On sparsity-inducing methods in system identification and state estimation

Laurent Bako

Received: 2021 | Revised: 2021 | Accepted: 2021

DOI: XXX/XXX

RESEARCH ARTICLE

On sparsity-inducing methods in system identification and state estimation

Laurent Bako

[1]Univ Lyon, Ecole Centrale Lyon, INSA Lyon, Université Claude Bernard Lyon 1, CNRS, Ampère, UMR 5005, 69130 Ecully, France

Correspondence
*Email: laurent.bako@ec-lyon.fr

Summary

The purpose of this paper is to survey some sparsity-inducing methods in system identification and state estimation. Such methods can be divided into two main categories: methods inducing sparsity in the parameters and those sparsifying the prediction error. In the last class we discuss in particular the Least Absolute Deviation (LAD) estimator and its robustness properties with respect to sparse noise in both cases of univariate and multivariate measurements. We also discuss the application of sparsity-inducing methods to switched system identification and to state estimation for linear systems in the presence sparse and dense measurement noises. While the presentation focuses essentially on bridging some existing results, some technical refinements and new features are also provided.

KEYWORDS:
system identification, robust regression, state estimation, switched linear systems, sparse optimization

1 | INTRODUCTION

Inferring a mathematical model from experimental data is a problem which is of fundamental interest in many engineering fields such as control theory, signal processing or machine learning. This process is called data-driven modeling or system identification. The focus of this paper is on the estimation problem that is, the step of the modelling procedure which is related to the learning of the best model (in the sense of an appropriate loss function) within a given family of candidate models. To be more specific, assume that we are given a set of N data samples $\left\{(y_t, x_t)\right\}_{t=1}^{N}$ generated by a system of the form

$$y_t = f(x_t, \theta^\circ) + v_t \tag{1}$$

where $\{f(\cdot, \theta) : \theta \in \mathbb{R}^n\}$ is a given family of functions parameterized by $\theta \in \mathbb{R}^n$ and $\left\{v_t\right\}$ is an unknown sequence, generally termed noise, but it can represent the combination of measurement errors and model mismatch. A critical step of the data-driven modelling procedure is the design of the estimation method in the presence of unknown disturbance $\left\{v_t\right\}$ affecting the data. Generally, the estimation method of a parametrized model intends to fit the available observations to the candidate model by optimizing a certain performance index. In this context of optimization-based estimation, the performance of the model strongly depends on the performance index which is optimized. To achieve a good estimation, one needs to design the to-be-optimized loss function (which is constituent of the performance index) based on the assumptions we set concerning the model uncertainty represented by $\left\{v_t\right\}$ in (1). A massively used loss function for designing the estimation scheme is the mean squares one (see e.g., prediction error methods [1]). While such estimators may be convenient when the error sequence $\left\{v_t\right\}$ is Gaussian, they are

known to perform very poorly when this is not the case. For example if the noise is Laplacian or has a heavy-tailed distribution, the popular least squares estimator is no more suitable. We will confine the following discussion to the scenario where the uncertainty $\{v_t\}$ may be of both dense and impulsive nature. That is, the elements of $\{v_t\}$ are bounded most of time but may assume arbitrarily large values showing up intermittently over time.

In this paper, we intend to present a picture of the use of sparsity-inducing optimization methods in system identification and state estimation. The driving principle of these methods is to minimize a sparsity measure (the number of nonzero elements of a given set). Such methods have been vastly promoted by many recent developments in the field of compressed sensing[2,3,4,5]. One can divide the existing sparsity-inducing methods into two categories: the parameter vector sparsification methods and the error sparsification methods. The first type of methods mainly searches for a model with the minimum number of prediction variables. In single model regression methods, the estimators with parameter sparsification properties rely largely on an ℓ_0/ℓ_1 regularization of a least squares cost function. A popular scheme of this type is the Lasso estimator[6,7,8] which admits many applications in signal processing[9,10] and system identification[11,12,13]. But other variants of this class of estimator exist in multiple models regression problems whose goal is to reduce the number of expected models out of the estimation process[14,15,16,17]. For example, considering the problem of identifying a dynamical switched system, one is confronted with the challenge that the switching signal is generally unknown. In this context, we are given a mixture of data generated by different interacting subsystems and we should infer a model for each individual subsystem without knowing which subsystem has generated which data. One natural approach to overcome this problem is to estimate a vector-valued sequence of parameters from the data under some sparsity constraints, the rationale of which is to force as many parameter vectors as possible to be equal. This idea has been proposed e.g., in[14,15,17].

The second class of estimation methods (i.e., those inducing prediction error sparsity) search for a parameter vector such that the associated model achieves a vector of prediction errors which is as sparse as possible. The basic frame for these methods is the Least Absolute Deviation (LAD) estimator initially proposed in the field of robust statistics[18]. There has been a recent surge of interest in the robustness properties of this estimator, due perhaps on the one hand to the new perspectives of analysis opened by the field of compressed sensing for ℓ_1 decoders[19] and on the other hand, to emerging applications such as the monitoring of cyber-physical systems which require robust estimation schemes. More generally, robustness properties are desired in estimation scenarios where the data may have been corrupted by adversarial attacks or loss of data packets (for example, if they are transmitted over a communication network), intermittent sensor failures, etc.

A well-known property of the LAD estimator is that of exact recovery in the presence of sparse noise $\{v_t\}$. If the noise sequence is sparse enough then, the LAD estimator is able to recover exact parameter vector regardless of the amplitudes of the nonzero instances of the (unknown) sequence $\{v_t\}$[20]. Moreover, the estimation error can be shown to be bounded when $\{v_t\}$ is a combination of dense and sparse components. The current paper will focus essentially on the formal characterization, the properties and the implementation of the LAD estimator and some of its variants for robust regression, hybrid system identification and resilient state estimation. The robustness properties of the LAD estimator make it applicable to hybrid system identification[21,22,23,24], subspace clustering[25,26] and secure state estimation[27,28,29,30,31,32].

Outline and contributions of this paper. The outline of this paper is as follows. We start by presenting the LAD estimator in Section 2 which, by adopting the perspective of compressive sampling, can be viewed as resulting from a convex relaxation of the ℓ_0 norm based estimator. The technical results presented therein are essentially refinements of existing findings along with complementary comments. A few exceptions though concern Theorems 3 and 4 which state the resilience properties of the LAD estimator with respect to outliers in finite time and infinite time respectively. These two analysis results are new. From a computational perspective, a new iterative algorithm is introduced in Section 2.4 to approach the k-smallest absolute deviation estimator. The principle of the proposed algorithm is to write the k-smallest objective function as a difference of two convex functions and then to iteratively approximate (locally) the second convex function by a linear one. Building on the properties derived in Section 2, Section 3 illustrates the application of sparsity-inducing optimization in hybrid system identification. It is

shown that the parameter vectors of a switched linear system can be identified incrementally, one after another, using a sparsity-inducing (robust) estimator. The approach was originally proposed in[22] but here, less restrictive conditions of exact recovery are discussed. An extension of the LAD estimator to multivariate regression is then discussed in Section 4 together with an important application to resilient state estimation in Section 4.2. The main new results of this section are Theorems 8 and 9 establishing resilience of a multivariate parameter estimator and a state estimator respectively. A few numerical illustrations are provided in Section 5 on some particular features. Finally, some concluding remarks are presented in Section 6.

Notation. We use N to denote the number of data points available for estimation, and $\mathbb{I} = \{1, \dots, N\}$ to represent the index set of the measurements. For a vector $v = [v_1 \ \cdots \ v_N]^\top \in \mathbb{R}^N$, $\mathrm{Supp}(v)$ denotes the support of v defined by $\mathrm{Supp}(v) = \{i \in \mathbb{I} : v_i \neq 0\}$.

Cardinality of a finite set. Throughout the paper, 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. Let $X = \begin{bmatrix} x_1 & x_2 & \cdots & x_N \end{bmatrix} \in \mathbb{R}^{n\times N}$ be the matrix formed with the available regressors $\{x_t\}_{t=1}^N$. 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 $\boldsymbol{y} = \begin{bmatrix} y_1 & y_2 & \cdots & y_N \end{bmatrix}^\top \in \mathbb{R}^N$, $\boldsymbol{y}_I$ 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. $\|\cdot\|_p$, $p = 1, 2, \dots, \infty$, denote the usual p-norms for vectors defined for any vector $z = \begin{bmatrix} z_1 & \cdots & z_N \end{bmatrix}^\top \in \mathbb{R}^N$, by $\|z\|_p = \left(|z_1|^p + \cdots + |z_N|^p\right)^{1/p}$. Note that in the limiting case where $p \to \infty$, we get $\|z\|_\infty = \max_{i=1,\dots,N} |z_i|$. The r-max norm of z denoted[1] $\|z\|_{1,[r]} = |z_{[1]}| + |z_{[2]}| + \cdots + |z_{[r]}|$ is the sum of the r largest entries of z in absolute value. The ℓ_0 norm[2] of z is defined to be the number of nonzero entries in z, i.e., $\|z\|_0 = \left|\{i : z_i \neq 0\}\right|$.

Matrix norms. For a matrix $A = \begin{bmatrix} a_1 & \cdots & a_N \end{bmatrix} \in \mathbb{R}^{n\times N}$ with $a_i \in \mathbb{R}^n$, we consider the following norms

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

2 | SPARSITY-INDUCING PARAMETER ESTIMATORS

We consider a MISO data-generating system described by an equation of the form

$$y_t = x_t^\top \theta^\circ + v_t \tag{2}$$

where $y_t \in \mathbb{R}$ is the output of the system at time $t \in \mathbb{Z}_+$, $x_t \in \mathbb{R}^n$ is the regressor and v_t represents some uncertainty (measurement noise, mismatch, etc). Here, $\theta^\circ \in \mathbb{R}^n$ is an unknown parameter vector. The system (2) can be static (i.e., x_t is entirely measured at time t) or dynamic in which case, the so-called vector of predictor variables x_t may assume a structure of the form

$$x_t = \begin{bmatrix} y_{t-1} & \cdots & y_{t-n_a} & u_{t-1}^\top & \cdots & u_{t-n_b}^\top \end{bmatrix}^\top. \tag{3}$$

with $u_t \in \mathbb{R}^{n_u}$ denoting the input of the dynamic system at time t and n_a and n_b being (known) integers (often called the orders of the system).

[1] This expansion is made with the assumption that the entries of z are ordered such that $|z_{[1]}| \geq |z_{[2]}| \geq \cdots \geq |z_{[N]}|$.

[2] This terminology is used with some abuse of language: 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.

The estimation problem of interest in this paper can be informally stated as follows:
Given a set $\varpi^N = \{(x_t, y_t) \in \mathbb{R}^n \times \mathbb{R} : t = 1, \dots, N\}$ of N input-output data points generated by the system (2) with $N \gg n$, find an estimate of the parameter vector θ°. More specifically, we wish to find a map

$$\Psi : (\mathbb{R}^n \times \mathbb{R})^N \to 2^{\mathbb{R}^n}, \quad \varpi^N \mapsto \Psi(\varpi^N) \tag{4}$$

which maps the data ϖ^N to a subset $\Psi(\varpi^N)$ of the parameter space $\mathbb{R}^n$. Here, $2^{\mathbb{R}^n}$ refers to the set collecting all subsets of $\mathbb{R}^n$.

The main challenge of this estimation problem resides in the fact that the sequence $\{v_t\}$ in (2) is unknown. If this sequence is completely arbitrary, then the generated data loose any informativity concerning the data-generating system. In this case we cannot hope for an accurate estimate of θ° from the data ϖ^N. Hence, for the estimation problem to make sense, we need to make some minimal assumption regarding the uncertainty $\{v_t\}$. The construction of the estimator is based on this assumption concerning the structure of $\{v_t\}$.

Let $\boldsymbol{v} = [v_1 \ v_2 \ \cdots \ v_N]^\top$ and $\boldsymbol{y} = [y_1 \ y_2 \ \cdots \ y_N]^\top$ denote some vectors collecting respectively the noise and output samples from (2). Form also a matrix $X = [x_1 \ x_2 \ \cdots \ x_N] \in \mathbb{R}^{n \times N}$ with all the available regressors x_t. Finally, for a given candidate parameter vector $\theta \in \mathbb{R}^n$, consider, for future reference, the vector of prediction errors induced by θ,

$$\phi(\theta) = \boldsymbol{y} - X^\top \theta. \tag{5}$$

Note that $\phi(\theta)$ reduces to $\boldsymbol{v}$ when $\theta = \theta^\circ$.
The next section discusses the problem of designing an appropriate estimator for the parameter vector θ° in (2) when the noise vector $\boldsymbol{v} \in \mathbb{R}^N$ is assumed to be sparse, that is, a certain proportion of its entries is equal to zero. More specifically, we will be interested in a class of estimators having the property of being insensitive to sparse noise. Such estimators are called resilient.

Definition 1 (Resilience of an estimator)**.** Consider the system (2) and denote with $\boldsymbol{v} \in \mathbb{R}^N$ the noise vector. A parameter estimator Ψ defined as in (4) is called *resilient* with respect to the r-sparse noise vector set $\mathcal{S}_r^N = \{\boldsymbol{z} \in \mathbb{R}^N : \|\boldsymbol{z}\|_0 \leq r\} \subset \mathbb{R}^N$, r being a positive integer, if there exists $\gamma > 0$ (depending on the data ϖ^N) such that

$$\forall \hat{\theta} \in \Psi(\varpi^N), \ \left\|\hat{\theta} - \theta^\circ\right\| \leq \gamma d(\boldsymbol{v}, \mathcal{S}_r^N) \tag{6}$$

for some distance function d defined by $d(\boldsymbol{v}, \mathcal{S}_r^N) = \inf_{w \in \mathcal{S}_r^N} \|\boldsymbol{v} - w\|$. Here, $\|\cdot\|$ is a generic notation for norms regardless of the space on which they are acting.

According to this definition, an estimator is resilient with respect to a set of disturbances if the estimation error induced by this estimator is completely insensitive to each instance of such disturbances. In effect, by Eq. (6) we have $\hat{\theta} = \theta^\circ$ whenever $\boldsymbol{v} \in \mathcal{S}_r^N$, that is the estimation error is zero for each $\boldsymbol{v} \in \mathcal{S}_r^N$.

2.1 | Estimation in the presence of sparse noise

2.1.1 | An ℓ_0 estimator for regression

If we assume that the uncertainty $\boldsymbol{v}$ in (2) is r-sparse for some positive integer r, that is, $\boldsymbol{v} \in \mathcal{S}_r^N \triangleq \{\boldsymbol{z} \in \mathbb{R}^N : \|\boldsymbol{z}\|_0 \leq r\}$, then a natural solution of the estimation problem can be obtained by sparsifying the error $\phi(\theta)$, i.e., searching for the set of parameter vectors which make this error the sparsest possible. We can therefore define an estimator $\Psi_0 : (\mathbb{R}^n \times \mathbb{R})^N \to 2^{\mathbb{R}^n}$ through the minimizing set the ℓ_0 norm of the function $\theta \mapsto \phi(\theta)$

$$\Psi_0(\varpi^N) = \arg\min_{\theta \in \mathbb{R}^n} \|\phi(\theta)\|_0 \tag{7}$$

We will refer to Ψ_0 as the ℓ_0 norm estimator. Note that the range of the function $\theta \mapsto \|\phi(\theta)\|_0$ for all $\theta \in \mathbb{R}^n$ is the finite set $\{0, 1, \dots, N\}$. Hence the estimator Ψ_0 is well-defined in the sense that the minimizing set in (7) is always non empty as long as the data set ϖ^N is formed of samples with bounded amplitudes. However, there is a priori no guarantee that $\Psi_0(\varpi^N)$ will

be a singleton nor that this singleton, when achievable, will coincide with the true parameter vector θ° of the data-generating system. For the sake of the analysis, assume for now that we can solve the ℓ_0 optimization problem appearing in (7). Then a first question of interest is whether the set $\Psi_0(\varpi^N)$ of estimates may contain θ°.

An answer is given by the following proposition.

Proposition 1. Consider the regressor matrix $X \in \mathbb{R}^{n\times N}$ generated by the system (2). Let $\xi_r^\circ(X)$ be the number defined by

$$\xi_r^\circ(X) = \max_{\substack{I^c\subset\mathbb{I}:\\ |I^c|=r}} \sup_{\substack{\eta\in\mathbb{R}^n\\ \eta\neq 0}} \left[\frac{\left\|X_{I^c}^\top\eta\right\|_0}{\left\|X^\top\eta\right\|_0}\right]. \tag{8}$$

Then $\theta^\circ \in \Psi_0(\varpi^N)$ for all $\boldsymbol{v} = \phi(\theta^\circ) \in \mathcal{S}_r^N$ if $\xi_r^\circ(X) \leq 1/2$. Moreover, $\Psi_0(\varpi^N) = \{\theta^\circ\}$ if $\xi_r^\circ(X) < 1/2$.

Proof. Let $I^c = \mathrm{Supp}(\boldsymbol{v})$ and $I^0 = \mathbb{I} \setminus I^c$. Then $|I^c| \leq r$ whenever $\boldsymbol{v} \in \mathcal{S}_r^N$. Now we observe that θ° lies in $\Psi_0(\varpi^N)$ if and only if for all $\eta \in \mathbb{R}^n$,

$$\|\boldsymbol{v}\|_0 = \left\|\boldsymbol{y} - X^\top\theta^\circ\right\|_0 \leq \left\|\boldsymbol{y} - X^\top(\theta^\circ+\eta)\right\|_0 = \left\|\boldsymbol{v} - X^\top\eta\right\|_0.$$

This is equivalent to $\|\boldsymbol{v}_{I^c}\|_0 - \left\|\boldsymbol{v}_{I^c} - X_{I^c}^\top\eta\right\|_0 \leq \left\|X_{I^c}^\top\eta\right\|_0$. Note that by the triangle inequality property of the ℓ_0 norm, $\|\boldsymbol{v}_{I^c}\|_0 - \left\|\boldsymbol{v}_{I^c} - X_{I^c}^\top\eta\right\|_0 \leq \left\|X_{I^c}^\top\eta\right\|_0$. Hence, for the above inequality to hold, it suffices that $\left\|X_{I^c}^\top\eta\right\|_0 \leq \left\|X_{I^0}^\top\eta\right\|_0$ for all $\eta \in \mathbb{R}^n$ or, equivalently, that $\left\|X_{I^c}^\top\eta\right\|_0 \leq 1/2\left\|X^\top\eta\right\|_0$. With $|I^c| \leq r$, it suffices indeed that $\xi_r^\circ(X) \leq 1/2$ with $\xi_r^\circ(X)$ defined as in (8). The proof of the second statement follows exactly the same steps as above by changing the large inequality symbol to a strict one and restricting η to be nonzero. □

For further analysis of the ability of problem (7) to solve the identification problem, we introduce the following measure of informativity (richness) of the regression data[22].

Definition 2 (An integer measure of genericity)**.** Let $X \in \mathbb{R}^{n\times N}$ be a data matrix satisfying $\mathrm{rank}(X) = n$. 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,

$$\nu_n(X) = \min\left\{m : \forall\, \mathcal{S} \subset \mathbb{I} \text{ with } |\mathcal{S}| = m,\ \mathrm{rank}(X_{\mathcal{S}}) = n\right\}. \tag{9}$$

For any $X \in \mathbb{R}^{n\times N}$ satisfying $\mathrm{rank}(X) = n$, the index $\nu_n(X)$ satisfies $n \leq \nu_n(X) \leq N$. The smaller $\nu_n(X)$, the more generic the data matrix X. In case the data are in *general position*, we have $\nu_n(X) = n$. It can be shown that

$$\xi_r^\circ(X) \leq \frac{r}{N - \nu_n(X) + 1}$$

for all $r \in \left\{0, \ldots, N - \nu_n(X) + 1\right\}$.

Proposition 2 ([22] Sufficient condition for ℓ_0 recovery)**.** Consider data $(\boldsymbol{y}, X)$ generated by the system (2) under the assumption that $\mathrm{rank}(X) = n$ and $\boldsymbol{v} = \phi(\theta^\circ) \in \mathcal{S}_r^N$, i.e., $\boldsymbol{v}$ is r-sparse. Then

$$r \leq \frac{N - \nu_n(X)}{2} \quad \Rightarrow \quad \Psi_0(\varpi^N) = \{\theta^\circ\}. \tag{10}$$

Algorithmic considerations. Unfortunately, solving directly or exactly the optimization problem in (7) is known to be NP-hard[33]. Nevertheless, there exist a number of heuristics which attempt to find indirectly the solution though rarely with theoretical guarantees of finding it. These methods are originally derived and more often applied to searching for the sparsest solution to an underdetermined set of linear equations. For the sake of the discussion, let us rewrite the estimator Ψ_0. For this purpose, assume that $\mathrm{rank}(X) = n$. Let $B_x \in \mathbb{R}^{N\times(N-n)}$ be an orthogonal matrix whose columns form a basis of $\mathrm{im}(X^\top)^\perp = \ker(X)$ (the orthogonal complement of the range space of $X^\top$). Then Ψ_0 defined in (7) can be re-expressed as

$$\Psi_0(\varpi^N) = \left\{(XX^\top)^{-1}X(\boldsymbol{y} - z^\star) : z^\star \text{ solves } (11)\right\}$$

That is, $\Psi_0(\varpi^N)$ consists of a collection of all vectors of the form $(XX^\top)^{-1}X(\boldsymbol{y}-z^\star)$ where $z^\star$ is a (any) solution of the equality constrained ℓ_0 problem

$$\min_{z\in\mathbb{R}^N}\left\{\ \|z\|_0\ :\ \tilde{\boldsymbol{y}}=Dz\right\} \tag{11}$$

with $D=B_x^\top$ and $\tilde{\boldsymbol{y}}=B_x^\top\boldsymbol{y}$. As already mentioned, the ℓ_0 problem has a combinatorial characteristic which makes it generically hard to solve at a reasonable computational cost. Hence, the basic principle of all existing numerical algorithms is to approximate the solution of (11) using different kinds of functions. Examples of methods which try to solve (11) in a somewhat greedy manner are the following:

- *iteratively reweighted least squares* (IRLS)[34]: In the objective of gaining computational efficiency, the principle of this approach is to replace the $\|z\|_0$ in (11) by a weighted quadratic function $\|Wz\|_2^2$ with W being a diagonal matrix with positive elements on its diagonal. However, since W is not known a priori it is iteratively selected through a specific mechanism which relies on the solution obtained at the previous step.

- *ℓ_p quasi-norm*[35,36]: In this approach the ℓ_0 norm is approximated by ℓ_p quasi-norms defined by $\|z\|_p=(|z_1|^p+\cdots+|z_N|^p)^{1/p}$ with $0<p<1$. This in turn is approximately solved through different heuristics. The smaller p in the range $]0,1[$, the better the approximation.

- *smoothed ℓ_0 norm*[37,38]: To overcome the combinatorial nature of the ℓ_0 norm, the idea of this class of methods is to replace it by a continuous and differentiable function, for example $\|z\|_0\approx\sum_{i=1}^N g_\sigma(z_i)$ with $g_\sigma(z_i)=1-\exp(-z_i^2/(2\sigma^2))$ for a small enough value of σ. The advantage here is that the to-be-minimized cost function becomes differentiable so that algorithms such as the gradient descent can be applied. Note however that the loss function g_σ is nonconvex with the consequence that there is no guarantee to achieve the global minimum.

- *(orthogonal) matching pursuit*[39,40]: The problem (11) is viewed as one of decomposing the signal $\tilde{\boldsymbol{y}}$ over the dictionary D with the sparsest weight vector z. Assuming that D has normalized columns (in the sense of the ℓ_2 norm), the algorithm computes incrementally the nonzero entries of z as the maximum (algebraic) projection of the signal $\tilde{\boldsymbol{y}}$ (or its residuals) onto the atoms (columns) of the dictionary D.

We will not discuss all such methods here for the regression problem stated above (see the beginning of Section 2). Instead, we will put the focus on the popular ℓ_1 (convex) relaxation of the ℓ_0 loss. In particular, we will discuss in Section 2.4 two iterative algorithms which rely on the ℓ_1 approximation.

2.1.2 | The Least Absolute Deviation estimator

The most successful approach to the ℓ_0 estimation problem consists in replacing the ℓ_0 norm in (7) by an ℓ_1 norm, the advantage being that the latter is convex[41,42]. Doing so, we replace the estimator Ψ_0 by Ψ_1 defined by

$$\Psi_1(\varpi^N)=\arg\min_{\theta\in\mathbb{R}^n}\|\phi(\theta)\|_1 \tag{12}$$

where $\|z\|_1=\sum_{i=1}^N|z_i|$ for any vector $z=[z_1\ \cdots\ z_N]^\top\in\mathbb{R}^N$. The underlying optimization problem in (12) corresponds to what is classically referred to as sparse error correction problem in the compressed sensing literature[43,44,19,45] or Least Absolute Deviation estimator in robust statistics[46,47]. Contrary to (7), the defining optimization problem of Ψ_1 above 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[48]. Furthermore, it can be observed that the cost $\theta\mapsto\|\phi(\theta)\|_1$ is continuous and coercive, i.e., $\lim_{\|\theta\|\to\infty}\|\phi(\theta)\|_1=+\infty$, if $\operatorname{rank}(X)=n$. Hence, (12) effectively admits a minimizer when X is full row rank, implying that the estimator Ψ_1 is well-defined in this case, that is, the minimizing set of $\theta\mapsto\|\phi(\theta)\|_1$ is non empty.

Characterization of the LAD estimator. We start by characterizing the set $\Psi_1(\varpi^N)$. For this purpose, let us introduce some notation. For any candidate estimate $\theta \in \mathbb{R}^n$, consider a partition $(\mathbb{I}^-(\theta), \mathbb{I}^+(\theta), \mathbb{I}^0(\theta))$ of the set of indices $\mathbb{I} \triangleq \{1, \dots, N\}$ defined by

$$\begin{aligned}\mathbb{I}^-(\theta) &= \left\{t \in \mathbb{I} : y_t - \theta^\top x_t < 0\right\}\\ \mathbb{I}^+(\theta) &= \left\{t \in \mathbb{I} : y_t - \theta^\top x_t > 0\right\}\\ \mathbb{I}^0(\theta) &= \left\{t \in \mathbb{I} : y_t - \theta^\top x_t = 0\right\}.\end{aligned}$$

Theorem 1 (Characterization of Ψ_1 [49])**.**
Consider the data ϖ^N generated by the system (2) and the estimator Ψ_1 defined in (12). Then the following conditions are equivalent:

S0. $\theta^\star \in \Psi_1(\varpi^N)$

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

$$\sum_{t\in\mathbb{I}^+(\theta^\star)} x_t - \sum_{t\in\mathbb{I}^-(\theta^\star)} x_t = \sum_{t\in\mathbb{I}^0(\theta^\star)} \lambda_t x_t \tag{13}$$

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

$$\Big|\sum_{t\in\mathbb{I}^+(\theta^\star)} \eta^\top x_t - \sum_{t\in\mathbb{I}^-(\theta^\star)} \eta^\top x_t\Big| \le \sum_{t\in\mathbb{I}^0(\theta^\star)} \left|\eta^\top x_t\right| \tag{14}$$

The theorem characterizes $\Psi_1(\varpi^N)$ as the subset of $\mathbb{R}^n$ containing all parameter vectors $\theta^\star$ which satisfy (13) or (14). In particular, the true θ° lies in $\Psi_1(\varpi^N)$ if it satisfies those conditions, i.e., if

$$\inf_\alpha \left\{\|\alpha\|_\infty : X_{\mathbb{I}^0(\theta^\circ)}\alpha = z^\circ\right\} \le 1 \tag{15}$$

with $z^\circ = \sum_{t\in\mathbb{I}^+(\theta^\circ)} x_t - \sum_{t\in\mathbb{I}^-(\theta^\circ)} x_t$. Intuitively, by assuming that all the regressors x_t have the same order of magnitude, if the noise vector $\boldsymbol{v}$ is sparse enough, i.e. if the cardinality of $\mathbb{I}^0(\theta^\circ)$ is large enough and $\operatorname{rank}(X_{\mathbb{I}^0(\theta^\circ)}) = n$, then (13) is very likely to hold. Below we provide more strict conditions on the data ϖ^N which guarantee that $\Psi_1(\varpi^N)$ is a singleton. The result is indeed a reformulation of Theorem 4 in [49].

Corollary 1 (Uniqueness of the solution [49])**.**
Under the conditions of Theorem 1, the following statements are equivalent:

S0'. $\theta^\star$ is the unique element of $\Psi_1(\varpi^N)$

S1'. (13) holds and $\operatorname{rank}(X_S) = n$ where $S = \left\{t \in \mathbb{I}^0(\theta^\star) : \left|\lambda_t\right| < 1\right\}$.

S2'. (14) holds with strict inequality symbol for all $\eta \in \mathbb{R}^n$, $\eta \neq 0$.

Although Theorem 1 and Corollary 1 give complete formal characterizations of the estimator Ψ_1, they do not propose an explicit closed-form expression for the estimates (i.e., members of $\Psi_1(\varpi^N)$). Indeed, thanks to the convexity of the objective function $\theta \mapsto \|\phi(\theta)\|_1$, the estimates can be efficiently approximated through numerical algorithms.

As already discussed, the recoverability conditions S1' or S2' are likely to hold for the true parameter vector θ° if the noise $\boldsymbol{v}$ from (2) is sparse enough. In case these conditions do not hold naturally for a given set ϖ^N of data, it is possible, in principle, to reinforce them by appropriately weighting the data.

Corollary 2. Consider the data ϖ^N generated by the system (2). Assume that the disturbance $\boldsymbol{v} = \phi(\theta^\circ)$ has a sign pattern such that $\operatorname{rank}(X_{\mathbb{I}^0(\theta^\circ)}) = n$. Then there exist infinitely many different weighting matrices $W = \operatorname{diag}(w_1, \cdots, w_N)$ with $\sum_{t=1}^N w_t = 1$,

such that[3]

$$\Psi_1(W\boldsymbol{y}, XW) = \underset{\theta\in\mathbb{R}^n}{\arg\min} \left\| W\phi(\theta) \right\|_1 = \{\theta^\circ\}\,. \tag{16}$$

Proof. Since $X_{\mathbb{I}^0(\theta^\circ)}$ is assumed to have full row rank, there exists a vector α such that $z^\circ = X_{\mathbb{I}^0(\theta^\circ)}\alpha$. Let $\epsilon > 0$. Consider a diagonal weighting matrix $W(\epsilon) = \mathrm{diag}(w_1, \ldots, w_N)$ where $w_t = w'_t / \sum_{t=1}^N w'_t$ with the w'_t being defined by $w'_t = 1$ for $t \in \mathbb{I}^0(\theta^\circ)$ and $w'_t = 1/(\|\alpha\|_\infty + \epsilon)$ for $t \in \mathbb{I}^-(\theta^\circ) \cup \mathbb{I}^+(\theta^\circ)$. By invoking Condition S1' of Corollary 1, $\Psi_1(W(\epsilon)\boldsymbol{y}, XW(\epsilon)) = \{\theta^\circ\}$ if and only if there exists $\tilde{\alpha}$ such that $\|\tilde{\alpha}\|_\infty < 1$ and $1/(\|\alpha\|_\infty + \epsilon)z^\circ = X_{\mathbb{I}^0(\theta^\circ)}\tilde{\alpha}$. Note that this condition is fulfilled with $\tilde{\alpha} = \alpha/(\|\alpha\|_\infty + \epsilon)$. □

A problem is still that it is not possible to define a priori appropriate weighting which will favor the recovery of θ° unless the sign pattern of the error vector $\boldsymbol{v}$ is known. Nevertheless, a greedy weighting like the one described in Section 2.4 can be helpful[42].

An interesting question one might ask is whether the surrogate optimization problem arising in (12) can ever yield the solution to the original problem (7). In the event of such an equivalence, under which conditions does it occur? An answer is given in Proposition 3 below. To state this result we need to introduce the concept of ℓ_1 norm concentration ratio[50].

Definition 3 (r-th concentration ratio)**.** Let $X \in \mathbb{R}^{n\times N}$ be a matrix such that $\mathrm{rank}(X) = n$. We call *r-th concentration ratio* of the matrix X with respect to the ℓ_1 norm, the number $\xi_r^1(X)$ defined by

$$\xi_r^1(X) = \max_{\substack{I^c\subset\mathbb{I}:\\ |I^c|\leq r}} \sup_{\substack{\eta\in\mathbb{R}^n\\ \eta\neq 0}} \left[\frac{\|X_{I^c}^\top \eta\|_1}{\|X^\top\eta\|_1} \right] \tag{17}$$

Note that the supremum in (17) is indeed attainable under the condition that $\mathrm{rank}(X) = n$ and so, the supremum can be replaced by a maximum symbol. Moreover, $\xi_r^1(X)$ can be re-expressed as

$$\xi_r^1(X) = \max_{\substack{\eta\in\mathbb{R}^n\\ \eta\neq 0}} \left[\frac{\|X^\top\eta\|_{1,[r]}}{\|X^\top\eta\|_1} \right],$$

where $\|\cdot\|_{1,[r]}$ denotes the r-max norm on $\mathbb{R}^N$ which associates to each $z = [z_1 \ \cdots \ z_N]^\top$ the sum of the r largest entries of z in absolute value.

Proposition 3. Let $(\boldsymbol{y}, X) \in \mathbb{R}^N \times \mathbb{R}^{n\times N}$ be data generated by the system (2) such that $\mathrm{rank}(X) = n$ and the set $\left\{\theta \in \mathbb{R}^n : \|\phi(\theta)\|_0 \leq r\right\}$ is non empty for some $r \in \{1, \ldots, N\}$. Consider the definition (17) of $\xi_r^1(X)$ and the estimators Ψ_0 and Ψ_1 defined in (7) and (12) respectively. Then the following two statements hold:

(a) If $\xi_r^1(X) \leq 1/2$, then $\Psi_0(\varpi^N) \subset \Psi_1(\varpi^N)$

(b) If $\xi_r^1(X) < 1/2$, then there exists a unique $\theta^\star \in \mathbb{R}^n$ satisfying $\left\|\phi(\theta^\star)\right\|_0 \leq r$ and $\Psi_0(\varpi^N) = \Psi_1(\varpi^N) = \left\{\theta^\star\right\}$.

Proof. See Appendix A. □

According to Proposition 3, if we let $\pi_1^c(X)$ denote the maximum integer r for which $\xi_r^1(X) < 1/2$,

$$\pi_1^c(X) = \max\left\{r : \xi_r^1(X) < 1/2\right\}, \tag{18}$$

then $\Psi_0(\varpi^N) = \Psi_1(\varpi^N)$ whenever $\left\{\theta \in \mathbb{R}^n : \|\phi(\theta)\|_0 \leq \pi_1^c(X)\right\} \neq \emptyset$. We call $\pi_1^c(X)$ the number of worst-case outliers that the LAD estimator is able to correct. For matrices X of small sizes the threshold $\pi_1^c(X)$ can be exactly computed using an algorithm described in[43]. Unfortunately the numerical complexity of that algorithm is combinatorial and therefore grows quickly with the dimensions of X to an unaffordable level. We will discuss in Section 2.3 alternative methods for overestimating $\pi_1^c(X)$.

[3]For convenience, we will sometimes write $\Psi(\boldsymbol{y}, X)$ instead of $\Psi(\varpi^N)$.

Indeed, the statement (b) of Proposition 3 is necessary and sufficient in the following sense.

Theorem 2 (Necessary and Sufficient Condition[20])**.** Consider the data $(\boldsymbol{y}, X) \in \mathbb{R}^N \times \mathbb{R}^{n\times N}$ under the assumption that $\operatorname{rank}(X) = n$. Let r be an integer. Then the following two statements are equivalent:

(i)

$$\forall(\boldsymbol{y}, \theta) \in \mathbb{R}^N \times \mathbb{R}^n,\ \|\phi(\theta)\|_0 \leq r \quad \Rightarrow \quad \Psi_1(\varpi^N) = \left\{\theta\right\} \tag{19}$$

(ii)

$$\xi_r^1(X) < \frac{1}{2} \tag{20}$$

A proof of Theorem 2 can be found in[20]. By this theorem we can see that for any positive integer r, if the cardinality of $\operatorname{Supp}(\boldsymbol{\upsilon})$ is smaller than r with r satisfying $\xi_r^1(X) < 1/2$, then $\Psi_1(\varpi^N) = \left\{\theta^\circ\right\}$ that is, the estimator recovers exactly θ°.

2.2 | Estimation in the presence of both dense and sparse noise

Consider now the scenario where the noise sequence $\left\{\upsilon_t\right\}$ contains both dense and sparse components. To analyze the performance of the LAD estimator in this case, consider the set $\mathcal{S}_r^N$ of r-sparse sequences defined earlier. Based on the ℓ_1 norm we define the distance d_1 from a point $\boldsymbol{z} \in \mathbb{R}^N$ to a set $\mathcal{S} \subset \mathbb{R}^n$ by

$$d_1(\boldsymbol{z}, \mathcal{S}) = \inf_{w\in\mathcal{S}} \|\upsilon - w\|_1 . \tag{21}$$

For any $\theta \in \mathbb{R}^n$, let us define the distance from $\phi(\theta)$ to $\mathcal{S}_r^N$,

$$\delta_r(\theta) = d_1(\phi(\theta), \mathcal{S}_r^N). \tag{22}$$

It can be observed that $\delta_r(\theta)$ is equal to the sum of the $N - r$ smallest entries of $\phi(\theta)$ in absolute value. Based of this notation, consider the following lemma[50].

Lemma 1. Let $(\boldsymbol{y}, X) \in \mathbb{R}^N \times \mathbb{R}^{n\times N}$ be data generated by the system (2) under the assumption that $\operatorname{rank}(X) = n$. Consider the definitions (17) and (22) of $\xi_r^1(X)$ and $\delta_r(\theta)$ respectively. If $\xi_r^1(X) < 1/2$, then

$$\forall(\theta, \theta'), \quad \left\|\phi(\theta') - \phi(\theta)\right\|_1 \leq \frac{1}{1 - 2\xi_r^1(X)}\Big[\left\|\phi(\theta')\right\|_1 - \|\phi(\theta)\|_1 + 2\delta_r(\theta)\Big]. \tag{23}$$

A different version of this lemma can be found in[34]. By relying now on Lemma 1, we obtain the following result.

Theorem 3. Let $(\boldsymbol{y}, X) \in \mathbb{R}^N \times \mathbb{R}^{n\times N}$ be data generated by the system (2) under the assumption that $\operatorname{rank}(X) = n$. If $\xi_r^1(X) < 1/2$, then

$$\forall\hat{\theta} \in \Psi_1(\varpi^N), \quad \left\|X^\top\left(\hat{\theta} - \theta^\circ\right)\right\|_1 \leq \frac{2}{1 - 2\xi_r^1(X)} d_1(\boldsymbol{\upsilon}, \mathcal{S}_r^N) \tag{24}$$

where $d_1(\boldsymbol{\upsilon}, \mathcal{S}_r^N)$ is the sum of the $N - r$ smallest entries of $\boldsymbol{\upsilon}$ in absolute value.

Proof. The proof of this theorem follows directly by applying Lemma 1 with $\theta' = \hat{\theta}$ and $\theta = \theta^\circ$. Then since $\hat{\theta} \in \Psi_1(\varpi^N)$, we have $\left\|\phi(\hat{\theta})\right\|_1 - \|\phi(\theta^\circ)\|_1 \leq 0$. As a consequence, we obtain directly from (23) that

$$\forall\hat{\theta} \in \Psi_1(\varpi^N), \quad \left\|\phi(\hat{\theta}) - \phi(\theta^\circ)\right\|_1 \leq \frac{1}{1 - 2\xi_r^1(X)}\Big[\left\|\phi(\hat{\theta})\right\|_1 - \|\phi(\theta^\circ)\|_1 + 2\delta_r(\theta)\Big] \leq \frac{2}{1 - 2\xi_r^1(X)}\delta_r(\theta^\circ).$$

By observing that $\phi(\hat{\theta}) - \phi(\theta^\circ) = X^\top\left(\hat{\theta} - \theta^\circ\right)$, the above inequality means that

$$\forall\hat{\theta} \in \Psi_1(\varpi^N), \quad \left\|X^\top\left(\hat{\theta} - \theta^\circ\right)\right\|_1 \leq \frac{2}{1 - 2\xi_r^1(X)}\delta_r(\theta^\circ).$$

Finally, by invoking the definition of $\delta_r(\theta)$ in (22), the result follows from the observation that $\delta_r(\theta^\circ) = d_1(\phi(\theta^\circ), \mathcal{S}_r^N) = d_1(\boldsymbol{\upsilon}, \mathcal{S}_r^N)$. The last equality is a consequence of the system equation (2) by which $\boldsymbol{\upsilon} = \phi(\theta^\circ)$. □

Theorem 3 establishes indeed a property of resilience for the LAD estimator in the sense of Definition 1. The parametric error in (24) is measured in term of a data-dependent norm. It is possible to replace this norm by one which is independent of the data but at the price of some conservatism. To see this, introduce the number $\gamma_{1,p}(X)$ defined by

$$\gamma_{1,p}(X) = \inf_{\substack{\eta\in\mathbb{R}^n \\ \eta\neq 0}} \frac{\left\|X^\top \eta\right\|_1}{\|\eta\|_p} \tag{25}$$

with $\|\cdot\|_p$ referring to the vector p-norm, $p \in \{1, 2, \ldots, \infty\}$. The so-defined $\gamma_{1,p}(X)$ is guaranteed to be strictly positive if $\mathrm{rank}(X) = n$.

$$\forall \hat{\theta} \in \Psi_1(\varpi^N), \quad \left\|\hat{\theta} - \theta^\circ\right\|_p \leq \frac{2}{\gamma_{1,p}(X)(1 - 2\xi_r^1(X))} d_1(\boldsymbol{v}, \mathcal{S}_r^N). \tag{26}$$

An interest of a bound such as (26) is that it can serve, for example, as a basis for experiment design. In effect, one can envision to select the excitation input $\{u_t\}$ in (3) so that the error bound in (26) is as small as possible. This can be achieved if $\gamma_{1,p}(X)$ is made large and $\xi_r^1(X)$ is made as small as possible for a large N.

2.3 | On the numerical computation of $\xi_r^1(X)$

Most of the previous results are conditioned by an assumption on the ℓ_1 concentration ratio $\xi_r^1(X)$ defined in (17). Hence, to be able to check this condition, it may be desirable to numerically assess the value of the parameter $\xi_r^1(X)$. However, an exact computation of this quantity involves a nonconvex optimization problem for which there is no generically efficient algorithm. Hence, we discuss here some overestimates of $\xi_r^1(X)$ which are obtainable through convex optimization.

Definition 4 (self-decomposability amplitude[20,49])**.** Let $X \in \mathbb{R}^{n\times N}$ be such that $\nu_n(X) \leq N - 1$. We call *self-decomposability amplitude* of X, the number $\kappa(X)$ defined by

$$\kappa(X) = \max_{k\in\mathbb{I}} \min_{\gamma_k\in\mathbb{R}^{N-1}} \left\{ \left\|\gamma_k\right\|_\infty \,:\, x_k = X_{\neq k}\gamma_k \right\}. \tag{27}$$

The condition $\nu_n(X) \leq N - 1$ guarantees that $\mathrm{rank}(X_{\neq k}) = n$ for all $k \in \mathbb{I}$, with $X_{\neq k} \triangleq X_{\mathbb{I}\setminus\{k\}}$ being the matrix obtained from X by removing its k-th column. This in turn ensures that the constraint involved in the defining optimization problem of $\kappa(X)$ in (27) is always feasible. Note that (27) can be reformulated in a more compact form as

$$\kappa(X) = \min_{\Lambda\in\mathbb{R}^{N\times N}} \left\{ \|\Lambda\|_{\max} \,:\, X = X\Lambda,\ \mathrm{diag}(\Lambda) = 0 \right\} \tag{28}$$

with $\|\cdot\|_{\max}$ referring to the entrywise maximum norm of matrices. Achieving the condition $\mathrm{rank}(X_{\neq k}) = n$ for all $k \in \mathbb{I}$ in practice seems easy provided that the number $N$of measurements is large enough compared to the dimension n of X.

Lemma 2 (Estimation of ℓ_1 concentration ratio)**.**
Let $X \in \mathbb{R}^{n\times N}$.

- If $\mathrm{rank}(X) = n$ then

$$\xi_r^1(X) \leq \frac{r}{\beta_1^\star}, \tag{29}$$

where

$$\beta_1^\star = \min_{t=1,\ldots,N} \min_{\eta\in\mathbb{R}^n} \left\{ \left\|X^\top \eta\right\|_1 \,:\, x_t^\top \eta = 1 \right\}.$$

- If $\nu_n(X) \leq N - 1$, then

$$\xi_r^1(X) \leq \frac{r}{2T(\kappa(X))}, \tag{30}$$

where $T(\alpha) = 1/2(1 + 1/\alpha)$.

- If $\nu_n(X) \leq N - 1$, then

$$\xi_r^1(X) \leq \min_{\Lambda\in\mathbb{R}^{N\times N}} \left\{ \|\Lambda\|_{\infty,[r]} \,:\, X = X\Lambda,\ \mathrm{diag}(\Lambda) = 0 \right\} \tag{31}$$

where $\|\Lambda\|_{\infty,[r]}$ is the sum of the r largest infinity norms of the columns of Λ.

Proof. See Appendix B. □

The equations (29)-(31) provide us with numerically computable overestimates of $\xi_r^1(X)$. Therefore, to check whether $\xi_r^1(X)$ is smaller than one half, it suffices that these bounds be smaller than $1/2$. Note also that plugging these overestimates of $\xi_r^1(X)$ in the error bounds of the form (24), we obtain overestimates of these bounds as well.

2.4 | Some computational aspects

As we have seen in Theorem 3, the LAD estimator is robust in the sense that it is capable of returning an estimate with bounded error even in the presence of virtually unbounded noise $\boldsymbol{v}$. However this requires that the number of arbitrarily large values in $\boldsymbol{v}$ be small enough. The question we ask now is whether one can enhance the performance of this estimator even when this condition would not be naturally satisfied by the data. We discuss below two heuristics for achieving this objective.

Reweighted ℓ_1 (RWℓ_1). The first one is the ℓ_1-reweighted heuristic proposed in[42] which solves iteratively a weighted ℓ_1 optimization problem, where the weights are updated at each iteration using the previous estimate. More precisely, the algorithm generates iterates according to

$$\theta^{(k)} \in \arg\min_{\theta\in\mathbb{R}^n} \left\| W^{(k)}\phi(\theta)\right\|_1, \tag{32}$$

with $W^{(k)} = \mathrm{diag}(\tilde{w}_1^{(k)}, \cdots, \tilde{w}_N^{(k)})$, $\tilde{w}_t^{(k)} = w_t^{(k)} / \sum_i w_i^{(k)}$, with

$$w_t^{(k)} = \begin{cases} 1 & \text{if } k = 0 \\ \dfrac{1}{|y_t - x_t^\top \theta^{(k-1)}| + \epsilon} & \text{if } k \geq 1 \end{cases} \tag{33}$$

for $t = 1, \ldots, N$ and ϵ being a small positive number. The basic idea of this reweighting algorithm is as follows: push further to zero those entries of $\phi(\theta)$ which are seemingly close to zero (in the light of the previous estimate) by assigning larger weights to them in the next iteration. Other weighting strategies exist, e.g., the one described in[23].

Iterative k-smallest. The second heuristic attempts to minimize, for a given positive integer r, the sum of the $N - r$ smallest entries (in absolute value) of the prediction error vector $\phi(\theta)$. The starting point for deriving the algorithm is to write the sum-of-smallest-entries cost function as a difference of two convex functions. More specifically, this cost function can be written in the form

$$J_{dc}(\theta) = \|\phi(\theta)\|_1 - \|\phi(\theta)\|_{1,[r]} \tag{34}$$

with $\|\phi(\theta)\|_{1,[r]}$ denoting, as already specified, the sum of the r largest entries (in absolute value) of $\phi(\theta)$. To derive a robust estimator which is completely insensitive to gross errors, this is the ideal cost function one would like to minimize. Unfortunately, J_{dc} is a nonconvex function (indeed it is concave), hence making the numerical search for a minimizer challenging. A very simple algorithm can be obtained by linearizing locally the second function $\theta \mapsto \|\phi(\theta)\|_{1,[r]}$. If $\theta^{(k)}$ is the estimate obtained at iteration k, then we can approximate the second function $\|\phi(\theta)\|_{1,[r]}$ about $\theta^{(k)}$ by the linear function $\theta \mapsto \left\|\phi(\theta^{(k)})\right\|_{1,[r]} + \left(\theta - \theta^{(k)}\right)^\top h^{(k)}$, where $h^{(k)} \in \partial\left\|\phi(\theta^{(k)})\right\|_{1,[r]}$ is a subgradient of the function $\theta \mapsto \|\phi(\theta)\|_{1,[r]}$ at $\theta^{(k)}$. If all the r largest entries of $\phi(\theta^{(k)})$ in absolute value are nonzero, then $h^{(k)}$ is uniquely defined as

$$h^{(k)} = \sum_{t\in\mathbb{I}^-(\theta^{(k)})\cap\Sigma_r(\theta^{(k)})} x_t - \sum_{t\in\mathbb{I}^+(\theta^{(k)})\cap\Sigma_r(\theta^{(k)})} x_t,$$

where $\Sigma_r(\theta)$ is the index set for the r largest entries in absolute value of $\phi(\theta)$. Therefore, the iterative scheme is given by

$$\theta^{(k+1)} \in \arg\min_{\theta\in\mathbb{R}^n} \left[\|\phi(\theta)\|_1 - \left(\theta - \theta^{(k)}\right)^\top h^{(k)}\right] \tag{35}$$

The initial value $\theta^{(0)}$ of this algorithm can be possibly selected as $\theta^{(0)} \in \arg\min_{\theta\in\mathbb{R}^n} \|\phi(\theta)\|_1$ or totally at random.

We will compare the performances of the two algorithms (RWℓ_1 and k-smallest) in Section 5.

2.5 | Asymptotic analysis of the LAD estimator

One question one may ask for example is how the estimator Ψ_1 may behave when the number N of data ϖ^N goes to infinity. We provide in Theorem 4 below sufficient conditions under which the estimation error is guaranteed to be bounded. To state this result we will need the following concept of persistence of excitation.

Definition 5. The sequence $\{x(t)\}$ is said to be persistently exciting (PE) if there exist two positive numbers α and β and a fixed time horizon $T > 0$ such that

$$\alpha \|\eta\| \leq \sum_{k=t+1}^{t+T} |x_k^\top \eta| \leq \beta \|\eta\| \quad \forall (t, \eta) \in \mathbb{Z}_+ \times \mathbb{R}^n \tag{36}$$

for some norm $\|\cdot\|$ on $\mathbb{R}^n$.

For any $t \in \mathbb{Z}_+$, denote with q_t the integer part of t/T, i.e., $q_t = \lfloor t/T \rfloor$ with $\lfloor \cdot \rfloor$ referring to the floor function. Consider the $r^\star$ largest values of the noise signal magnitude $\{|v_t|\}$ in any interval $[t+1, t+T]$ of length T and pose $r_t = q_t r^\star$ and $p_t = \lfloor q_t r^\star / T \rfloor$. As defined before, let $\mathcal{S}^t_{r_t}$ be the set all r_t-sparse signals of length t. Also, let $\boldsymbol{v}_{1:t}$ be the vector formed with the first t entries of the vector $\boldsymbol{v}$.

Theorem 4. Assume that $\{x_t\}$ is persistently exciting in the sense of Definition 5 with a given time horizon T. Consider the above notations $r^\star$, q_t and p_t. If there exists $T_0 > 0$ such that for all $t \geq T_0$,

$$\frac{\beta}{\alpha} \frac{p_t + 1}{q_t} < \frac{1}{2}, \tag{37}$$

then

$$\left\| \hat{\theta}_t - \theta^\circ \right\| \leq \frac{2}{\alpha q_t \left(1 - 2\dfrac{\beta(p_t+1)}{\alpha q_t}\right)} d_1(\boldsymbol{v}_{1:t}, \mathcal{S}^t_{r_t}) \quad \forall \hat{\theta}_t \in \Psi_1(\varpi^t) \quad \forall t \geq T_0 \tag{38}$$

where $\|\cdot\|$ denotes the norm involved in the PE definition 5 and d_1 is defined as in (21).

Proof. The idea of the proof is to apply Theorem 3. For this purpose we start by estimating $\xi^1_{r_t}(X_{1:t})$. Note that by the PE assumption, we can write

$$\left\| X_{1:t}^\top \eta \right\|_1 \geq \sum_{i=1}^{q_t} \sum_{k=(i-1)T+1}^{iT} \left| x_k^\top \eta \right| \geq q_t \alpha \|\eta\|$$

and

$$\left\| X_{1:t}^\top \eta \right\|_{1,[r_t]} \leq \sum_{i=1}^{p_t+1} \sum_{k=(i-1)T+1}^{iT} |x_k^\top \eta| \leq (p_t + 1)\beta \|\eta\|$$

It follows that

$$\xi^1_{r_t}(X_{1:t}) = \sup_{\eta \neq 0} \frac{\left\| X_{1:t}^\top \eta \right\|_{1,[r_t]}}{\left\| X_{1:t}^\top \eta \right\|_1} \leq \frac{p_t + 1}{q_t} \frac{\beta}{\alpha}$$

which shows that $\xi^1_{r_t}(X_{1:t}) < 1/2$ when (37) holds. By virtue of Theorem 3 (see Eq. (24)), we then have

$$q_t \alpha \left\| \hat{\theta}_t - \theta^\circ \right\| \leq \left\| X_{1:t}^\top (\hat{\theta}_t - \theta^\circ) \right\|_1 \leq \frac{2}{1 - 2\xi^1_{r_t}(X_{1:t})} d(\boldsymbol{v}_{1:t}, \mathcal{S}^t_{r_t}) \quad \forall \hat{\theta}_t \in \Psi_1(\varpi^t).$$

Now the result follows by noticing from the above upper-bound on $\xi^1_{r_t}(X_{1:t})$ that $1 - 2\xi^1_{r_t}(X_{1:t}) \geq 1 - 2\dfrac{p_t+1}{q_t}\dfrac{\beta}{\alpha}$. □

The condition (37) of the theorem constrains the frequency of appearance of the large values in $\{v_t\}$ with regards to the richness of the regression data. In words, the claim is that the estimation error does not depend on the $r^\star$ largest values of the true noise sequence showing up in any interval of length T provided that (37) holds.

Under the conditions of Theorem 4, if the noise sequence is strictly sparse and is such that the $t - r_t$ smallest entries of $\boldsymbol{v}_{1:t}$ are zero, then $\hat{\theta}_t = \theta^\circ$ for all $t \geq T_0$. More generically, when the noise $\boldsymbol{v}$ has some dense component, we can state the following result:

Corollary 3. Under the conditions of Theorem 4 assume further that

$$\sup_{t \geq T_0} \frac{\beta}{\alpha} \frac{p_t + 1}{q_t} < \delta,$$

for some constant δ obeying $0 < \delta < 1/2$. If the $T - r^\star$ smallest entries (in absolute value) of $\boldsymbol{v}_{t+1:t+T}$ in any time interval $[t+1, t+T]$ are uniformly bounded by $\epsilon \geq 0$, then any $\hat{\theta}_t \in \Psi_1(\varpi^t)$ satisfies

$$\limsup_{t \to +\infty} \left\| \hat{\theta}_t - \theta^\circ \right\| \leq \frac{2}{\alpha(1-2\delta)} (T - r^\star)\epsilon.$$

Proof. Departing from (38), it is immediate that

$$\left\| \hat{\theta}_t - \theta^\circ \right\| \leq \frac{2}{\alpha(1-2\delta)} \frac{d_1(\boldsymbol{v}_{1:t}, \mathcal{S}^t_{r_t})}{q_t}$$

As already remarked in Section 2.2 (see, e.g., the comment on Eq. (24), Theorem 3), $d_1(\boldsymbol{v}_{1:t}, \mathcal{S}^t_{r_t})$ is equal to the sum in absolute value of the $t - r_t$ smallest entries of $\boldsymbol{v}_{1:t}$ (which, by the assumption of the corollary, are all bounded by ϵ). Therefore,

$$\frac{d_1(\boldsymbol{v}_{1:t}, \mathcal{S}^t_{r_t})}{q_t} \leq \frac{(t - r_t)\epsilon}{q_t} = \left(\frac{t}{q_t} - r^\star\right)\epsilon \leq \left(\frac{T}{q_t} + T - r^\star\right)\epsilon$$

The last inequality is indeed a consequence of the definition of q_t by which we have $q_t \leq t/T < q_t + 1$ and consequently that $T \leq t/q_t < T + T/q_t$. Combining with the previous step gives

$$\left\| \hat{\theta}_t - \theta^\circ \right\| \leq \frac{2}{\alpha(1-2\delta)} \left(\frac{T}{q_t} + T - r^\star\right)\epsilon$$

□

Now the claim of the corollary follows by observing that $T/q_t \to 0$ as $t \to +\infty$.

3 | SWITCHED ARX SYSTEM IDENTIFICATION

3.1 | The SARX identification problem

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

$$y_t = a^1_{\sigma(t)} y_{t-1} + \cdots + a^{n_a}_{\sigma(t)} y_{t-n_a} + (b^1_{\sigma(t)})^\top u_{t-1} + \cdots + (b^{n_b}_{\sigma(t)})^\top u_{t-n_b} + e_t \tag{39}$$

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) are the output and input lags (also called the orders of the system). $\{e_t\}$ models potential model mismatch and measurement noise. $\sigma(t) \in \mathbb{S} \triangleq \{1, \ldots, s\}$ 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 (40) is called a Switched Auto-Regressive eXogenous (SARX) model. For convenience, we rewrite (39) in the form

$$y_t = x_t^\top \theta^\circ_{\sigma(t)} + e_t, \tag{40}$$

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

$$\theta^{\circ}_{\sigma(t)} = \left[a^1_{\sigma(t)} \ \cdots \ a^{n_a}_{\sigma(t)} \ (b^1_{\sigma(t)})^\top \ \cdots \ (b^{n_b}_{\sigma(t)})^\top\right]^\top \tag{41}$$

and $x_t \in \mathbb{R}^n$ is the regressor at time $t \in \mathbb{Z}_+$ defined as in (3).

We consider the problem of inferring a model of the form (40) from a finite collection of measurements $\left\{(x_t, y_t)\right\}_{t=1}^N$ under the assumption that the switching signal $\{\sigma(t)\}$ is unknown. This means that we do not know beforehand which data pair is associated with which parameter vector. We will assume 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.
- The parameter vectors $\left\{\theta^{\circ}_i\right\}_{i\in\mathbb{S}}$ defining the subsystems of the SARX (40) are pairwise distinct, that is, for all $(i, j) \in \mathbb{S}^2$ with $i \neq j$, we have $\theta^{\circ}_i \neq \theta^{\circ}_j$.
- Each individual ARX subsystem is minimal in the ordinary sense.[4]

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[51,52], the second assumption implies minimality of the SARX system. The interested reader is referred to the papers[52,53] 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.

A geometrical interpretation. From a geometrical perspective, the switched system identification problem formulated above is equivalent to that of subspace clustering[54,26,25], i.e., the problem of estimating subspaces from unlabeled data that lie in the union of those subspaces. In effect, if we neglect the noise and introduce the notations,

$$\bar{\theta}_i = [1 \ \ \theta_i^\top]^\top \text{ and } \bar{x}_t = [y_t \ \ -x_t^\top]^\top, \tag{42}$$

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 $\left\{\bar{x}_t\right\}_{t=1}^N$ 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 data set.

3.2 | The sparse optimization approach

One approach to solve the switched system identification problem consists in viewing the equation (40) as that of a single linear model affected by sparse noise[22]. To discuss the rationale of this approach, consider an arbitrary parameter vector θ°_i of the SARX system. Then (40) can be written as

$$y_t = x_t^\top \theta^{\circ}_i + v_{it} + e_t \tag{43}$$

where $v_{it} = x_t^\top(\theta^{\circ}_{\sigma(t)} - \theta^{\circ}_i)$. The so-defined sequence $\left\{v_{it}\right\}$ is sparse to some degree since $v_{it} = 0$ whenever the subsystem i is activated, i.e., whenever $\sigma(t) = i$. For the sake of clarity, assume for now that the noise sequence $\left\{e_t\right\}$ is identically null. Let $\theta \in \mathbb{R}^n$ denote a candidate parameter vector and consider the notation $\phi(\theta)$ introduced in (5) for the vector of prediction errors. Then we can observe that if $\theta = \theta_i$ for some $i \in \{1, \dots, s\}$, then $\phi(\theta)$ is a sparse vector. More precisely, if we denote with N_i the number of data (x_t, y_t) generated by the subsystem indexed by i, then $\phi(\theta)$ contains at least N_i zero entries. By relying on our earlier discussion, we can naturally search for one parameter vector θ°_i of the SARX system (40) by solving the sparse

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

optimization problem

$$\min_{\theta} \|\phi(\theta)\|_0 . \tag{44}$$

Trying to solve problem (44) is equivalent to attempting to find a homogeneous hyperplane (or a vector $\bar{\theta}$) that contains (that is orthogonal to) as many data vectors $\bar{x}_t$ as possible.
If all the submodels are sufficiently excited within the data $\{x_t\}_{t=1}^N$ then, as suggested by the following proposition, the solution to problem (44) is a PV representing one of the constituent submodels of system (40).

Proposition 4 (Noise-free data[22])**.** Let ϖ^N be data generated by the SARX system (40) under noise-free assumption ($e = 0$). Assume that each subsystem has generated a sufficiently large number of data in the sense that $\left|\mathbb{I}^0(\theta_i^\circ)\right| \geq s\nu_n(X)$ for all $i \in \mathbb{S}$ with s being the number of subsystems in (40). Then

$$\Psi_0(\varpi^N) \subset \{\theta_1^\circ, \ldots, \theta_s^\circ\} \tag{45}$$

Remark 1. Note that Proposition 4 above is indeed less restrictive than its analogue Lemma 7 of[22]. It can be viewed as a special case of Proposition 5 whose proof is given in Appendix C.

Next, we characterize the uniqueness of the minimizer of (44) in terms of the n-genericity index of the data matrix X.

Theorem 5 ([22])**.** Let $\varpi^N = (y, X) \in \mathbb{R}^N \times \mathbb{R}^{n\times N}$ be data generated by the SARX system (40) under noise-free assumption ($e = 0$) and pose $\phi(\theta) = y - X^\top\theta$. Then the following statements hold true.

1. If there is a $\theta^\star \in \mathbb{R}^n$ such that $\left\|\phi(\theta^\star)\right\|_0 \leq (N - \nu_n(X))/2$, then
$$\Psi_0(\varpi^N) = \{\theta^\star\}. \tag{46}$$
2. If in addition, $\left|\mathbb{I}^0(\theta_i^\circ)\right| \geq s\nu_n(X)$ for all $i \in \mathbb{S}$ and $N \geq (2s-1)\nu_n(X)$, then
$$\theta^\star \in \{\theta_1^\circ, \ldots, \theta_s^\circ\}.$$

We observe that Theorem 5, as stated above, is indeed a refinement of the one in[22]. Its proof follows directly from Propositions 2 and 4.

Noise-aware sparse optimization. In case the noise is not equal to zero in the data-generating system (40), then solving problem (44) 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

$$\begin{aligned} &\min_{(\theta,\xi)\in\mathbb{R}^n\times\mathbb{R}_+^N} \|\xi\|_0 \\ &\text{s.t.} \quad |y_t - x_t^\top\theta| \leq \varepsilon + \xi_t,\ t = 1, \ldots, N. \end{aligned} \tag{47}$$

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 $\theta \in \{\theta_1^\circ, \ldots, \theta_s^\circ\}$, 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 (47) can be written in a more compact form as

$$\min_{\theta\in\mathbb{R}^n} \|\phi(\theta)\|_{0,\varepsilon} \tag{48}$$

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

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

for any $a = [a_1 \ \cdots \ a_N]^\top \in \mathbb{R}^N$. In other 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 (48) with respect to the goal of recovering the parameter vectors of system (40). This is discussed next. For notational convenience, let us introduce the number $\delta(X)$ defined for $X \in \mathbb{R}^{n\times N}$ by

$$\delta(X) = \max_{\substack{I\subset\mathbb{I}\\ |I|\geq \nu_n(X)}} \frac{\sqrt{|I|}}{\lambda_{\min}^{1/2}\left(X_I X_I^\top\right)} \tag{49}$$

where $\mathbb{I} = \{1, \dots, N\}$ refers to the column index set of X. 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 5 (Noisy data)**.** Let $(\boldsymbol{y}, X) \in \mathbb{R}^N \times \mathbb{R}^{n\times N}$ be data generated by the SARX system (40) 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 $\left|\mathbb{I}^{\leq\varepsilon}(\theta_i^\circ)\right| \geq s\nu_n(X)$ for all $i \in \mathbb{S}$, where $\mathbb{I}^{\leq\varepsilon}(\theta_i^\circ) = \left\{t \in \mathbb{I} : |y_t - x_t^\top\theta_i^\circ| \leq \varepsilon\right\}$. Then with $\phi(\theta) = \boldsymbol{y} - X^\top\theta$, it holds that

$$\forall \hat{\theta} \in \arg\min_{\theta\in\mathbb{R}^n} \|\phi(\theta)\|_{0,\varepsilon}\,, \exists i^* \in \mathbb{S}, \left\|\hat{\theta} - \theta_{i^*}^\circ\right\|_2 \leq 2\varepsilon\delta(X). \tag{50}$$

Proof. See Appendix C. □

It is interesting to note that (45) is a special case of (50) corresponding to the scenario when the noise is absent ($\varepsilon = 0$).

3.3 | Recoverability of the true parameter vectors

As already alluded to in Section 2.1.2, it is computationally preferable to implement a convex surrogate of the ℓ_0 estimator. Therefore, after having motivated in Section 3.2 the application of sparse optimization to switched system identification, we turn now to the LAD estimator for computing numerically the estimates.

Recoverability of the true PVs through solving a sequence of ℓ_1 problems. We now propose some conditions on the data generated by (40) which allow for an exact recovery of all the true PVs by convex optimization.
For $\theta \in \mathbb{R}^n$, let $\mathbb{I}^c(\theta) = \left\{t \in \mathbb{I} : y_t - x_t^\top\theta \neq 0\right\}$ collect all the data indices $t \in \mathbb{I}$ at which the prediction error induced by θ is nonzero. Define $X^1 = X$ and for any $j = 2, \dots, s$, let $X^j = X_{\mathbb{I}^c(\theta_{j-1}^\circ)\cap\cdots\cap\mathbb{I}^c(\theta_1^\circ)}$ be the matrix formed with the columns x_t of X which are indexed by $\mathbb{I}^c(\theta_{j-1}^\circ) \cap \cdots \cap \mathbb{I}^c(\theta_1^\circ)$ with the θ_j° representing the true parameter vectors. Similarly, we define the column vectors $\{\boldsymbol{y}^j\}$ by: $\boldsymbol{y}^1 = \boldsymbol{y}$ and $\boldsymbol{y}^j = \boldsymbol{y}_{\mathbb{I}^c(\theta_{j-1}^\circ)\cap\cdots\cap\mathbb{I}^c(\theta_1^\circ)}$ for $j = 2, \dots, s$. With these notations, we present below an immediate corollary to Theorem 2, which is relevant to the linear switched identification problem.

Theorem 6. Consider the data $(\boldsymbol{y}, X) \in \mathbb{R}^N \times \mathbb{R}^{n\times N}$ generated by the SARX system (40) under the assumption that the noise $\{e_t\}$ is zero. Let $\left\{(\boldsymbol{y}^j, X^j)\right\}_{j=1}^s$ be defined as above. Consider the notation π_1^c introduced in (18) and assume that:

- For all $j = 1, \dots, s$, the matrix X^j satisfies $\operatorname{rank}(X^j) = n$,
- For all $j = 1, \dots, s$, $\left|\mathbb{I}^c(\theta_j^\circ) \cap \cdots \cap \mathbb{I}^c(\theta_1^\circ)\right| \leq \pi_1^c(X^j)$.

Then

$$\Psi_1(\boldsymbol{y}^j, X^j) = \arg\min_{\theta\in\mathbb{R}^n} \left\|\boldsymbol{y}^j - (X^j)^\top\theta\right\|_1 = \left\{\theta_j^\circ\right\} \ \forall i = 1, \dots, s$$

i.e., all the true parameter vectors $\left\{\theta_1^\circ, \dots, \theta_s^\circ\right\}$ can be extracted one after another by solving ℓ_1 minimization problems of the form (12).

Proof. The theorem is structurally similar to Theorem 14 in[22]; the two theorems only differ in their respective assumptions. Hence, their proofs are quite similar as well. Note that the full row rank condition for the matrices X^j ensures that the numbers $\pi_1^c(X^j)$ are well-defined through the concentration ratios $\xi_r^1(X^j)$ (see Definition 3). As to the conclusion of the theorem, it

follows from Proposition 3 and the definition of π_1^c. In effect, if $\left|\mathbb{I}^c(\theta_1^\circ)\right| = \left\|\boldsymbol{y}^1 - (X^1)^\top\theta_1^\circ\right\|_0 \leq \pi_1^c(X^1)$ then $\xi_r^1(X^1) < 1/2$ with $r = \left|\mathbb{I}^c(\theta_1^\circ)\right|$ and hence, by Proposition 3, we have $\Psi_1(\boldsymbol{y}^1, X^1) = \left\{\theta_1^\circ\right\}$. Repeating this reasoning for all the X^j gives the conclusion. □

To illustrate the condition of Theorem 6, 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 j, $\pi_1^c(X^j)$ is about one third of the number of columns in X^j. Then $(134, 45, 21)$ 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 is empirically discussed in[22], recovery of the PVs is still possible beyond the theoretical conditions thanks to the ℓ_1 reweighted scheme (see Section 2.4).

Summary of the identification algorithm. We have seen that we can identify one of the s parameter vectors of a switched system such as (40) from the whole dataset by applying an appropriate (robust) sparsity-inducing identifier. If there is one submodel i satisfying $\left|\mathbb{I}^c(\theta_i^\circ)\right| \leq \pi_1^c(X)$, the RW$\ell_1$ algorithm (see Section 2.4) will find (after only one iteration) a vector θ^* in the set $\left\{\theta_1^\circ, \ldots, \theta_s^\circ\right\}$ of the true parameter vectors. If this condition is not fulfilled, the RWℓ_1 algorithm may not converge towards a point in $\left\{\theta_1^\circ, \ldots, \theta_s^\circ\right\}$. However, as argued in[42,22] and suggested by different experiments reported therein, the algorithm is likely to find the vector θ^* that realizes the sparsest error $\phi(\theta)$. According to Proposition 4 and Theorem 5, such a point θ^* is likely to be in $\left\{\theta_1^\circ, \ldots, \theta_s^\circ\right\}$ when enough rich data are available.

Without loss of generality, we can denote with $\hat{\theta}_1$, i.e., the estimate of θ_1°, the point of $\left\{\theta_1^\circ, \ldots, \theta_s^\circ\right\}$ to which the algorithm converges when it is run over the entire mixed dataset. 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 first the data generated by that submodel. The indices of such data can be determined as

$$\mathcal{I}(\hat{\theta}_1) = \left\{ t \in \{1, \ldots, N\} \,:\, \frac{|\bar{x}_t^\top \hat{\bar{\theta}}_1|}{\|\bar{x}_t\|_2 \cdot \|\hat{\bar{\theta}}_1\|_2} \leq \text{Thres} \right\} \tag{51}$$

where it is assumed that Tresh $\in [0,\ 1]$ is a tolerance threshold and $\hat{\bar{\theta}}_1 = [1\ \ \hat{\theta}_1^\top]^\top$. From the data indexed by $\mathbb{I} \setminus \mathcal{I}(\hat{\theta}_1)$, we estimate θ_2°. We can repeat this procedure until all the PVs are identified (see Algorithm 1 for a summary of all the steps). Note any robust estimator can be used in Step 3.1 of Algorithm 1.

Algorithm 1 Identification of all PVs

1. **Inputs:** $(\boldsymbol{y}, X) \in \mathbb{R}^N \times \mathbb{R}^{n \times N}$
2. **Initialization:** $\mathcal{S} \leftarrow \emptyset$, $J \leftarrow \{1, \ldots, N\}$
3. **While** $|J| \neq 0$, repeat:
   - 3.1: Estimate a submodel by a robust/sparse identifier (e.g., RW-ℓ_1 of k-smallest algorithms) based on the data $(\boldsymbol{y}_J, X_J)$
   - 3.2: Record the identified PV: $\mathcal{S} \leftarrow \mathcal{S} \cup \left\{\hat{\theta}\right\}$
   - 3.4: Remove indices of data associated with the identified submodel:
     $$J \leftarrow J \setminus (J \cap \mathcal{I}(\hat{\theta})),$$
     with $\mathcal{I}(\hat{\theta})$ defined as in Eq. (51).
4. **Return** $\mathcal{S}$ and $s = |\mathcal{S}|$.

3.4 | Uncertainty sets induced by noise

Consider now the more realistic situation where the dense noise sequence $\{e_t\}$ in (40) is nonzero but is bounded. In this case, the identification process is unlikely to return the true parameter vectors. Instead, each PV estimate will come out with an associated uncertainty set. This is typically due to the fact that the dense noise sequence is only known to be bounded.

A theoretical characterization of the uncertainty. Assume that the noise $\{e_t\}$ acting in the SARX system (40) is bounded. Let $\varepsilon \geq 0$ be such that $|e_t| \leq \varepsilon$ for all $t = 1, \dots, N$. Assume that the two conditions of Theorem 6 are satisfied with each $\mathbb{I}^c(\theta_i^\circ)$ replaced now by $\mathbb{I}^{>\varepsilon}(\theta_i^\circ) = \{t \in \mathbb{I} : y_t - x_t^\top \theta_i^\circ > \varepsilon\}$. Let r_i be the cardinality of $\mathbb{I}^{>\varepsilon}(\theta_i^\circ)$. Denote with $\hat{\theta}_i$ the estimate (by the approach discussed earlier) of θ_i°, $i = 1, \dots, s$ 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. Then according to Theorem 3, if $\xi^1_{r_i}(X^i) < 1/2$, we have

$$\left\|\hat{\theta}_i - \theta_i^\circ\right\|_2 \leq R_i \triangleq \frac{2}{\gamma_{1,2}(X^i)(1 - 2\xi^1_{r_i}(X^i))} |\mathbb{I}^{\leq\varepsilon}(\theta_i^\circ)|\varepsilon \tag{52}$$

where $\gamma_{1,2}(X^i)$ is defined as in (25). This means that for all $i = 1, \dots, s$, the estimate $\hat{\theta}_i$ lies in the ball centered at the true PV θ_i° and having a radius R_i. The size of these balls increases naturally with the magnitude of the noise. It is desirable that the s uncertainty balls defined around the different PVs do not intersect. This requires that we put a distinguishability condition on the true parameter vectors $\{\theta_i^\circ\}$,

$$\min_{i \neq j} \left\|\theta_i^\circ - \theta_j^\circ\right\|_2 > 2 \max_{i \in \mathbb{S}} R_i.$$

We close this section by observing that the principle of the sparse optimization-based method discussed here for switched system identification can be extended to some other problems involving hybrid systems: state estimation and control for switched linear systems[55,56].

4 | EXTENSION TO MIMO SYSTEMS

4.1 | Mutivariable regression

We now consider an estimation scenario where the output y_t is multivariate. More precisely, consider a data-generating system described by an equation of the form

$$y_t = A^\circ x_t + v_t \tag{53}$$

where $y_t \in \mathbb{R}^m$, $x_t \in \mathbb{R}^n$, $v_t \in \mathbb{R}^m$ are respectively the output, the regressor and the measurement noise at time t; A° is an unknown parameter matrix to be determined. We make the assumption that the vector-valued sequence $\{v_t\}$ is block-sparse in the sense that the scalar-valued sequence $\{\|v_t\|\}$ contains a relatively large proportion of zeros. Suppose that we have collected N noisy measurements $Y = [y_1 \ y_2 \ \cdots \ y_N] \in \mathbb{R}^{m \times N}$ of the output and $X = [x_1 \ x_2 \ \cdots \ x_N] \in \mathbb{R}^{n \times N}$ of corresponding regressors. The estimation problem is then to infer an estimate of the matrix A°.

To solve this problem we define, similarly to (12), a nonsmooth optimization-based estimator Ψ_2 by

$$\Psi_2(Y, X) = \arg\min_{A \in \mathbb{R}^{m \times n}} \|Y - AX\|_{2,\text{col}} \tag{54}$$

where $\|Y\|_{2,\text{col}} = \sum_{t=1}^N \|y_t\|_2$.

Let $\mathbb{I} = \{1, \dots, N\}$ be the index set for the data and for a matrix $A \in \mathbb{R}^{m \times n}$, define $\mathbb{I}^0(A) = \{t \in \mathbb{I} : y_t - Ax_t = 0\}$ and $\mathbb{I}^c(A) = \{t \in \mathbb{I} : y_t - Ax_t \neq 0\}$. Using these notations, a formal characterization of the estimator Ψ_2 is given as follows.

Theorem 7 ([49]). Let $A^\star \in \mathbb{R}^{m \times n}$. Consider the data (Y, X) generated by (53). Then the following statements are equivalent:

T0. $A^\star \in \Psi_2(Y, X)$

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

$$\sum_{t\in\mathbb{I}^c(A^\star)} v_t^\star x_t^\top + \sum_{t\in\mathbb{I}^0(A^\star)} \beta_t x_t^\top = 0, \tag{55}$$

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

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

$$\Big|\sum_{t\in\mathbb{I}^c(A^\star)} v_t^{\star\top}\Lambda x_t\Big| \leq \sum_{t\in\mathbb{I}^0(A^\star)} \left\|\Lambda x_t\right\|_2. \tag{56}$$

T3. The condition

$$\inf_{Z\in\mathbb{R}^{m\times p}} \left\{\|Z\|_{2,\infty} : V^\star X_{\mathbb{I}^c(A^\star)}^\top = Z X_{\mathbb{I}^0(A^\star)}^\top\right\} \leq 1 \tag{57}$$

holds with $p = \left|\mathbb{I}^0(A^\star)\right|$ and $V^\star$ being a matrix formed with the unit 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 two conditions holds:

T1'. (55) holds and $\operatorname{rank}(X_{\mathcal{T}}) = n$ where $\mathcal{T} = \left\{t \in \mathbb{I}^0(A^\star) : \left\|\beta_t\right\|_2 < 1\right\}$.

T2'. (56) holds with strict inequality symbol for all $\Lambda \in \mathbb{R}^{m\times n}$, $\Lambda \neq 0$.

For A° to lie in $\Psi_2(Y,X)$ it must satisfy (55)-(57). Again, similarly as in the case of the estimator Ψ_1 we see that the required conditions are all the more likely to be satisfied as (a) all the regressors have the same order of magnitude; (b) the matrix $X_{\mathbb{I}^0(A^\circ)}$ is generic (full row rank) ; (c) the cardinality of $\mathbb{I}^0(A^\circ)$ is large enough compared to that of $\mathbb{I}^c(A^\circ)$.

Now, similarly as in Definition 3 let us introduce a new measure of genericity of the data matrix X. It can be viewed as a generalization of the one in Definition 3.

Definition 6 (r-th concentration ratio)**.** Let $X \in \mathbb{R}^{n\times N}$ be a matrix such that $\operatorname{rank}(X) = n$. We call *r-th concentration ratio* of the matrix X with respect to the sum of ℓ_2 norm, the number $\xi_r^2(X)$ defined by

$$\xi_r^2(X) = \max_{\substack{I^c\subset\mathbb{I}\\ |I^c|=r}} \max_{\substack{\Lambda\in\mathbb{R}^{m\times n}\\ \Lambda\neq 0}} \frac{\left\|\Lambda X_{I^c}\right\|_{2,\mathrm{col}}}{\|\Lambda X\|_{2,\mathrm{col}}} \tag{58}$$

Note in passing that by following a similar reasoning as in the proof of Lemma 2, it is possible to show that $\xi_r^2(X)$ satisfies the upper bound in (31) provided that $\nu_n(X) \leq N-1$.

Corollary 4 ([49])**.** Let r be an integer and $Y = \begin{bmatrix} y_1 & \cdots & y_N\end{bmatrix} \in \mathbb{R}^{m\times N}$ be the output matrix generated by system (53). Then the following three statements are equivalent.

(j)

$$\forall (A,Y) \in \mathbb{R}^{m\times n}\times\mathbb{R}^{m\times N},\ \left|\mathbb{I}^0(A)\right| \leq r \quad \Rightarrow \quad A \in \Psi_2(Y,X) \tag{59}$$

(jj)

$$\xi_r^2(X) \leq 1/2 \tag{60}$$

(jjj)

$$\max_{\substack{(I,I^c):\\ |I^c|=r}} \max_{V\in\mathbb{B}^{m\times|I^c|}} \min_{Z\in\mathbb{R}^{m\times|I|}} \left\{\|Z\|_{2,\infty} :\ X_{I^c}V^\top = X_I Z^\top\right\} \leq 1 \tag{61}$$

with $\mathbb{B}^{m\times q} = \left\{\begin{bmatrix} b_1 & \cdots & b_q\end{bmatrix} \in \mathbb{R}^{m\times q},\ b_i \in \mathcal{B}_2^m(0,1)\right\}$ and the outer maximization being taken over all partitions (I, I^c) of $\mathbb{I}$ such that $|I^c| = r$.

For a matrix $V = [v_1 \cdots v_N] \in \mathbb{R}^{m\times N}$, let $I^c(V) = \left\{t \in \mathbb{I} : v_t \neq 0\right\}$ refer to the index set of the nonzero columns of V. Let $\mathcal{S}_r^{m\times N} = \left\{V \in \mathbb{R}^{m\times N} : |I^c(V)| \leq r\right\}$. Consider also the notation $d_2(V, \mathcal{S}_r^{m\times N})$ for the distance from V to $\mathcal{S}_r^{m\times N}$,

$$d_2(V, \mathcal{S}_r^{m\times N}) = \inf_{W\in\mathcal{S}_r^{m\times N}} \|V - W\|_{2,\mathrm{col}} \tag{62}$$

Note that if $r = 0$ then $\mathcal{S}_r^{m\times N} = \{0\}$ so that $d_2(V, \mathcal{S}_r^{m\times N}) = \|V\|_{2,\text{col}}$.

Theorem 8. Let $(Y, X) \in \mathbb{R}^{m\times N} \times \mathbb{R}^{n\times N}$ be data generated by system (53). Consider the definition (58) of $\xi_r^2(X)$ under the assumption that $\text{rank}(X) = n$. If $\xi_r^2(X) < 1/2$ then the following holds:

$$\forall \hat{A} \in \Psi_2(Y, X), \quad \left\|(\hat{A} - A^\circ)X\right\|_{2,\text{col}} \leq \frac{2}{1 - 2\xi_r^2(X)} d_2(V, \mathcal{S}_r^{m\times N}). \tag{63}$$

The proof of this theorem is quite similar to that of Theorem 3 and therefore omitted.

Theorem 8 implies that if the true noise matrix V lies in $\mathcal{S}_r^{m\times N}$, i.e., if it is $r-$(block) sparse, then $\Psi_2(Y, X) = \{A^\circ\}$. Otherwise the estimation error bound depends on the distance $d_2(V, \mathcal{S}_r^{m\times N})$ from V to $\mathcal{S}_r^{m\times N}$. As mentioned earlier (see Section 2.2) this distance is equal to the sum of the $N - r$ smallest elements of the set $\left\{\|v_i\|_2 : t = 1, \dots, N\right\}$.

If we define

$$\gamma_{2,p}(X) = \inf_{\substack{\Lambda\in\mathbb{R}^{m\times n} \\ \Lambda\neq 0}} \frac{\|\Lambda X\|_{2,\text{col}}}{\|\Lambda\|_p}$$

then the parametric estimation error can be bounded with respect to the matrix p-norm as

$$\left\|\hat{A} - A^\circ\right\|_p \leq \frac{2}{\gamma_{2,p}(X)(1 - 2\xi_r^2(X))} d_2(V, \mathcal{S}_r^{m\times N}) \quad \forall \hat{A} \in \Psi_2(Y, X). \tag{64}$$

Remark 1. The discussion of Section 2.4 on sparsity enhancing through iterative algorithms can be extended to the multivariable case. Both the reweighted and the difference of convex algorithms are still applicable with some adjustments, see e.g.,[57] for an application of this framework to switched state-space model identification.

4.2 | Resilient state estimation

In this section we illustrate how sparse optimization can be used to design resilient state estimators, see, e.g.,[29,27,28]. For this purpose, let us consider a linear time-invariant system subject to disturbances

$$\Sigma : \begin{cases} x_{t+1} &= Ax_t + w_t \\ y_t &= Cx_t + f_t \end{cases} \tag{65}$$

where $x_t \in \mathbb{R}^n$ is the state of the system, y_t is the output, w_t and f_t are the process and measurement noises respectively. (A, C) are constant real matrices with appropriate dimensions. We consider the problem of estimating the state matrix $\mathcal{X} = [x_0 \ \cdots \ x_N]$ from a collection $Y = [y_0 \ \cdots \ y_N]$ of $N + 1$ measurements over a finite time horizon under the assumptions that the noise sequences $\{w_t\}$ and $\{f_t\}$ are somewhat (block) sparse. More specifically, we may view each of these uncertainty sequences as the sum of a dense component and a sparse one.

Let $\mathcal{X} = [x_0 \ x_1 \ \cdots \ x_N]$ be the state matrix on the considered time horizon; use the notation $\mathcal{X}_{i:j} = [x_i \ \cdots \ x_j]$ to denote the submatrix of $\mathcal{X}$ from column i through j. Define similarly W and F from the sequences $\{w_t\}$ and $\{f_t\}$ respectively. Then Eq. (65) gives $\mathcal{X}_{1:N} = A\mathcal{X}_{0:N-1} + W$ and $Y = C\mathcal{X}_{0:N} + F$. By exploiting the structure of these relations we formulate the state estimator as the set-valued map $\mathcal{S}$ defined by

$$\mathcal{S}(Y) = \underset{Z\in\mathbb{R}^{n\times(N+1)}}{\arg\min} \ V(Z), \tag{66}$$

where

$$V(Z) = \left\|Z_{1:N} - AZ_{0:N-1}\right\|_{2,\text{col}} + \gamma \left\|Y - CZ_{0:N}\right\|_{2,\text{col}} \tag{67}$$

with $\gamma > 0$ being a regularization parameter intended here to tradeoff the contribution of each of the two terms.

Next we present an analysis of the performance of the so-defined estimator $\mathcal{S}$. For this purpose, introduce the notations $\mathbb{I} = \{1, \dots, N\}$ and $\mathbb{J} = \{0, \dots, N\}$ for the index sets of the state matrix columns and the output measurements respectively. For any $I \subset \mathbb{I}$ and $J \subset \mathbb{J}$, define $f_{I,J} : \mathbb{R}^{n\times(N+1)} \to \mathbb{R}_+$ by

$$f_{I,J}(E) = \left\|(E_{1:N} - AE_{0:N-1})_I\right\|_{2,\text{col}} + \gamma \left\|(CE_{0:N})_J\right\|_{2,\text{col}},$$

where $(CE_{0:N})_J$ is the matrix formed by the columns of $CE_{0:N}$ indexed by J. With this rationale we have

$$f_{\mathbb{I},\mathbb{J}}(E) = \left\|E_{1:N} - AE_{0:N-1}\right\|_{2,\text{col}} + \gamma \left\|CE_{0:N}\right\|_{2,\text{col}} \tag{68}$$

A property of $f_{\mathbb{I},\mathbb{J}}$ that will be useful in the following derivations is the one of positive-definiteness. Indeed as stated in the lemma below, $f_{\mathbb{I},\mathbb{J}}$ is a norm on $\mathbb{R}^{n\times(N+1)}$ under observability assumption.

Lemma 3. Consider the system (65) with order n and assume that the number N of data satisfies $N > n$. Then the function $f_{\mathbb{I},\mathbb{J}}$ defined in (68) is a norm if and only if (A, C) is observable.

Proof. The proof of this lemma is immediate. In effect, the property of homogeneity and that of the triangle inequality of $f_{\mathbb{I},\mathbb{J}}$ are directly inherited from those of the norm $\|\cdot\|_{2,\text{col}}$. Therefore, we just need to illustrate how the positive-definiteness is related to observability. This in turn is straightforward since $f_{\mathbb{I},\mathbb{J}}(E) = 0$ if and only if $e_{t+1} = Ae_t$ for $t = 0, \dots, N-1$ and $\mathcal{O}e_0 = 0$ with $\mathcal{O} = [C^\top\ (CA)^\top\ \cdots\ (CA^{N-1})^\top]^\top$ being the observability matrix of Σ and e_0 the first column of E. Hence the proof is concluded. □

Let us use $I_r(E) \subset \mathbb{I}$ to denote the index set of the r largest entries of $\left\{\left\|e_t - Ae_{t-1}\right\|_2 : t \in \mathbb{I}\right\}$ and $J_{r'}(E) \subset \mathbb{J}$ to represent the index set of the r' largest entries of the set $\left\{\left\|Ce_t\right\|_2 : t \in \mathbb{J}\right\}$. Using these notations we introduce the following number

$$\mu_{r,r'}(\Sigma) = \sup_{\substack{E\in\mathbb{R}^{n\times N}\\ E\neq 0}} \frac{f_{I_r,J_{r'}}(E)}{f_{\mathbb{I},\mathbb{J}}(E)} \tag{69}$$

where $I_r(E)$ and $J_{r'}(E)$ are replaced by I_r and $J_{r'}$ for notational simplicity. $\mu_{r,r'}(\Sigma)$ reflects somehow a quantitative observability measure of the system Σ. The more observable Σ the smaller $\mu_{r,r'}(\Sigma)$.

Theorem 9. Consider the system (65) and assume it to be observable. Let $W \in \mathbb{R}^{n\times N}$ and $F \in \mathbb{R}^{m\times(N+1)}$denote matrices formed from the noise sequences $\left\{w_t\right\}$ and $\left\{f_t\right\}$ respectively. Let $Y \in \mathbb{R}^{m\times(N+1)}$ be the output measurement matrix with the assumption that $N > n$. If $\mu_{r,r'}(\Sigma) < 1/2$ for some integers r and r', then for all $\hat{\mathcal{X}} \in \mathcal{S}(Y)$, the error $E = \hat{\mathcal{X}} - \mathcal{X}$ is bounded as follows

$$f_{\mathbb{I},\mathbb{J}}(E) \leq \frac{2}{1 - 2\mu_{r,r'}(\Sigma)}\left[d_2(W, \mathcal{S}_r^{n\times N}) + \gamma d_2(F, \mathcal{S}_{r'}^{m\times(N+1)})\right] \tag{70}$$

Proof. Pose $E = \hat{\mathcal{X}} - \mathcal{X}$. By applying Lemma 5 in Appendix D to the two functions $Z \mapsto \left\|Z_{1:N} - AZ_{0:N-1}\right\|_{2,\text{col}}$ and $Z \mapsto \gamma\left\|Y - CZ_{0:N}\right\|_{2,\text{col}}$ with $V' = \hat{\mathcal{X}}$ and $V = \mathcal{X}$, it is straightforward to arrive at the following inequality

$$\begin{aligned} f_{\mathbb{I},\mathbb{J}}(E) - 2f_{I_r,J_{r'}}(E) \leq\ & V(\hat{\mathcal{X}}) - V(\mathcal{X}) \\ & + 2\inf_{(R,S)\in\mathcal{S}_r^{n\times N}\times\mathcal{S}_{r'}^{m\times(N+1)}}\left[\left\|R - (\mathcal{X}_{1:N} - A\mathcal{X}_{0:N-1})\right\|_{2,\text{col}} + \gamma\left\|S - (Y - C\mathcal{X}_{0:N})\right\|_{2,\text{col}}\right] \end{aligned}$$

By referring to (65), we note that the true noises matrices W and F satisfy $W = \mathcal{X}_{1:N} - A\mathcal{X}_{0:N-1}$ and $F = Y - C\mathcal{X}_{0:N}$. Moreover, since $V(\hat{\mathcal{X}}) - V(\mathcal{X}) \leq 0$, we have

$$f_{\mathbb{I},\mathbb{J}}(E) - 2f_{I_r,J_{r'}}(E) \leq 2\left[d_2(W, \mathcal{S}_r^{n\times N}) + d_2(F, \mathcal{S}_{r'}^{m\times(N+1)})\right]$$

By now invoking the definition (69), we see that $f_{\mathbb{I},\mathbb{J}}(E) - 2f_{I_r,J_{r'}}(E) \geq (1 - 2\mu_{r,r'})f_{\mathbb{I},\mathbb{J}}(E)$ and hence the previous inequality becomes

$$(1 - 2\mu_{r,r'})f_{\mathbb{I},\mathbb{J}}(E) \leq 2\left[d_2(W, \mathcal{S}_r^{n\times N}) + d_2(F, \mathcal{S}_{r'}^{m\times(N+1)})\right].$$

This is the desired result. □

In reference to Definition 1, Theorem 9 establishes the resilience property for the estimator $\mathcal{S}$. According to this theorem if $\mu_{r,r'}(\Sigma) < 1/2$ for some positive integers r and r' and if the true disturbances W and F lie in $\mathcal{S}_r^{n\times N}$ and $\mathcal{S}_{r'}^{m\times(N+1)}$, then $f_{\mathbb{I},\mathbb{J}}(E) = 0$ which, under the observability of Σ (see Lemma 3); implies $E = 0$, that is, the true state is exactly recovered despite

the presence of sparse noises in the dynamics and measurement equations. Moreover, the estimation error remained bounded if W and F do not lie in $\mathcal{S}_r^{n\times N}$ and $\mathcal{S}_{r'}^{m\times(N+1)}$ but are situated at a bounded distance from these sets. Finally, the theorem applies to situations where $\{w_t\}$ and $\{f_t\}$ are just viewed as dense noises. In that case the conditions are fulfilled with $r = r' = 0$ so that $\mu_{r,r'}(\Sigma) = 0$, $\mathcal{S}_r^{n\times N} = \{0\}$ and $\mathcal{S}_{r'}^{m\times(N+1)} = \{0\}$. The bound in (70) then becomes

$$f_{\mathbb{I},\mathbb{J}}(E) \leq 2\big[\,\|W\|_{2,\text{col}} + \gamma\,\|F\|_{2,\text{col}}\,\big].$$

If we make the assumption that gross errors are present solely in the output noise $\{f_t\}$, then we can take $r = 0$ so that the condition of Theorem 9 becomes $\mu_{0,r'}(\Sigma) < 1/2$. The following lemma shows that $\mu_{0,r'}(\Sigma)$ can be overestimated by solving a convex optimization.

Lemma 4. Consider the doubly indexed function $\mu_{r,r'}$ defined in (69) under the assumption that the system Σ is observable. Then

$$\mu_{0,r'}(\Sigma) \leq \frac{\gamma r'\sqrt{m}}{\beta_2^\star}, \tag{71}$$

where

$$\beta_2^\star = \inf_{t,j}\inf_{E}\left\{ f_{\mathbb{I},\mathbb{J}}(E) \,:\, c_j^\top e_t = 1 \right\}. \tag{72}$$

Proof. Recall that

$$\mu_{0,r'}(\Sigma) = \sup_{\substack{E\in\mathbb{R}^{n\times N}\\ E\neq 0}} \frac{\gamma\left\|(CE_{0:N})_{J_{r'}}\right\|_{2,\text{col}}}{f_{\mathbb{I},\mathbb{J}}(E)}$$

Moreover,

$$\gamma\left\|(CE_{0:N})_{J_{r'}}\right\|_{2,\text{col}} \leq \gamma r' \sup_{t\in\mathbb{J}}\|Ce_t\|_{2,\text{col}} \leq \gamma r'\sqrt{m}\sup_{(t,j)\in\mathbb{J}\times[\![m]\!]}|c_j^\top e_t|,$$

with $[\![m]\!] = \{1,\dots,m\}$. Since the set $\mathbb{J}\times[\![m]\!]$ is finite, the supremum in the above chain of inequalities is attainable, i.e., the sup can be replaced with max. Without loss of generality, we can assume $\max_{(t,j)\in\mathbb{J}\times[\![m]\!]}|c_j^\top e_t| \neq 0$. Now we use similar arguments as in the proof of Lemma 2 to obtain the result:

$$\begin{aligned}
\mu_{0,r'}(\Sigma) &\leq \gamma r'\sqrt{m}\sup_{\substack{E\in\mathbb{R}^{n\times N}\\ E\neq 0}}\sup_{(t,j)\in\mathbb{J}\times[\![m]\!]}\frac{|c_j^\top e_t|}{f_{\mathbb{I},\mathbb{J}}(E)}\\
&= \frac{\gamma r'\sqrt{m}}{\inf_{E,t,j}\left\{\dfrac{f_{\mathbb{I},\mathbb{J}}(E)}{|c_j^\top e_t|} \,:\, c_j^\top e_t \neq 0\right\}}\\
&= \frac{\gamma r'\sqrt{m}}{\inf_{t,j}\inf_E\left\{ f_{\mathbb{I},\mathbb{J}}(E) \,:\, c_j^\top e_t = 1\right\}}
\end{aligned}$$

□

Example 1. For a system with matrices defined by

$$A = \begin{bmatrix} 0.75 & 0.65 \\ 0.65 & -0.75 \end{bmatrix}, \quad C = \begin{bmatrix} -0.15 & -1 \end{bmatrix}$$

an estimation horizon $N = 100$ and a regularization parameter $\gamma = 0.01$, we get $\beta_2^\star = 0.348$ for the number defined in (72). Hence, from (71) we can infer that the state estimator (66) will be resilient if the number r' of nonzero columns in the output noise matrix F satisfies $r' \leq 17$. Of course this threshold (which is indeed an underestimate of the true one) for resilience of the estimator $\mathcal{S}$ depends on the properties of the to-be-observed system Σ, the length of the estimation horizon and the value of the regularization parameter γ. For example, if we change γ to 0.1 then the threshold of tolerated large values is reduced to 4.

5 | SOME SIMULATION RESULTS

5.1 | Identification of a linear model in the face of sparse noise

We first consider a linear model of the form (2) with $\theta^\circ = \begin{bmatrix} -0.40 & 0.25 & -0.15 & 0.08 \end{bmatrix}^\top$ driven by a white normal input $\{u_t\}$. Assume that the model error $\{v_t\}$ is strictly sparse and that its nonzero instances are sampled from a Gaussian distribution of variance 20^2. Considering a Monte-Carlo simulation of size 100, we present in Figure 1 the results obtained for $N = 200$ with the LAD, the RWℓ_1 (by setting $\epsilon = 0.1$ in (33)) and the k-smallest (with $r = 0.6N$ in (34)) algorithms in term of empirical probabilities of exact recovery. We measure here the sparsity level of the vector $\boldsymbol{v}$ as the ratio $1 - \|\boldsymbol{v}\|_0 / N$ of zero instances in it. What the results show is that the sparser $\{v_t\}$, the easier it is to retrieve the true parameter vector. As already shown in different works, RWℓ_1 effectively enhances the probability of obtaining the true PV even for small sparsity levels of $\boldsymbol{v}$. Finally the approximate k-smallest algorithm presents similar performance as the RWℓ_1 at least on this example.

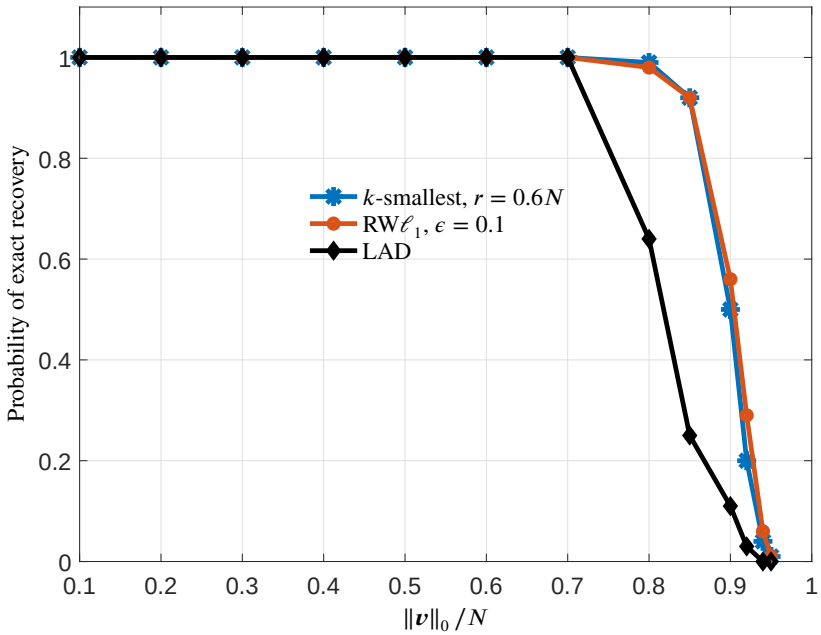

FIGURE 1 Empirical probabilities of exact recovery of a single dynamic ARX model in the presence of sparse noise versus level of sparsity of the noise. The sparsity is expressed in term of the fraction of nonzero in $\{v_t\}$. Comparison of LAD estimator (black); RWℓ_1 (red) and the k-smallest (blue) algorithms.

Considering the k-smallest algorithm in particular, it may be instructive to study the influence of the parameter r in (34) on its performance. Figure 2 displays the average estimation error over 100 independent runs of the algorithm for different values of the parameter r. It turns out that the performance is best when r is equal to the true number of gross errors (here, a ratio of 60%).

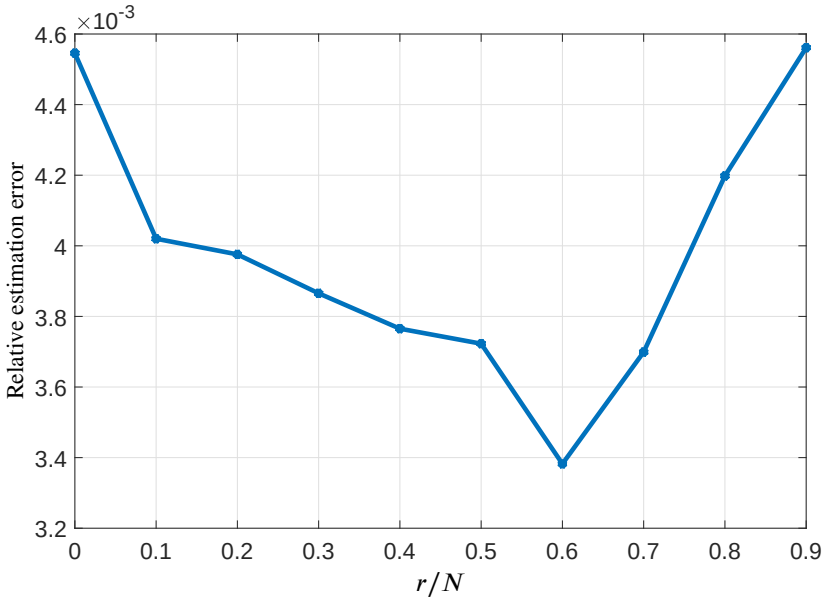

FIGURE 2 Performance of the k-smallest algorithm: average relative parametric estimation error $\|\hat{\theta} - \theta^\circ\|_2 / \|\theta^\circ\|_2$ over 100 independent runs versus the ratio r/N (see (34)). Conditions of the experiment: $N = 500$ data points generated by the linear system with a noise sequence $\{v_t\}$ defined by $v_t = s_t + d_t$ with $\{s_t\}$ being a sparse noise such that 60% of its values are nonzero (with these nonzero values being sampled from $\mathcal{N}(0, 20^2)$) and $\{d_t\}$ is a dense noise such that its magnitude satisfies a Signal to Noise Ratio (SNR) of about 30 dB.

5.2 | Identification of a Switched ARX system

To illustrate some of the methods presented above, we consider data generated by a SISO switched linear system of the form (40) with $s = 6$ subsystems of orders $(n_a, n_b) = (2, 2)$ and described by the following parameter vectors:

$$\theta_1^\circ = \begin{bmatrix} -0.40 \\ 0.25 \\ -0.15 \\ 0.08 \end{bmatrix}, \quad \theta_2^\circ = \begin{bmatrix} 1.55 \\ -0.58 \\ -2.10 \\ 0.96 \end{bmatrix}, \quad \theta_3^\circ = \begin{bmatrix} 1 \\ -0.24 \\ -0.65 \\ 0.30 \end{bmatrix},$$

$$\theta_4^\circ = \begin{bmatrix} 1.20 \\ -0.35 \\ 1.40 \\ -0.90 \end{bmatrix}, \quad \theta_5^\circ = \begin{bmatrix} -0.05 \\ 0.50 \\ -1.3 \\ 0.5 \end{bmatrix}, \quad \theta_6^\circ = \begin{bmatrix} 1.15 \\ -0.35 \\ 0.80 \\ -0.15 \end{bmatrix}.$$

The input signal $\{u_t\}$ is drawn as a realization of a zero-mean white Gaussian noise with unit variance. As to the switching signal, it is uniformly sampled over $\mathbb{S} = \{1, \ldots, 6\}$. As a consequence, the number of data pertaining to each subsystem is approximately the same. This is supposed to be the most challenging scenario for the error sparsification method presented in Section 3 with regards to its principle of identifying one submodel at a time. In effect, none of the submodels achieves enough sparsity over the entire dataset in the sense that for any $i \in \mathbb{S}$ the number of nonzeros in the error vector $\phi(\theta_i^\circ)$ is quite large.

Noise-free identification of the SARX system. The objective of this experiment is to study numerically the capacity of the sparse optimization approach to recover the true switched system in ideal conditions (no noise) but when the number of subsystems increases. For this purpose, we consider a simulation scenario where the noise sequence $\{e_t\}$ in (40) is equal to zero. We test the sparse optimization-based identification algorithm described in Section 3.3. Recall that this method estimates the

parameter vectors of the switched system one after another. For the identification of each individual subsystem at each step, one can, in principle, employ any sparsity-inducing (or robust) identifier. Here, we implement the ℓ_1 reweighted algorithm (RWℓ_1) and the one relying on iterative approximation of difference of convex functions (k-smallest), see Section 2.4. We then count over 100 realizations of the data, the number of times each algorithm successfully recovers all the s true parameter vectors. Such an experiment is repeated for different values of s ranging from 1 to 6 such that the ratio N/s of data points (y_t, x_t) with respect to the number of subsystems is kept constant and equal to 100. The results depicted in Figure 3 confirm the intuition that the identification of the SARX system gets increasingly challenging as the number of subsystems grows. A second teaching of this experiment is that the k-smallest algorithm tends to perform better than RWℓ_1.

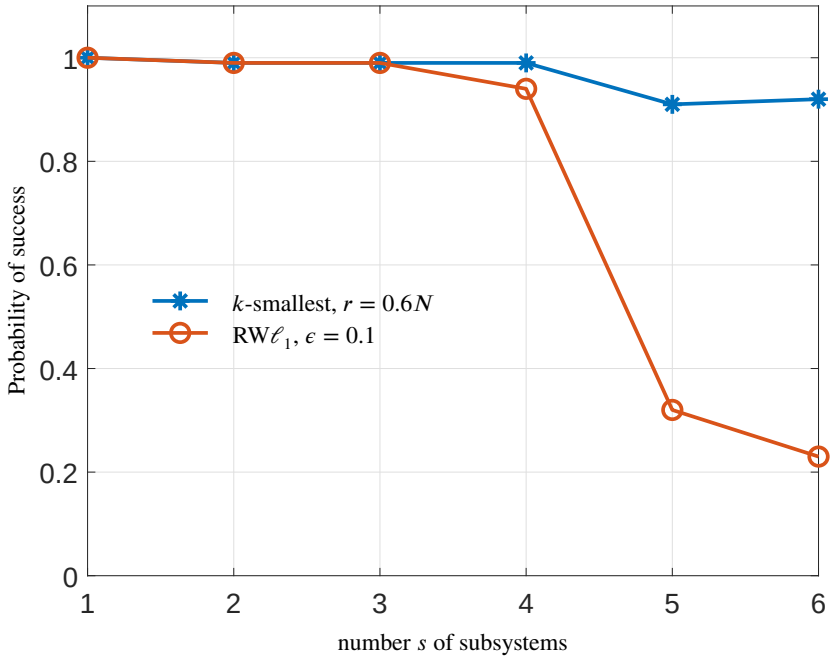

FIGURE 3 Empirical probabilities of successes in recovering the true switched system parameters in a noise-free scenario. Comparison of the RWℓ_1(red) and the k-smallest (blue) algorithms (Section 2.4) when there are both applied for the identification of the switched system along the incremental strategy (Algorithm 1).

Identification of the SARX system from noisy data. We consider now a more realistic scenario where the data are affected by noise. The noise sequence $\{e_t\}$ in (40) is independently sampled from zero-mean Gaussian distribution with a variance such that the Signal to Noise Ratio (SNR) is kept equal to 30 dB.

We measure the performance of each estimation algorithm through the parametric relative error defined by

$$E_r = \frac{1}{s}\sum_{i=1}^{s}\frac{\left\|\hat{\theta}_i - \theta_i^\circ\right\|_2}{\left\|\theta_i^\circ\right\|_2} \tag{73}$$

where $\hat{\theta}_i$ is the estimate of θ_i°. Note that the computation of the performance index (73) requires an appropriate reordering of the estimated parameter vectors. This reordering makes sense only if we can distinguish which estimate corresponds to which true parameter vector, a property which may be hard to guarantee if the noise level is high or the algorithms do not identify correctly all the parameters. Here, with the proportion of 30 dB of noise this reordering is still possible. Figure 4 presents average relative errors E_r achieved by the k-smallest and the RWℓ_1 algorithms on the SARX example when the number of subsystems increases

from 1 to 6. This result reveals that the k-smallest algorithm tends to be more stable than RWℓ_1 as the number of subsystems goes up hence confirming the trend suggested by the results depicted in Figure 3.

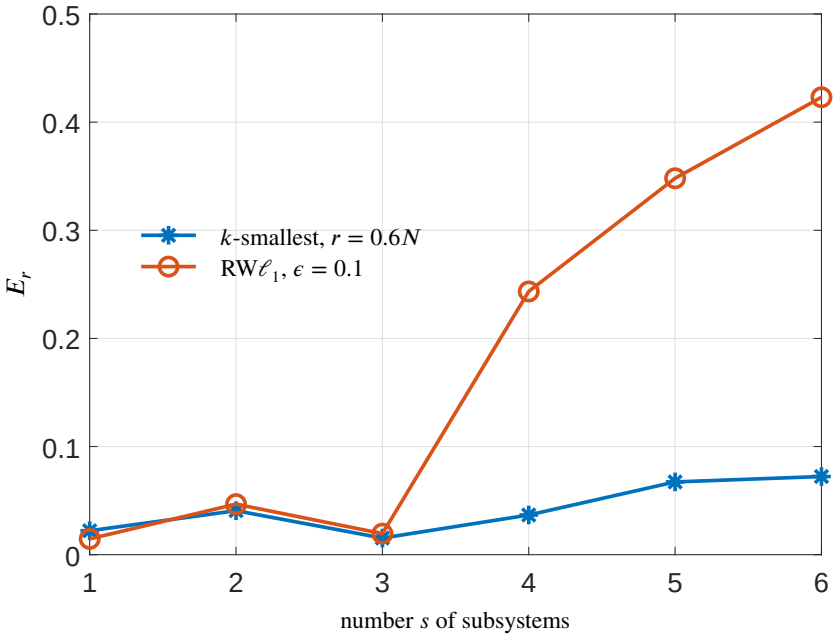

FIGURE 4 Averaged (over 100 realizations) relative estimation errors E_r delivered by the k-smallest (blue) and the RWℓ_1 (red) algorithms for the SARX system. $N = 800$ noisy data points with SNR = 30 dB.

5.3 | On the genericity properties of the regression data

Given a data matrix $(\boldsymbol{y}, X) \in \mathbb{R}^N \times \mathbb{R}^{n \times N}$ it may be desirable to evaluate numerically how many outliers the estimator $\Psi_1(\boldsymbol{y}, X)$ may be robust to in the worst case. As we have seen in Section 2.3, such a numerical certificate can be derived by (over-)estimating $\xi_r^1(X)$. Pessimistic conclusions can be drawn as to the capacity of Ψ_1 to handle sparse noise by relying on the upper bounds established in Lemma 2 for $\xi_r^1(X)$. Although they are efficiently computable (through the resolution of convex problems), their computational cost is still high. For this reason, we illustrate their values on a static data matrix X of small sizes $(n, N) = (2, 100)$. Two cases are studied for a data-generating system of the form (2):

- static model: the regression vectors $x_t \in \mathbb{R}^2$ are sampled independently from a Gaussian $\mathcal{N}(0, \sigma^2 I_2)$ with $\sigma^2 = 100$.
- dynamic ARX model: the regression vectors $x_t \in \mathbb{R}^2$ are structured as $x_t = [y_{t-1}\ u_{t-1}]^\top$ and $\theta^\circ = [-0.75\ 1]^\top$ with y_t and u_t denoting the output and input samples at time t. $\{u_t\} \subset \mathbb{R}$ is selected to be the realization of a white zero-mean Gaussian noise with variance 100.

We consider two cases: one where the columns of X have been normalized so that they have unit ℓ_2 norm and a second one where no normalization is applied. It appears from Table 1 that in the first case, the ratio of worst-case correctable large errors by the LAD estimator is significantly higher than in the second. We also observe quite logically, that static data are more generic than the dynamic data as they allow for more outliers to be handled. When the columns of X are normalized and the data are very generic then the over-estimates proposed in Lemma 2 are very close to the true concentration ratio $\xi_1(X)$.

	Normalized		Unnormalized	
Estimates of π_1^c	Static	Dynamic	Static	Dynamic
Eq. (30)	0.29	0.27	0.12	0.12
Eq. (31)	0.28	0.26	0.16	0.16
True π_1^c	0.31	0.28	0.21	0.22

TABLE 1 Average estimates (over 100 realizations) of the ratios of worst-case correctable outliers by the LAD estimator. $X \in \mathbb{R}^{n\times N}$ with $(n, N) = (2, 100)$ is a matrix containing (1) static Gaussian data and (2) dynamic ARX regression data.

6 | CONCLUSION

Summary of this paper. A sparse optimization problem can be viewed as one which involves the minimization of the cardinality of a set (number of nonzero entries in a vector, number of nonzero singular values of a matrix, i.e., its rank). In this paper we have discussed the potential of application of the sparse optimization paradigm to a sample of illustrative control-related problems:

- robust estimation: In a regression problem, since sparsity-inducing methods only care about minimizing the number of nonzero errors, the noise affecting the data can indeed take on arbitrarily large values without affecting very much the performance of the estimator provided the number of such large values is limited. Hence sparsity-inducing estimation methods are naturally robust against outliers.
- hybrid system identification: As illustrated in Section 3, sparsity-inducing optimization is a valuable approach for identifying switched systems and piecewise affine systems from data when the switching signal is not known.
- resilient state estimation: Considering a linear dynamic system subject to potentially large errors in the state and/or measurement equations, we have shown in Section 4.2 how a state estimator can be designed which may be insensitive to these errors.

Beyond these three classes of problems, sparsity-inducing optimization can be a methodological ingredient in many other problems such as regressor selection, estimation in the conditions where the data sequences suffer some missing points[58], maximum hands-off control[59], time optimal control[60,61], control of hybrid systems[56], fault-tolerant control, state estimation for switched system[55], subspace clustering[26,25], signal recovery, image denoising, etc.

Discussions. From a computational perspective it is fair to recognize that solving directly the sparse optimization problem is challenging since its strict formulation is nonconvex and generically NP-hard. To get around this difficulty, some efficient convex relaxations exist, though the conditions under which such relaxed formulations can recover the solution of the original sparse problem are restrictive. For example, the condition of exact recoverability by the ℓ_1 surrogate problem (convex) turns out to be more restrictive than its ℓ_0 counterpart (nonconvex) as it demands a higher level of sparsity of the error. In the robust regression problem, the number of outliers the LAD (or ℓ_1) estimator can handle is smaller than the outlier-tolerance capacity of the ℓ_0 estimator. But an encouraging fact is that the number of outliers handled by the LAD estimator is all the larger as the regression matrix X is generic. Hence, a key to enhance the robustness of the LAD to outliers would be to generate, by an appropriate selection of the excitation signal, the identification data so that X is sufficiently generic, for example in the sense of a small concentration ratio $\xi_r^1(X)$ for large enough r. While this richness enhancing procedure is possible for the identification problem, it appears to be more difficult for the state estimation problem where the counterpart of the regression matrix of interest is the observability matrix. The problem one is facing in this case is that the observability matrix is structured and hard to transform by other enriching ways than output feedback.

On the other hand, nonconvex approaches to the sparse optimization problem, even though not supported by strong theoretical guarantees, can yield good estimation when the noise sequence (in the regression for example) is not very sparse. Examples of

such methods are the RWℓ_1 and the k-smallest algorithms discussed and implemented in the present paper or those listed in Section 2.1.1.

Possible directions for future research on the topic of sparsity-inducing optimization methods in control theory may concern further investigations of the application potential of such methods to more control-related problems. In particular, it is of interest to investigate extensions of the current results to continuous-time systems along, for example, the framework of the book[62]. From the practical standpoint, further demonstration is needed concerning the applicability and efficiency of such methods to real-life systems.

How to cite this article: L. Bako (2021), On sparsity-inducing methods in system identification and state estimation, *Int J Robust Nonlinear Control*, *2020;00:1–17*.

APPENDIX

A PROOF OF PROPOSITION 3

Proof. The proofs of statements (a) and (b) are quite similar. Hence we will just prove (a). Consider $\theta^\star \in \arg\min_{\theta\in\mathbb{R}^n} \|\phi(\theta)\|_0$. Then it follows from the assumption that $\left\{\theta \in \mathbb{R}^n : \|\phi(\theta)\|_0 \leq r\right\}$ is non empty that $\left|\mathbb{I}^c(\theta^\star)\right| = \left\|\phi(\theta^\star)\right\|_0 \leq r$ with $\mathbb{I}^c(\theta^\star) = \mathbb{I} \setminus \mathbb{I}^0(\theta^\star)$. So, since $\left|\mathbb{I}^c(\theta^\star)\right| \leq r$, $\xi_r^1(X) \leq 1/2$ implies that

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

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

$$\|X^\top_{\mathbb{I}^c(\theta^\star)}\eta\|_1 \leq \|X^\top_{\mathbb{I}^0(\theta^\star)}\eta\|_1 \tag{A1}$$

We now make two remarks. First, since $\boldsymbol{y}_{\mathbb{I}^0(\theta^\star)} - X^\top_{\mathbb{I}^0(\theta^\star)}\theta^\star = 0$, we have $\|X^\top_{\mathbb{I}^0(\theta^\star)}\eta\|_1 = \|\boldsymbol{y}_{\mathbb{I}^0(\theta^\star)} - X^\top_{\mathbb{I}^0(\theta^\star)}(\theta^\star + \eta)\|_1$ and hence (A1) reads as

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

Second, by using the triangle inequality, we observe that the right hand side term of (A1) can be lower-bounded as follows

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

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

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

Therefore $\theta^\star \in \arg\min_{\theta\in\mathbb{R}^n} \|\phi(\theta)\|_1$ hence proving the statement (a). The proof of (b) follows the same line of reasoning by just changing some large inequalities into strict inequalities. □

B PROOF OF LEMMA 2

Proof. For the proof of (29), we can write immediately

$$\xi_r^1(X) \leq r \max_t \max_{\substack{\eta\in\mathbb{R}^n\\ \eta\neq 0}} \left[\frac{|x_t^\top\eta|}{\left\|X^\top\eta\right\|_1}\right] = r \max_{t=1,\ldots,N} \frac{1}{\beta_t} = \frac{r}{\min_t \beta_t}$$

with

$$\begin{aligned}\beta_t &= \min_{\substack{\eta\in\mathbb{R}^n\\ \eta\neq 0}} \frac{\left\|X^\top\eta\right\|_1}{|x_t^\top\eta|}\\ &= \min_{\eta\in\mathbb{R}^n}\left\{\left\|X^\top\eta\right\|_1 : |x_t^\top\eta| = 1\right\}\\ &= \min_{\eta\in\mathbb{R}^n}\left\{\left\|X^\top\eta\right\|_1 : x_t^\top\eta = 1\right\}\end{aligned}$$

As to the proof of (30), it can be found in[20] in a more general framework. We repeat it here for completeness on the current particular case. For any t, consider writing x_t as a linear combination of the other vectors of X, namely write $x_t = X_{\neq t}\gamma_t$. This decomposition is always possible thanks to the assumption that $\nu_n(X) \leq N-1$. Then

$$|x_t^\top \eta| = |\gamma_t^\top X_{\neq t}^\top \eta| \leq \|\gamma_t\|_\infty \left\|X_{\neq t}^\top \eta\right\|_1 \leq \kappa(X)\left(\left\|X^\top \eta\right\|_1 - |x_t^\top \eta|\right)$$

It follows that

$$|x_t^\top \eta| \leq \frac{\kappa(X)}{1+\kappa(X)}\left\|X^\top \eta\right\|_1$$

which gives

$$\xi_r^1(X) \leq r \max_t \max_{\substack{\eta\in\mathbb{R}^n \\ \eta\neq 0}} \left[\frac{|x_t^\top \eta|}{\left\|X^\top \eta\right\|_1}\right] \leq \frac{r\kappa(X)}{1+\kappa(X)}$$

Hence (30) is proved.

It remains to prove (31). Note that the optimization problem on the right hand side of (31) is always feasible under the assumption that $\nu_n(X) \leq N-1$. Consider an arbitrary feasible matrix $\Lambda \in \mathbb{R}^{N\times N}$. Then

$$\xi_r^1(X) = \max_{\substack{\eta\in\mathbb{R}^n \\ \eta\neq 0}} \left[\frac{\left\|X^\top \eta\right\|_{1,[r]}}{\left\|X^\top \eta\right\|_1}\right] = \max_{\substack{\eta\in\mathbb{R}^n \\ \eta\neq 0}} \left[\frac{\left\|\Lambda^\top X^\top \eta\right\|_{1,[r]}}{\left\|X^\top \eta\right\|_1}\right]$$

By noting that $\left\|\Lambda^\top X^\top \eta\right\|_{1,[r]} \leq \|\Lambda\|_{\infty,[r]} \left\|X^\top \eta\right\|_1$ we see that $\xi_r^1(X) \leq \|\Lambda\|_{\infty,[r]}$. □

C PROOF OF PROPOSITION 5

Proof. The proof is similar to that of Lemma 1 in[22]. Because the data are generated by the system (40), 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 $|y_t - x_t^\top \theta_i^\circ| \leq \varepsilon$. It follows that $\mathbb{I} = \mathbb{I}^{\leq\varepsilon}(\theta_1^\circ) \cup \cdots \cup \mathbb{I}^{\leq\varepsilon}(\theta_s^\circ)$. Let $\hat{\theta} \in \arg\min_{\theta\in\mathbb{R}^n} \|\phi(\theta)\|_{0,\varepsilon}$. Then

$$\left|\mathbb{I}^{\leq\varepsilon}(\hat{\theta})\right| \leq \sum_{i=1}^{s} \left|\mathbb{I}^{\leq\varepsilon}(\hat{\theta}) \cap \mathbb{I}^{\leq\varepsilon}(\theta_i^\circ)\right| \tag{C2}$$

We then claim that there is an $i^* \in \mathbb{S}$ such that $\left|\mathbb{I}^{\leq\varepsilon}(\hat{\theta}) \cap \mathbb{I}^{\leq\varepsilon}(\theta_{i^*}^\circ)\right| \geq \nu_n(X)$. To see this, assume that the opposite holds, meaning that for all $i \in \mathbb{S}$, $\left|\mathbb{I}^{\leq\varepsilon}(\hat{\theta}) \cap \mathbb{I}^{\leq\varepsilon}(\theta_i^\circ)\right| < \nu_n(X)$. Then by applying (C2) and using the definition of $\hat{\theta}$, we immediately obtain $|\mathbb{I}^{\leq\varepsilon}(\theta_i^\circ)| \leq |\mathbb{I}^{\leq\varepsilon}(\hat{\theta})| < s\nu_n(X)$ 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 $I^* \triangleq \mathbb{I}^{\leq\varepsilon}(\hat{\theta}) \cap \mathbb{I}^{\leq\varepsilon}(\theta_{i^*}^\circ)$ 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 $|y_t - x_t^\top \theta_{i^*}^\circ| \leq \varepsilon$. As a result,

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

by the triangle inequality, so that

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

The conclusion follows naturally from this. □

D A USEFUL TECHNICAL LEMMA

Lemma 5. Let $\Psi : \mathcal{M} \to \mathbb{R}_+$ be a positive function defined on the whole set of real matrices such that for a specific matrix $V = [v_1 \ \ldots \ v_N] \in \mathbb{R}^{n\times N}$, $\Psi(V) = \sum_{i=1}^N \psi(v_i)$ with $\psi : \mathbb{R}^n \to \mathbb{R}_+$ being a norm. Let $I_r(V) \subset \mathbb{I} = \{1, \ldots, N\}$ be the index set of

the r largest elements of $\{\psi(v_i) : i \in \mathbb{I}\}$. Denote with $S_r^{n\times N}$ the set of matrices in $\mathbb{R}^{n\times N}$ with at most r nonzero columns. Then

$$\Psi(V' - V) - 2\Psi((V' - V)_{I_r(V)}) \leq \Psi(V') - \Psi(V) + 2 \inf_{W \in S_r^{n\times N}} \Psi(V - W) \tag{D3}$$

where $(V - V')_{I_r(V)}$ is a matrix formed with the columns of $V - V'$ which are indexed by $I_r(V)$.

Proof. For notational simplicity, we denote $I_r(V)$ with I and its complement in $\mathbb{I}$ with I^c. By applying the triangle inequality, we have

$$\begin{aligned}\Psi(V' - V) &= \Psi_I(V' - V) + \Psi_{I^c}(V' - V)\\ &\leq \Psi(V'_I - V_I) + \Psi(V_{I^c}) + \Psi(V'_{I^c})\end{aligned}$$

Now we can find an upper bound of $\Psi(V'_{I^c})$ as follows:

$$\begin{aligned}\Psi(V'_{I^c}) &= \Psi(V') - \Psi(V'_I)\\ &= \Psi(V_I) - \Psi(V'_I) + \Psi(V') - \big(\Psi(V) - \Psi(V_{I^c})\big)\\ &\leq \Psi(V'_I - V_I) + \Psi(V') - \Psi(V) + \Psi(V_{I^c})\end{aligned}$$

The last inequality is a consequence of the triangle inequality property by which $\Psi(V_I) - \Psi(V'_I) \leq \Psi(V'_I - V_I)$. Combining both inequalities yields

$$\Psi(V' - V) - 2\Psi((V' - V)_I) \leq \Psi(V') - \Psi(V) + 2\Psi(V_{I^c})$$

Noting that $\Psi(V_{I^c}) = \inf_{W \in S_r^{n\times N}} \Psi(V - W)$, the result follows. □

References

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

2. Foucart S, Rauhut H. *A mathematical introduction to compressive sensing*. Birkhäuser . 2013.

3. Eldar YC, (Eds) GK. *Compressed Sensing: Theory and Applications*. ? Cambridge University Press . 2012.

4. Rani M, Dhok SB, Deshmukh RB. A Systematic Review of Compressive Sensing:Concepts, Implementations and Applications. *IEEE Access* 2018; 6: 4875-4894.

5. Candès EJ, Wakin MB. An Introduction To Compressive Sampling. *IEEE Signal Processing Society* 2008; 25: 21-30.

6. Tibshirani R. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B* 1996; 58: 267-288.

7. Tibshirani RJ. The lasso problem and uniqueness. *Electronic Journal of Statistics* 2013; 7: 1456-1490.

8. Wainwright MJ. Sharp thresholds for High-Dimensional and noisy sparsity recovery using ℓ_1-Constrained Quadratic Programming (Lasso). *IEEE Transactions on Information Theory* 2009; 55(5): 2183-2202.

9. Zhang Z, Xu Y, Yang J, Li X, Zhang D. A survey of sparse representation: algorithms and applications.. *IEEE Access* 2015; 3: 490-530.

10. Tan M, Tsang IW, Wang L. Matching pursuit LASSO part I: Sparse recovery over big dictionary. *IEEE Transactions on Signal Processing* 2014; 63: 727-741.

11. Fattahi S, Sojoudi S. Data-driven sparse system identification. In: 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton). ; 2018.

12. Wang DQ, Yan YR, Liu YJ, Ding JH. Model recovery for Hammerstein systems using the hierarchical orthogonal matching pursuit method. *Journal of Computational and Applied Mathematics* 2019(345): 135-145.

13. You JY, Liu YJ, Chen J, Ding F. Iterative identification for multiple-input systems with time-delays based on greedy pursuit and auxiliary model. *Journal of the Franklin Institute* 2019(356): 5819-5833.

14. Ohlsson H, Ljung L, Boyd S. Segmentation of ARX-models using sum-of-norms regularization. *Automatica* 2010; 46: 1107-1111.

15. Ohlsson H, Ljung L. Identification of switched linear regression models using sum-of-norms regularization. *Automatica* 2013; 49: 1045-1050.

16. Piga D, Tóth R. An SDP approach for ℓ_0-minimization: Application to ARX model segmentation. *Automatica* 2013; 49: 3646-3653.

17. Ozay N, Sznaier M, Lagoa C, Camps O. A Sparsification Approach to Set Membership Identification of Switched Affine Systems. *IEEE Transactions on Automatic Control* 2012; 57: 634-648.

18. Huber PJ. The place of the L_1 norm in robust estimation. *Computational Statistics & Data Analysis* 1987; 5: 255-262.

19. Candès E, Randall PA. Highly robust error correction by convex programming. *IEEE Transactions on Information Theory* 2006; 54: 2829-2840.

20. Bako L. On a Class of Optimization-Based Robust Estimators. *IEEE Transactions on Automatic Control* 2017; 62: 5990-5997.

21. Maruta I, Sugie T. Identification of PWA models via data compression based on ℓ_1 optimization. In: IEEE Conference on Decision and Control and European Control Conference (CDC-ECC) Orlando, FL, USA. ; 2011.

22. Bako L. Identification of switched linear systems via sparse optimization. *Automatica* 2011; 47: 668-677.

23. Le VL, Lauer F, Bloch G. Selective ℓ_1 minimization for sparse recovery. *IEEE Transactions on Automatic Control* 2013; 59: 3008-3013.

24. Bako L, Yahya O. Piecewise Affine System identification: A least harmonic mean approach. In: IEEE Conference on Decision and Control, Nice, France. ; 2019.

25. Bako L. Subspace clustering through parametric representation and sparse optimization. *IEEE Signal Processing Letters* 2014; 21: 356-360.

26. Elhamifar E, Vidal R. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE transactions on pattern analysis and machine intelligence* 2013; 35: 2765-2781.

27. Fawzi H, Tabuada P, Diggavi S. Secure Estimation and Control for Cyber-Physical Systems Under Adversarial Attacks. *IEEE Transactions on Automatic Control* 2014; 59(6): 1454–1467. doi: 10.1109/TAC.2014.2303233

28. Han D, Mo Y, Xie L. Convex optimization based state estimation against sparse integrity attacks. *IEEE Transactions on Automatic Control* 2019; 64: 2383-2395.

29. Kircher A, Bako L, Blanco E, Benallouch M. An optimization framework for resilient batch estimation in Cyber-Physical Systems. *IEEE Transactions on Automatic Control (To appear : 10.1109/TAC.2021.3121223)* 2021.

30. Mishra S, Shoukry Y, Karamchandani N, Diggavi SN, Tabuada P. Secure State Estimation Against Sensor Attacks in the Presence of Noise. *IEEE Transactions on Control of Network Systems* 2017; 4: 49–59. doi: 10.1109/TCNS.2016.2606880

31. Shoukry Y, Tabuada P. Event-Triggered State Observers for Sparse Sensor Noise/Attacks. *IEEE Transactions on Automatic Control* 2016; 61: 2079–2091. doi: 10.1109/TAC.2015.2492159

32. Farahmand S, Giannakis GB, Angelosante D. Doubly robust smoothing of dynamical processes via outlier sparsity constraints. *IEEE Transactions on Signal Processing* 2011(59): 4529-4543.

33. Natarajan BK. Sparse Approximate Solutions to Linear Systems. *SIAM Journal on Computing* 1995; 24: 227-234.

34. Daubechies I, DeVore R, Fornasier M, Güntürk CS. Iteratively reweighted least squares minimization for sparse recovery. *Communications on Pure and Applied Mathematics* 2010; 63: 1-38.

35. Chartrand R. Exact reconstruction of sparse signals via nonconvex minimization. *IEEE Signal Processing Letters* 2007; 14: 707-710.

36. Ashour ME, Lagoa CM, Aybat NS. Lp Quasi-norm Minimization. In: 53rd Asilomar Conference on Signals, Systems, and Computers. ; 2019.

37. Mohimani H, Babaie-Zadeh M, Jutten C. A fast approach for overcomplete sparse decomposition based on smoothed ℓ^0 norm. *IEEE Transactions on Signal Processing* 2008; 57: 289-301.

38. Hyder M, Mahata K. An improved smoothed ℓ^0 approximation algorithm forsparse representation. *EEE Transactions on Signal Processing* 2010; 58: 2194-2205.

39. Mallat SG, Zhang Z. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing* 1993; 41: 3397-3415.

40. Tropp J. Greed is good: Algorithmic results for sparse approximation. *IEEE Transactions on Information Theory* 2004; 50: 2231-2242.

41. Bruckstein AM, Donoho DL, Elad M. From Sparse Solutions of Systems of Equations to Sparse Modeling of Signals and Images. *SIAM Review* 2009; 51: 34-81.

42. Candès EJ, Wakin M, Boyd S. Enhancing sparsity by reweighted ℓ_1 minimization. *Journal Fourier Analysis and Applications* 2008; 14: 877-905.

43. Sharon Y, Wright J, Ma Y. Computation and Relaxation of Conditions for Equivalence between ℓ^1 and ℓ^0 Minimization. *UIUC Technical Report UILU-ENG-07-2008* 2007.

44. Candès EJ, Rudelson M, Tao T, Vershynin R. Error correction via linear programming. In: Proceedings of the 46th Annual IEEE Symposium on Foundations of Computer Science (FOCS), 295-308. ; 2005.

45. Mitra K, Veeraraghavan A, Chellappa R. Analysis of sparse regularization based robust regression approaches. *IEEE Transactions on Signal Processing* 2012; 61: 1249-1257.

46. Rousseeuw PJ, Leroy AM. *Robust Regression and Outlier Detection*. John Wiley & Sons, Inc. . 2005.

47. Huber PJ, Ronchetti EM. *Robust Statistics*. A. John Wiley & Sons, Inc. Publication (2nd Ed) . 2009.

48. Boyd S, Vandenberghe L. *Convex Optimization*. Cambridge University Press . 2004.

49. Bako L, Ohlsson H. Analysis of A Nonsmooth Optimization Approach to Robust Estimation. *Automatica* 2016; 66: 132-145.

50. Bako L. Analysis of the least sum-of-minimums estimator for switched systems. *IEEE Transactions on Automatic Control* 2021; 66: 3733-3740.

51. Petreczky M, Bako L, Lecoeuche S. Minimality and identifiability of SARX systems. In: IFAC Symposium on System Identification, Brussels, Belgium. ; 2012.

52. Petreczky M, Bako L, Lecoeuche S, Motchon K. Minimality and identifiability of discrete-time SARX systems. *International Journal of Robust and Nonlinear Control* 2020; 30: 5871-5891.

53. Petreczky M, Bako L, Schuppen vJH. Identifiability of discrete-time linear switched systems. In: Hybrid Systems: Computation and Control, Stockholm, Sweden. ; 2010.

54. Vidal R. Subspace clustering. *IEEE Signal Processing Magazine* 2011; 28: 52-68.

55. Bako L, Lecoeuche S. A sparse optimization approach to state observer design for switched linear systems. *Systems & Control Letters* 2013; 62: 143-151.

56. Kreiss J, Bako L, Blanco E. Optimal control of discrete-time switched linear systems via continuous parameterization. In: IFAC World Congress, Toulouse, France. ; 2017.

57. Bako L, Le VL, Lauer F, Bloch G. Identification of MIMO switched state-space systems. In: American Control Conference, Washington DC. ; 2013.

58. Liu Z, Hansson A, Vandenberghe . Nuclear norm system identification with missing inputs and outputs. *Systems & Control Letters* 2013; 62: 605-612.

59. Nagahara M, Quevedo DE, Nesic D. Maximum Hands-Off Control: A Paradigm of Control Effort Minimization. *IEEE Transactions on Automatic Control* 2016; 61: 735-747.

60. Chen D, Bako L, Lecoeuche S. The minimum-time problem for discrete-time linear systems: A non-smooth optimization approach. In: IEEE Multi-Conference on Systems and Control, Dubrovnik, Croatia. ; 2012.

61. Ikeda T, Nagahara M. Time-optimal hands-off control for linear time-invariant systems. *Automatica* 2019; 99: 54-58.

62. Nagahara M. *Sparsity Methods for Systems and Control*. Now Publishers . 2020.