsymbol{\phi}(\theta) = \boldsymbol{y} - X^\top \theta \tag{3.8}$$

where $\boldsymbol{y} = [y_1 \quad \cdots \quad y_N]^\top \in \mathbb{R}^N$ and $X = [x_1 \quad \cdots \quad x_N] \in \mathbb{R}^{n \times N}$. Likewise define $\boldsymbol{e} = [e_1 \quad \cdots \quad e_N]^\top \in \mathbb{R}^N$. Let us denote with N_i the number of data (x_t, y_t) generated by the subsystem indexed by i. Then we can observe that if $\theta = \theta_i$ for some $i \in \{1, \ldots, s\}$, then $\boldsymbol{\phi}(\theta)$ is a sparse vector, i.e., a vector where many entries are equal to zero. More precisely, $\boldsymbol{\phi}(\theta)$ contains N_i zero entries.
In order to avoid ambiguity in the definition of the number N_i, we make the following formal assumption throughout all the chapter. This assumption will remain implicitly in force whenever the notation N_i is invoked.

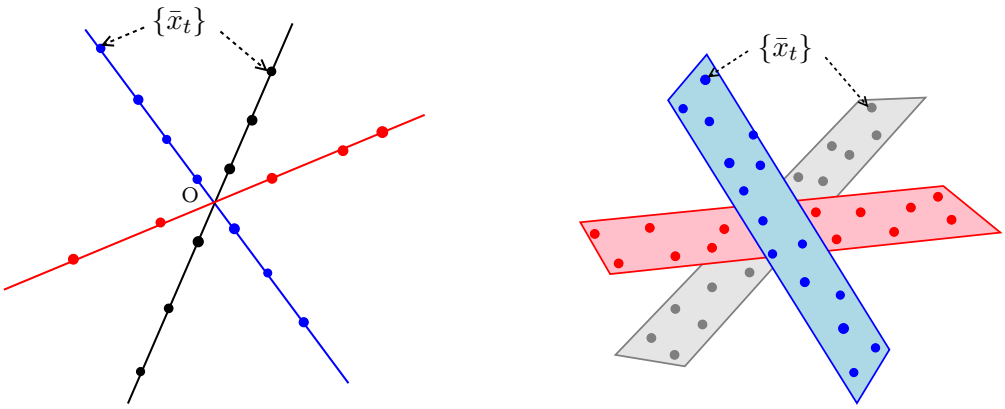


Figure 3.1: Switched system identification as a subspace clustering problem. Illustrative representation of homogenous hyperplanes in one and two dimensions.

Assumption 3.1. *There is no data pair (x_t, y_t) that fits two different submodels of the SARX (3.2), i.e., $y_t = \theta_i^\top x_t = \theta_j^\top x_t \Rightarrow i = j$.*

To determine the PV θ_i that achieves the sparsest error $\boldsymbol{\phi}(\theta_i)$, we can in principle solve the sparse optimization problem

$$\min_{\theta} \|\boldsymbol{\phi}(\theta)\|_0 , \tag{3.9}$$

where $\|z\|_0$ denotes the ℓ_0 norm[2] of z, that is, the number of nonzero entries of z, $\|z\|_0 = |\{i : z_i \neq 0\}|$. Trying to solve problem (3.9) is equivalent to attempting to find a homogeneous hyperplane (or a vector $\bar{\theta}$) that contains (that is orthogonal to) as many data $\bar{x}_t$ as possible. For the purpose of discussing the ability of problem (3.9) to solve the identification the problem, we introduce the following measure of informativity (richness) of the regression data [1].

Definition 3.1 (An integer measure of genericity)**.**
Let $X \in \mathbb{R}^{n\times N}$ be a data matrix satisfying $\operatorname{rank}(X) = n$ *and let $\mathbb{I} = \{1, \ldots, N\}$ be the index set of the data. The n-genericity index of X denoted $\nu_n(X)$, is defined as the minimum integer m such that any $n \times m$ submatrix of X has rank n,*

$$\nu_n(X) = \min \{m : \forall\, \mathcal{S} \subset \mathbb{I} \text{ with } |\mathcal{S}| = m,\ \operatorname{rank}(X_{\mathcal{S}}) = n\}. \tag{3.10}$$

If all the submodels are sufficiently excited within the data $\{x_t\}_{t=1}^N$ then, as suggested by the following lemma, the solution to problem (3.9) is a PV representing one of the constituent submodels of system (3.2).

Proposition 3.1 (Noise-free data)**.** *Let $(\boldsymbol{y}, X) \in \mathbb{R}^N \times \mathbb{R}^{n\times N}$ be data generated by the SARX system (3.2) under noise-free assumption ($\boldsymbol{e} = 0$) and pose $\boldsymbol{\phi}(\theta) = \boldsymbol{y} - X^\top\theta$. Assume that each subsystem has generated a sufficiently large number of data in the sense that $|\mathbb{I}^0(\theta_i^o)| \geq s\nu_n(X)$ for all $i \in \mathbb{S}$ with s being the number of subsystems in (3.2). Then*

$$\arg\min_{\theta\in\mathbb{R}^n} \|\boldsymbol{\phi}(\theta)\|_0 \subset \{\theta_1^o, \ldots, \theta_s^o\} \tag{3.11}$$

Proposition 3.1 says that the set of parameter vectors minimizing $\|\boldsymbol{\phi}(\theta)\|_0$ is included in $\{\theta_1^o, \ldots, \theta_s^o\}$. Next, we characterize the uniqueness of the minimizer of (3.9) in terms of the n-genericity index of the data matrix X.

Theorem 3.1. *Let $(\boldsymbol{y}, X) \in \mathbb{R}^N \times \mathbb{R}^{n\times N}$ be data generated by the SARX system (3.2) under noise-free assumption ($\boldsymbol{e} = 0$) and pose $\boldsymbol{\phi}(\theta) = \boldsymbol{y} - X^\top\theta$. Then the following statements hold true.*

1. *If there is a $\theta^\star \in \mathbb{R}^n$ such that $\|\boldsymbol{\phi}(\theta^\star)\|_0 \leq (N - \nu_n(X))/2$, then*

$$\arg\min_{\theta\in\mathbb{R}^n} \|\boldsymbol{\phi}(\theta)\|_0 = \{\theta^\star\}. \tag{3.12}$$

2. *If in addition Assumption 3.1 holds and $N \geq (2s-1)\nu_n(X)$, then*

$$\theta^\star \in \{\theta_1^o, \ldots, \theta_s^o\}.$$

[2]Strictly speaking, ℓ_0 is not a norm as it does not satisfy the property of positive scalability, i.e., $\|\lambda z\|_0 = |\lambda|\,\|z\|_0$ does not hold in general.

Noise-aware sparse optimization. In case the noise is not equal to zero in the data-generating system (3.2), then solving problem (3.9) is unlikely to return a true parameter vector. This observation prompts us to reformulate the search query. To this end, assume that the noise sequence $\{e_t\}$ is bounded by a given positive number ε. Then consider the alternative formulation

$$\begin{aligned} \min_{(\theta,\xi)\in\mathbb{R}^n\times\mathbb{R}^N_+} & \|\xi\|_0 \\ \text{s.t.} \quad & |y_t - x_t^\top\theta| \leq \varepsilon + \xi_t,\ t = 1,\dots,N. \end{aligned} \tag{3.13}$$

The decision variables here are the PV $\theta \in \mathbb{R}^n$ and the positive slack variable $\xi \in \mathbb{R}^N_+$. The rationale behind this formulation is that if $\theta \in \{\theta_1^o,\dots,\theta_s^o\}$, then $|y_t - x_t^\top\theta| \leq \varepsilon$ whenever $\sigma(t) = i$. Consequently, the corresponding entry ξ_t of ξ can be set equal to zero hence yielding a sparse vector ξ. Indeed (3.13) can be written in a more compact form as

$$\min_{\theta\in\mathbb{R}^n} \|\boldsymbol{\phi}(\theta)\|_{0,\varepsilon} \tag{3.14}$$

where the notation $\|\cdot\|_{0,\varepsilon}$ is defined by

$$\|a\|_{0,\varepsilon} = \left|\left\{i = 1,\dots,N : \max(0,|a_i| - \varepsilon) \neq 0\right\}\right|$$

for any $a = [a_1 \quad \cdots \quad a_N]^\top \in \mathbb{R}^N$. In others words $\|a\|_{0,\varepsilon}$ is the number of entries in a which have absolute value strictly larger than ε. Hence when $\varepsilon = 0$, $\|a\|_{0,\varepsilon}$ coincides with $\|a\|_0$.

Now we ask the question of what is the significance of the solutions to problem (3.14) with respect to the goal of recovering the parameter vectors of system (3.2). This is discussed next. For notational convenience, let us introduce the number $\delta(X)$ defined by

$$\delta(X) = \max_{\substack{I\subset\mathbb{I}\\ \nu_n(X)\leq|I|}} \frac{\sqrt{|I|}}{\lambda_{\min}^{1/2}\left(X_I X_I^\top\right)} \tag{3.15}$$

The maximum is taken here over all subsets of $\mathbb{I}$ having cardinality at least equal to $\nu_n(X)$. The notation $\lambda_{\min}^{1/2}\left(X_I X_I^\top\right)$ refers to the square root of the minimum eigenvalue of $X_I X_I^\top$, that is, the minimum singular value of $X_I^\top$ which is guaranteed to be strictly positive by the fact that $|I| \geq \nu_n(X)$.

Proposition 3.2 (Noisy data)**.** *Let $(\boldsymbol{y}, X) \in \mathbb{R}^N \times \mathbb{R}^{n\times N}$ be data generated by the SARX system (3.2) where the noise sequence $\{e_t\}$ is assumed to be bounded: there is $\varepsilon > 0$ such that $\max_t |e_t| \leq \varepsilon$. Assume that each subsystem has generated a sufficiently large number of data in the sense that $|\mathbb{I}^\varepsilon(\theta_i^o)| \geq s\nu_n(X)$ for all $i \in \mathbb{S}$, where $\mathbb{I}^\varepsilon(\theta_i^o) = \left\{t \in \mathbb{I} : |y_t - x_t^\top\theta_i^o| \leq \varepsilon\right\}$. Then with $\boldsymbol{\phi}(\theta) = \boldsymbol{y} - X^\top\theta$, it holds that*

$$\forall\hat{\theta} \in \arg\min_{\theta\in\mathbb{R}^n} \|\boldsymbol{\phi}(\theta)\|_{0,\varepsilon}, \exists i^* \in \mathbb{S}, \left\|\hat{\theta} - \theta_{i^*}^o\right\|_2 \leq 2\varepsilon\delta(X). \tag{3.16}$$

Proof. The proof is similar to that of Lemma 1 in [1]. Because the data are generated by the system (3.2), it is clear, under the boundedness assumption on the noise, that for any $t \in \mathbb{I}$, there exists $i \in \mathbb{S}$ such that $|y_t - x_t^\top\theta_i^o| \leq \varepsilon$. It follows that $\mathbb{I} = \mathbb{I}^\varepsilon(\theta_1^o) \cup \cdots \cup \mathbb{I}^\varepsilon(\theta_s^o)$. Let

$\hat{\theta} \in \arg\min_{\theta\in\mathbb{R}^n} \|\boldsymbol{\phi}(\theta)\|_{0,\varepsilon}$. Then

$$\left|\mathbb{I}^\varepsilon(\hat{\theta})\right| \leq \sum_{i=1}^{s} \left|\mathbb{I}^\varepsilon(\hat{\theta}) \cap \mathbb{I}^\varepsilon(\theta_i^o)\right| \tag{3.17}$$

We then claim that there is an $i^* \in \mathbb{S}$ such that $\left|\mathbb{I}^\varepsilon(\hat{\theta}) \cap \mathbb{I}^\varepsilon(\theta_{i^*}^o)\right| \geq \nu_n(X)$. To see this, assume that the opposite holds, meaning that for all $i \in \mathbb{S}$, $\left|\mathbb{I}^\varepsilon(\hat{\theta}) \cap \mathbb{I}^\varepsilon(\theta_i^o)\right| < \nu_n(X)$. Then by applying (3.17) and using the definition of $\hat{\theta}$, we immediately obtain $|\mathbb{I}^\varepsilon(\theta_i^o)| \leq |\mathbb{I}^\varepsilon(\hat{\theta})| < s\nu_n(X)$ which is in contradiction with the assumption of the proposition. Hence i^* exists as stated. Denote with $\boldsymbol{y}_{I^*}$ a vector formed with the outputs indexed by $I^* \triangleq \mathbb{I}^\varepsilon(\hat{\theta}) \cap \mathbb{I}^\varepsilon(\theta_{i^*}^o)$ and with X_{I^*} the matrix formed with the regressors indexed by I^*. For all $t \in I^*$, we have $|y_t - x_t^\top \hat{\theta}| \leq \varepsilon$ and $|y_t - x_t^\top \theta_{i^*}^o| \leq \varepsilon$. As a result,

$$\left\|X_{I^*}^\top \left(\hat{\theta} - \theta_{i^*}^o\right)\right\|_2 \leq \|\boldsymbol{y}_{I^*} - X_{I^*}\theta_{i^*}^o\|_2 + \|\boldsymbol{y}_{I^*} - X_{I^*}\hat{\theta}\|_2 \leq 2\sqrt{|I^*|}\varepsilon$$

so that

$$\left\|\hat{\theta} - \theta_{i^*}^o\right\|_2 \leq \frac{2\sqrt{|I^*|}\varepsilon}{\lambda_{\min}^{1/2}(X_{I^*}X_{I^*}^\top)}.$$

The conclusion follows naturally from this. □

It is interesting to note that (3.11) is a special case of (3.16) corresponding to the scenario when noise is absent ($\varepsilon = 0$).

3.3.2 Convex relaxation

Note that the problem (3.9) is NP-hard in general, see e.g., [42]. As a consequence, minimizing directly the cost function in (3.9) is in general intractable. A popular alternative [18, 21] is to consider a convex relaxation of problem (3.9) based on the ℓ_1 norm.

$$\min_{\theta\in\mathbb{R}^n} \|\boldsymbol{\phi}(\theta)\|_1, \tag{3.18}$$

where $\|z\|_1 = \sum_{i=1}^N |z_i|$ for any vector $z \in \mathbb{R}^N$. This latter problem corresponds to what is classically referred to as sparse error correction problem in [56] and [20]. Contrary to the problem (3.9), the problem (3.18) is convex and can even be transformed into a classical linear program. It can therefore be efficiently solved by standard convex optimization techniques such as interior points methods [16].

An interesting question one might ask is whether the surrogate problem (3.18) can ever yield the solution to the original problem (3.9). In the event of such an equivalence, under which conditions it occurs. An answer is given in Lemma 3.1 and Theorem 3.2 below.

Lemma 3.1. *Let $X \in \mathbb{R}^{n\times N}$ and d be an integer such that $\operatorname{rank}(X) = n$ and*

$$\max_{\substack{I\subset\mathbb{I}\\|I|=d}} \max_{\substack{\eta\in\mathbb{R}^n\\\eta\neq 0}} \frac{\|X_I^\top \eta\|_1}{\|X^\top \eta\|_1} \leq \frac{1}{2} \tag{3.19}$$

with $\mathbb{I} = \{1, \ldots, N\}$. Consider $\boldsymbol{y} \in \mathbb{R}^N$ such that the set $\{\theta \in \mathbb{R}^n : \|\boldsymbol{\phi}(\theta)\|_0 \leq d\}$, with $\boldsymbol{\phi}(\theta) =$

$\boldsymbol{y} - X^\top \theta$, *is non empty. Then*

$$\arg\min_{\theta\in\mathbb{R}^n} \|\boldsymbol{\phi}(\theta)\|_0 \subset \arg\min_{\theta\in\mathbb{R}^n} \|\boldsymbol{\phi}(\theta)\|_1 . \tag{3.20}$$

Proof. The proof is quite similar to the second part of the proof of Theorem 2.1. It is provided for completeness. Consider $\theta^\star \in \arg\min_{\theta\in\mathbb{R}^n} \|\boldsymbol{\phi}(\theta)\|_0$. Then it follows $|\mathbb{I}^c(\theta^\star)| = \|\boldsymbol{\phi}(\theta^\star)\|_0 \leq d$ with $\mathbb{I}^c(\theta^\star) = \mathbb{I} \setminus \mathbb{I}^0(\theta^\star)$. Note that the inequality in (3.19) still holds if we replace the equality $|I| = d$ by the inequality $|I| \leq d$. So, since $|\mathbb{I}^c(\theta^\star)| \leq d$, it is true that

$$\max_{\substack{\eta\in\mathbb{R}^n \\ \eta\neq 0}} \frac{\|X^\top_{\mathbb{I}^c(\theta^\star)}\eta\|_1}{\|X^\top\eta\|_1} \leq \frac{1}{2}$$

This means that for any $\eta \in \mathbb{R}^n$,

$$\|X^\top_{\mathbb{I}^c(\theta^\star)}\eta\|_1 \leq \|X^\top_{\mathbb{I}^0(\theta^\star)}\eta\|_1$$

We now make two remarks. First, since $\boldsymbol{y}_{\mathbb{I}^0(\theta^\star)} - X^\top_{\mathbb{I}^0(\theta^\star)}\theta^\star = 0$, $\|X^\top_{\mathbb{I}^0(\theta^\star)}\eta\|_1 = \|\boldsymbol{y}_{\mathbb{I}^0(\theta^\star)} - X^\top_{\mathbb{I}^0(\theta^\star)}(\theta^\star + \eta)\|_1$. Second, by using the triangle inequality, it can be observed that

$$\|\boldsymbol{y}_{\mathbb{I}^c(\theta^\star)} - X^\top_{\mathbb{I}^c(\theta^\star)}\theta^\star\|_1 - \|\boldsymbol{y}_{\mathbb{I}^c} - X^\top_{\mathbb{I}^c(\theta^\star)}(\theta^\star + \eta)\|_1 \leq \|X^\top_{\mathbb{I}^c(\theta^\star)}\eta\|_1.$$

It follows that

$$\|\boldsymbol{y}_{\mathbb{I}^c(\theta^\star)} - X^\top_{\mathbb{I}^c(\theta^\star)}\theta^\star\|_1 \leq \|\boldsymbol{y}_{\mathbb{I}^0(\theta^\star)} - X^\top_{\mathbb{I}^0(\theta^\star)}(\theta^\star+\eta)\|_1 + \|\boldsymbol{y}_{\mathbb{I}^c} - X^\top_{\mathbb{I}^c(\theta^\star)}(\theta^\star+\eta)\|_1 = \|\boldsymbol{y} - X^\top(\theta^\star+\eta)\|_1.$$

Finally adding $\|\boldsymbol{y}_{\mathbb{I}^0(\theta^\star)} - X^\top_{\mathbb{I}^0(\theta^\star)}\theta^\star\|_1$ (which is indeed equal to zero) to the left hand side member of the inequality symbol gives

$$\|\boldsymbol{\phi}(\theta^\star)\|_1 \leq \|\boldsymbol{\phi}(\theta^\star + \eta))\|_1 \quad \forall \eta \in \mathbb{R}^n.$$

Hence $\theta^\star \in \arg\min_{\theta\in\mathbb{R}^n} \|\boldsymbol{\phi}(\theta)\|_1$ hence proving the lemma. □

According to this lemma, if we let

$$\pi_1^c(X) = \max\{d : \text{ Eq. (3.19) holds}\},$$

then (3.20) holds whenever $\{\theta \in \mathbb{R}^n : \|\boldsymbol{\phi}(\theta)\|_0 \leq \pi_1^c(X)\} \neq \emptyset$. As already discussed earlier in Chapter 2, a number of the form $\pi_1^c(X)$ is expensive to evaluate numerically. So, we provide a stronger version of the lemma as follows.

Theorem 3.2 (Equivalence ℓ_0/ℓ_1)**.** *Let $X \in \mathbb{R}^{n\times N}$ be a self-decomposable matrix in the sense of Definition 2.1 and let $\xi(X)$ be defined as in (2.12). Then the following holds true:*

$$\begin{aligned} \forall \theta^\star \in \mathbb{R}^n,\ \forall \boldsymbol{y} \in \mathbb{R}^N, \|\boldsymbol{\phi}(\theta^\star)\|_0 < T(\xi(X)) \\ \Rightarrow \quad \arg\min_{\theta\in\mathbb{R}^n} \|\boldsymbol{\phi}(\theta)\|_0 = \arg\min_{\theta\in\mathbb{R}^n} \|\boldsymbol{\phi}(\theta)\|_1 = \{\theta^\star\} \end{aligned} \tag{3.21}$$

with $\boldsymbol{\phi}(\theta) = \boldsymbol{y} - X^\top\theta$.

The theorem suggests that if there is one subsystem indexed by some i that has generated

a much larger number of data than all the others, then the corresponding PV θ_i^o is the unique solution to problem (3.18).

Recoverability of the true PVs through solving a sequence of ℓ_1 problems. We now propose some conditions on the data generated by (3.2) which allow for an exact recovery of all the true PVs by convex optimization. Without loss of generality, we can assume in this subsection that the subsystems of system (3.2) are indexed in such a way that $N_1 \geq N_2 \geq \ldots \geq N_s$ with $N_i = N - \|\boldsymbol{\phi}(\theta_i^o)\|_0$ being the number of data points pertaining to subsystem i. Define $X_1 = X$, and for any $j = 2, \ldots, s$, let X_j be the matrix X_{j-1} from which all the data vectors x_t related to the subsystem $j-1$ have been deleted. This way, X_1 contains $N = N_1 + \ldots + N_s$ columns, X_2 contains $N - N_1$ columns, X_3 contains $N - N_1 - N_2$ columns and so forth. Let $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s$ be defined similarly. With these notations, we present below an immediate corollary to Theorem 2.2, which is relevant to the linear switched identification problem.

Theorem 3.3. *Consider the data $(\boldsymbol{y}, X) \in \mathbb{R}^N \times \mathbb{R}^{n \times N}$ generated by the SARX system (3.2). Let $\{N_i\}_{i=1}^s$ and $\{(\boldsymbol{y}_i, X_i)\}_{i=1}^s$ be defined as above. Assume that:*

- *For all $i = 1, \ldots, s$, each matrix X_i is self-decomposable,*
- *For all $i = 1, \ldots, s$, $N_i > N - \sum_{k=1}^{i-1} N_k - T(\xi(X_i))$.*

Then

$$\arg\min_{\theta \in \mathbb{R}^n} \left\| \boldsymbol{y}_i - X_i^\top \theta \right\|_1 = \{\theta_i^o\} \; \forall i = 1, \ldots, s.$$

i.e., all the true parameter vectors $\{\theta_1^o, \ldots, \theta_s^o\}$ can be extracted one after another by solving ℓ_1 minimization problems of the form (3.18).

To illustrate the condition of Theorem 3.3, consider an SARX system with $s = 3$ modes. Assume for example that the total number of data points collected from this SARX system is $N = 200$. For the sake of simplicity, let us assume that for any i, $T(\xi(X_i))$ is about one third of the number of columns in X_i. Then $(N_1, N_2, N_3) = (134, 45, 21)$ is an example of distribution (of the data samples per subsystem) that fulfills the condition of the theorem. Hence the conditions appear to be strong unless one has the possibility in practice to control somehow the switching signal. Note however that these conditions suffer from some degree of pessimism since they are only sufficient. As will be empirically discussed in the sequel, recovery of the PVs is still possible beyond the theoretical conditions thanks to the reweighted scheme described in Algorithm 3.1.

3.3.3 Summary of the identification algorithm

We have seen in the previous subsections that by applying Algorithm 3.1, we can identify one of the s parameter vectors of a switched system such as (3.2) from the whole dataset. If there is one submodel i satisfying $N_i > N - T(\xi(X))$ (see Theorem 2.13) then Algorithm 3.1 will find (after only one iteration) a vector θ^* in the set $\{\theta_1^o, \ldots, \theta_s^o\}$. If this condition is not fulfilled, Algorithm 3.1 may not converge towards a point in $\{\theta_1^o, \ldots, \theta_s^o\}$. However, as argued in [21, 1] and suggested by different experiments reported therein, the algorithm is likely to find the vector θ^* that realizes the sparsest error $\boldsymbol{\phi}(\theta)$. According to Proposition 3.1 and Theorem 3.1, such a point θ^* is in $\{\theta_1^o, \ldots, \theta_s^o\}$ when enough rich data are available.
Without loss of generality, we will denote with $\hat{\theta}_1$, i.e. the estimate of θ_1^o, the point of $\{\theta_1^o, \ldots, \theta_s^o\}$ to which the algorithm converges when it is run over all the data. Observe that $\hat{\theta}_1$ can be obtained from the whole mixed data without any prior clustering. Given $\hat{\theta}_1$, we need now to estimate

Algorithm 3.1 Reweighted ℓ_1 minimization

Inputs: Data $\{(x_t, y_t)\}_{t=1}^N$

Initialization: Set the initial weights as: $w_t^{(0)} = 1,\ t = 1, \dots, N$ and $W^{(0)} = \text{diag}(w_1^{(0)}, \dots, w_N^{(0)})$; Initialize a counter, $r \leftarrow 0$.

Repeat

1. Solve the convex problem
$$\theta^{(r)} = \arg\min_{\theta \in \mathbb{R}^n} \left\| W^{(r)} \boldsymbol{\phi}^o(\theta) \right\|_1$$
where $\boldsymbol{\phi}^o(\theta)$ is an ℓ_2-normalized version of $\boldsymbol{\phi}(\theta)$ defined as
$$\boldsymbol{\phi}^o(\theta) = \begin{bmatrix} \dfrac{\bar{x}_1^\top \bar{\theta}}{\|\bar{x}_1\|_2} & \dots & \dfrac{\bar{x}_N^\top \bar{\theta}}{\|\bar{x}_N\|_2} \end{bmatrix}^\top.$$
with $\bar{x}_t$ and $\bar{\theta}$ defined as in (3.7).
2. Update the weights as
$$w_t^{(r+1)} = \frac{1}{|\boldsymbol{\phi}_t^o(\theta^{(r)})| + \varepsilon},\ t = 1, \dots, N$$
with $\boldsymbol{\phi}_t^o(\theta^{(r)})$ denoting the t-th entry of the vector $\boldsymbol{\phi}^o(\theta^{(r)})$.
3. $r \leftarrow r + 1$

Until r attains a pre-specified maximum number of iterations $r_{\max}$ or until convergence (for example when $\|\theta^{(r)} - \theta^{(r-1)}\|_2 < \text{Tol}$, where $r > 2$ and Tol is a threshold).

Return $\theta^{(r)}$

the rest of the PVs. However we cannot proceed this time with the whole dataset because the algorithm may still converge to the same PV θ_1. Therefore it is preferable to remove the data generated by that submodel. The indices of such data can be determined as

$$I(\hat{\theta}_1) = \left\{ t \in \{1, \dots, N\} : \frac{|\bar{x}_t^\top \hat{\bar{\theta}}_1|}{\|\bar{x}_t\|_2 \cdot \|\hat{\bar{\theta}}_1\|_2} \leq \text{Thres} \right\} \tag{3.22}$$

where it is assumed that Tresh $\in [0,\ 1]$ is a tolerance threshold and $\hat{\bar{\theta}}_1 = [1 \quad \hat{\theta}_1^\top]^\top$. From the data indexed by $I \setminus I(\hat{\theta}_1)$, we estimate θ_2. We can repeat this procedure until all the PVs are identified. A pseudo-code of the method is summarized in Algorithm 3.2.

3.4 Uncertainty sets induced by noise

Consider now the more realistic situation where the dense noise sequence $\{e_t\}$ in (3.2) is nonzero but is bounded. In this case, the identification process is unlikely to return the true parameter vectors irrespective of whether the conditions[3] of Theorem 3.3 hold or not. Instead, each PV estimate will come up with an associated uncertainty set. This is typically due to the fact that the dense noise sequence is only known to be bounded.

[3] In the presence of noise the number of data points generated by a subsystem i can be defined as $N_i = |\mathbb{I}^\varepsilon(\theta_i^o)|$.

Algorithm 3.2 Identification of all PVs

1. **Inputs:** $\{(x_t, y_t)\}_{t=1}^N$
2. **Initialization:** $\mathcal{S} \leftarrow \emptyset$, $J \leftarrow \{1, \ldots, N\}$
3. **While** $|J| \neq 0$
   - Estimate a submodel by the reweighted ℓ_1 minimization method (See Algorithm 3.1) based on the data whose indices are contained in J
   - Record the identified PV: $\mathcal{S} \leftarrow \mathcal{S} \cup \{\theta\}$
   - Remove from J indices of data generated by the submodel obtained:

     $$J \leftarrow J \setminus (J \cap I(\theta)),$$

     with $I(\theta)$ defined as in Eq. (3.22).
4. **EndWhile**
5. **Return** $\mathcal{S}$ and $s = |\mathcal{S}|$.

3.4.1 A theoretical characterization of the uncertainty

Let $\varepsilon \geq 0$ be such that $|e_t| \leq \varepsilon$ for all $t = 1, \ldots, N$. Assume that the conditions of Theorem 3.3 are satisfied. Denote with $\hat{\theta}_i$ the estimate (by the approach discussed earlier) of θ_i^o, $i = 1, \ldots, s$ that is, $\hat{\theta}_i \in \arg\min_{\theta \in \mathbb{R}^n} \|\boldsymbol{y_i} - X_i^\top \theta\|_1$ where $\boldsymbol{y_i}$ and X_i are defined as in Section 3.3.2. Then according to Corollary 2.2,

$$\|\hat{\theta}_i - \theta_i^o\|_2 \leq r_i \varepsilon \tag{3.23}$$

where

$$r_i = \frac{2N_i}{\sigma_{1,2}(X_i)\Big[1 - \dfrac{N - \sum_{k=1}^i N_k}{T(\xi(X_i))}\Big]},$$

$$\sigma_{1,2}(X_i) = \inf_{\eta \neq 0} \frac{\left\|X_i^\top \eta\right\|_1}{\|\eta\|_2}$$

This means that for all $i = 1, \ldots, s$, the estimate $\hat{\theta}_i$ lies in the ball centered at θ_i^o and having a radius of $r_i\varepsilon$. The size of these balls increases naturally with the magnitude of the noise. Note that r_i is decreasing as a function of the proportion $N_i / \sum_{k=1}^{i-1} N_k$ of data points generated by the subsystem i. Hence the parametric error relative to submodel i is all the smaller as the number of data points pertaining to this mode is large. To avoid any ambiguity, one may require that the s uncertainty balls defined around the different PVs do not intersect, a requirement which translates into the condition $\|\theta_i^o - \theta_j^o\|_2 > (r_i + r_j)\varepsilon$ for all pairs $(i, j) \in \mathbb{S}^2$ with $i \neq j$. In other words, the true PVs θ_i^o must be distinguishable enough.

A discussion on a noise-aware convex formulation. Problem (3.18) or its reweighted versions as formulated in Algorithm 3.1 do not take the noise explicitly into account. An

alternative to this would be to consider a convex relaxation of (3.13) in the form

$$\begin{aligned} \min_{(\theta,\xi)\in\mathbb{R}^n\times\mathbb{R}^N_+} & \|\xi\|_1 \\ \text{s.t.} \quad & |y_t - x_t^\top\theta| \leq \varepsilon + \xi_t,\ t = 1,\dots,N. \end{aligned} \tag{3.24}$$

Noting that this is equivalent to

$$\min_{\theta\in\mathbb{R}^n} \varphi(\boldsymbol{y}^\top - \theta^\top X)$$

with φ defined by $\varphi(A) = \sum_{t=1}^N \max(0, |a_i| - \varepsilon)$ for any matrix $A = [a_1 \quad \cdots \quad a_N] \in \mathbb{R}^{1\times N}$, we can apply Theorem 2.3 to find a bound on the estimation error associated with (3.24). This leads to a bound of the form (2.22). It then turns out that applying Corollary 2.2 leads to the same bound as the one in (3.23).

Remark 3.1. *The sparse optimization approach presented here for a MISO switched system* (3.2) *can be extended to MIMO systems. We have discussed such an extension in, for example, [8] for switched state-space models with measured state. Note that the analysis of Chapter 2 is readily applicable to this estimator. In case the continuous state is not measured, the identification problem is far more challenging.*

3.5 Applications

In this section, we apply the proposed identification algorithm to a SISO SARX model composed of three linear submodels of order two. The SARX model is defined by

$$y_t = x_t^\top \theta^o_{\sigma(t)} + e_t \tag{3.25}$$

with $x_t = [y_{t-1} \quad y_{t-2} \quad u_{t-1} \quad u_{t-2}]^\top$, $\sigma(t) \in \{1, 2, 3\}$ and

$$\begin{aligned} \theta^o_1 &= [-0.40 \quad 0.25 \quad -0.15 \quad 0.08]^\top, \\ \theta^o_2 &= [1.55 \quad -0.58 \quad -2.10 \quad 0.96]^\top, \\ \theta^o_3 &= [1 \quad -0.24 \quad -0.65 \quad 0.30]^\top. \end{aligned} \tag{3.26}$$

Using this switched model, we generate the identification data under the following conditions:

- The excitation input $\{u_t\}$ is a centered signal with normal distribution and variance unity.
- The noise $\{e_t\}$ is a white Gaussian noise whose magnitude is such that the Signal to Noise Ratio (SNR) is equal to 30 dB with respect to the output signal.
- The switching sequence $\{\sigma(t)\}$ is uniformly distributed in $\{1, 2, 3\}$.

3.5.1 Performance of the (reweighted) ℓ_1 relaxation

In a first experiment, we test the ability of the ℓ_1-relaxation to exactly recover the solution of the sparse optimization problem. With regard to this goal we can set the noise sequence $\{e_t\}$ to be identically null. We assign a fixed value to the sparsity of the error $\boldsymbol{\phi}(\theta^o_3)$ (expressed in terms of the number $\|\boldsymbol{\phi}(\theta^o_3)\|_0$ of nonzero components in $\boldsymbol{\phi}(\theta^o_3)$) and then solve problem (3.18) 100 times on different independent simulations of input-output data of length $N = 100$ each.

This procedure is repeated for different values of $\|\boldsymbol{\phi}(\theta_3^o)\|_0$ reported in Table 3.1. In this table we display for each given value of $\|\boldsymbol{\phi}(\theta_3^o)\|_0$, the percentage of successes in attempting to compute the solution of (3.9) by solving (3.18). Since the data (x_t, y_t), $t = 1, \ldots, N$, is generated from (3.25) with white noise as input and uniformly distributed discrete mode, it holds with overwhelming probability that the columns of X are in general position. Consequently, it can be reasonably assumed that $\nu_n(X) = n$, i.e., $\nu_n(X)$ is equal to the dimension of x_t. It can be observed from the results of Table 3.1 that (3.18) effectively solves (3.9) successfully with a score of 100% over 100 trials (on randomly generated data) once $\|\boldsymbol{\phi}(\theta_3^o)\|_0$ falls under 48.

$\|\boldsymbol{\phi}(\theta_3^o)\|_0$	58%	55%	53%	50%	48%	45%
# succ.	46%	76%	94%	99%	100%	100%

Table 3.1: Equivalence between ℓ_0 and ℓ_1 minimizations versus the sparsity of $\boldsymbol{\phi}(\theta_3^o)$. The ℓ_0 norm of $\boldsymbol{\phi}(\theta_3^o)$ is expressed as a fraction of the nonzero entries over the total length of the vector $\boldsymbol{\phi}(\theta_3^o)$.

Now we propose, in the same conditions as the first experiment, to approach the solution of the sparse optimization problem (3.9) by instead running the reweighted ℓ_1 optimization technique described in Algorithm 3.1. The related results are presented in Table 3.2. It turns out that the reweighted ℓ_1 optimization approach significantly improves the ℓ_1-relaxation. When $\|\boldsymbol{\phi}(\theta_3^o)\|_0$ starts getting much larger than $\|\boldsymbol{\phi}(\theta_1^o)\|_0$ and $\|\boldsymbol{\phi}(\theta_2^o)\|_0$, Algorithm 3.1 may not keep on converging towards θ_3^o any longer. Instead, it is likely to converge towards θ_1 or θ_2 since $\|\boldsymbol{\phi}(\theta_1)\|_0$ and $\|\boldsymbol{\phi}(\theta_2)\|_0$ decrease as $\|\boldsymbol{\phi}(\theta_3^o)\|_0$ increases.

$\|\boldsymbol{\phi}(\theta_3^o)\|_0$	58%	55%	53%	50%	48%	45%
# succ.	94%	100%	100%	100%	100%	100%

Table 3.2: Approximation of ℓ_0 by reweighted ℓ_1 minimization versus the sparsity of $\boldsymbol{\phi}(\theta_3^o)$. The ℓ_0 norm of $\boldsymbol{\phi}(\theta_3^o)$ is expressed as a fraction of the nonzero entries over the total length of the vector $\boldsymbol{\phi}(\theta_3^o)$.

3.5.2 Identification of the PVs

The second objective of the numerical experiments is to test the statistical robustness of the identification algorithm. For this purpose, we use 100 different independent realizations of the input, the discrete state and the output noise (SNR=30 dB) to generate 100 data sequences of length $N = 600$ each. The identification algorithm (Algorithm 2 indeed) is then run on each of these different 100 data sequences. At each run, the first 300 points are used to identify a model and the whole sequence of length 600 is used to validate the estimated model, i.e., to verify its ability to reconstruct the system output from the true input and an estimated discrete state. The performance is measured in term of the FIT criterion [38]

$$\text{FIT} = \left(1 - \frac{\|\hat{\boldsymbol{y}} - \boldsymbol{y}\|_2}{\|\boldsymbol{y} - \bar{y}\mathbf{1}_N\|_2}\right) \times 100\% \tag{3.27}$$

which measures the fitting error between the true output sequence $\boldsymbol{y}$ and the estimated model output sequence $\hat{\boldsymbol{y}}$. In this formula, $\bar{y}$ stands for the mean of the true output sequence and $\mathbf{1}_N$

$$\hat{\theta}_1 = \begin{bmatrix} -0.3914 & \pm & 0.0115 \\ 0.2452 & \pm & 0.0106 \\ -0.1666 & \pm & 0.0201 \\ 0.0875 & \pm & 0.0200 \end{bmatrix}, \hat{\theta}_2 = \begin{bmatrix} 1.5360 & \pm & 0.0549 \\ -0.5706 & \pm & 0.0337 \\ -2.0680 & \pm & 0.1421 \\ 0.9434 & \pm & 0.0728 \end{bmatrix}, \hat{\theta}_3 = \begin{bmatrix} 0.9909 & \pm & 0.0128 \\ -0.2365 & \pm & 0.0124 \\ -0.6727 & \pm & 0.0263 \\ 0.3102 & \pm & 0.0271 \end{bmatrix}$$

(a) – Average estimates over 100 independent runs of the identification algorithm: $\varepsilon = 0.1$, $\text{Tol} = 10^{-3}$, $\text{Thres} = 0.05$.

$$\hat{\theta}_1^{LS} = \begin{bmatrix} -0.3989 & \pm & 0.0044 \\ 0.2490 & \pm & 0.0042 \\ -0.1511 & \pm & 0.0107 \\ 0.0829 & \pm & 0.0122 \end{bmatrix}, \hat{\theta}_2^{LS} = \begin{bmatrix} 1.5458 & \pm & 0.0069 \\ -0.5769 & \pm & 0.0071 \\ -2.0978 & \pm & 0.0167 \\ 0.9543 & \pm & 0.0174 \end{bmatrix}, \hat{\theta}_3^{LS} = \begin{bmatrix} 0.9974 & \pm & 0.0060 \\ -0.2391 & \pm & 0.0062 \\ -0.6493 & \pm & 0.0129 \\ 0.2961 & \pm & 0.0137 \end{bmatrix}$$

(b) – Least squares average estimates if the discrete state were known.

Table 3.3: Comparison of the proposed identification algorithm to standard least squares (if the discrete state were known) over 100 independent runs: $\varepsilon = 0.1$, $\text{Tol} = 10^{-3}$, $\text{Thres} = 0.05$.

is an N-dimensional vector with all entries equal to one.

For simplicity, it is assumed that the number of submodels is fed into the identification algorithm.[4] We present in Table 3.3 the average values of the estimated PVs together with their standard deviations over 100 independent runs of the algorithm. Along with those results are provided, for comparison purpose, the PVs' estimates the standard least squares would yield if the discrete mode sequence were fully known. By comparing the averaged estimates $\{\hat{\theta}_i\}_{i=1}^s$ of the PVs displayed in Table 3.3 to the true values $\{\theta_i\}_{i=1}^s$ given in (3.26), we can see that the proposed algorithm has effectively recovered the true PVs with a relatively good precision despite the presence of noise. Moreover, by judging from the standard deviations, we are prompted to conclude that the algorithm performs well. Of course, when the data is noise-free, the parameters are exactly recovered by the algorithm.
Figure 3.2 represents the distribution of the FIT over 100 runs of the identification algorithm on independent input-output data. This plot shows that most of the runs of the algorithm yield a FIT greater than 90%. In fact 98% of the runs produce a FIT measure larger than 87% (which means 100% if there were no noise in the data) on both identification and validation data. It can therefore be concluded that 98% of the runs yield the correct PVs.

3.6 Conclusion

We have presented in this chapter a summary of our contributions to switched ARX system identification. Before closing this part we make the following remarks:

- The discussion has started by a sparse optimization paradigm which is later shown to be relaxable to a convex nonsmooth problem. The latter formulation in addition of being easier to solve, enjoys some sparsity-inducing properties. An important feature of the proposed method is that, contrary to the majority of hybrid system identification methods, it can be theoretically analyzed. For example, under some worse-case conditions pertaining to the proportion of data generated by each subsystem, it is proved that the parametric estimation error is stable and bounded.

[4] Note that the structure of the identification algorithm allows for the estimation of the number of submodels (see [1] for more details).

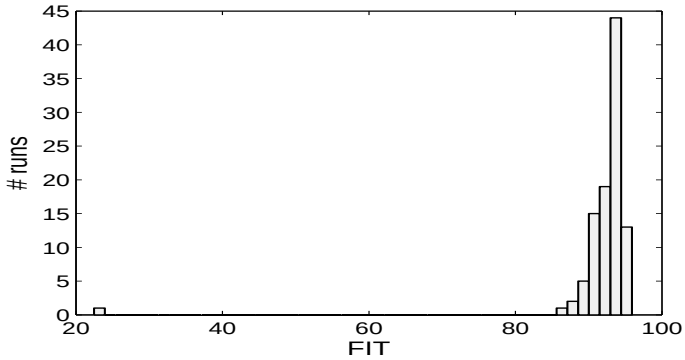


Figure 3.2: Distribution of the FIT measure over the 100 independent runs of the algorithm. $\varepsilon = 0.1$, Tol $= 10^{-3}$, Thres $= 0.05$.

- Note that the number of subsystems need not be known a priori here. Also the discussed sparse optimization approach can be extended to switched nonlinear systems in which each subsystem is nonlinear and described as a weighted expansion over a dictionary of basis functions, see e.g. [5].
- A weakness of the method however is that the (sufficient) correctness conditions might be frequently violated in practice. Fortunately, different conclusive experiments on simulated data tend to show that, thanks to the ℓ_1-reweighted scheme (see Algorithm 3.1), the performance of the method can be boosted far more beyond the established theoretical conditions. Another problem is related to the one-by-one extraction of the PVs which imposes that a threshold be defined for removing data at each step of the identification process (see (3.22)). Without prior knowledge concerning the magnitude of the noise, selecting this threshold may become increasingly challenging as the level of noise gets higher. One way to overcome this problem would be to estimate all PVs simultaneously. For this purpose, a procedure like the one described in Section 4.2.3 can be applied.

Chapter 4

Identification of piecewise affine systems

Contents

This chapter is mainly based on our papers [5, 6, 4].

4.1 General idea of nonlinear system modeling

A general approach for modeling smooth nonlinear systems consists in expanding the nonlinearity over an appropriate dictionary of elementary (known) basis functions. The resulting model structure has the advantage, from an estimation perspective, of being linear with respect to the parameters. The estimation can then be efficiently addressed through solving a convex optimization problem. For an overview on black-box nonlinear modeling we refer to the survey papers [58, 35] and also to some more recent works such as e.g., [50, 61, 52]. Examples of frequently used basis functions include polynomials, wavelets, radial basis functions (RBF), kernel functions, linear functions.

4.1.1 Expansion of nonlinearity on basis functions

We first consider the problem of identifying a nonlinear system represented by a Nonlinear AutoRegressive eXogenous (NARX) model of the form

$$y_t = f(x_t) + e_t \tag{4.1}$$

with output $y_t \in \mathbb{R}$, input $u_t \in \mathbb{R}^{n_u}$ and regressor vector

$$x_t = \begin{bmatrix} y_{t-1} & \cdots & y_{t-n_a} & u_t^\top & u_{t-1}^\top & \ldots & u_{t-n_b}^\top \end{bmatrix}^\top, \tag{4.2}$$

where the integers n_a and n_b are called the orders of the NARX model. The function f is assumed to be a smooth nonlinear function defined on a bounded set $\mathscr{X} \subset \mathbb{R}^n$, with $n = n_a + (n_b+1)n_u$; it is completely unknown. $\{e_t\}$ is a sequence of modeling errors or noise terms; $\{x_t\}$ is measured. Note that from an estimation viewpoint, the vector x_t need not be structured as in Eq. (4.2). It can be unstructured or take a very arbitrary form. Therefore the discussions to follow apply to static nonlinear regression as well.

Given a finite collection $\{(y_t, x_t)\}_{t=1}^N$ of data generated by a system of the form (4.1), the nonlinear system identification problem aims at finding a model which explains best the data. Finding a solution to this problem typically requires performing two steps: (1) choose a model structure which consists of a family of parametrized models of a certain form; (2) instantiate the parameters of that model structure by optimizing a performance function defined on the available measurements. In the conventional nonlinear approximation framework, the nonlinear function f is approximated with a model of the form

$$\hat{f}(x) = \sum_{i=1}^{s} w_i \varphi_i(x) \tag{4.3}$$

where the functions $\varphi_i : \mathbb{R}^n \to \mathbb{R}$, $i = 1, \ldots, s$, are some known basis functions and $\{w_i\}_{i=1}^s$ are the associated parameters (also called weights). Although many choices are possible for the basis functions[1], a widely adopted choice in practice is the basis of Gaussian radial basis functions (RBF) which are of the form

$$\varphi_i(x) = \exp\Big(-\frac{\|c_i - x\|^2}{2\sigma_i^2}\Big) \tag{4.4}$$

with $\sigma_i > 0$ being the so-called widths of the RBFs and the c_i are called their centers. A very appealing feature of model (4.3) is that if the basis functions φ_i's are assumed to be entirely known, then $\hat{f}(x)$ depends linearly on the weights w_i's. This eases considerably the estimation step. For example, it follows naturally from the linear structure that the associated least squares criterion (on the difference between measured output and model prediction) is convex. However since s is generally finite in practice, the approximation capability of the model (4.3) depends strongly on the choice of the parameters $\{(c_i, \sigma_i)\}$. The selection of these parameters is hard to optimize since finding both the weights and the parameters of the basis functions is a nonconvex program. One way to ease manual tuning, is to choose the σ_i's to be all equal and use directly the training regression data $\{x_t\}_{t=1}^N$ as the centers. As a consequence, we obtain in model (4.3) as many basis functions as data samples with $\varphi_t = k(x_t, \cdot)$ and $k(\cdot, \cdot)$ denoting the Gaussian kernel function [55] defined by $k(x, y) = \exp(-\|x - y\|_2^2/(2\sigma^2))$. The model (4.3) is then said to be a non parametric model because its structure depends directly on the training dataset. By letting

$$\varphi(x) = \begin{bmatrix} \varphi_1(x) & \cdots & \varphi_N(x) \end{bmatrix}^\top \in \mathbb{R}^N \tag{4.5}$$

and

$$\theta = \begin{bmatrix} w_1 & \cdots & w_N \end{bmatrix}^\top. \tag{4.6}$$

[1] Other possibilities include polynomials, Fourier basis functions, general radial basis functions, wavelets, ...

one can interpret model (4.3) as a linear model obtained by lifting the original n-dimensional regression space to a high (and possibly infinite) dimensional space. The mapping $\varphi : \mathbb{R}^n \to \mathbb{R}^N$ is then the transformation map. In this space, the model is conceptually linear with respect to the parameter vector θ to be estimated. This is more apparent by rewriting (4.1) as

$$y_t = \theta^\top \varphi(x_t) + \mu(x_t) + e_t \tag{4.7}$$

where $\mu(x) = f(x) - \hat{f}(x)$ refers to model mismatch accounting for error incurred by the approximation of f with (4.3). It is to be noticed that some of the basis components of the model (4.7) may actually be redundant or irrelevant. For this reason one can require that the weight vector θ being estimated be as sparse as possible i.e., θ must have most of its entries equal to zero. This sparsity requirement on the weights θ presents the advantage of reducing the complexity of the identified model while enhancing its generalization capacity.

4.1.2 Piecewise affine models for nonlinear systems

The general nonlinear model (4.3) can achieve great accuracy if its parameters are appropriately selected. This comes however at the price of some structural complexity as the number of basis functions can be large. As a result the model can be difficult to exploit in practice especially for control design. A simpler model can be obtained by considering the special case where

- the basis functions $\{\varphi_i\}$ are linear/affine with respect to x and are defined by $\varphi_i(x) = a_i^\top x + b_i$ with $(a_i, b_i) \in \mathbb{R}^n \times \mathbb{R}$
- the weights $w_i : \mathscr{X} \to \{0, 1\}$ are functions of x which take binary values
- for any $x \in \mathscr{X}$, $\sum_{i=1}^s w_i(x) = 1$ that is, only one weight is nonzero for a given x
- the sets $\{\mathscr{X}_i\}_{i=1}^s$ defined by $\mathscr{X}_i = \{x \in \mathscr{X} : w_i(x) = 1\}$ form a complete partition of $\mathscr{X}$.

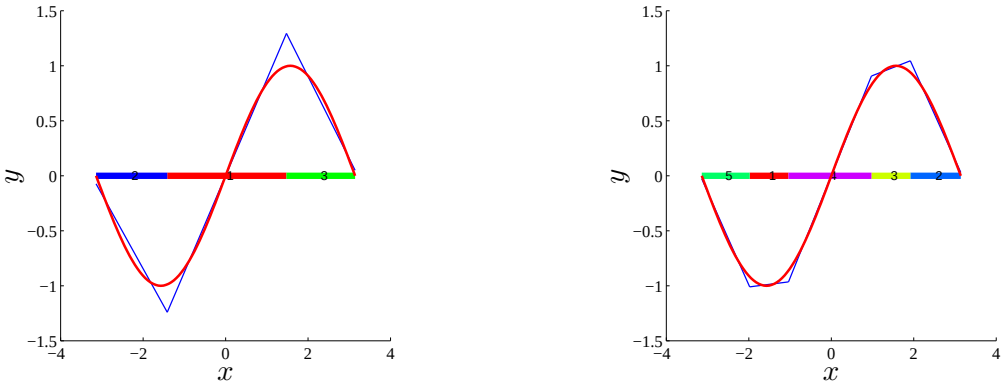
Under the above listed conditions, Eq. (4.3) reads as

$$\hat{f}(x) = \sum_{i=1}^{s} w_i(x)(a_i^\top x + b_i) = a_j^\top x + b_j \quad \text{if } w_j(x) = 1 \tag{4.8}$$

Hence with the above configuration, the nonlinear system is modeled by a finite number of linear/affine submodels, each of which is related to a region of the system operating space. For illustration purpose, an example of such a piecewise affine modeling of a static sinus function is depicted in Figure 4.1. The main advantage in doing so is that linear/affine models are simpler than general nonlinear models and there exists an abundant theory for controlling and analyzing them. One can therefore reasonably hope for a somewhat easier extension of the available expertise in linear system theory to nonlinear systems through the PWA modeling.

Indeed it can be shown via simple arguments from differential calculus that any smooth nonlinear system can be approximated to an arbitrary precision using a PWA map. This property is known as a universal approximation capability [60, 17] of PWA systems.
To see this assume that the nonlinear function f in (4.1) is continuously differentiable ($f \in \mathcal{C}^1$). Consider s nominal points $\{c_i\}_{i=1}^s$ in $\mathscr{X}$. Then the first order Taylor series expansion of f in the neighborhood of c_i is given as follows

$$f(x) = f(c_i) + (x - c_i)^\top \nabla_x f(c_i) + \varepsilon(c_i, x), \tag{4.9}$$



(a) PWA with 3 submodels (in blue)

(b) PWA with 5 submodels (in blue)

Figure 4.1: Illustration of PWA approximation of a static nonlinear function on the interval $[-\pi,\ \pi]$ of $\mathbb{R}$. This example is generated with the HIT toolbox [27].

where $\nabla_x f(c_i)$ is the gradient of f at c_i and $\varepsilon(c_i, x)$ accounts for the higher order terms and satisfies $\lim_{x\to c_i} \varepsilon(c_i, x) = 0$. When x lies in a neighboring ball $B(c_i, \delta_i) = \{x \in \mathscr{X} : \|x - c_i\| \leq \delta_i\}$ of c_i (with $\delta_i > 0$ small), the nonlinear term $\varepsilon(c_i, x)$ can be considered negligible so that f can be approximated by the affine part of (4.9). Following this idea, $\mathscr{X}$ can be decomposed into a set of a large enough number s of regions $B(c_i, \delta_i)$, $i = 1, \ldots, s$ on which an affine approximation holds. Then, by denoting

$$a_i = \nabla f(c_i) \quad \text{and} \quad b_i = f(c_i) - c_i^\top \nabla f(c_i), \tag{4.10}$$

the function $\hat{f}$ defined by

$$\hat{f}(x) = a_i^\top x + b_i \quad \text{if} \quad x \in B(c_i, \delta_i) \tag{4.11}$$

approaches f. As such the function $\hat{f}$ is not well defined as the set of balls $B(c_i, \delta_i)$, $i = 1, \ldots, s$, is not necessarily a disjoint cover of $\mathscr{X}$. To overcome this problem we can redefine δ_i as a function of x in the form $\delta_i(x) = \min_{j=1,\ldots,s:j\neq i} \|x - c_j\|_2$ so that the euclidean ball $B(c_i, \delta_i)$ becomes

$$\mathscr{X}_i = \left\{x \in \mathscr{X} : \forall j = 1, \ldots, s,\ \|x - c_i\|_2 \leq \|x - c_j\|_2\right\}. \tag{4.12}$$

With this rearrangement, we have $\cup_{i=1}^s \mathscr{X}_i = \mathscr{X}$ and $\text{int}(\mathscr{X}_i) \cap \text{int}(\mathscr{X}_j) = \emptyset \forall i \neq j$, where $\text{int}(\mathscr{X}_i)$ refers to the interior of $\mathscr{X}_i$. One can easily recognize that $\mathscr{X}_i$ corresponds indeed to a polyhedron (which is more common in the definition of PWA models [28]) of the form

$$\mathscr{X}_i = \{x \in \mathscr{X} : H_i x \preceq h_i\}, \tag{4.13}$$

where

$$\begin{aligned} H_i &= \begin{bmatrix} c_1 - c_i & \cdots & c_{i-1} - c_i & c_{i+1} - c_i & \cdots & c_s - c_i \end{bmatrix}^\top \\ h_i &= \begin{bmatrix} \beta_{1,i} & \cdots & \beta_{i-1,i} & \beta_{i+1,i} & \cdots & \beta_{s,i} \end{bmatrix}^\top, \end{aligned} \tag{4.14}$$

with the $\beta_{j,i}$ being defined as $\beta_{j,i} = \frac{1}{2}(c_j^\top c_j - c_i^\top c_i)$. The so-defined partition $\{\mathscr{X}_i\}_{i=1}^s$ is called a Voronoi partition associated with the centers (or seeds) $\{c_i\}_{i=1}^s$. An example of such a partition for a two-dimensional subset is depicted in Figure 4.2. Then $\hat{f}$ is modified to be

$$\hat{f}(x) = \theta_i^\top \begin{bmatrix} x \\ 1 \end{bmatrix} \quad \text{if} \quad x \in \mathscr{X}_i. \tag{4.15}$$

Thus, Eq. (4.15) represents a PWA approximation of the nonlinear continuous system defined in (4.1). Note that in the definition (4.15), ambiguity may still occur on the boundaries of $\mathscr{X}_j$. This issue can be avoided by arbitrarily assigning the points that lie on the common boundaries of any two different sets $\mathscr{X}_i$ and $\mathscr{X}_j$, to the one with the smallest index.

Theoretical approximation error. It can be shown that the approximation error induced by the approximation of a nonlinear smooth function is bounded. To see this assume that $\mathscr{X}$ is a compact set. Then by assuming further that $f \in \mathcal{C}^2$ and denoting with $\nabla^2 f(x) \in \mathbb{R}^{n\times n}$ the Hessian matrix of f at any point $x \in \mathscr{X}$, $\left\|\nabla^2 f(x)\right\|_2$ is upper bounded over $\mathscr{X}$. It follows from the Taylor theorem (for multivariable functions) that

$$\left|f(x) - f(c) - \nabla f(c)^\top (x-c)\right| \leq M \left\|x-c\right\|_2^2 \quad \forall\, x \in \mathscr{X}, \tag{4.16}$$

where

$$M = \sup_{x\in\mathscr{X}} \left\|\nabla^2 f(x) \odot W\right\|_2. \tag{4.17}$$

Here, $\|\cdot\|_2$ stands for matrix 2-norm (also called the spectral norm), $\odot$ refers to the Hadamard product (elementwise matrix product) and $W \in \mathbb{R}^{n\times n}$ is a constant matrix defined by $W_{ij} = 1$ if $i \neq j$ and $W_{ij} = 1/2$ if $i = j$. In a particular cell $\mathscr{X}_i$, the inequality (4.16) becomes

$$\left|f(x) - a_i^\top x - b_i\right| \leq M_i r_i^2 \quad \forall x \in \mathscr{X}_i$$

with $r_i = \max_{x\in\mathscr{X}_i} \left\|x - c_i\right\|_2$ and M_i defined as in (4.17) with $\mathscr{X}$ replaced by the euclidean ball $B(c_i, r_i)$. M_i is an intrinsic characteristic of the function f which reflects its smoothness on the cell $\mathscr{X}_i$. As for the number r_i, it characterizes the size of $\mathscr{X}_i$ as a neighborhood of c_i.

The discussion of this section has shown that, through some modeling abstraction, a quite large class of smooth nonlinear systems can be viewed as PWA systems. It can indeed be shown that a much larger class of nonlinear systems including those where the function f is discontinuous can be represented as PWA systems. In the next section we will formally present the class of (strictly) piecewise affine ARX systems.

4.1.3 Piecewise Affine Systems

A PieceWise Affine ARX system is a dynamical system described by an equation of the form

$$y_t = f_{\text{PWA}}(x_t) + e_t \tag{4.18}$$

where $y_t \in \mathbb{R}$ is the output of the system, e_t is accounting for potential noise or model mismatch and $x_t \in \mathbb{R}^n$ is the regressor vector defined as in (4.2). The map $f_{\text{PWA}} : \mathscr{X} \to \mathbb{R}$ is a piecewise

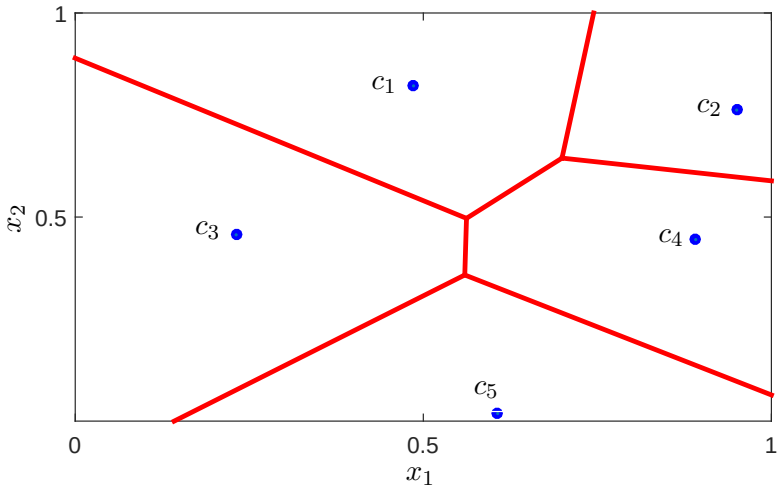


Figure 4.2: Example of Voronoi-type partition of $[0\ 1] \times [0\ 1] \subset \mathbb{R}^2$.

affine function defined by

$$f_{\text{PWA}}(x) = \begin{cases} a_1^\top x + b_1 & \text{if} \quad x \in \mathscr{X}_1 \\ \vdots & \vdots \\ a_s^\top x + b_s & \text{if} \quad x \in \mathscr{X}_s \end{cases} \tag{4.19}$$

where the $\{\mathscr{X}_i\}_{i=1}^s$ are convex polyhedra which form a partition of the regression space $\mathscr{X} \subset \mathbb{R}^n$ i.e., $\mathscr{X} = \cup_{i=1}^s \mathscr{X}_i$, $\text{int}(\mathscr{X}_i) \cap \text{int}(\mathscr{X}_j) = \emptyset$ for $i \neq j$. As a polyhedron, each $\mathscr{X}_i$ is defined by

$$\mathscr{X}_i = \{x \in \mathbb{R}^n : H_i x \preceq h_i\} \tag{4.20}$$

for some $H_i \in \mathbb{R}^{d_i \times n}$ and $h_i \in \mathbb{R}^{d_i}$ with d_i denoting the number of hyperplanes delimiting the polyhedron and $\preceq$ standing for component-wise inequality. For convenience, we may write

$$y_t = \tilde{x}_t^\top \theta^o_{\sigma(t)} + e_t \tag{4.21}$$

with $\tilde{x}_t = [x_t^\top \quad 1]^\top$, $\theta_i^o = [a_i^\top \quad b_i]^\top$ and $\sigma(t)$ being the discrete state defined by $\sigma(t) = i \in \{1, \ldots, s\}$ if $x_t \in \mathscr{X}_i$.

On the difference between PWARX and SARX models. It is worth observing the following from the definition of the PWA map:

(i) The PVs θ_i are not necessarily all distinct. An example of situation where two PVs are equal is represented in Figure 4.3. In this case, the pairs $(\theta_1, \mathscr{X}_1)$ and $(\theta_4, \mathscr{X}_4)$ of the PWA map differ only in their validity regions.

(ii) It may happen that some pairs (x_t, y_t) fit more than one different affine submodels. That is, it is possible (if noise is assumed nonexistent) to have $y_t = \theta_i^\top \tilde{x}_t = \theta_j^\top \tilde{x}_t$, while $x_t \in \mathscr{X}_i$, $x_t \notin \mathscr{X}_j$ and $\theta_i \neq \theta_j$.

If one disregards the correspondence between the regions and the affine subsystems, then a PWA system is structurally equivalent to a switched system where, by definition, the nature of the switched mechanism is unspecified. One tiny difference is that case (ii) above can occur in a

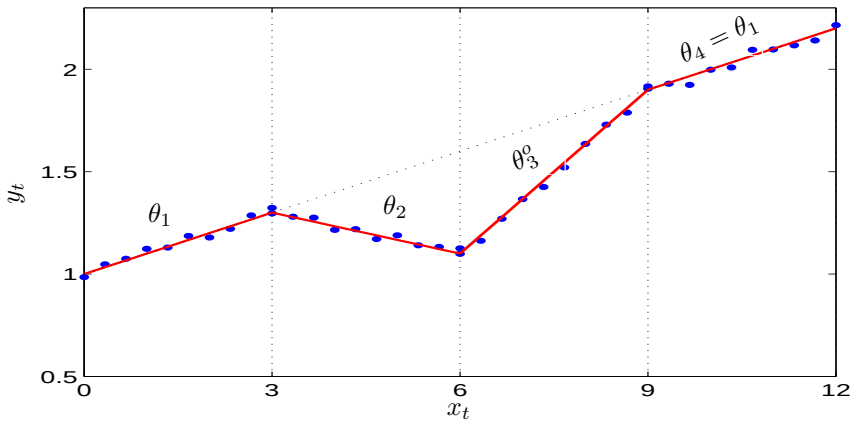


Figure 4.3: One-dimensional example of a PWA map where two PVs are equal.

switched system but not case (i). However if one concentrates only on the identification of the PVs, a method which is designed for switched systems is readily applicable to PWA as well. In the latter case, information about local linearity can be exploited in order to ease the estimation process.

4.2 Identification of PWA systems

Given observations $\{(y_t, x_t)\}_{t=1}^N$ generated by a PWA model of the form (4.18), with x_t defined as in (4.2), we are interested here in estimating the parameter vectors $\{\theta_i\}_{i=1}^s$ together with the associated validity regions $\{\mathscr{X}_i\}_{i=1}^s$. Note that estimation of the regions boils down to the determination of separating hyperplanes once the data are correctly partitioned. This can be efficiently handled via many quite standard SVM-based classification tools [63, 55, 12]. Therefore we will not treat explicitly the region estimation problem here.
To proceed with the estimation of the PVs, we will assume that the orders n_a and n_b are finite, equal for all submodels and known a priori. However the number of submodels may be unknown. It is further assumed that each individual affine submodel is minimal in the ordinary sense.

The major challenge in inferring PWA models from data lies in the fact that the switching law defined by the partition $\{\mathscr{X}_i\}_{i=1}^s$ is not available. For this reason, the classical idea for tackling the PWA system identification problem is to start by partitioning the data according to their most presumably generating submodels. This is achieved in general through data clustering algorithms which rely on the principle that two close regressors (in the sense of the euclidean norm) are likely to have been generated by the same affine submodel [28, 41]. Given such a partition, standard linear identification techniques such as the least squares method can be applied to recover each of the constituent submodels of (4.18). One problem however in that conventional approach is that data clustering algorithms are rarely guaranteed to be optimal. In effect, most of those algorithms iteratively re-assign the data to the different submodels, starting from a certain initial guess. They are hence very sensitive to initialization.

4.2.1 A first introductory idea: overparameterization

One way to overcome the combinatorial nature of the PWA regression problem is by employing overparameterization. This consists in indexing the PVs in (4.21) with time instead of the unknown discrete mode. With this idea in mind, the problem becomes that of searching for as many PVs as data points while trying to reduce the number of the identified PVs. The reduction can be achieved either a posteriori through a second-stage clustering procedure [28] or simultaneously as the estimation step through a sparsity-inducing regularization [43]. We discuss here a special formulation of the overparameterization technique given by

$$\min_{\theta_1,\ldots,\theta_N} \sum_{t=1}^{N} \sum_{k=1}^{t} w_{t,k} d_{t,k} \tag{4.22}$$

with

$$d_{t,k} = |y_k - \theta_t^\top \tilde{x}_k| + |y_t - \theta_k^\top \tilde{x}_t| + \gamma \|\theta_t - \theta_k\|_2 \tag{4.23}$$

and

$$w_{t,k} = \exp\big[-\frac{1}{\sigma_x^2} \|x_t - x_k\|_2^2 - \frac{1}{\sigma_y^2} \|y_t - y_k\|_2^2 \big]. \tag{4.24}$$

The rationale behind this formulation is as follows. For each pair (x_t, x_k) of data points, we would like to find a pair of PVs θ_k and θ_t to fit (as far as possible) the data points $(\tilde{x}_t, y_t)$ and $(\tilde{x}_k, y_k)$ respectively in such a way that both PVs are close whenever x_t and x_k are close in the sense of the euclidean norm. This last requirement justifies the presence of the quantity $\|\theta_t - \theta_k\|_2$ in the cost function. The fitting error is measured here in terms of the ℓ_1-norm which is known to offer a robust minimization capacity (see the earlier chapters 2 and 3) with respect to potential mismatches. The weight $w_{t,k} \in [0,\ 1]$ attached to the term $d_{t,k}$ is intended for reinforcing the fact that θ_t and θ_k must be all the closer as x_t and x_k are spatially close. The parameters σ_x^2 and σ_y^2 appearing in the expression of $w_{t,k}$ can be empirically chosen as the average of the pairwise squared-distances, i.e.,

$$\sigma_x^2 = \frac{2}{N(N-1)} \sum_{t=1}^{N} \sum_{k=1}^{t} \|x_t - x_k\|_2^2 ,$$
$$\sigma_y^2 = \frac{2}{N(N-1)} \sum_{t=1}^{N} \sum_{k=1}^{t} \|y_t - y_k\|_2^2 .$$

When the data are noise-free, σ_x^2 and σ_y^2 can be chosen such that the solution $\{\theta_1, \ldots, \theta_N\}$ to problem (4.22) enjoys the ideal property that $\theta_k = \theta_t$ whenever $\sigma(k) = \sigma(t)$. This property can be enhanced by increasing the regularization parameter γ in (4.23).

Observing that the optimization problem (4.22) is convex, it can be efficiently solved by existing numerical tools [16]. A sequence $\{\theta_t\}_{t=1}^N$ of PVs can hence be computed from the data. Ideally, the PVs that correspond to data points generated by the same submodel of (4.18) are expected to be equal. However strict equality may not hold between such vectors if for example the training data are contaminated by noise. In order to reduce the number of PVs, one can cluster the set of identified PVs by using for example the K-means algorithm [26]. However, as already mentioned above, convergence properties of such clustering algorithms are not well established.

Example 4.1. For illustration purpose we use the following numerical example [11]

$$y_t = \begin{cases} \theta_1^\top \tilde{x}_t + e_t, \text{ if } [4 \quad -1 \quad 10]\tilde{x}_t < 0 \\ \theta_2^\top \tilde{x}_t + e_t, \text{ if } \begin{bmatrix} -4 & 1 & 10 \\ 5 & 1 & -6 \end{bmatrix} \tilde{x}_t \preceq 0 \\ \theta_3^\top \tilde{x}_t + e_t, \text{ if } [-5 \quad -1 \quad 6]\tilde{x}_t < 0 \end{cases} \tag{4.25}$$

with $\tilde{x}_t = [y_{t-1}\ u_{t-1}\ 1]^\top$ and the PVs are given by

$$\theta_1 = \begin{bmatrix} -0.4 \\ 1.0 \\ 1.5 \end{bmatrix}, \quad \theta_2 = \begin{bmatrix} 0.5 \\ -1.0 \\ -0.5 \end{bmatrix}, \quad \theta_3 = \begin{bmatrix} -0.3 \\ 0.5 \\ -1.7 \end{bmatrix}.$$

Solving (4.22) respectively with noise-free and noisy (SNR=30 dB) data yield the sequences $\{\theta_t\}$ depicted in Figure 4.4. The outcome of this experiment is that when the data are completely noise-free, such ideal result as $\theta_t \in \{\theta_1^o, \ldots, \theta_s^o\}$ can be reasonably expected. In the presence of noise however, the story is slightly different, hence prompting the necessity of clustering the estimated PVs.

4.2.2 The nonsmooth optimization approach

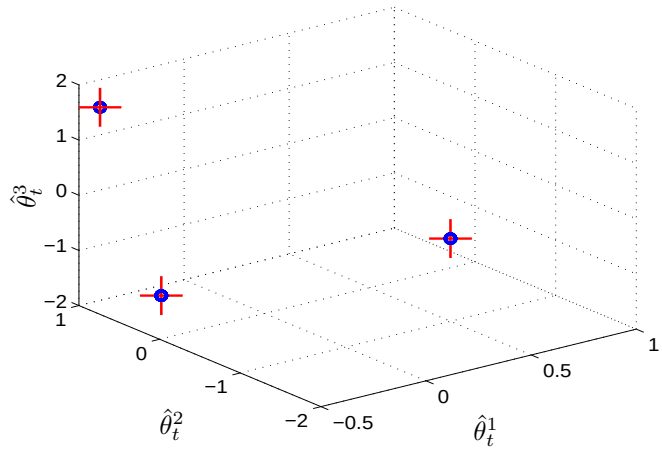
The method presented above has the inconvenience of necessitating a subsequent clustering step after a large number of affine submodels have been identified first. In this section, a slightly different approach is taken. While this approach keeps the main basic features of the one developed in Subsection 4.2.1, it provides directly a number of submodels that is as small as possible. For clarity of presentation it is perhaps better to focus on the estimation of a first single[2] PV from the entire mixed data set.
Based on the considerations of Chapter 3 and the well-known sparsity-promoting property of the ℓ_1-norm, one can search for one of the θ_is by minimizing the cost function

$$J_{\text{SARX}}(\theta) = \sum_{t=1}^{N} |y_t - \theta^\top \tilde{x}_t|. \tag{4.26}$$

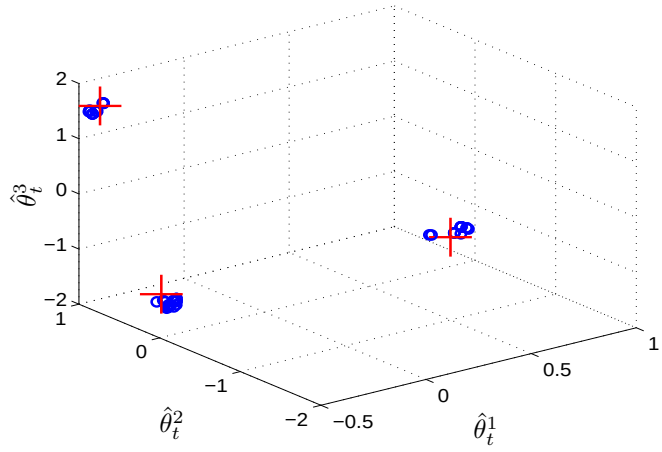
This solution derived in [1] applies to general switched systems where no additional information is available about the nature of the switching mechanism. Therefore this solution is, in principle, directly applicable to PWA systems. However since the model is affine, the effectiveness of a direct minimization of (4.26) can be questioned in the light of Corollary 2.3. Nevertheless we can still exploit the sparse optimization paradigm together with the continuity of the map f_{PWA} to design an identifier. By the continuity and the local linearity (affine) of f_{PWA} we are prompt to consider that two close regressors are very likely to pertain to the same affine subsystem. Taking advantage of this, we introduce the following cost function

$$J_{\text{pwa}}(\theta) = \sum_{t=1}^{N} \sum_{k=1}^{t} w_{t,k}(|y_k - \theta^\top \tilde{x}_k| + |y_t - \theta^\top \tilde{x}_t|) \tag{4.27}$$

[2]Similarly as in Chapter 3, the remaining PVs can be obtained by repeating the estimation process over a modified data set where the data pairs pertaining to the firstly identified PVs have been removed.



(a) No noise, $\gamma = 1$.



(b) Small amount of noise, $\gamma = 1$.

Figure 4.4: Estimated sequence $\hat{\theta}_t = [\hat{\theta}_t^1 \ \hat{\theta}_t^2 \ \hat{\theta}_t^3]^\top$.

in which the weights $w_{t,k}$ are defined as in (4.24). This last objective function resembles the one in (4.22) but with the fondamental difference that it can, in contrast to (4.22), provide directly a single θ in $\{\theta_1^o, \ldots, \theta_s^o\}$ without requiring parameter clustering. To further enhance sparsity, we can call upon the type of iterative reweighting procedure described in Section 2.5. The iterative scheme consists in generating a a finite sequence of parameter vectors according to

$$\theta^{(r)} = \underset{\theta \in \mathbb{R}^{n+1}}{\arg\min}\, J_{\text{pwa}}^{(r)}(\theta) \tag{4.28}$$

for an iteration index r ranging in $0, \ldots, r_{\max}$ with

$$J_{\text{pwa}}^{(r)}(\theta) = \sum_{t=1}^{N} \sum_{k=1}^{t} w_{t,k} (\alpha_k^{(r)} |y_k - \theta^\top \tilde{x}_k| + \alpha_t^{(r)} |y_t - \theta^\top \tilde{x}_t|) \tag{4.29}$$

and $\left\{\alpha_t^{(r)}\right\}_{r=1}^{r_{\max}}$ being a sequence defined by $\alpha_t^{(0)} = 1/N$ for all t,

$$\begin{cases} \alpha_t^{(r)} = \dfrac{\xi_t^{(r)}}{\sum_{t=1}^{N} \xi_t^{(r)}}, & \text{if } r \geq 1, \\ \xi_t^{(r)} = \dfrac{1}{|y_t - \tilde{x}_t^\top \theta^{(r-1)}| + \delta} \end{cases}$$

Here, the number $\delta > 0$ is a small positive number intended essentially here for preventing division by zero.

4.2.3 An iterative sparsity-promoting scheme

In this section we consider a PWARX system where the partition of the regression space has the particular form described in (4.13)-(4.14). This is called the Voronoi partition. If we adopt such a partition for the regression space $\mathscr{X}$, the PWA identification problem reduces to that of identifying the centers $\{c_i\}_{i=1}^s$ and the PVs $\{\theta_i\}_{i=1}^s$. Note that some of the approaches discussed earlier are still applicable in this scenario: for example, overparameterization, combined k-means/k-subspaces, one-by-one estimation based on sparsity-inducing optimization. A special implementation of the latter is discussed here for PWA identification. But it is worth noting that the principle of this scheme is applicable to the general scope of the sparse optimization approach for hybrid system identification (including the one presented in Chapter 3 for switched systems).

The goal is to address some shortcomings of the one-by-one estimation strategy which was implemented for example by Algorithm 3.2. Recall that in the sparse optimization approach, the parameter vectors are not estimated all at once; instead they are incrementally identified one after another through an ℓ_1-norm convex relaxation technique. After each parameter vector (PV) is identified, the data points pertaining to the corresponding submodel need to be removed before proceeding with the identification of the next PV. The data removal steps require however the definition of appropriate thresholds, the choice of which gets increasingly delicate as the level of noise becomes high. We discuss an efficient implementation of the sparsity ideas of [1] (see also Chapter 3) such that all the PVs can be identified all at once. Being able to identify simultaneously the parameter vectors has an essential advantage. It allows us to get rid of the necessity to extract the PVs in an incremental manner and thereby dropping the requirement

of thresholds for data elimination. By so doing however, the method looses it convex feature. The main device for estimating all the PVs simultaneously is the introduction of some implicit discriminatory weights on the fitting errors involved in the objective function to be minimized.

Intuition of the method. Inspired by the results of Chapter 3 (which formulates the identification of a single PV), one could consider formulating the objective function

$$h(\{(\theta_i, c_i)\}_{i=1}^s) = \sum_{i=1}^{s}\sum_{t=1}^{N} r_t(\theta_i, c_i). \tag{4.30}$$

with $r_t : \mathbb{R}^{n+1} \times \mathbb{R}^n \to \mathbb{R}_+$ is a time indexed function defined by

$$r_t(\theta, c) = |y_t - \theta^\top \tilde{x}_t| + \gamma \|c - x_t\|_2 .$$

However, minimizing simply the convex function h will not yield the searched parameter vectors $\{\theta_i^o, c_i^o\}$. The reason for this is as follows.

Proposition 4.1. *Let $\{\theta_i^\star, c_i^\star\}_{i=1}^s$ be a minimizer of the objective function in (4.30). Then it holds that*

$$\sum_{t=1}^{N} r_t(\theta_i^\star, c_i^\star) = \min_{\theta,c} \sum_{t=1}^{N} r_t(\theta, c) \ \ \forall i \in \mathbb{S}. \tag{4.31}$$

Proof. Suppose, for contradiction purpose, that the statement (4.31) is not true. Consider a pair $(\theta^\star, c^\star)$ in $\arg\min_{(\theta,c)} \sum_{t=1}^N r_t(\theta, c)$. Then there is at least one pair $(\theta_q^\star, c_q^\star)$, $q \in \mathbb{S}$, such that $\sum_{t=1}^N r_t(\theta^\star, c^\star) < \sum_{t=1}^N r_t(\theta_q^\star, c_q^\star)$. It follows that

$$h((\theta^\star, c^\star), \ldots, (\theta^\star, c^\star)) = s\sum_{t=1}^{N} r_t(\theta^\star, c^\star) < h(\{(\theta_i^\star, c_i^\star)\}_{i=1}^s).$$

The first term on the right hand side refers to the value taken by h when all pairs (θ_i, c_i) are set equal to $(\theta^\star, c^\star)$. This contradicts the assumption that $\{(\theta_i^\star, c_i^\star)\}_{i=1}^s$ is a minimizer of (4.30). In conclusion, (4.31) holds as claimed. □

The proposition says that any minimizing sequence $\{(\theta_i, c_i)\}_{i=1}^s$ of (4.30) is such that $(\theta_i, c_i) \in \arg\min_{(\theta,c)} \sum_{t=1}^N r_t(\theta, c)$ for all $i = 1, \ldots, s$. This suggests that there is very little hope of estimating correctly the parameter vectors and the regions centers by minimizing (4.30). For example if $\arg\min_{(\theta,c)} \sum_{t=1}^N r_t(\theta, c) = \{(\theta^\star, c^\star)\}$, then any minimizing sequence $\{(\theta_i, c_i)\}_{i=1}^s$ of h is such that $(\theta_i, c_i) = (\theta^\star, c^\star)$ for all $i = 1, \ldots, s$. As a result, it is impossible to obtain s distinct pairs by minimizing (4.30). It is therefore necessary to impose an additional constraint on the allowable solution set if we want the solution to coincide with $\{(\theta_i^o, c_i^o)\}_{i=1}^s$. To this end, we will introduce a discriminative weighting strategy. The idea is to assign some weights to the different terms of the objective function (4.30). By doing so, we introduce an implicit discriminatory treatment of the data points i.e., the data samples are no longer fitted similarly to the parameter vectors.

Derivation of the algorithm. To proceed, consider a weighted version of (4.30) in the form

$$\sum_{i=1}^{s}\sum_{t=1}^{N} w_i(t) r_t(\theta_i, c_i).$$

As an initialization step, let us minimize this cost by considering that $w_j(t) = 1$ for all $(t, j) \in \mathbb{I} \times \mathbb{S}$. Denote with $\left\{(\theta_i^{(1)}, c_i^{(1)})\right\}_{i=1}^{s}$ the resulting solution. Introduce the notation

$$I_i^{(1)} = \{t \in \mathbb{I} : i \in \underset{j=1,\dots,s}{\arg\min}\, r_t(\theta_j^{(1)}, c_j^{(1)})\} \tag{4.32}$$

for the collection of all time indices t for which $r_t(\theta_i^{(1)}, c_i^{(1)})$ is minimum among all others. Let $X_{I_i^{(1)}}$ be a matrix formed by the regressors $\left\{x(t) : t \in I_i^{(1)}\right\}$; similarly, let $\boldsymbol{y}_{I_i^{(1)}}$ be a vector formed with the outputs indexed by $I_i^{(1)}$. The collection $\left\{I_i^{(1)}\right\}$ defines a partition of the data set. Using this partition the prior estimates are updated as follows:

$$\begin{aligned} \theta_i^{(1)} &\leftarrow \underset{\theta\in\mathbb{R}^n}{\arg\min} \left[\left\|\theta - \theta_i^{(1)}\right\|_2^2 + \eta\|\boldsymbol{y}_{I_i^{(1)}} - X_{I_i^{(1)}}^{\top}\theta\|_2^2\right] \\ c_i^{(1)} &\leftarrow \underset{c\in\mathbb{R}^n}{\arg\min} \left[\left\|c - c_i^{(1)}\right\|_2^2 + \eta \sum_{t\in I_i^{(1)}} \|c - x_t\|_2^2\right]. \end{aligned} \tag{4.33}$$

where $\eta > 0$.

Now we need to update the weights before performing the next iteration. From the values $\left\{(\theta_i^{(1)}, c_i^{(1)})\right\}$, define for any $t \in \mathbb{I}$, $j \in \mathbb{S}$, the "fitting error" $r_t(\theta_j^{(1)}, c_j^{(1)}) = |y_t - \tilde{x}_t^{\top}\theta_j^{(1)}| + \eta\left\|c_j^{(1)} - x_t\right\|_2$, and the weights

$$w_j^{(1)}(t) = \frac{\prod_{i\neq j} r_t(\theta_i^{(1)}, c_i^{(1)})}{[r_t(\theta_j^{(1)}, c_j^{(1)})]^{s-1} + \delta} \tag{4.34}$$

with $\delta > 0$ a small number destined for preventing division by 0. The so-defined $w_j^{(1)}(t)$ will be large when $r_t(\theta_j^{(1)}, c_j^{(1)})$ is small compared to all the other errors $r_t(\theta_i^{(1)}, c_i^{(1)})$, $i \neq j$. A large fitting error $r_t(\theta_j^{(1)}, c_j^{(1)})$ on the contrary will indicate a discrepancy between the available estimate $(\theta_j^{(1)}, c_j^{(1)})$ and the data pair (x_t, y_t). This results in a small weight $w_j^{(1)}(t)$ being assigned to the sample (x_t, y_t). Therefore, minimizing the weighted cost $\sum_{i=1}^{s}\sum_{t=1}^{N} w_i^{(1)}(t) r_t(\theta_i, c_i)$ is expected to decrease further the smallest error $r_t(\theta_j, c_j)$ in the next iteration.

The overall algorithm proceeds as follows. Let $w_j^{(0)}(t) = 1$ for any $(t, j) \in \mathbb{I} \times \mathbb{S}$, define a small number $\delta^0 > 0$. Then apply for any iteration number $k \geq 0$,

$$\left\{(\theta_i^{(k+1)}, c_i^{(k+1)})\right\}_{i=1}^{s} \in \underset{\substack{(\theta_i,c_i)\in\mathbb{R}^{n+1}\times\mathbb{R}^n \\ i=1,\dots,s}}{\arg\min} h^{(k)}\left(\{(\theta_i, c_i)\}_{i=1}^{s}\right) \tag{4.35}$$

$$\theta_i^{(k+1)} \leftarrow \left(I + \eta X_{I_i^{(k+1)}} X_{I_i^{(k+1)}}^{\top}\right)^{-1}\left(\theta_i^{(k+1)} + \eta X_{I_i^{(k+1)}} \boldsymbol{y}_{I_i^{(k+1)}}\right),\ i \in \mathbb{S} \tag{4.36}$$

$$c_i^{(k+1)} \leftarrow \frac{1}{1 + \eta|I_i^{(k+1)}|}\left(c_i^{(k+1)} + \eta \sum_{k\in I_i^{(k+1)}} x_t\right),\ i \in \mathbb{S} \tag{4.37}$$

until some stopping condition is satisfied. Such a condition can be of the form

$$\left\|\chi^{(k+1)} - \chi^{(k)}\right\|_2 < \epsilon$$

with $\chi^{(k)} = \text{col}(\theta_1^{(k)}, \ldots, \theta_s^{(k)}, c_1^{(k)}, \ldots, c_s^{(k)})$ being a vector formed by vectorizing all parameters. In (4.36)-(4.37) the set $I_i^{(k+1)}$ is defined as in (4.32) from $\left\{\theta_i^{(k+1)}, c_i^{(k+1)}\right\}_{i=1}^{s}$. Concerning the function $h^{(k)}$ in (4.35), it is a weighted version of h in (4.30) defined by

$$h^{(k)}(\{(\theta_i, c_i)\}_{i=1}^{s}) = \sum_{i=1}^{s}\sum_{t=1}^{N} w_i^{(k)}(t) r_t(\theta_i, c_i),$$

where $\{w_j^{(k)}(t)\}$ form a sequence of weights which is updated as

$$w_j^{(k+1)}(t) = \frac{q_j(t)}{\sum_{j=1}^{s} q_j(t)},\ j \in \mathbb{S},\ t \in \mathbb{I}, \tag{4.38}$$

$$q_j(t) = \frac{\prod_{i \neq j} r_t(\theta_i^{(k+1)}, c_i^{(k+1)})}{[r_t(\theta_j^{(k+1)}, c_j^{(k+1)})]^{s-1} + \delta^{(k)}},\ j \in \mathbb{S},\ t \in \mathbb{I}, \tag{4.39}$$

$$\delta^{(k+1)} = \alpha\delta^{(k)},\ 0 < \alpha \leq 1. \tag{4.40}$$

After the algorithm is completed, one can consider refining the estimates by performing s least squares estimations based on the last partitioning $\left\{I_i^{(k+1)}\right\}_{i=1}^{s}$ of the data set.

To test the algorithm (4.35)-(4.37), we consider two examples: two static functions and a nonlinear dynamic system.

Example 4.2. Applying the method to a set of data generated by the two static systems defined by

$$y_t = \sin(u_t) + e_t \tag{4.41}$$

$$y_t = -3/2u_t^2 - 1/2u_t^3 + e_t, \tag{4.42}$$

yields the results depicted in Figure 4.5. It can be seen that the algorithm returns estimates of reasonable accuracy. Note that the performance of the estimated PWA model can be much better if there were no noise in the identification data and if the number of submodels is increased. See for example Figure 4.6 for an illustration.

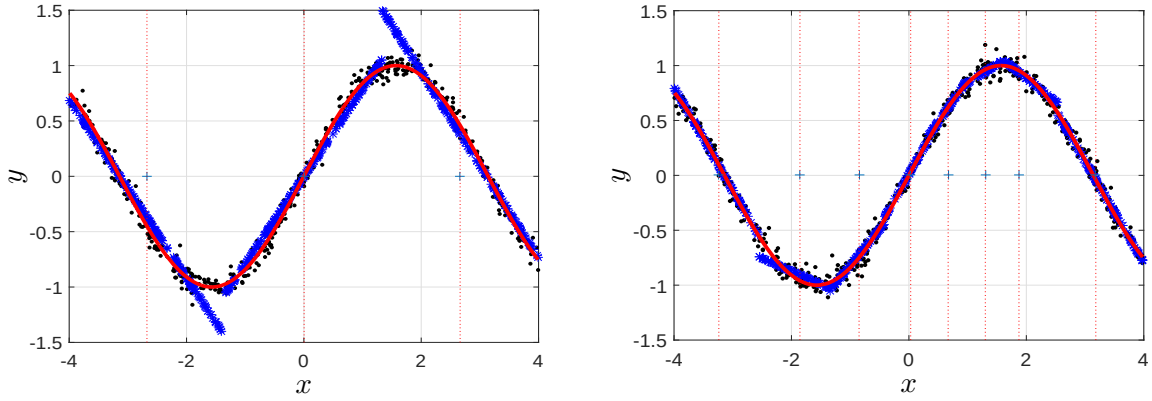
Example 4.3. The second example concerns the PWA modeling of a nonlinear dynamical system defined by

$$y_t = \frac{\sin(0.1u_{t-1})(y_{t-1}+3)y_{t-1}y_{t-2}}{1 + y_{t-1}^2 + y_{t-2}^2} + u_{t-1} + e_t. \tag{4.43}$$

The excitation signal is chosen to be the multisine signal represented in Figure 4.7; the noise sequence $\{e_t\}$ is generated as the realization of a white Gaussian noise so as to achieve an SNR of 20 dB. Applying the algorithm (4.35)-(4.37) to $N = 600$ samples, we can estimate a PWA model with the Voronoi type of partition. The results displayed in Figure 4.8 seem to yield a "good" fit as the number of submodels goes up although the increase rate of the fit slows down considerably after the second submodel.

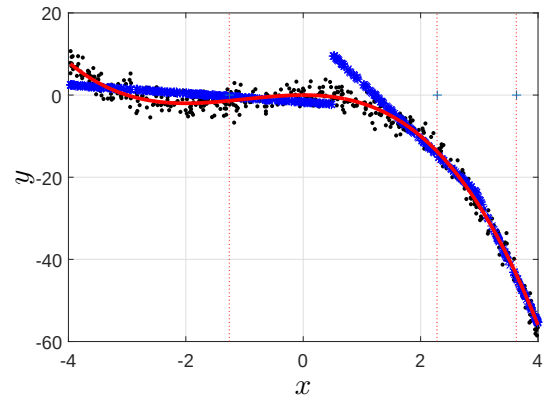
4.3 Adaptive identification of PWA models

We now discuss a simple adaptive scheme for the estimation of PWA models with the Voronoi type of partition. We present here a simple recursive approach. The data are assumed to be

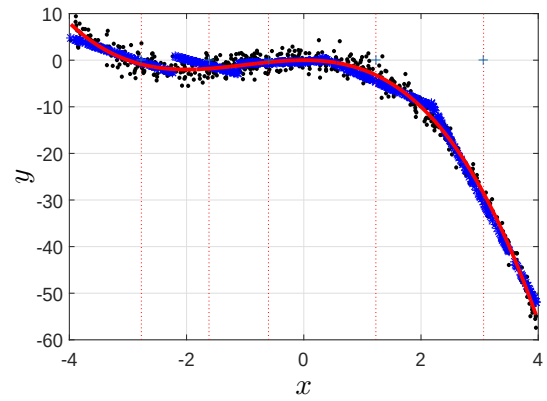


(a) 3 submodels

(b) 8 submodels

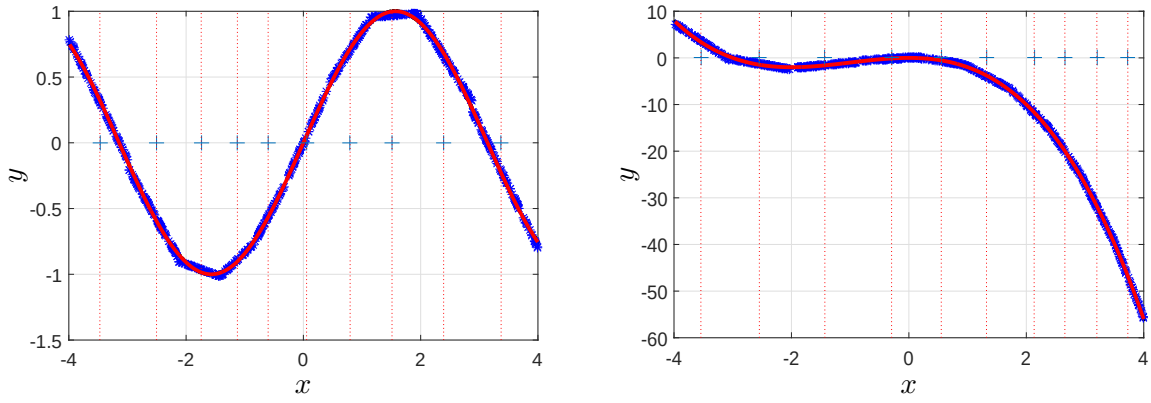


(c) 3 submodels



(d) 5 submodels

Figure 4.5: PWA modeling of static nonlinear functions from noisy data with the proportion of noise being such that the SNR is equal to 20 dB: noisy data (black dots); true nonlinear function (red solid line); PWA function (blue stars). The two top subfigures are related to system (4.41) while the two other pertain to system (4.42).



(a) 10 submodels (b) 10 submodels

Figure 4.6: PWA modeling of static nonlinear functions from noise-free data: observations (black dots); true nonlinear function (red solid line); PWA function (blue stars). The crosses indicate the positions of the centers $\{c_i\}$.

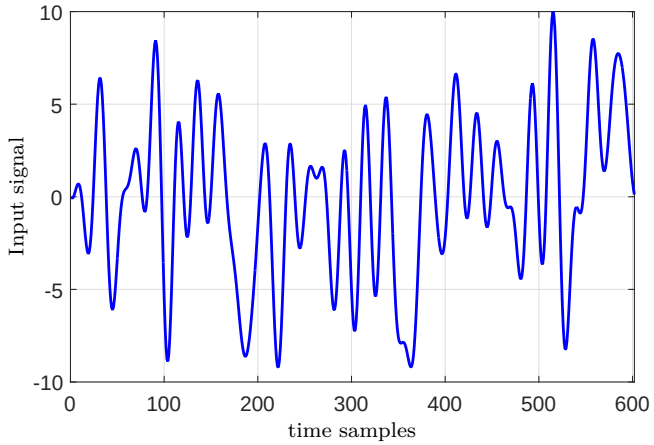
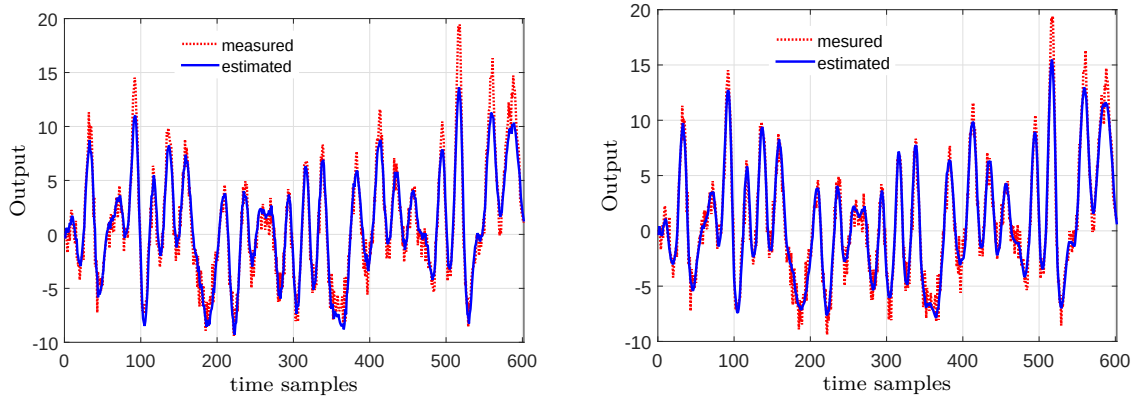


Figure 4.7: Excitation signal for the identification of the nonlinear system (4.43).

sequentially acquired. Then, starting from some initial values, we proceed alternately to data clustering and parameter update on-line. At a given time, the discrete state is inferred based on the information available up to that time and the PVs are accordingly refined via, for example, recursive least squares. In comparison with batch mode methods, it is fair to say that our method, though possibly less effective than some of them, makes it easier to effectively handle higher dimensional data or larger amounts of data. Furthermore, it can be used for on-line identification of hybrid systems.

To proceed with the description of the method, let us assume that the regressors $\{x_t\}_{\sigma(t)=i}$ and the noise sequence $\{e_t\}_{\sigma(t)=i}$ pertaining to each submodel i have Gaussian distributions $\mathcal{N}(c_i, \Lambda_i)$ and $\mathcal{N}(0, \lambda_i^e)$ respectively. The parameters of the distributions are updated thanks to the empirical formula for expected value while the submodels' parameters are updated via recursive least squares. More precisely, if $c(t-1)$ and $\Lambda(t-1)$ designate the prior values for the center and the covariance matrix respectively, then we compute $c(t)$ and $\Lambda(t)$ when x_t becomes



(a) PWA with 1 submodel: FIT=71.30

(b) PWA with 2 submodels: FIT=77.60

Figure 4.8: PWA modeling of a nonlinear system: comparison of system and model outputs. The model is identified with 60% of the data and validation is done on the entire data set.

available as follows

$$c(t) = \arg\min_{c\in\mathbb{R}^n} \left[\gamma \left\|c - c(t-1)\right\|_2^2 + (1-\gamma)\left\|c - x_t\right\|_2^2\right]$$
$$\Lambda(t) = \gamma\Lambda(t-1) + (1-\gamma)(x_t - c(t))(x_t - c(t))^\top$$

where $\gamma \in [0,\ 1]$ is a forgetting factor. In a quite symmetric way, by letting $\theta(t-1)$ denote the prior value of the parameter vector, $\theta(t)$ is obtained via a proximal operator

$$\theta(t) = \arg\min_{\theta\in\mathbb{R}^n} \left[\gamma(\theta - \theta(t-1))^\top P^{-1}(t-1)(\theta - \theta(t-1)) + (y_t - \theta^\top \tilde{x}_t)^2\right]$$
$$P^{-1}(t) = \gamma P^{-1}(t-1) + x_t x_t^\top.$$

Here, $P^{-1}(t)$ can be viewed as a kind of correlation matrix of the sequence $\{x_t\}$.
Let $(c_i(t), \Lambda_i(t))$ and $\lambda_i^e(t))$ denote the estimates of the parameters of the distributions (introduced above) obtained empirically from observations up to time t. Likewise, denote with $\theta_i(t)$ the estimate of the PV θ_i^o at time t. Then the recursive identification algorithm consists of the following steps:

- At the initial time $t = 0$, generate vectors $\{c_i(0), \theta_i(0)\}$ at random; select the matrices $\left\{\Lambda_i^{-1}(0), P_i(0)\right\}$ to be proportional to the identity matrix. Finally, set $\lambda_i^e(0)$ to be a small positive scalar.

- At each subsequent time $t = 1, 2, \ldots$, find an estimate of the discrete state as

$$\hat{\sigma}(t) = \arg\min_{i=1,\ldots,s} \mathcal{J}_i(t),$$

  where

$$\mathcal{J}_i(t) = (c_i(t-1) - x_t)^\top \Lambda_i^{-1}(t)(c_i(t-1) - x_t) + \lambda_i^e(t)^{-1}\left(y_t - \theta_i(t-1)^\top \tilde{x}_t\right)^2.$$

  Then update the variables as follows:

– center $c_{\hat{\sigma}(t)}$ and associated covariance $\Lambda^{-1}_{\hat{\sigma}(t)}$:

$$c_{\hat{\sigma}(t)}(t) = \gamma c_{\hat{\sigma}(t)}(t-1) + (1-\gamma)x_t \tag{4.44}$$
$$v_{\hat{\sigma}(t)}(t) = \sqrt{\gamma}(x_t - c_{\hat{\sigma}(t)}(t-1)) \tag{4.45}$$
$$\Lambda^{-1}_{\hat{\sigma}(t)}(t) = \gamma^{-1}\Big[I_n - \frac{\Lambda^{-1}_{\hat{\sigma}(t)}(t-1)v_{\hat{\sigma}(t)}(t)v_{\hat{\sigma}(t)}(t)^\top}{\gamma + v_{\hat{\sigma}(t)}(t)^\top \Lambda^{-1}_{\hat{\sigma}(t)}(t-1)v_{\hat{\sigma}(t)}(t)}\Big]\Lambda^{-1}_{\hat{\sigma}(t)}(t-1) \tag{4.46}$$
$$c_i(t) = c_i(t-1), \quad \Lambda_i^{-1}(t) = \Lambda_i^{-1}(t-1) \ \forall i \in \{1,\dots,s\}, i \neq \hat{\sigma}(t) \tag{4.47}$$

– output noise variance $\lambda^e_{\hat{\sigma}(t)}$:

$$\varepsilon_{\hat{\sigma}(t)}(t|t-1) = y_t - \theta^\top_{\hat{\sigma}(t)}(t-1)\tilde{x}_t \tag{4.48}$$
$$\lambda^e_{\hat{\sigma}(t)}(t) = \gamma\lambda^e_{\hat{\sigma}(t)}(t-1) + (1-\gamma)\varepsilon^2_{\hat{\sigma}(t)}(t|t-1) \tag{4.49}$$
$$\lambda^e_i(t) = \lambda^e_i(t-1) \ \forall i \in \{1,\dots,s\}, i \neq \hat{\sigma}(t) \tag{4.50}$$

– and $(\theta_{\hat{\sigma}(t)}, P_{\hat{\sigma}(t)})$ by recursive least squares:

$$q(t) = \frac{P_{\hat{\sigma}(t)}(t-1)\tilde{x}_t}{\gamma + \tilde{x}_t^\top P_{\hat{\sigma}(t)}(t-1)\tilde{x}_t} \tag{4.51}$$
$$\theta_{\hat{\sigma}(t)}(t) = \theta_{\hat{\sigma}(t)}(t-1) + q(t)\varepsilon_{\hat{\sigma}(t)}(t|t-1) \tag{4.52}$$
$$P_{\hat{\sigma}(t)}(t) = \gamma^{-1}\left(I_n - q(t)\tilde{x}_t^\top\right)P_{\hat{\sigma}(t)}(t-1) \tag{4.53}$$
$$\theta_i(t) = \theta_i(t-1), \quad P_i(t) = P_i(t-1) \ \forall i \in \{1,\dots,s\}, i \neq \hat{\sigma}(t) \tag{4.54}$$

In the interest of computational load one can consider simplifying the algorithm by forcing for example $\Lambda_j(t)$ to the identity matrix and by dropping the parameters related to the distributions of the output noise $\{e_t\}$. More precisely, removing Eqs (4.45)-(4.50) yields the algorithm described in our paper [4]. Interested readers are referred to this latter paper for an application of the recursive identifier (4.44)-(4.54) to the modeling of an open channel system.
Let us close this section by giving a few comments on the estimation method:

- The proposed recursive algorithm has the property of being able to operate online while allowing for the simultaneous estimation of the validity regions and the affine submodels by the same procedure. Still, it has a very low complexity at each update (about $O(n^2)$).

- The true number of submodels need not be known in advance. Given an upper bound $\bar{s} \ll N$ on the number s of submodels, the algorithm is able to operate with $\bar{s}$ initially presumed submodels and determine s in the end, as the number of modes to which a significant number of data have been affected. After the algorithm converges, the $\bar{s} - s$ unnecessary submodels should stop being assigned data so that in the end, only a negligible amount of data have been affected to those submodels. They can therefore be recognized and eliminated.

- Note that the concept of the method is not restricted to the recursive least squares (RLS) as an identifier. In fact any linear and fast recursive algorithm can be used for the estimation of the parameters. For example we can use algorithms such as the stochastic gradient algorithm, the equation error identifier, the projection algorithm, the Kalman filter based

identifier [31]. It might be advisable to endow those routines with forgetting capabilities in such a way that effects of wrong prior decisions concerning the discrete mode can be quickly removed.

- An open problem regarding the recursive identifier is convergence analysis. In this respect, it should be observed that the algorithm is trying to solve a nonconvex optimization problem. As a consequence, global convergence is very hard to guarantee. In practice however, the algorithm performs surprisingly well. The performance can even be enhanced via multiple random initializations in parallel.

4.4 Conclusion

This chapter provides an overview of our reflections on piecewise affine system identification from input-output measurements. In particular, two groups of approaches have been discussed. The first group is based on sparsity-inducing optimization. The second relies on an adaptive clustering/identification scheme. We have contributed to the development of some other approaches which are not presented here but they can be found in full detail in [15, 14]. Also, an application of PWA identification to the modeling of the NOx emission of diesel engines can be found in [64].

Chapter 5

Conclusions and Perspectives

Contents

5.1 Summary

This report is intended as an overview of my research results over the period 2009-2016. The problem dealt with is that of estimating (learning) models from data. Globally speaking, my research concerns the development of methods and algorithms for extracting relevant features/patterns contained in informative observations collected, in principle, from real-world processes. This way of constructing models is sometimes called data-based modeling with regards to its principle of attempting to generically mimic (hidden) laws of nature by performing quantitative experiments. Different types of model structures can be considered: static, dynamic, linear, nonlinear, switched, piecewise affine. Given a model structure parameterized with unknown parameters and a collection of data which are assumed to be informative enough, our estimation approach consists in formalizing appropriately an optimization problem whose solution (a vector/matrix of parameters) (1) produces a model that fits best the data (2) is easy to compute numerically. The current manuscript describes a few specific approaches to switched and piecewise affine systems identification but the applicability scope of the methods is much larger. In particular, the scope goes beyond the traditional view of system identification (which is confined to the estimation of dynamic models) and embraces the much larger topic of machine learning. Given the nature of the challenge posed by the hybrid system identification problem, a common thread of our methods is their sparsity-inducing capability which, as discussed in Chapter 2, is closely relevant to robustness (discriminatory insensitivity to outliers). Due to the central role of this sparsity-inducing property, Chapter 2 has received a more detailed treatment than the other chapters. We have then shown how this property can be exploited for the identification of switched and PWA systems. In the case of switched systems one of the main results has

the following flavor: if the regression data are sufficiently rich and if each subsystem has been visited a sufficient number of times, then the parameters of the submodels can be estimated as the solutions of a sparse optimization problem. For computational efficiency, a sparsity-inducing convex relaxation has then been considered. Correctness of this latter procedure is proved under informativity assumptions and an additional requirement related to the proportion of data generated by each subsystem.

5.2 Perspectives

In this section, we describe our plan for future research. This is articulated around four main points:

- capitalize on the developed identification methods by providing further analysis for understanding them and packaging them in a software
- formalize goal-oriented identification of PWA models for nonlinear systems
- generalize and extend sparsity-inducing optimization techniques to the design of control policies and other problems in control theory and machine learning
- analyze performance and stability of nonlinear systems through a PWA differential embedding.

5.2.1 Analysis and application of hybrid system identification methods

Analysis of algorithms. It is fair to observe that after about fifteen years of intensive research, the topic of hybrid system identification is not completely mature yet. In effect, the research community working on this topic is still facing many theoretical challenges the main of which being convergence and correctness analysis of most of the developed algorithms. The expected outcome of such an analysis is to provide a priori some guarantees as to the reliability of a given identification algorithm under specific assumptions. From a practical standpoint, it is important to provide some guidance for users who are interested in applying existing methods to modeling problems. The purpose of the analysis is to derive the necessary elements for appreciating a priori which method among the existing ones would perform (or is likely to perform) better in which circumstances. It should however be kept in mind that due to the nontrivial (and generally nonconvex) structure of the hybrid system computation algorithms, the analysis is a very challenging task.

Identifiability and persistence of excitation. Closely related issues to the convergence analysis are the study of structural identifiability of hybrid system models and persistence of excitation of the regression data. Identifiability is an injectivity property of the parameterization map defined from a parameter space to a set of (hybrid) dynamical models. As such, it is related to the well-posedness of the identification problem. If the to-be-identified system lies in the range space of the parameterization map and this latter has the property of identifiability, then it is possible, in principle, to find uniquely the parameter vector associated with the true system.
As for the persistence of excitation[1] (PE), it is a property that reflects sufficient informativity

[1]Note that this terminology is reserved by some authors to the specific situation where the estimation schemes are adaptive.

for a specific collection of the input-output data. Recall that the goal of system identification is to infer from a *single* input-output trajectory (of possibly infinite-length) of the system *all* its possible trajectories. Persistence of excitation then characterizes the minimal condition the regression data set generated by a single experiment on the system must satisfy for this inference to be possible. There are two levels of analysis here: (i) characterize the PE condition that the regression data must satisfy (this is expressed in Chapter 3 for example in term of the n-genericity index); (ii) characterize sufficient richness of inputs, i.e., find the set of input signals (and possibly the set of switching signals) which, if fed to the to-be-identified hybrid system, will generate regressors that satisfy the PE condition. While these aspects are well-understood for linear model sets, there are still much obscure for hybrid models. In the current literature, only the level (i) is formalized for a few methods. We have already made some progress over these subjects in the papers [48, 47, 46] but the derived conditions are generic in the sense that they do not depend of any specific estimation algorithm. Therefore, it still remains to connect these results to the analysis of concrete algorithms.

Developing a unified toolbox. It can be observed from the rapidly growing literature that there is an increasing interest for hybrid system identification on both the fronts of theory and application (see e.g. the surveys [45, 29]). The large number of references covered by the above cited survey papers also show how vast the scope of applications may be. The consequence of the dynamism of the research on hybrid system identification is the production of a large amount of methods supported by different mathematical frameworks. This diversity of methods and algorithms prompts the necessity of providing a proper guidance for potential users. So, in the next few years we intend to devote some effort in promoting the field of hybrid system identification. Perhaps one way to ease the applicability of the methods to real problems is to build a unified toolbox[2] with the most efficient methods for each class of models. There are two possibilities for achieving such a goal: The first is a unilateral coding of the existing methods on the same platform; the second possibility is a collaborative development through launching a call for contributions to the respective authors.

5.2.2 Optimal control of switched systems

An ongoing research project is that of optimal control of switched systems. The control law in this case is constituted of two components: a continuous input which controls the evolution of the system in each discrete mode and a switching control signal which selects over time which subsystem should be activated. The optimal control design therefore is concerned with finding a joint optimal policy for both types of control components. This is a hard combinatorial problem. For the time being it has not been completely addressed yet. For example, no exact solution is available for the infinite horizon quadratic control problem. For the finite horizon version, the solution has been formally characterized (see e.g., [69, 51]) but its implementation still suffers from an exponential complexity even in finite horizon. Judging from the nature of the challenge posed by this problem, the sparse optimization approach together with its nonsmooth convex relaxation seem applicable to it. The huge numerical complexity of the solution is indeed due to the fact that an exponential number of switching sequences have to be explored. One possible idea to reduce the computational complexity is to parameterize the system equations so as to replace the switching sequence with auxiliary continuous input variables which therefore should

[2] To our knowledge, the only accessible toolbox is the HIT package [27] developed by Ferrari-Trecate. But it is based on a single PWA identification method, the one in [28].

satisfy a sparsity constraint. The hope in doing so is to be able to transform the optimal (quadratic) control problem whose natural formulation suffers from a combinatorial complexity, to a continuous (but potentially nonconvex) optimization problem. We think that sparsity inducing-optimization methods treated in Chapters 2 and 3 of the current manuscript might help derive efficient convex relaxations of this problem.

5.2.3 PWA modeling for nonlinear systems control and analysis

Another ongoing research project concerns the control design and the analysis for nonlinear systems based on piecewise affine systems (thesis of S. Waitman). Our ultimate goal is to reach a rich set of systematic design and analysis tools for general nonlinear systems. A first step towards this goal is to describe nonlinear systems with model structures which are simple enough to apprehend and yet able to capture the essential behavior of the nonlinear system with respect to the application of interest. In this respect, a good candidate of model class is, as discussed in Chapter 4, that of PWA models. A PWA model consists of a partition of the state-input space of the system into local regions, each of which is associated with an affine time-invariant model. The advantage of such models, is that they arise naturally from a basic intuition of control engineering practitioners, the notion of operating point. They are therefore more easily amenable to interpretation and analysis by judiciously using the available knowledge on linear systems. Note that for the analysis of nonlinear systems based on the PWA model to be conclusive, it is necessary to embed the approximation error into an uncertainty set that has to be taken into account during the analysis. There are two sources of uncertainty: the theoretical approximation error resulting from the approximation of nonlinear system by a PWA system ; the uncertainty/bias induced by the estimation algorithm and the noise contained in the measurements. The pair formed by a nominal PWA model and the uncertainty set is termed PWA Differential Inclusions (PWA-DI). We would like to develop systematic procedures for constructing PWA-DI for nonlinear systems either from the nonlinear equations of the system (when available) or from experimental data using system identification techniques or a combination of both approaches.

5.2.4 PWA modeling for nonlinear control

Another aspect of our future research plan is to tie the search for a PWA model to the application it is intended for. For example if the intended use of the model is control design, it would be interesting to connect the model derivation step and the controller design step as parts of a single design chain for nonlinear control systems. A widespread practice is to treat these steps as two totally independent tasks. However, the quality of a model is closely related to its purpose. Therefore, finding pragmatically the "best" model requires setting a quantitative "measure" of model quality which is consistent with the expected/achievable control performance. Optimizing such an application-dependent "measure" is indeed the only way to make sure that the searched model, in addition to describing the process with an acceptable accuracy, is structurally suitable for its purpose. For example, depending on the intended application, some dynamics or behaviors of the true system might not be relevant; they can therefore be neglected, thereby resulting in a rough model and yet still pertinent in regards to the envisaged application. The system identification and control communities tend to grow more or less independently and without much interactions. In nonlinear control theory, it is generally assumed that an analytic model is available while in system identification, the custom is to compute a general purpose black-box model by minimizing only a prediction error, that is, without taking explicitly into

account any specified purpose of the model. As a consequence, the model obtained via system identification might be unnecessarily complex and thus not directly suitable for a specific control design technique or the control laws designed under exact model assumption (which does not hold in reality) might not be applicable to the true system. These observations have driven an intense research activity referred to as *identification for control* [30, 13]. However, the models considered in most existing methods are essentially linear. We envisage reflecting on the extension of these methods to PWA models when the system of interest is nonlinear.

Bibliography

[1] L. Bako. Identification of switched linear systems via sparse optimization. *Automatica*, 47:668–677, 2011.

[2] L. Bako. Subspace clustering through parametric representation and sparse optimization. *IEEE Signal Processing Letters*, 21:356–360, 2014.

[3] L. Bako. Adaptive identification of linear systems subject to gross errors. *Automatica*, 67:192–199, 2016.

[4] L. Bako, K. Boukharouba, E. Duviella, and S. Lecoeuche. A recursive identification algorithm for switched linear/affine models. *Nonlinear Analysis:Hybrid Systems*, 5:242–253, 2011.

[5] L. Bako, K. Boukharouba, and S. Lecoeuche. An $\ell_0 - \ell_1$ norm based optimization procedure for the identification of switched nonlinear systems. In *IEEE Conference on Decision and Control, Atlanta, GA, USA*, 2010.

[6] L. Bako, K. Boukharouba, and S. Lecoeuche. Recovering piecewise affine maps by sparse optimization. In *IFAC Symposium on System Identification, Brussels, Belgium*, 2012.

[7] L. Bako, D. Chen, and S. Lecoeuche. A numerical solution to the minimum-time control problem for linear discrete-time systems. Technical report, `http://arxiv.org/abs/1109.3772`, 2011.

[8] L. Bako, V. L. Le, F. Lauer, and G. Bloch. Identification of mimo switched state-space systems. In *American Control Conference, Washington DC*, 2013.

[9] L. Bako and S. Lecoeuche. A sparse optimization approach to state observer design for switched linear systems. *Systems & Control Letters*, 62:143–151, 2013.

[10] L. Bako and H. Ohlsson. Analysis of a nonsmooth optimization approach to robust estimation. *Automatica*, 66:132–145, 2016.

[11] A. Bemporad, A. Garulli, S. Paoletti, and A. Vicino. A bounded-error approach to piecewise affine system identification. *IEEE Transactions on Automatic Control*, 50:1567–1580, 2005.

[12] K. P. Bennett and O. L. Mangasarian. Multicategory discrimination via linear programming. *Optimization Methods and Software*, 4:27–39, 1994.

[13] X. Bombois, M. Gevers, G. Scorletti, and B. Anderson. Robustness analysis tools for an uncertainty set obtained by prediction error identification. *Automatica*, 37(10):1629–1636, 2001.

[14] K. Boukharouba. *Modélisation et classification de comportements dynamiques des systèmes hybrides*. PhD thesis, Université Lille 1, 2011.

[15] K. Boukharouba, L. Bako, and S. Lecoeuche. Identification of piecewise affine systems based on dempster-shafer theory. In *IFAC Symposium on System Identification, Saint Malo, France*, pages 1662–1667, 2009.

[16] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[17] L. Breiman. Hinging hyperplanes for regression, classification and function approximation. *IEEE Transactions on Information Theory*, 39:999–1013, 1993.

[18] A. M. Bruckstein, D. L. Donoho, and M. Elad. From sparse solutions of systems of equations to sparse modeling of signals and images. *SIAM Review*, 51:34–81, 2009.

[19] E. Candès and P. A. Randall. Highly robust error correction by convex programming. *IEEE Transactions on Information Theory*, 54:2829–2840, 2006.

[20] E. J. Candès, M. Rudelson, T. Tao, and R. Vershynin. Error correction via linear programming. In *Proceedings of the 46th Annual IEEE Symposium on Foundations of Computer Science (FOCS), 295-308.*, 2005.

[21] E. J. Candès, M. Wakin, and S. Boyd. Enhancing sparsity by reweighted ℓ_1 minimization. *Journal Fourier Analysis and Applications*, 14:877–905, 2008.

[22] D. Chen, L. Bako, and Lecoeuche. The minimum-time problem for discrete-time linear systems: A non-smooth optimization approach. In *IEEE International Conference on Control Applications*, 2012.

[23] D. Chen, L. Bako, and S. Lecoeuche. A recursive sparse learning method: Application to jump markov linear systems. In *18th IFAC World Congress, Milano, Italy*, 2011.

[24] D. Chen, L. Bako, and S. Lecoeuche. Estimation récursive et robuste en présence d'erreurs éparses inconnues. *Journal Européen des Systèmes Automatisés*, 46:763–777, 2012.

[25] P. L. dos Santos, T. P. Azevedo-Perdicoulis, C. Novara, J. A. Ramos, and D. E. R. (Eds). *Linear Parameter-Varying System Identification: New Developments and Trends*. World Scientific Publishing, 2012.

[26] R. O. Duda and P. E. Hart. *Pattern classification and scene analysis*. New York, Wiley, 1973.

[27] G. Ferrari-Trecate. Hybrid identification toolbox (HIT), 2005.

[28] G. Ferrari-Trecate, M. Muselli, D. Liberati, and M. Morari. A clustering technique for the identification of piecewise affine systems. *Automatica*, 39:205–217, 2003.

[29] A. Garulli, S. Paoletti, and A. Vicino. A survey on switched and piecewise affine system identification. In *IFAC Symposium on System Identification, Brussels, Belgium*, 2012.

[30] M. Gevers. Identification for control: From the early achievements to the revival of experiment design. *European Journal of Control*, 11:1–18, 2005.

[31] G. C. Goodwin and K. S. Sin. *Adaptive Filtering Prediction and Control*. Dover Publications, Inc. New York, NY, USA, 2009.

[32] M. Grant and S. Boyd. CVX: Matlab software for disciplined convex programming, version 1.2, june 2009, (build 711). June 2009, Build 711.

[33] P. J. Huber. The place of l_1-norm in robust estimation. *Computational Statistics and Data Analysis*, 5:255–262, 1987.

[34] P. J. Huber and E. M. Ronchetti. *Robust Statistics*. A. John Wiley & Sons, Inc. Publication (2nd Ed), 2009.

[35] A. Juditsky, H. Hjalmarsson, A. Benveniste, B. Delyon, L. Ljung, J. Sjoberg, , and Q. Zhang. Nonlinear black box models in system identification: Mathematical foundations. *Automatica*, 31:1725–1751, 1995.

[36] A. L. Juloski, S. Weiland, and W. Heemels. A bayesian approach to identification of hybrid systems. *IEEE Transactions on Automatic Control*, 50:1520–1533, 2005.

[37] L. Ljung. *System Identification: Theory for the user (2nd Ed.)*. PTR Prentice Hall., Upper Saddle River, USA, 1999.

[38] L. Ljung. *System Identification Toolbox User's Guide*. 7th ed. Natick, MA:The MathWorks Inc., 2009.

[39] R. A. Maronna, R. D. Martin, and V. J. Yohai. *Robust Statistics: Theory and Methods*. John Wiley & Sons, Inc., 2006.

[40] K. Mitra, A. Veeraraghavan, and R. Chellappa. Analysis of sparse regularization based robust regression approaches. *IEEE Transactions on Signal Processing*, 61:1249–1257, 2013.

[41] H. Nakada, K. Takaba, and T. Katayama. Identification of piecewise affine systems based on statistical clustering technique. *Automatica*, 41:905–913, 2005.

[42] B. K. Natarajan. Sparse approximate solutions to linear systems. *SIAM Journal on Computing*, 24:227–234, 1995.

[43] H. Ohlsson and L. Ljung. Identification of switched linear regression models using sum-of-norms regularization. *Automatica*, 49:1045–1050, 2013.

[44] N. Ozay and M. Sznaier. Hybrid system identification with faulty measurements and its application to activity analysis. In *IEEE Conference on Decision and Control and European Control Conference, Orlando, FL, USA*, 2011.

[45] S. Paoletti, A. Juloski, G. Ferrari-Trecate, and R. Vidal. Identification of hybrid systems: A tutorial. *European Journal of Control*, 13:242–260, 2007.

[46] M. Petreczky and L. Bako. On the notion of persistence of excitation for linear switched systems. In *IEEE Conference on Decision and Control and European Control Conference, Orlando, FL, USA*, 2011.

[47] M. Petreczky, L. Bako, and S. Lecoeuche. Minimality and identifiability of switched arx systems. In *IFAC Symposium on System Identification, Brussels, Belgium*, 2012.

[48] M. Petreczky, L. Bako, and J. H. van Schuppen. Identifiability of discrete-time linear switched systems. In *Proceedings of the 13th ACM international conference on Hybrid systems: computation and control, Stockholm, Sweden*, pages 141–150, 2010.

[49] M. Petreczky, L. Bako, and J. H. van Schuppen. Realization theory of discrete-time linear switched systems. *Automatica*, 49:3337–3344, 2013.

[50] G. Pillonetto, F. Dinuzzo, T. Chen, G. De Nicolao, and L. Ljung. Kernel methods in system identification, machine learning and function estimation: A survey. *Automatica*, 50:657–682, 2014.

[51] P. Riedinger and J. C. Vivalda. Dynamic output feedback for switched linear systems based on a lqg design. *Automatica*, 54:235–245, 2015.

[52] J. Roll, A. Nazin, and L. Ljung. Nonlinear system identification via direct weight optimization. *Automatica*, 41:475–490, 2005.

[53] P. J. Rousseeuw. Least median of squares regression. *Journal of the American Statistical Association*, 79:871–880, 1984.

[54] P. J. Rousseeuw and A. M. Leroy. *Robust Regression and Outlier Detection*. John Wiley & Sons, Inc., 2005.

[55] B. Schölkopf and A. J. Smola. *Learning with Kernels Support Vector Machines, Regularization, Optimization and Beyond*. MIT Press, Cambridge, MA, 2002.

[56] Y. Sharon, J. Wright, and Y. Ma. Computation and relaxation of conditions for equivalence between ℓ^1 and ℓ^0 minimization. *UIUC Technical Report UILU-ENG-07-2008*, 2007.

[57] Y. Sharon, J. Wright, and Y. Ma. Minimum sum of distances estimator: robustness and stability. In *American Control Conference, St. Louis, Missouri, USA*, 2009.

[58] J. Sjoberg, Q. Zhang, L. Ljung, A. Benveniste, B. Delyon, P. Glorennec, H. Hjalmarsson, and A. Juditsky. Non linear black box modeling in system identification: an unified overview. *Automatica*, 33:1691–1724, 1997.

[59] T. Soderstrom and P. Stoica. *System identification*. Prentice Hall, Upper Saddle River, USA, 1989.

[60] E. D. Sontag. Nonlinear regulation: The piecewise linear approach. *IEEE Transactions on Automatic Control*, 26:346–357, 1981.

[61] J. A. K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle. *Least Squares Support Vector Machines*. World Scientific, 2002.

[62] R. Toth. *Modeling and identification of linear parameter-varying systems*. Springer, 2010.

[63] V. Vapnik. *Satistical Learning theory*. New-York: Wiley, 1998.

[64] Y. Vereshchaga, S. Stadlbauer, L. Bako, and L. del Re. Piecewise affine modeling of nox emission produced by a diesel engine. In *European Control Conference, Zurich, Switzerland*, 2013.

[65] R. Vidal, A. Chiuso, and S. Soatto. Observability and identifiability of jump linear systems. In *Proceedings of the IEEE Conference on Decision and Control, Las Vegas, USA*, volume 4, pages 3614–3619, 2002.

[66] R. Vidal, S. Soatto, Y. Ma, and S. Sastry. An algebraic geometric approach to the identification of a class of linear hybrid systems. In *Proceedings of the IEEE Conference on Decision and Control, Maui, Hawaii, USA*, volume 1, pages 167–172, 2003.

[67] W. Xu, E.-W. Bai, and M. Cho. System identification in the presence of outliers and random noises: A compressed sensing approach. *Automatica*, 50:2905–2911, 2014.

[68] G. Zames. Feedback and optimal sensitivity: Model reference transformations, multiplicative seminorms, and approximate inverses. *IEEE Transactions on Automatic Control*, 26:301–320, 1981.

[69] W. Zhang, A. Abate, J. Hu, and M. Vitus. Exponential stabilization of discrete-time switched linear systems. *Automatica*, 45:2526–2536, 2009.