

Identification of switched linear systems via sparse optimization [★]

Laurent Bako ^{a,b}

^aUniv Lille Nord de France, F-59000 Lille, France.

^bEMDouai, IA, F-59500 Douai, France.

Abstract

The work presented in this paper is concerned with the identification of switched linear systems from input-output data. The main challenge with this 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. To overcome this difficulty, we present here a sparse optimization approach inspired by very recent developments from the community of compressed sensing. We formally pose the problem of identifying each submodel as a combinatorial ℓ_0 optimization problem. This is indeed an NP-hard problem which can interestingly, as shown by recent literature, be relaxed into a (convex) ℓ_1 -norm minimization problem. We present sufficient conditions for this relaxation to be exact. The whole identification procedure allows us to extract the parameter vectors (associated with the different subsystems) one after another without any prior clustering of the data according to their respective generating-submodels. Some simulation results are included to support the potentialities of the proposed method.

Key words: switched linear systems; system identification; sparse optimization; hybrid systems.

1 Introduction

We consider in this paper the problem of identifying a switched linear system from a collection of input-output data. A switched linear system corresponds to the behavior that results from the interaction (mainly switching) between a finite set of linear dynamical subsystems. Examples of such systems arise in many different engineering fields, typically genetic regulatory networks study, air traffic management, nonlinear systems control, manufacturing processes modeling, computer vision, etc. (see e.g., [19,29] for more examples). Mathematically speaking, a switched linear system can be viewed as a relation (model) of the form

$$y(t) = \theta_{\lambda_t}^\top x(t) + e(t), \quad (1)$$

relating a vector $x(t) \in \mathbb{R}^n$ called the regressor vector and a signal $y(t) \in \mathbb{R}$ designated as the output of the

system being modeled at time t . Here, $\lambda_t \in \{1, \dots, s\}$ is the discrete state or the discrete mode, i.e., the index of the active submodel at time t and $\theta_{\lambda_t} \in \mathbb{R}^n$ is the associated parameter vector (PV). The sequence $\{e(t)\}$ of errors refers to potential mismatch or noise; it is assumed to be bounded. In a general situation, the vector $x(t)$ appearing in (1) need not be structured but when dealing with the input-output behavior of switched dynamical systems, it sometimes takes the form

$$x(t) = \begin{bmatrix} y(t-1) & \dots & y(t-n_a) \\ u(t-1)^\top & \dots & u(t-n_b)^\top \end{bmatrix}^\top, \quad (2)$$

where $u(t) \in \mathbb{R}^{n_u}$ and $y(t) \in \mathbb{R}$ are respectively the input and output of the considered system, n_a and n_b are its orders. The dimension of $x(t)$ is therefore $n = n_a + n_b n_u$ and the model (1) is designated as a Switched Auto-Regressive eXogenous (SARX) model.

1.1 The switched system identification problem

Given observations $\{x(t), y(t)\}_{t=1}^N$ generated by a switched linear model of the form (1), with $x(t)$ defined

[★] This paper was not presented at any IFAC meeting. Corresponding author L. Bako. Tel. +33 327 712 127. Fax +33 327 712 917.

Email address: laurent.bako@mines-douai.fr (Laurent Bako).

as in (2), we are interested here in estimating the parameter vectors $\{\theta_j\}_{j=1}^s$.

We start by recalling from [31] that the problem of inferring a switched model such as (1) from a set of finite measurements, admits multiple solutions so that the identification problem is not well-posed. If the structural indexes 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 dataset. 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 the data. In order to remove the identifiability issue, we will assume in this paper that the orders n_a and n_b are finite, equal for all submodels and known a priori. With this setting for the structural indexes n_a and n_b , the SARX of interest here will be viewed as the one that, among all switched linear models consistent with the data, has a number of submodels that is as small as possible. The interested reader is referred to the paper [26] for a more complete treatment of the identifiability problem in the framework of switched linear state space models.

1.2 Prior work

During the last ten years a number of interesting results have been achieved in the field of hybrid system identification. Examples of such works include, in the case of switched linear models, the algebraic-geometric method [32,21,30], the product-of-errors based method [18]. Other methods such as the mixed integer programming approach [27], the bounded-error approach [3], the bayesian learning based procedure [17], the clustering-based strategies [13,14,22,4] apply to piecewise affine systems, i.e., particular switched linear/affine systems where the switching surfaces are the faces of a set of non-overlapping polyhedra. An excellent survey can be found in [25] where most of the methods developed prior to 2007 have been summarized. Despite the clear merits of all these pioneering contributions, one can fairly observe that the subject of hybrid system identification is still open on many challenging issues such as computational complexity reduction, optimality and convergence analysis of the proposed methods. Recently, a promising idea has emerged as to what extent some results from sparse optimization based signal recovery can be applied to hybrid system identification. This idea may indeed be an expedient for tackling simultaneously both the clustering and estimation problems that are inherent to hybrid system computation. The work of Ozay et al. [24] exploits successfully such an approach. The identification of the parameter vectors is formulated as the problem of recovering a sparse vector-valued sequence, the instances of which sequence are subsequently agglomerated to reach a minimum number of submodels. The work of Elhamifar and Vidal [12] also suggests sparse representation as a possible alternative for solving the

problem of subspace clustering. More precisely, the authors of [12] consider the problem of estimating bases for a set of linear/affine subspaces from data lying in the union of these subspaces. A limitation of their work however is that the mixed subspaces need to be linearly independent, an assumption which is violated when dealing with the union of more than one hyperplanes.

1.3 Contributions of this paper

The contribution of the paper consists in the development of a new 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 [9]. 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. A number of results from the field of compressed sensing [7,8,10] can be insightful for this purpose.

1.4 Outline of this paper

The remainder of the paper is organized as follows. We start by presenting in Section 2 the main mathematical terminology used in the paper. We then describe in Section 3 the proposed algorithm for the identification of switched linear systems with arbitrary switchings. Section 4 contains some numerical results that confirm the potential of our method. Section 5 concludes the paper.

2 Mathematical preliminaries

In this preliminary section we introduce some mathematical concepts and notations that will be extensively used throughout the paper. We first introduce a notion of k -genericity index.

Definition 1 For a given data matrix

$X = [x(1) \dots x(N)] \in \mathbb{R}^{n \times N}$ with $n \leq N$, and for any

integer k verifying $0 < k \leq \text{rank}(X)$, we define the k -genericity index $\nu_k(X)$ of X to be the minimum integer m such that any $n \times m$ submatrix of X has rank k :

$$\nu_k(X) = \min \left\{ m : \forall (t_1, \dots, t_m) \text{ with } t_i \neq t_j \text{ for } i \neq j, \right. \\ \left. \text{rank} [x(t_1) \cdots x(t_m)] = k \right\}. \quad (3)$$

If $k > \text{rank}(X)$, we set by convention $\nu_k(X) = +\infty$ and if $k = 0$ we set $\nu_0(X) = 0$ for all X .

For an overview on the function $\nu_k(\cdot)$, we quickly mention the following two obvious properties.

- (1) If $\nu_k(X) = k$ then $\nu_p(X) = p$ for all $p \leq k$.
- (2) For any $k \leq \text{rank}(X)$, it holds that $k \leq \nu_k(X) \leq N$.

Observe additionally that when the data $\{x(t)\}_{t=1}^N$ are in *general position*, i.e., when any subset $\{x(t_1), \dots, x(t_n)\}$ of n data vectors are linearly independent, we have $\nu_n(X) = n$. Hence, the number $\nu_n(X)$ characterizes a property of richness (or linear independence) of the columns of X . For $\nu_n(X)$ to be finite, we need to assume that $\text{rank}(X) = n$. It will be so in all the paper. Other possible indexes for measuring linear independence between the columns of a matrix are the so-called *spark* and *mutual coherence* whose formal definitions are recalled below from [6] and [11].

Definition 2 The *spark* of a given matrix X denoted $\text{spark}(X)$, is the smallest number σ such that there exists a set of σ columns of X that are linearly dependent. In fact we have

$$\text{spark}(X) = \min_{\substack{z \in \ker(X) \\ z \neq 0}} \|z\|_0 \quad (4)$$

where $\ker(X)$ refers to the kernel subspace of matrix X and $\|z\|_0$ stands for the number of nonzero entries in z . If $\ker(X) = \{0\}$, then $\text{spark}(X)$ will be conventionally set to $+\infty$.

Definition 3 The *mutual coherence* of a given matrix $X = [x(1), \dots, x(N)]$ denoted $\mu(X)$, is the largest absolute value of the cosine between different columns of X . More precisely,

$$\mu(X) = \max_{\substack{1 \leq t, k \leq N \\ t \neq k}} \frac{|x(t)^\top x(k)|}{\|x(t)\|_2 \cdot \|x(k)\|_2} \quad (5)$$

where, without loss of generality¹, it is assumed that all the columns of X are nonzero.

¹ One can always remove all zero columns from X .

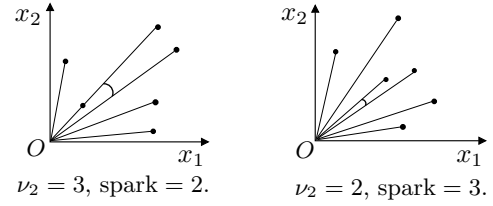


Figure 1. Illustration of the k -genericity index, the spark and the mutual coherence for the case $n = 2$. The data are viewed as vectors originating from the origin O . The mutual coherence corresponds to the cosine of the smallest angle between two pairs of data vectors. For example, the mutual coherence of the data represented on the left is smaller than the one of the data on the right.

Note that the n -genericity index, the spark and the mutual coherence (see Figure 1 for a graphical illustration) are all some measures of how rich (we will say also generic) the columns of the data matrix $X = [x(1) \cdots x(N)]$ are. For example, the smaller the index $\nu_n(X)$ is, the richer the data contained in X are. Similarly, small mutual coherence (or large spark) indicates highly linearly independent columns. Given a data matrix X , we may, due to the extensive use of the genericity index and the spark throughout the paper, be interested in evaluating $\nu_n(X)$ or $\text{spark}(X)$. However, using directly Definitions 1 and 2 to this purpose seems computationally difficult. We present in Lemma 4 an alternative sufficient condition for checking whether $\nu_n(X) = n$ or not. Lemma 5 gives a rough idea of how large is the spark by looking at the mutual coherence, which is much easier to compute.

Lemma 4 Assume that all the columns of $X \in \mathbb{R}^{n \times N}$ are nonzero. Then for any integer p satisfying $p < 1 + \frac{1}{\mu(X)}$, it holds that $\nu_p(X) = p$. In particular, if $n < 1 + \frac{1}{\mu(X)}$ then $\nu_n(X) = n$.

PROOF. Let \tilde{X} be defined from X by normalizing its columns with ℓ_2 norm, i.e., the columns of \tilde{X} are defined as $\tilde{x}(t) = x(t)/\|x(t)\|_2$. Such an operation does not affect the rank properties of X so that \tilde{X} and X have the same genericity indexes. Our method of proof follows similar arguments as in a lemma stated in [6]. Let I be a subset of $\{1, \dots, N\}$ with cardinality $|I| = p$. Denote by \tilde{X}_I the matrix formed with the columns of \tilde{X} whose indexes are in I and define $G_I = \tilde{X}_I^\top \tilde{X}_I$. According to the Gershgorin disk theorem [16], if γ is an eigenvalue of G_I , then there is an $i \in \{1, \dots, p\}$ such that

$$|\gamma - g_{ii}| \leq \sum_{\substack{k=1 \\ k \neq i}}^p |g_{ik}|,$$

where the g_{ik} are the entries of G_I . From the definition of G_I , we know that $g_{ii} = \tilde{x}(i)^\top \tilde{x}(i) = 1$ for all i and $|g_{ik}| = |\tilde{x}(i)^\top \tilde{x}(k)| \leq \mu(X)$ for all (i, k) verifying $i \neq k$. We can hence write, for any eigenvalue γ of G_I ,

$$|\gamma - 1| \leq (p - 1)\mu(X),$$

which is equivalent to

$$-(p - 1)\mu(X) + 1 \leq \gamma \leq (p - 1)\mu(X) + 1.$$

Therefore if $-(p - 1)\mu(X) + 1 > 0$ or equivalently, if $p < 1 + 1/\mu(X)$, all the eigenvalues of G_I are strictly positive. In other words G_I is positive definite and $\text{rank}(\tilde{X}_I) = \text{rank}(X_I) = p$. We have hence shown that any subset of p columns of X has rank p provided $p < 1 + 1/\mu(X)$. This is equivalent to the claim of the lemma.

The following lemma is a restatement of Lemma 4 and will be of interest in the next developments.

Lemma 5 ([11]) *For any matrix $X \in \mathbb{R}^{n \times N}$ it holds that*

$$1 + \frac{1}{\mu(X)} \leq \text{spark}(X). \quad (6)$$

3 The proposed solution to the identification problem

In this section we present the main results of the paper. We begin with a short overview on the switched system identification problem. The switched system identification problem can be posed as the problem of finding a set $\{\theta_j\}_{j=1}^s \subset \mathbb{R}^n$ of parameter vectors and an associated switching sequence $\{\lambda_t\}_{t=1}^N \subset \{1, \dots, s\}$ by solving the optimization problem

$$\min_{\theta_1, \dots, \theta_s; \lambda_1, \dots, \lambda_N} \sum_{t=1}^N (y(t) - \theta_{\lambda_t}^\top x(t))^2 \quad (7)$$

which, by eliminating the discrete mode, is equivalent to

$$\min_{\theta_1, \dots, \theta_s} \sum_{t=1}^N \min_{i=1, \dots, s} (y(t) - \theta_i^\top x(t))^2. \quad (8)$$

Finding an optimal solution to this problem may involve searching exhaustively over the set $\{1, \dots, s\}^N$ of all possible discrete state paths, an ideal path for which there is a set of corresponding PVs that fits the data. However, such a procedure is computationally very hard. Therefore, a driving principle of many approaches is to first overcome the combinatorial nature of the problem. The algebraic-geometric method [21] eliminates it by transforming the identification problem into the one of

fitting the data with one homogeneous polynomial. In the same vein, a regularized product-of-errors criterion is formed in [18] and optimized through continuous non-linear optimization techniques. The recursive approach discussed in [1] alternates between data assignment to submodels and parameter vector update. The approach of the present paper is based on a formulation of the switched system identification as a sparse optimization problem which can subsequently be relaxed into a convex program.

3.1 The method

Given the data $\{(x(t), y(t))\}_{t=1}^N$ generated by the system (1), we consider the error vector

$$\phi(\theta) = \mathbf{y} - X^\top \theta - \mathbf{e} \quad (9)$$

where $\mathbf{y} = [y(1) \dots y(N)]^\top$, $\mathbf{e} = [e(1) \dots e(N)]^\top$ and $X = [x(1) \dots x(N)]$. Let us denote with N_i the number of data $(x(t), y(t))$ generated exclusively by the subsystem indexed by i . Then we can observe that if $\theta = \theta_i$ for some $i \in \{1, \dots, s\}$, then $\phi(\theta)$ is a sparse vector, i.e., a vector where many entries are equal to zero. More precisely, $\phi(\theta)$ contains exactly N_i zero entries and $N - N_i$ nonzero entries.

In order to avoid any ambiguity in the definition of the number N_i , we make the following formal assumption throughout all the paper.

Assumption 6 *There is no data pair $(x(t), y(t))$ that fits two different submodels of the SARX (1), i.e., $y(t) = \theta_i^\top x(t) + e(t) = \theta_j^\top x(t) + e(t) \implies i = j$.*

The method to be presented relies on exploiting the sparsity of the vector $\phi(\theta)$ when $\theta \in \mathbb{R}^n$ is a PV associated with a submodel of (1). For the sake of clarity, assume for now that the noise sequence $\{e(t)\}$ is identically null and introduce the notations

$$\bar{\theta}_i = \begin{bmatrix} 1 & \theta_i^\top \end{bmatrix}^\top \quad \text{and} \quad \bar{x}(t) = \begin{bmatrix} y(t) & -x(t)^\top \end{bmatrix}^\top. \quad (10)$$

Then for any time instant t , there is $i \in \{1, \dots, s\}$ 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 $\bar{\theta}_i$, $i = 1, \dots, s$. 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 dataset. 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.

The vector $\phi(\theta)$ defined in (9) can be interpreted as the projections of the data $\bar{x}(t)$ onto a given vector $\bar{\theta} = [1 \ \theta^\top]^\top \in \mathbb{R}^{n+1}$,

$$\phi(\theta) = \begin{bmatrix} \bar{x}(1)^\top \bar{\theta} \\ \vdots \\ \bar{x}(N)^\top \bar{\theta} \end{bmatrix}. \quad (11)$$

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

$$\min_{\theta} \|\phi(\theta)\|_0, \quad (12)$$

where $\|z\|_0$ denotes the ℓ_0 norm² of z , that is, the number of nonzero entries of z , $\|z\|_0 = |\{i : z_i \neq 0\}|$. Trying to solve problem (12) 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.

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 (12) is a PV representing one of the constituent submodels of system (1).

Lemma 7 *Assume that $N_i \geq s\nu_n(X)$ for all i , where s is the number of submodels in (1). If Assumption 6 holds, then*

$$\arg \min_{\theta} \|\phi(\theta)\|_0 = \theta_{i_o} \quad (13)$$

where $i_o \in \{i : N_i \geq N_j \ \forall j = 1, \dots, s\}$ is one of the indices of submodels that has generated the most number of data.

PROOF. We just need to show that the solution $\theta \neq 0$ to (12) lies in $\{\theta_1, \dots, \theta_s\}$, the true set of PVs. Suppose by contradiction that this is not the case. Then θ achieves a sparser³ error vector $\phi(\theta)$ than all the θ_i 's. In other words, if we let $I(\theta) = \{t : \bar{x}(t)^\top \bar{\theta} = 0\}$, then $|I(\theta)| \geq N_i \geq s\nu_n(X)$ for all $i = 1, \dots, s$, where $|I(\theta)|$ is the cardinality of $I(\theta)$. Denote with n_i the number of data generated by submodel i , and whose indices are contained in $I(\theta)$. If we had $n_i < \nu_n(X)$ for all i , then we would get $|I(\theta)| = \sum_{i=1}^s n_i < s\nu_n(X)$ which would

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.

³ We say that a vector x is sparser than another vector y of the same dimension if $\|x\|_0 \leq \|y\|_0$.

clearly violate one of the Lemma's assumptions. Therefore there is an index j such that $n_j \geq \nu_n(X)$, i.e., $I(\theta)$ contains at least $\nu_n(X)$ indices of data vectors $\bar{x}(t)$ generated by the same submodel j . Such data vectors form a matrix $\mathcal{X}_j = [\bar{x}(t_1^j) \ \dots \ \bar{x}(t_{n_j}^j)]^\top$ in $\mathbb{R}^{n_j \times (n+1)}$. It follows that both $\bar{\theta}$ and $\bar{\theta}_j$ lie in the null space of \mathcal{X}_j , which is one dimensional by evoking the fact that $n_j \geq \nu_n(X)$. As a consequence, $\bar{\theta}$ can be written as $\bar{\theta} = \lambda \bar{\theta}_j$ with λ a nonzero scalar. Since by definition, the first entries of $\bar{\theta}$ and $\bar{\theta}_j$ are equal to 1, we have necessarily $\lambda = 1$, $\bar{\theta} = \bar{\theta}_j$ and hence $\theta = \theta_j$, which contradicts the initial thesis.

While Lemma 7 says that the set of parameter vectors minimizing $\|\phi(\theta)\|_0$ is included in $\{\theta_1, \dots, \theta_s\}$, it does not give any further information on whether the minimizer can be unique or not. Next, we characterize the uniqueness of the minimizer of (12) in terms of the n -genericity index of the data matrix X .

Theorem 8 *If there is a vector θ satisfying*

$$\|\phi(\theta)\|_0 \leq \frac{N - \nu_n(X)}{2}, \quad (14)$$

then θ is necessarily the unique vector that achieves the sparsest possible error $\phi(\theta)$.

If in addition, Assumption 6 holds and $N \geq (2s - 1)\nu_n(X)$, then $\theta \in \{\theta_1, \dots, \theta_s\}$.

PROOF. We proceed by contradiction. Assume that there is $\gamma \in \mathbb{R}^n$, $\gamma \neq \theta$, achieving an error $\phi(\gamma)$ that is at least as sparse as $\phi(\theta)$, i.e., $\|\phi(\gamma)\|_0 \leq \|\phi(\theta)\|_0 \leq \frac{N - \nu_n(X)}{2}$. By introducing the notation $I(\theta) = \{t : \bar{x}(t)^\top \bar{\theta} = 0\}$, we can write

$$\begin{aligned} |I(\theta)| &= N - \|\phi(\theta)\|_0 \geq \frac{N + \nu_n(X)}{2} \\ |I(\gamma)| &= N - \|\phi(\gamma)\|_0 \geq \frac{N + \nu_n(X)}{2}. \end{aligned}$$

Using these inequalities, we can now bound the cardinality of $I(\theta) \cap I(\gamma)$ as follows

$$\begin{aligned} |I(\theta) \cap I(\gamma)| &= |I(\theta)| + |I(\gamma)| - |I(\theta) \cup I(\gamma)| \\ &\geq N + \nu_n(X) - |I(\theta) \cup I(\gamma)| \\ &\geq \nu_n(X) \end{aligned} \quad (15)$$

because $N \geq |I(\theta) \cup I(\gamma)|$. Let H designate the matrix formed with the vectors $\bar{x}(t) \in \mathbb{R}^{n+1}$ whose indices t are in $I(\theta) \cap I(\gamma)$. It follows from (15) and the definition of $\nu_n(X)$ that $\text{rank}(H) = n$ and both $\bar{\theta}$ and $\bar{\gamma}$ lie in the nullspace of H^\top which is of dimension one. This, together with the fact that the first entries of $\bar{\theta}$ and $\bar{\gamma}$

are all equal to one (see e.g., Eq. (10) for the definition of $\bar{\theta}$ from θ), lead to $\bar{\theta} = \bar{\gamma}$, which in turn implies that $\theta = \gamma$. We thus obtain a contradiction and so, the claim of the theorem holds.

The proof of the second statement follows similar steps as the proof of Lemma 7. Using the same notations, if n_i were strictly less than $\nu_n(X)$ for all $i = 1, \dots, s$, then it would hold that $1/2(N + \nu_n(X)) \leq |I(\theta)| = \sum_{i=1}^s n_i < s\nu_n(X)$. It is easy to see that this is incompatible with the assumption $N \geq (2s - 1)\nu_n(X)$. There is therefore at least one j such that $n_j > \nu_s(X)$. We can now follow the same line of arguments as in the proof of Lemma 7 to conclude that θ is necessarily in $\{\theta_1, \dots, \theta_s\}$.

3.2 Relaxation of the ℓ_0 problem through the basis pursuit method

Note however that the problem (12) is a hard non-convex optimization problem which is NP-hard in general, see e.g., [23]. As a consequence, minimizing directly the cost function in (12) is in general intractable. A popular alternative [6,9] is to consider a convex relaxation of problem (12) based on the ℓ_1 norm. This relaxation strategy is known as the *basis pursuit (BP)* method and leads to the problem

$$\min_{\theta} \|\phi(\theta)\|_1, \quad (16)$$

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 [28] and [8]. Contrary to the problem (12), the convex problem (16) can be transformed into a classical linear program which is efficiently solvable by standard convex optimization techniques [5].

A first natural question is, under which conditions solving the convex problem (16) can lead to the solution of the combinatorial problem (12). Second, can uniqueness be guaranteed for that solution. In order to investigate these important questions, we will slightly reformulate the optimization problem (16). The objective is to recast it as a more standard problem known as sparse signal representation to which some results from the literature of compressed sensing [6,11] might apply. Multiply Eq. (9) by the orthogonal projection matrix

$$P_X = I_N - X^\top (X X^\top)^{-1} X. \quad (17)$$

Because $P_X X^\top = 0$, this yields⁴

$$P_X \mathbf{y} = P_X \phi(\theta).$$

⁴ Remember that the noise vector \mathbf{e} is assumed to be zero for now.

For the purpose of the analysis to be presented, we now set $\mathbf{z} = \phi(\theta)$ and replace the problem (16) with the following constrained one,

$$\begin{aligned} & \min_{\theta, \mathbf{z}} \|W_X \mathbf{z}\|_1 \\ & \text{subject to } P_X \mathbf{y} = P_X \mathbf{z} \\ & \quad \quad \quad \mathbf{z} = \mathbf{y} - X^\top \theta \end{aligned} \quad (18)$$

where we have introduced a weighting matrix $W_X = \text{diag}(\|p(1)\|_2, \dots, \|p(N)\|_2)$ with the $p(i)$ referring to columns of P_X . It is assumed here that P_X has no zero column so that W_X is a positive definite diagonal matrix. Hence the problem (18) above can be viewed as an ℓ_1 relaxation of the weighted ℓ_0 problem $\min_{\theta} \|W_X \phi(\theta)\|_0$ which, thanks to the nonzero scale-invariance of the ℓ_0 norm, is strictly equivalent to (12). Notice also that the role of W_X in (18) is to normalize the columns of P_X .

Remark 9 Note that the (N, N) -matrix P_X in (17) can be replaced by any full row rank matrix $P \in \mathbb{R}^{(N-n) \times N}$ spanning the orthogonal complement of the column space $\text{im}(X^\top)$ of X^\top . Preferably, such a matrix should be selected such that it has orthogonal rows (i.e., satisfying $P P^\top = I$). The above matrix W_X needs then to be defined from the columns of P .

In the following we present sufficient conditions for the convex optimization problem (18) to uniquely recover the minimizer of (12). We finally apply these results in Theorem 14 to the special context of switched linear system identification and derive sufficient conditions for determining exactly all the PVs by means of ℓ_1 norm minimization.

Theorem 10 *If there is a vector θ achieving an error $\phi(\theta)$ such that*

$$\|\phi(\theta)\|_0 < \frac{1}{2} \left(1 + \frac{1}{m(X)} \right), \quad (19)$$

with

$$\begin{aligned} m(X) &= \max_{\substack{1 \leq t, k \leq N \\ t \neq k}} \frac{|M_X(t, k)|}{\sqrt{(1 - M_X(t, t))(1 - M_X(k, k))}} \\ M_X &= X^\top (X X^\top)^{-1} X, \end{aligned} \quad (20)$$

then θ is the unique solution to the ℓ_1 minimization problem (18). In Eq. (20), $|M_X(t, k)|$ stands for the absolute value of the (t, k) -entry of M_X .

PROOF. The theorem follows directly as a consequence of Theorem 7 in [6]. To see this, note that by

letting $z = \phi(\theta)$, problem (18) is equivalent to solving

$$\begin{aligned} & \min_z \|W_X z\|_1 \\ & \text{subject to } P_X \mathbf{y} = P_X z \end{aligned} \quad (21)$$

for z and then computing θ from the linear equation $z = \mathbf{y} - X^\top \theta$. With $\text{rank}(P_X) = N - n < N$, this corresponds to a problem studied in [6]. According to Theorem 7 of that paper, the vector z can be uniquely recovered if $\|z\|_0 = \|\phi(\theta)\|_0 < 1/2(1 + 1/\mu(P_X))$ where $\mu(P_X)$ denotes the mutual coherence of P_X , see Definition 3. We are now left with showing that $m(X) = \mu(P_X)$, which follows by straightforward calculations.

Theorem 11 *If there is a vector θ obeying the condition (19), then θ is the unique solution to both problems (12) and (18).*

For the proof of Theorem 11, we need the following lemma.

Lemma 12 *The orthogonal projection matrix P_X defined in (17) satisfies*

$$\text{spark}(P_X) = N - \nu_n(X) + 1. \quad (22)$$

PROOF. From the definition (4) of the spark, we know that

$$\begin{aligned} \text{spark}(P_X) &= \min_{\substack{z \in \ker(P_X) \\ z \neq 0}} \|z\|_0 \\ &= \min_{\substack{z \in \text{im}(X^\top) \\ z \neq 0}} \|z\|_0. \end{aligned}$$

The second equality follows from the fact that $\ker(P_X) = \text{im}(X^\top)$. To prove (22), we will, in view of the previous equalities, just show that

$$N - \nu_n(X) + 1 = \min_{\substack{z \in \text{im}(X^\top) \\ z \neq 0}} \|z\|_0.$$

By definition of the n -genericity index $\nu_n(X)$, there is a subset $\{t_1, \dots, t_{q-1}\}$ of $\{1, \dots, N\}$, with $q = \nu_n(X) \geq n$, such that the rank of $[x(t_1), \dots, x(t_{q-1})]$ is strictly less than n . As a consequence, we can find $w^* \in \mathbb{R}^n$, $w^* \neq 0$, such that $x(t_1)^\top w^* = \dots = x(t_{q-1})^\top w^* = 0$. For such a w^* , we must have $x(t)^\top w^* \neq 0$ for all $t \notin \{t_1, \dots, t_{q-1}\}$. If this were not the case, i.e., if there existed a $t \notin \{t_1, \dots, t_{q-1}\}$ such that $x(t)^\top w^* = 0$, then we would have $w^* = 0$. This follows from the definition of $\nu_n(X)$ which guarantees that the matrix $[x(t_1), \dots, x(t_{q-1}), x(t)]^\top \in \mathbb{R}^{q \times n}$ has full column rank (since it contains $\nu_n(X)$ columns). Let us now take $z^* = X^\top w^* \in \text{im}(X^\top)$. Then, $\|z^*\|_0 = \|X^\top w^*\|_0$ is exactly

equal to $N - (q - 1) = N - \nu_n(X) + 1$. It follows that $N - \nu_n(X) + 1 \geq \min_{\substack{z \in \text{im}(X^\top) \\ z \neq 0}} \|z\|_0 = \text{spark}(P_X)$.

Now let us assume by contradiction that the last inequality is strict. Then there exists at least one $z \in \text{im}(X^\top)$, $z \neq 0$, such that $N - \nu_n(X) + 1 > \|z\|_0$ or equivalently, $N - \|z\|_0 \geq \nu_n(X)$. In words, this means that the number of zero entries in z (which is equal to $N - \|z\|_0$) is greater than $\nu_n(X)$. As $z \in \text{im}(X^\top)$, there is $w \in \mathbb{R}^n$, $w \neq 0$ such that $z = X^\top w$. One can therefore find some time indices k_1, \dots, k_q , with $q \geq \nu_n(X)$ verifying $x(k_i)^\top w = 0$, $i = 1, \dots, q$. However from Definition 1 of $\nu_n(X)$, $q \geq \nu_n(X)$ implies that $\text{rank}([x(k_1) \cdots x(k_q)]) = n$, which in turn leads to the conclusion $w = 0$, contradicting the fact that $z \neq 0$.

PROOF. [Proof of Theorem 11] From Lemma 5 we know that $1 + \frac{1}{\mu(P_X)} \leq \text{spark}(P_X)$. We have also seen in the proof of Theorem 10 that $\mu(P_X) = m(X)$ and in Lemma 12 that $\text{spark}(P_X) = N - \nu_n(X) + 1$. This leads to $1 + \frac{1}{m(X)} \leq N - \nu_n(X) + 1$. Hence if there is a vector θ obeying condition (19), i.e. if there is θ verifying $\|\phi(\theta)\|_0 < \frac{1}{2} \left(1 + \frac{1}{m(X)}\right)$, then it also holds that $\|\phi(\theta)\|_0 < \frac{1}{2} (N - \nu_n(X) + 1)$. Since $\|\phi(\theta)\|_0$ is an integer, it can then be verified that $\|\phi(\theta)\|_0 \leq \frac{N - \nu_n(X)}{2}$ whether $N - \nu_n(X) + 1$ is odd or even. Now, the claim of Theorem 11 follows by direct application of Theorems 8 and 10.

Remark 13 *If instead of the specific matrix P_X we use an arbitrary full row rank matrix $P \in \mathbb{R}^{(N-n) \times N}$ verifying $PX^\top = 0$, then Theorems 10 and 11 still hold with the number $m(X)$ appearing in (19) replaced by $\mu(P)$. Also the equality (22) is still true if P_X is replaced by P .*

Theorem 11 says that if there is a vector θ producing a sufficiently sparse error $\phi(\theta)$, then (18) is equivalent to (12) in the sense that they have the same (unique) solution. Since (18) is convex, we can therefore recover exactly the solution to the combinatorial problem (12) by means of convex optimization.

Without loss of generality, we can assume in this subsection that the subsystems of system (1) are indexed in such a way that $N_1 \geq N_2 \geq \dots \geq N_s$. Define $X_1 = X$, and for any $j = 2, \dots, s$, 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 $N = N_1 + \dots + N_s$ columns, X_2 contains $N - N_1$ columns, X_3 contains $N - N_1 - N_2$ columns and so forth. With these notations, we present below an immediate corollary to Theorem 11, which is relevant to the linear switched identification problem.

Theorem 14 Consider the data matrix $X \in \mathbb{R}^{n \times N}$ generated by the SARX system (1) and assume that

$$\begin{aligned} N_1 &> N - \vartheta(X_1) > 0, \\ N_2 &> N - N_1 - \vartheta(X_2) > 0, \\ &\vdots \\ N_{s-1} &> N - N_1 - \dots - N_{s-2} - \vartheta(X_{s-1}) > 0, \end{aligned} \quad (23)$$

where $\vartheta(X_j) = 1/2(1 + 1/m(X_j))$ with $m(X_j)$ defined as in (20) for all $j = 1, \dots, s$. Then all the parameter vectors $\{\theta_1, \dots, \theta_s\}$ can be extracted one after another by solving ℓ_1 minimization problems of the form (18).

PROOF. Consider the whole dataset $X_1 = X \in \mathbb{R}^{n \times N}$. Then the inequality $N_1 > N - \vartheta(X_1)$, is equivalent to $\|\phi(\theta_1)\|_0 = N - N_1 < \vartheta(X_1)$, which by Theorem 11, implies that θ_1 solves (uniquely) both (12) and (18). The result follows by applying the same reasoning to X_2, \dots, X_{s-1} . Finally, X_s contains data generated only by the submodel with index s ; the PV θ_s of this submodel can therefore be immediately obtained by ℓ_1 or ℓ_2 minimization.

As it turns out, the main difficulty in applying the basis pursuit method to computing the submodels of a hybrid system is that it may happen that none of the submodels achieves a sufficiently sparse error vector $\phi(\theta)$. In other words, the conditions of Theorem 14 may not be easy to meet on arbitrary datasets. We therefore need, in more general cases, to find a way of increasing the possibility to effectively obtain, by convex ℓ_1 optimization, the solution θ which yields the sparsest error vector $\phi(\theta)$.

3.3 Improving sparsity

When the PVs θ_i do not realize sufficiently sparse errors $\phi(\theta_i)$, we may increase the capacity of the BP method to still recover them by instead solving a weighted variant

$$\min_{\theta} \|WW_X \phi(\theta)\|_1 \quad (24)$$

of problem (16) and (18), where $W = \text{diag}(w_1, \dots, w_N)$ is a weighting diagonal matrix with elements $w_t \geq 0$. While the role of the weighting matrix $W_X > 0$ defined in (18) is to compensate potential differences of magnitude in the columns of P_X ⁵, $W \geq 0$ in (24) designates an adjustable weighting matrix which is intended, if it is appropriately chosen, for reinforcing the sparsity-promoting ability of the ℓ_1 norm. To see this, suppose for example that we are seeking to estimate a particular PV $\theta = \theta_i$ for some $i = 1, \dots, s$. If by some means we could know (at least approximately) the entire discrete state

⁵ In what follows, we will not make W_X appear explicitly.

sequence $\{\lambda_t\}_{t=1}^N$, then by setting $w_t = 1$ when $\lambda_t = i$ and $w_t = 0$ when $\lambda_t \neq i$, $\theta = \theta_i$ could be recovered from (24). Hence, by appropriately adjusting the weights one can encourage the obtention of the sparsest possible error vector $\phi(\theta)$. Although the discrete state sequence is not known, the weights can be iteratively approached by solving a sequence of convex ℓ_1 norm optimization problems of the form (24). This idea is supported by the results of [9] which argue that sparsity of the solution can be enhanced through reweighting ℓ_1 optimization. For the sake of completeness we recall, with a slight adaptation to our setting, the reweighted ℓ_1 optimization technique [9] in Algorithm 1 (see below).

Algorithm 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

$$\theta^{(r)} = \arg \min_{\theta} \|W^{(r)} \phi^o(\theta)\|_1$$

where $\phi^o(\theta)$ is an ℓ_2 -normalized version of $\phi(\theta)$ defined as

$$\phi^o(\theta) = \begin{bmatrix} \frac{\bar{x}(1)^\top \bar{\theta}}{\|\bar{x}(1)\|_2} & \dots & \frac{\bar{x}(N)^\top \bar{\theta}}{\|\bar{x}(N)\|_2} \end{bmatrix}^\top.$$

(2) Update the weights as

$$w_t^{(r+1)} = \frac{1}{|\phi_t^o(\theta^{(r)})| + \varepsilon}, \quad t = 1, \dots, N$$

(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 $\theta^{(r)}$

3.4 Dealing with noisy data

We now turn to the case when the identification data are corrupted by a moderate amount of noise $\{e(t)\}$. Then recall from Eq. (9) that

$$\mathbf{y} = X^\top \theta + \phi(\theta) + \mathbf{e}.$$

If the error sequence $\{e(t)\}$ were known, the previous noise-free identification method could be readily applied

with $\tilde{\mathbf{y}} = \mathbf{y} - \mathbf{e} \in \mathbb{R}^N$ as the new output vector and $X \in \mathbb{R}^{n \times N}$ as the regressor matrix. If θ represents one submodel of the switched system (1), then

$$\tilde{\phi}(\theta, \mathbf{e}) = \tilde{\mathbf{y}} - X^\top \theta = \mathbf{y} - X^\top \theta - \mathbf{e} \quad (25)$$

should be sparse, which leads to the minimization of $\|\tilde{\phi}(\theta, \mathbf{e})\|_1$ as a possible way of recovering θ . Although $\{e(t)\}$ is not measurable, one can still look for the sparsest error by minimizing $\|\phi(\theta)\|_1$ and ignoring the noise \mathbf{e} . This is only possible up to what extent the ℓ_1 optimization can stand noise. In fact the ℓ_1 estimator can correct (be insensitive to) very gross error entries with arbitrary magnitude provided the considered error vector is sparse enough. However it is much less capable of removing the effect of non-sparse small noise \mathbf{e} than the ℓ_2 norm for example. Therefore a quite natural approach would be to minimize the ℓ_1 norm $\|\tilde{\phi}(\theta, \mathbf{e})\|_1$ of $\tilde{\phi}(\theta, \mathbf{e})$ while minimizing the ℓ_2 norm $\|\mathbf{e}\|_2$ of \mathbf{e} . There are two possible ways of implementing this idea.

A first possibility is to minimize $\|\tilde{\phi}(\theta, \mathbf{e})\|_1$ while constraining \mathbf{e} to be bounded with respect to a certain norm. In [7], Candès and Randall used this approach to correct errors occurring when decoding messages transmitted over communication channels. Their idea is to estimate, under sparsity of (25), the error \mathbf{e} together with the vector θ which we suppose here to represent one submodel of the switched system (1). To do so, one needs however to know a priori an upper bound η on the norm of the noise. More precisely, a somewhat tight bound η satisfying $\|\mathbf{e}\|_\ell \leq \eta$, with ℓ a certain norm in $\{2, \infty, \dots\}$, is required. If θ is a PV for the switched linear system, then θ may be computed from the convex program

$$\begin{aligned} \min_{\theta, \mathbf{e}} & \|W\tilde{\phi}(\theta, \mathbf{e})\|_1 \\ \text{subject to} & \|\mathbf{e}\|_\ell \leq \eta. \end{aligned} \quad (26)$$

The solution θ to (26) is expected to be in $\{\theta_1, \dots, \theta_s\}$; the vector $\mathbf{e} \in \mathbb{R}^N$ is the variable whose estimate is expected to be equal or close to the true error \mathbf{e} defined in (9). Here, $W = \text{diag}(w_1, \dots, w_N)$ is a positive semi-definite and diagonal weighting matrix. For both the cases $\ell = 2$ and $\ell = \infty$ (see [7] for more comments), problem (26) is convex and can be efficiently solved using for example interior-point or simplex methods [5].

A second alternative approach is to trade-off $\|\tilde{\phi}(\theta, \mathbf{e})\|_1$ and $\|\mathbf{e}\|_2^2$ i.e., to consider a convex optimization problem of the form

$$\min_{\theta, \mathbf{e}} \gamma \|\tilde{\phi}(\theta, \mathbf{e})\|_1 + \frac{1}{2} \|\mathbf{e}\|_2^2. \quad (27)$$

In this latter method, no a priori upper bound is required on the magnitude of the noise but a regularization term

γ needs to be set. Problem (27) can be solved for example by Iterated Shrinkage (see e.g., [6] for some comments on these methods).

3.5 Summary of the algorithm

We have seen in the previous subsections that by applying Algorithm 1, we can identify one of the s parameter vectors of a switched system such as (1) from the whole dataset. If the conditions of Theorem 11 are satisfied, then we know that Algorithm 1 will find (after only one iteration) a vector θ^* in the set $\{\theta_1, \dots, \theta_s\}$ such that both $\|\phi(\theta^*)\|_0$ and $\|\phi(\theta^*)\|_1$ are minimum. If these conditions are not fulfilled, Algorithm 1 may not converge towards a point in $\{\theta_1, \dots, \theta_s\}$. However, as argued in [9] and suggested by different experiments reported in [9] and in Section 4 of this paper, the algorithm is likely to find the vector θ^* that realizes the sparsest error $\phi(\theta)$. According to Lemma 7 and Theorem 8, such a point θ^* is in $\{\theta_1, \dots, \theta_s\}$ when enough rich data are available.

Without loss of generality, we will denote with $\hat{\theta}_1$, i.e. the estimate of θ_1 , the point of $\{\theta_1, \dots, \theta_s\}$ 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 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

$$I(\hat{\theta}_1) = \left\{ t \in \{1, \dots, N\} : \frac{|\bar{x}(t)^\top \hat{\theta}_1|}{\|\bar{x}(t)\|_2 \cdot \|\hat{\theta}_1\|_2} \leq \text{Thres} \right\} \quad (28)$$

where it is assumed that $\text{Tresh} \in [0, 1]$ is a tolerance threshold and $\hat{\theta}_1 = [1 \ \hat{\theta}_1^\top]^\top$. 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 2.

Observe that if the number of submodels is known beforehand, it can be fed into the identification Algorithm 2. In such a case the *while* loop can be replaced by a *for* loop. Also, the stopping test need not be that strict as the condition $|J| \neq 0$. The algorithm can also be stopped once the cardinality of J falls under a given number. Before pursuing this subsection, let us emphasize a few remarks.

Remark 15 (Number of submodels) *When the threshold Thres in Algorithm 2 is suitably chosen, the proposed identification scheme can automatically provide the number of submodels. However when dealing with highly noisy data, selecting a threshold may, as is well-known, be a delicate task (for more details, see the paragraph concerning the implementation issues).*

Algorithm 2 Identification of all PVs

- (1) **Inputs:** $\{(x(t), y(t))\}_{t=1}^N$
- (2) **Initialization:** $\mathcal{S} \leftarrow \emptyset, J \leftarrow \{1, \dots, N\}$
- (3) **While** $|J| \neq 0$
 - Estimate a submodel by the reweighted ℓ_1 minimization method (See Algorithm 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. (28).

- (4) **EndWhile**
 - (5) **Return** \mathcal{S} and $s = |\mathcal{S}|$.
-

Remark 16 (Unknown orders) *In case the orders n_a and n_b are unknown, assume that some upper bounds $\bar{n}_a \geq n_a$ and $\bar{n}_b \geq n_b$ are available a priori. We can then parameterize the regressor $x(t)$ with \bar{n}_a and \bar{n}_b i.e., define a regressor*

$$x_{\bar{n}_a, \bar{n}_b}(t) = \begin{bmatrix} y(t-1) \cdots y(t-\bar{n}_a) \\ u(t-1)^\top \cdots u(t-\bar{n}_b)^\top \end{bmatrix}^\top.$$

In this case, we can, instead of (18), solve the convex problem

$$\min_{\theta} [\|W_\phi \phi(\theta)\|_1 + \|W_\theta \theta\|_1]$$

where W_ϕ and W_θ are some diagonal weighting matrices with positive entries. This way, the solution θ is expected to be sparse; its sparsity can be enhanced by choosing a matrix W_θ so as to penalize more the entries of θ which correspond to the components $y(t-i)$ and $u(t-j)$ of $x_{\bar{n}_a, \bar{n}_b}(t)$ having the largest i and j .

Implementation issues. Implementation of the identification Algorithm 2 necessitates mainly three user-defined parameters to be set. The number ε in Algorithm 1 aims essentially to prevent division by zero; the tolerance Tol in the same algorithm is used to detect convergence. Finally, the number Thres appearing in Algorithm 2 is a decision parameter. As is apparent from their respective roles, the first two user-defined parameters are easy to tune and do not have much effect on the performance of the algorithm. The parameter Tol can even be replaced with a fixed maximum number of iterations (See Algorithm 1). The last parameter is probably the most delicate to set when the identification data are affected by noise. The difficulty in choosing Thres increases with the magnitude of the noise. First notice that the threshold Thres does not impact the identification of the first PV θ_1 (first is used here with respect to the numbering agreed in the beginning of this subsection) because all the dataset is processed at this first

step. However, it does have an influence on the identification of the subsequent PVs. For example the identification of the second PV θ_2 requires that we manipulate the dataset in the objective of promoting the convergence of the algorithm to that second PV. An intuitive way of achieving this goal is, as proposed in Algorithm 2, to remove the data generated by the first submodel, before proceeding further. These data need not be removed entirely. For the algorithm to choose θ_2 as a convergence point, we just need a sufficient number of them to be removed. What can happen however is that if Thres is too small with respect to the level of noise, only a very small number of the data related to submodel 1 will be removed so that the algorithm can still identify θ_1 in the second step. This will likely result in an over-estimation of the number of submodels with redundant PVs. Although the number of estimated submodels is not minimal, the input-output map can still be reconstructed from the identified PVs. If Thres is too large, more than the data pertaining to the first submodel may be removed, probably causing the algorithm to provide bad estimates for some of the remaining PVs.

4 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

$$y(t) = \theta_{\lambda_t}^\top [y(t-1) \ y(t-2) \ u(t-1) \ u(t-2)]^\top + e(t) \quad (29)$$

with $\lambda_t \in \{1, 2, 3\}$ and

$$\begin{aligned} \theta_1 &= [-0.40 \ 0.25 \ -0.15 \ 0.08]^\top, \\ \theta_2 &= [1.55 \ -0.58 \ -2.10 \ 0.96]^\top, \\ \theta_3 &= [1 \ -0.24 \ -0.65 \ 0.30]^\top. \end{aligned} \quad (30)$$

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 $\{\lambda_t\}$ is uniformly distributed in $\{1, 2, 3\}$.

For all the results to be presented in this section, all convex optimization instances occurring in our method have been numerically implemented with the Matlab based software developed by Grant and Boyd [15].

4.1 Tests on the BP and the reweighted ℓ_1 methods

In a first experiment, we verify the ability of the BP method to exactly recover the solution of the sparse optimization problem, under the condition derived in Theorem 11. 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 $\phi(\theta_3)$ (expressed in terms of the number $\|\phi(\theta_3)\|_0$ of nonzero components in $\phi(\theta_3)$) and then solve problem (18) 100 times on different independent simulations of input-output data of length $N = 100$ each. This procedure is repeated for different values of $\|\phi(\theta_3)\|_0$ reported in Table 1. In this table we display for each given value of $\|\phi(\theta_3)\|_0$, the percentage of successes in attempting to compute the solution of (12) by solving (18). Since the data $(x(t), y(t))$, $t = 1, \dots, N$, is generated from (29) 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 $\nu_4(X)$ is as small as 4 i.e., $\nu_4(X)$ is equal to the dimension of $x(t)$, see (29). According to Theorem 11, if $\|\phi(\theta_3)\|_0$ is roughly less than $\frac{N-\nu_4(X)}{2} = 48$, then equivalence holds between problems (12) and (18) and the unique solution to both of them is θ_3 . This can be verified from the results of Table 1 where we see that (18) effectively solves (12) successfully with a score of 100% over 100 trials (on randomly generated data) once $\|\phi(\theta_3)\|_0$ falls under 48.

$\ \phi(\theta_3)\ _0$	58%	55%	53%	50%	48%	45%
# succ.	46%	76%	94%	99%	100%	100%

Table 1
Equivalence between ℓ_0 and ℓ_1 minimizations versus the sparsity of $\phi(\theta_3)$. The ℓ_0 norm of $\phi(\theta_3)$ is expressed as a fraction of the nonzero entries over the total length of the vector $\phi(\theta_3)$.

Now we propose, in the same conditions as the first experiment, to solve the sparse optimization problem (12) with the reweighted ℓ_1 optimization technique described in Algorithm 1 [9]. The related results are presented in Table 2. It turns out that the reweighted ℓ_1 optimization approach significantly improves the basis pursuit technique. When $\|\phi(\theta_3)\|_0$ starts getting much larger than $\|\phi(\theta_1)\|_0$ and $\|\phi(\theta_2)\|_0$, Algorithm 1 may not keep on converging towards θ_3 any longer. Instead, it is likely to converge towards θ_1 or θ_2 since $\|\phi(\theta_1)\|_0$ and $\|\phi(\theta_2)\|_0$ decrease as $\|\phi(\theta_3)\|_0$ increases.

$\ \phi(\theta_3)\ _0$	58%	55%	53%	50%	48%	45%
# succ.	94%	100%	100%	100%	100%	100%

Table 2
Approximation of ℓ_0 by reweighted ℓ_1 minimization versus the sparsity of $\phi(\theta_3)$. The ℓ_0 norm of $\phi(\theta_3)$ is expressed as a fraction of the nonzero entries over the total length of the vector $\phi(\theta_3)$.

4.2 Identification of the PVs

The second objective 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. The user-specified parameters of Algorithm 1 and Algorithm 2 are set to $\varepsilon = 0.1$, Tol = 0.001 and Thres = 0.05. 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. This is evaluated with the criterion [20]

$$\text{FIT} = \left(1 - \frac{\|\hat{\mathbf{y}} - \mathbf{y}\|_2}{\|\mathbf{y} - \bar{y}\mathbf{1}_N\|_2}\right) \times 100\% \quad (31)$$

which measures the fitting error between the true output sequence \mathbf{y} and the estimated model output sequence $\hat{\mathbf{y}}$. In this formula, \bar{y} stands for the mean of the true output sequence and $\mathbf{1}_N$ is an N -dimensional vector with all entries equal to one. The reader can refer to Figure 2 for an insight into the form of the input-output signals. For better visualization purpose, the size of the observation window in that figure has been shortened.

The number of submodels is given a priori. We start by assuming that the number of submodels is fed into the identification algorithm. We present in Table 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 to the true values $\{\theta_i\}_{i=1}^s$ given in (30), 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 $\sigma(\hat{\theta}_i)$, we are prompted to conclude that the algorithm performs well on statistically independent realizations of the input-output data. Of course, when the data is noise-free, the parameters are exactly recovered by the algorithm.

In Figure 3 is represented 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.

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

Average estimates over 100 independent runs of the identification algorithm: $\varepsilon = 0.1$, Tol = 0.001, Thres = 0.05.

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

Least squares average estimates if the discrete state were known.

Table 3

Comparison of the proposed identification algorithm to standard least squares (if the discrete state were known) over 100 independent runs: $\varepsilon = 0.1$, Tol = 0.001, Thres = 0.05.

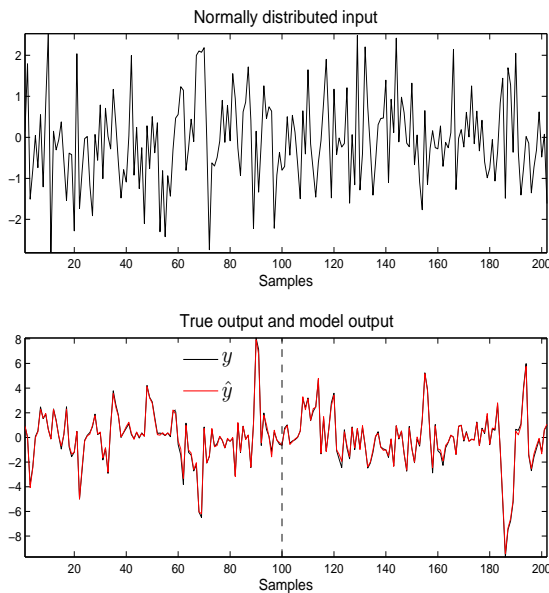


Figure 2. 100 identification and 100 validation data with an SNR of 30 dB. The algorithm is run with $\varepsilon = 0.1$, Tol = 0.001 and Thres = 0.05: Obtained FIT = 93%.

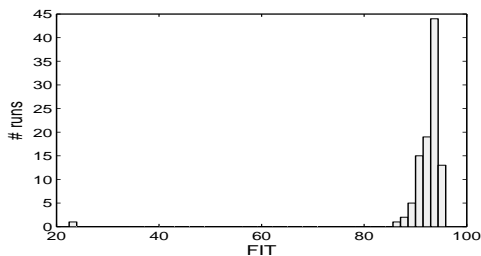


Figure 3. Distribution of the FIT measure over the 100 independent runs of the algorithm. $\varepsilon = 0.1$, Tol = 0.001, Thres = 0.05.

Thres	0.1	0.1	0.2	0.3
SNR (dB)	40	30	25	20
# successes	98%	90%	96%	87%
FIT	97.9%	93%	83.8%	63.7%

Table 4

Performance of our method in determining the correct number of submodels, under different proportions of noise. Here, # successes is the number of times the algorithm yields the correct number of submodels (which is equal to 3) over 100 runs using a fixed threshold Thres. The value of FIT displayed here is the average fit over 100 independent runs.

The number of submodels is unknown. We turn now to the case when the number s of submodels is not known a priori. In such a context, we need to estimate, as in Algorithm 2, the number of submodels together with the parameter vectors. As we have seen in Subsection 3.5, the ability of the algorithm to provide the correct number (in fact the minimum number) of submodels from noisy data depends heavily on the threshold Thres⁶. And selecting a threshold in a noisy environment is known to be challenging in general.

For different levels of noise, the number of submodels is estimated over 100 independent realizations of data. The results displayed in Table 4 reveal a rather good tendency of the proposed method in recovering the true number of submodels.

5 Conclusion

The paper discusses a new optimization-based method for the identification of linear switched systems. By exploiting some ideas from the field of sparse signal recovery, we have first formulated this challenging iden-

⁶ This problem is common to most existing methods that estimate the number of submodels, i.e. their capacity of determining the correct number of submodels depends on a user-specified parameter, see [3,2,4,24].

tification problem as a combinatorial ℓ_0 minimization problem. As such however, the problem is still computationally intractable so that practical implementation may involve relaxing it into an ℓ_1 norm based program, which is convex and therefore solvable with classical and well documented tools. Some sufficient conditions are derived for the convex relaxation strategy to exactly recover the solution of the initial ℓ_0 norm problem. In practice, different conclusive experiments on simulated data tend to show that the performance of the convex relaxation technique can be boosted far more beyond the established theoretical conditions of equivalence. Hence, the method for extracting the parameter vectors associated with the different subsystems, proves to be both computationally simple and satisfactorily efficient. As future work, we may envision to further study the properties of the method by (i) studying the convergence of the reweighted ℓ_1 minimization problem described in Algorithm 1 in the context of hybrid system identification, (ii) looking for tight bounds on the estimation error in highly noisy conditions. It would also be interesting to look for a way of efficiently extending the proposed method to switched systems in which the noise has a specific structure (for example, when the linear subsystems are described with ARMAX models).

Acknowledgements

The author would like to thank the anonymous reviewers, the editor and the associate editor whose constructive comments on an earlier version of this manuscript have been of great help in improving the presentation.

References

- [1] L. Bako, K. Boukharouba, E. Duviella, and S. Lecoeuche. A recursive identification algorithm for switched linear/affine models. *Nonlinear Analysis: Hybrid Systems (To appear)*, 2010.
- [2] L. Bako and R. Vidal. Algebraic identification of switched MIMO ARX models. In M. Egerstedt and B. Mishra, editors, *Hybrid Systems: Control and Computation*, volume 4981 of *LNCS*, pages 43–57. Springer Verlag, 2008.
- [3] 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.
- [4] 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.
- [5] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.
- [6] 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.
- [7] E. Candès and P. A. Randall. Highly robust error correction by convex programming. *IEEE Transactions on Information Theory*, 54:2829–2840, 2006.
- [8] 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.
- [9] E. J. Candès, M. Wakin, and S. Boyd. Enhancing sparsity by reweighted ℓ_1 minimization. *Journal Fourier Analysis and Applications*, 14:877–905, 2008.
- [10] D. L. Donoho. For most large underdetermined systems of linear equations the minimal ℓ^1 -norm solution is also the sparsest solution. *Communications on Pure and Applied Mathematics*, 59:797–829, 2006.
- [11] D. L. Donoho and M. Elad. Optimal sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization. *The Proc. Nat. Aca. Sci.*, 100:2197–2202, 2003.
- [12] E. Elhamifar and R. Vidal. Sparse subspace clustering. In *2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami beach, FL, USA*, pages 2790–2797, 2009.
- [13] 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.
- [14] G. Ferrari-Trecate and M. Schinkel. Conditions of optimal classification for piecewise affine regression. In O. Maler and A. Pnueli, editors, *Hybrid Systems: Computation and Control*, volume 2623 of *LNCS*, pages 188–202. Springer-Verlag, 2003.
- [15] M. Grant and S. Boyd. CVX: Matlab software for disciplined convex programming, version 1.2, june 2009, (build 711). June 2009, Build 711.
- [16] R. A. Horn and C. R. Johnson. *Matrix analysis*. Cambridge University Press, 1985.
- [17] 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.
- [18] F. Lauer, R. Vidal, and G. Bloch. A product-of-errors framework for linear hybrid identification problem. In *IFAC Symposium on System Identification, Saint Malo, France*, pages 563–568, 2009.
- [19] D. Liberzon. *Switching in systems and control*. Birkhauser, Boston, MA, 2003.
- [20] L. Ljung. *System Identification Toolbox User's Guide*. 7th ed. Natick, MA: The MathWorks Inc., 2009.
- [21] Y. Ma and R. Vidal. Identification of deterministic switched ARX systems via identification of algebraic varieties. In M. Morari and L. Thiele, editors, *Hybrid systems computation and control, Zurich, Switzerland*, volume 3414, pages 449–465. Springer-Verlag, 2005.
- [22] H. Nakada, K. Takaba, and T. Katayama. Identification of piecewise affine systems based on statistical clustering technique. *Automatica*, 41:905–913, 2005.
- [23] B. K. Natarajan. Sparse approximate solutions to linear systems. *SIAM Journal on Computing*, 24:227–234, 1995.
- [24] N. Ozay, M. Sznaier, C. Lagoa, and O. Camps. A sparsification approach to set membership identification of a class of affine hybrid systems. In *Proceedings of the IEEE Conference on Decision and Control, Cancun, Mexico*, pages 123–130, 2008.
- [25] S. Paoletti, A. Juloski, G. Ferrari-Trecate, and R. Vidal. Identification of hybrid systems: A tutorial. *European Journal of Control*, 13:242–260, 2007.
- [26] 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.

- [27] J. Roll, A. Bemporad, and L. Ljung. Identification of piecewise affine systems via mixed-integer programming. *Automatica*, 40:37–50, 2004.
- [28] 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.
- [29] Z. Sun and S. S. Ge. *Switched Linear Systems: Control and Design*. Springer, London, UK, 2005.
- [30] R. Vidal. Recursive identification of switched ARX systems. *Automatica*, 44:2274–2287, 2008.
- [31] 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.
- [32] 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.



Laurent Bako received a "Diplôme d'ingénieur" in Electrical Engineering from Ecole Nationale Supérieure d'Ingénieurs de Poitiers and the M.Sc. degree from Université de Poitiers, both in 2005. He was a visiting researcher in the Center for Imaging Sciences, at the Johns Hopkins University in the

Summer-Fall 2007. In 2008 he obtained the Ph.D. degree in Automatic Control and Computer Sciences from Université des Sciences et Technologies de Lille. He has been serving as an assistant professor at Ecole des Mines de Douai, in the Department of Computer Sciences and Automatic Control since December 2008. His research interests are mainly in control theory, system identification, hybrid systems, machine learning.