Analysis of the least sum-of-minimums estimator for switched systems

Laurent Bako

Analysis of the least sum-of-minimums estimator for switched systems

Laurent Bako

***Abstract*—This paper considers a particular parameter estimator for switched systems and analyzes its properties. The estimator in question is defined as the map from the data set to the solution set of an optimization problem where the to-be-optimized cost function is a sum of pointwise infima over a finite set of sub-functions. This is a hard nonconvex problem. The paper studies some fundamental properties of this problem such as uniqueness of the solution or boundedness of the estimation error regardless of computational considerations. The interest of the analysis is to lay out the main influential properties of the data on the performance of this (ideal) estimator.**

***Index Terms*—System identification, switched systems, sparsity, data richness, robustness to outliers.**

I. Introduction

A switched system is defined by a finite set of dynamic systems together with a map, called the switching law, which selects over time which system (subsystem) is activated [10], [17]. The switching law may be time-driven, event-driven or state-driven. Such systems can be viewed as formal descriptions of physical phenomena taking place in, for example, power converters [11], video sequences (from segmentation perspective) [18]. Finding mathematical representations of switched systems is fundamental for the purpose of control, analysis or diagnosis. In this paper we discuss the theoretical performances/properties of a particular method for identifying a switched model from measurements.

The problem of identifying switched systems directly from input-output data has been largely investigated in the recent literature. Examples of contributions include the works reported in [19], [1], [12], [16], [5] most of which rely on numerical optimization. Some surveys of the topic can be found in [9], [4], [13] (see the references therein). It is fair to remark that a large number of computational methods have been proposed for estimating the parameters of switched systems. However, an important aspect that is not well understood yet is how the properties of the data quantitatively impact the performance of estimation methods operating on those data. In other words, the necessary properties of informativity of the data which favor correct estimation is still to be further investigated. In the current work we take a step forward in the study of such informativity properties. Note that so far, only a very few works have considered the fundamental question of characterizing data informativity (richness) in the context of switched system identification [14], [18]. [14] sketches a broad purpose condition of persistence of excitation for estimating switched state-space realizations. As to the characterization formulated in [18], it can be interpreted as a rank condition in a lifted space (resulting from polynomial embedding of the regressors). However, neither of these contributions proposed a characterization of the parametric estimation error bound as an explicit function of the informativity degree of the regression data.

The goal of this paper is to analyze the properties of a particular estimator which we call here the Least Sum-of-Minimums estimator (LSM) for switched system identification. This estimator maps the data to the parameter space (of the constituent subsystems) by associating to a given data set the minimizing set of some data-dependent cost function. The cost function is formed as a sum of pointwise infima of the prediction errors associated to each subsystem. While the prediction errors may be measured in the LSM framework with multiple different loss functions, we focus specifically on the case of the absolute deviation loss function. We note that the LSM estimator is neither analytically expressible, nor numerically solvable directly at a reasonable computational price. Heuristics exist however that allow to approach the solution with, sometimes, guarantees of optimality. For a numerical approach to this problem we refer for example, to [8]. The perspective taken here is formal rather than computational, the goal being to lay out the properties the data should enjoy to allow for an adequate retrieval of the system parameters, at least in principle. In the wake of our previous work reported in [1], we first derive conditions on the data that guarantee exact recoverability of the true parameter matrix in the hypothetical scenario where the measurements would be essentially noise-free. A striking property of the absolute deviation loss (used in the framework of the LSM estimator) is that it allows for exact recovery even in the face of a *sparse noise*, provided that the number of nonzero values in the sparse noise sequence does not exceed a certain threshold prescribed by the informativity degree of the data. In the more realistic situations where the data are affected by both *dense and sparse noise*, we provide parametric error bounds for the estimates delivered by the estimator. The interest of our results reside in the fact that they reveal the impact of the data informativity on the attainable performance of the (ideal) switched system estimator. This feature makes them potentially useful for optimal experiment design, that is, the process of defining adequately the data-generating experimental conditions that would lead to the smallest (estimation) uncertainty bound.

Outline. We state the switched system identification problem in Section II and define the LSM estimator. We start the analysis by considering essentially the noiseless scenario in Section III and then the noisy one in Section IV. The main conclusions of our study are recapitulated in Section V.

Notation. $\mathbb{R}$ denotes the set of real numbers; $\mathbb{R}_+$ is the set of nonnegative real numbers. For a matrix $A = [a_1 \quad \cdots \quad a_s] \in \mathbb{R}^{n\times s}$, we use $\mathrm{Set}(A)$ to denote the finite set formed with the columns of A, i.e., $\mathrm{Set}(A) = \{a_1, \ldots, a_s\}$. If $\mathcal{S}$ is a finite set, then $|\mathcal{S}|$ denotes the cardinality of $\mathcal{S}$. If $x \in \mathbb{R}$ then $|x|$ is the absolute value of x. For $x = \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix} \in \mathbb{R}^n$, $\|x\|_0$ will refer to the ℓ_0 norm of x (namely the number of nonzero entries in the vector x); and $\|x\|_1 = \sum_i |x_i|$ will denote the ℓ_1 norm of x. If $X \in \mathbb{R}^{n\times N}$ is a matrix and $I \subset \{1, \ldots, N\}$ is a subset of the column index of X, then X_I denotes the submatrix of X formed with the columns of X which are indexed by I. Similarly, for a vector $v \in \mathbb{R}^N$, v_I refers to the subvector of v consisting in the entries of v indexed by I.

II. The switched system identification problem

A. The data-generating system

Consider a (possibly nonlinear) switched system described by an equation of the form

$$y_t = x_t^\top a^\circ_{\sigma(t)} + v_t, \tag{1}$$

L. Bako is with Laboratoire Ampère (UMR CNRS 5005) – Ecole Centrale de Lyon – Université de Lyon, 69134, Ecully, France. E-mail: laurent.bako@ec-lyon.fr.

where $t \in \mathbb{Z}_+$ refers to discrete-time, $y_t \in \mathbb{R}$ is the output of the system at time t, $x_t \in \mathbb{R}^n$ is the regressor. As to v_t, it refers to potential additive noise component. $\sigma : \mathbb{Z}_+ \to \mathbb{S} \triangleq \{1, \dots, s\}$ defines a switching signal and $a_i^\circ \in \mathbb{R}^n$, $i \in \mathbb{S}$, denote some distinct parameter vectors. Eq. (1) describes a switched system composed of s dynamical subsystems each of which is activated one after another in time by the switching signal σ.

The model (1) captures the situations where the regressor x_t is directly observed or obtained through an intermediary nonlinear mapping of some observable signal $z_t \in \mathbb{R}^d$. We will assume that

$$x_t = \varphi(z_t) \tag{2}$$

where $\varphi : \mathbb{R}^d \to \mathbb{R}^n$ is some (known) linear or nonlinear map. Hence, depending on the choice of the mapping φ, the model (1) can describe both linear and nonlinear switched systems.

We further observe that the system represented by (1) can be static, in which case z_t is an unstructured multivariate input vector, or dynamic. In this latter case, z_t in (2) may assume the form

$$z_t = [y_{t-1} \quad \cdots \quad y_{t-n_a} \quad u_t^\top \quad u_{t-1}^\top \quad \cdots \quad u_{t-n_b}^\top]^\top \in \mathbb{R}^d \tag{3}$$

with n_a and n_b being some integers and $u_t \in \mathbb{R}^{n_u}$ the input of the system. Note that n_a can be taken equal to zero in which case x_t reduces to $x_t = [u_t^\top \quad u_{t-1}^\top \quad \cdots \quad u_{t-n_b}^\top]^\top$ (hence yielding a switched nonlinear system of Finite Impulse Response type).

B. The least sum of minimums estimator

For convenience we collect the true parameter vectors $a_i^\circ \in \mathbb{R}^n$ from (1) in a matrix $A^\circ = [a_1^\circ \quad \cdots \quad a_s^\circ] \in \mathbb{R}^{n \times s}$ which we call the true parameter matrix. Given a collection of N data points

$$\varpi^N = ((x_1, y_1), \dots, (x_N, y_N)) \tag{4}$$

generated by the switched system (1), the estimation problem of interest here is to estimate the parameter matrix A°.

The focus of this paper is this estimation problem. We consider that the number s of subsystems and the structural parameters (n_a, n_b) entering the definition of x_t in (2)-(3) are known a priori. Our goal is to design a map, called estimator, which maps the data ϖ^N to the set of parameters describing the constituent subsystems of the switched system (1). To begin with the approach taken in this paper to such an estimation problem, let $\mathbb{T}$ and $\mathbb{S}$ denote the index sets of the data and the subsystems respectively, i.e., $\mathbb{T} = \{1, \dots, N\}$ and $\mathbb{S} = \{1, \dots, s\}$. Use the notation Σ to denote the set of all maps $\sigma : \mathbb{T} \to \mathbb{S}$ (called here switching signals). Consider the cost function $\mathcal{J}^\circ : \mathbb{R}^{n \times s} \times \Sigma \to \mathbb{R}_+$ defined by

$$\mathcal{J}^\circ(A, \sigma) = \sum_{t=1}^{N} \left| y_t - a_{\sigma(t)}^\top x_t \right|$$

where $A \in \mathbb{R}^{n \times s}$ and $\sigma \in \Sigma$. Then a natural estimator of A° can be defined as the set-valued map $\Psi : (\mathbb{R}^n \times \mathbb{R})^N \to \mathbb{R}^{n \times s}$,

$$\Psi(\varpi^N) = \left\{ \mathrm{Set}(\hat{A}) : \exists \hat{\sigma} \in \Sigma, (\hat{A}, \hat{\sigma}) \in \arg\min_{A, \sigma} \mathcal{J}^\circ(A, \sigma) \right\}$$

$\Psi(\varpi^N)$ is the set of all sets $\mathrm{Set}(\hat{A})$ for all $\hat{A} \in \mathbb{R}^{n \times s}$ such that $(\hat{A}, \hat{\sigma})$ is a minimizer of $\mathcal{J}^\circ(A, \sigma)$ for some switching signal $\hat{\sigma}$. If we let

$$\mathcal{J}(A) = \sum_{t=1}^{N} \min_{i=1,\dots,s} \left| y_t - a_i^\top x_t \right| \tag{5}$$

then it can be easily shown that

$$\Psi(\varpi^N) = \left\{ \mathrm{Set}(\hat{A}) : \hat{A} \in \arg\min_{A} \mathcal{J}(A) \right\}. \tag{6}$$

Hence, minimizing $\mathcal{J}^\circ(A, \sigma)$ is equivalent to minimizing $\mathcal{J}(A)$ in (5). The so defined Ψ will be called the least sum-of-minimums (LSM) estimator. Because the prediction error is measured here in term of the absolute value loss function, we may also refer to Ψ in the sequel as the absolute deviation LSM estimator. We start by observing that solving numerically any of these formulations of the switched identification problem is quite hard. The focus of this paper is not on this computational aspect but on the formal properties of the map Ψ. More precisely, we are interested in characterizing conditions (on the data-generating system (1) and on the properties of the data) under which $\Psi(\varpi^N)$ may contain a singleton (unique solution) or may be located at a bounded distance from the true parameter matrix A°. The primary interest of such conditions is to emphasize the main influential factors of the estimator's performance. From this perpective, we do not expect the intended properties to be necessarily numerically verifiable but to have a rather qualitative flavor which may serve for experiment design for instance.

III. Basic properties of the estimator

We start by introducing some definitions. For any matrix $A = [a_1 \quad \cdots \quad a_s] \in \mathbb{R}^{n \times s}$, let $\sigma_A \in \Sigma$ be a switching signal satisfying

$$\sigma_A(t) \in \arg\min_{i \in \mathbb{S}} \left| y_t - x_t^\top a_i \right| \tag{7}$$

for all $t \in \mathbb{T}$. The defining constraint (7) of the switching signal σ_A allows indeed for multiple choices of $\sigma_A(t)$ whenever $\arg\min_{i \in \mathbb{S}} \left| y_t - x_t^\top a_i \right| \subset \mathbb{S}$ is not a singleton. One simple choice to solve this issue would be to set arbitrarily $\sigma_A(t)$ to be equal to the smallest element of $\arg\min_{i \in \mathbb{S}} \left| y_t - x_t^\top a_i \right|$. However, for the purpose of our analysis we will define such $\sigma_A(t)$ in a more specific way. Consider the index set

$$I_i(A) = \{ t \in \mathbb{T} : \sigma_A(t) = i \}. \tag{8}$$

Then for all $A \in \mathbb{R}^{n \times s}$, we have $I_i(A) \cap I_j(A) = \emptyset$ for $i \neq j$ and $\mathbb{T} = \cup_{i=1}^{s} I_i(A)$. For reasons that will become clear in the rest of the paper, it is desired here that $\min_{i \in \mathbb{S}} |I_i(A)|$ be as large as possible. That is, we want the partition $\{I_i(A)\}_{i \in \mathbb{S}}$ of $\mathbb{T}$ to be as balanced as possible in term of the cardinalities of its members. Hence, it is of interest to use the possible extra-degree of freedom offered by Eq. (7) to select σ_A so as to maximize $\min_{i \in \mathbb{S}} |I_i(A)|$ subject to the constraint (7). In case the maximizing σ_A is still not unique, we can make it unique for a given A by selecting the one which assigns to each t, the smallest admissible index $i \in \mathbb{S}$. To sum up, given $A \in \mathbb{R}^{n \times s}$, σ_A can be selected uniquely by following the process described above.

Given σ_A, let us now define the vector $\phi(A)$ collecting the errors of the form $y_t - x_t^\top a_{\sigma_A(t)}$ for $t \in \mathbb{T}$,

$$\phi(A) = \left[y_1 - x_1^\top a_{\sigma_A(1)} \quad \cdots \quad y_N - x_N^\top a_{\sigma_A(N)} \right]^\top. \tag{9}$$

Then the cost function $\mathcal{J}(A)$ in (5) is the ℓ_1 norm of $\phi(A)$, $\mathcal{J}(A) = \|\phi(A)\|_1$. Note in passing that $\mathcal{J}(A)$ is invariant under column permutation of the matrix A. This property implies that $\mathcal{J}(A)$ is indeed a function of $\mathrm{Set}(A)$. Note that this is an intrinsic property of the multiple-regression problem. In other words, the invariance property of $\mathcal{J}(A)$ does not constitute any restriction on the switching mechanism of the to-be-identified data-generating system (1).

A. Informativity measure of data and exact recovery

For any $r \in \{0, \dots, N\}$, denote with $\mathcal{S}_r \subset \mathbb{R}^N$ the set of r-sparse vectors in $\mathbb{R}^N$, i.e.,

$$\mathcal{S}_r = \left\{ w \in \mathbb{R}^N : \|w\|_0 \leq r \right\}. \tag{10}$$

For $A \in \mathbb{R}^{n\times s}$, define the distance $\delta_r(A)$ from $\phi(A)$ to the set $\mathcal{S}_r$ by

$$\delta_r(A) = \inf_w \Big\{ \|\phi(A) - w\|_1 : w \in \mathcal{S}_r \Big\}. \tag{11}$$

The so-defined $\delta_r(A)$ represents in fact the sum of the $N-r$ smallest entries (in absolute value) of $\phi(A)$. In particular, $\delta_0(A) = \|\phi(A)\|_1$ and $\delta_N(A) = 0$.

For any subset $\mathcal{T}$ of $\mathbb{T}$, let $\phi_{\mathcal{T}}(A)$ refer to a subvector of $\phi(A)$ formed with the entries indexed by $\mathcal{T}$.

Definition 1 (Concentration ratio)**.** Consider the dataset ϖ^N and the associated map ϕ defined in (9). Let $r \in \{0,\dots,N\}$. We call r-th concentration ratio of ϕ on the dataset ϖ^N expressed in (4), the number defined by

$$\xi_r(\varpi^N) = \sup_{\substack{(A,A')\in(\mathbb{R}^{n\times s})^2 \\ \mathcal{T}\subset\mathbb{T}}} \Big\{ \frac{\|\phi_{\mathcal{T}}(A) - \phi_{\mathcal{T}}(A')\|_1}{\|\phi(A) - \phi(A')\|_1} : \phi(A) \neq \phi(A'), |\mathcal{T}| \leq r \Big\}. \tag{12}$$

The supremum is taken here with respect to any pair $(A, A') \in (\mathbb{R}^{n\times s})^2$ such that $\phi(A) \neq \phi(A')$ and over all subsets $\mathcal{T}$ of $\mathbb{T}$ whose cardinality does not exceed r. The supremum exists because it is applied to a set which is upper-bounded by 1.

We interpret the concentration ratio as a function which measures quantitatively different levels r of informativity of the data. For a given level r, the data ϖ^N are all the more informative as $\xi_r(\varpi^N)$ is small. Ideally, we would like $\xi_r(\varpi^N)$ to be as small as possible for the largest possible level r.

Computing numerically $\xi_r(\varpi^N)$ would require in general solving a hard combinatorial optimization problem, the complexity of which might not be affordable in practice. It can however be more cheaply over-approximated thanks to the direct observation that $\xi_r(\varpi^N) \leq r\xi_1(\varpi^N)$. This is because searching for $\xi_1(\varpi^N)$ instead of $\xi_r(\varpi^N)$ alleviates considerably the combinatorial nature of the problem. Note in passing that $\xi_r(\varpi^N)$ is an increasing function of r and satisfies $\xi_0(\varpi^N) = 0$ and $\xi_N(\varpi^N) = 1$.

Remark 1. *In the special case where $s = 1$ (i.e., the situation where* (1) *reduces to a single subsystem), the matrix A reduces to a single vector, say $A = a \in \mathbb{R}^n$. We recover the classical linear regression problem. Then*

$$\begin{aligned}\phi(A) &= \begin{bmatrix}(y_1 - x_1^\top a) & (y_2 - x_2^\top a) & \cdots & (y_N - x_N^\top a)\end{bmatrix}^\top \\ &= \mathbf{y} - X^\top a.\end{aligned}$$

where $X = [x_1 \quad \cdots \quad x_N] \in \mathbb{R}^{n\times N}$ is a matrix collecting all the regressors $\{x_t\}_{t\in\mathbb{T}}$ generated by (1) *and $\mathbf{y} = [y_1 \quad \cdots \quad y_N]$ is the vector of all output samples. In this case, $\xi_r(\varpi^N)$ in* (12) *takes the form*

$$\xi_r^\circ(\varpi^N) = \sup_{\substack{\eta\in\mathbb{R}^n \\ \mathcal{T}\subset\mathbb{T}}} \left\{ \frac{\|X_{\mathcal{T}}^\top \eta\|_1}{\|X^\top \eta\|_1} : \eta \neq 0, |\mathcal{T}| \leq r \right\} \tag{13}$$

where it is assumed that $\operatorname{rank}(X) = n$, that is, X is full row rank. The notation $X_{\mathcal{T}}$ refers to the matrix formed with the columns of X indexed by $\mathcal{T}$. We observe that in this case, $\xi_r^\circ(\varpi^N)$ depends only on the regressor matrix X. Moreover, it can be overestimated through the solution of a convex optimization, see [2].

Using the concentration ratio introduced in (12), we can now state a fundamental lemma for our analysis (see Lemma 2 below, which can be viewed as a special reformulation of Lemma 4.2 in [3]). To ease the proof, we start with a preliminary technical lemma.

Lemma 1. *Let $r \in \{0,\dots,N\}$ and $\mathcal{S}_r$ be defined as in* (10)*. Consider an arbitrary vector $v \in \mathbb{R}^N$ and define*[1] *$\mathcal{T}_r(v) \subset \mathbb{T}$ to be the index set of the r largest entries in absolute value of v. Then for all $v' \in \mathbb{R}^N$,*

$$\begin{aligned}\|v' - v\|_1 - 2\|(v' - v)_{\mathcal{T}_r(v)}\|_1 \\ \leq \|v'\|_1 - \|v\|_1 + 2\inf_{w\in\mathcal{S}_r} \|w - v\|_1\end{aligned} \tag{14}$$

Proof. See Appendix A. □

Lemma 2. *Let $r \in \{0,\dots,N\}$. Consider the dataset ϖ^N as in* (4) *and $\xi_r(\varpi^N)$ as defined in* (12)*. If $\xi_r(\varpi^N) < 1/2$, then*

$$\begin{aligned}\|\phi(A') - \phi(A)\|_1 \leq &\frac{1}{1 - 2\xi_r(\varpi^N)} \big(\mathcal{J}(A') - \mathcal{J}(A) + 2\delta_r(A)\big) \\ &\forall (A, A') \in \mathbb{R}^{n\times s} \times \mathbb{R}^{n\times s}\end{aligned} \tag{15}$$

with $\phi(A)$, $\mathcal{J}(A)$ and $\delta_r(A)$ defined in (9)*,* (5) *and* (11) *respectively.*

Proof. Let $\mathcal{T}$ be a subset of $\mathbb{T}$ containing the indices of the r largest entries of $\phi(A)$ in absolute value. We apply the result of Lemma 1 with $v = \phi(A)$ and $v' = \phi(A')$, which leads immediately to

$$\begin{aligned}\|\phi(A') - \phi(A)\|_1 - 2\|\phi_{\mathcal{T}}(A') - \phi_{\mathcal{T}}(A)\|_1 \\ \leq \|\phi(A')\|_1 - \|\phi(A)\|_1 + 2\delta_r(A)\end{aligned} \tag{16}$$

where $\delta_r(A)$ is defined as in (11). From the definition (12) of ξ_r, it can further be observed that

$$\|\phi_{\mathcal{T}}(A') - \phi_{\mathcal{T}}(A)\|_1 \leq \xi_r(\varpi^N) \|\phi(A') - \phi(A)\|_1,$$

which in turn implies that $\big(1 - 2\xi_r(\varpi^N)\big) \|\phi(A') - \phi(A)\|_1$ is smaller than the left hand side term of (16). We therefore get

$$\begin{aligned}\big(1 - 2\xi_r(\varpi^N)\big) \|\phi(A') - \phi(A)\|_1 \\ \leq \|\phi(A')\|_1 - \|\phi(A)\|_1 + 2\delta_r(A)\end{aligned}$$

and the result follows. □

Remark 2. *In the scenario of Remark 1, the result of Lemma 2 would read as*

$$\|X^\top (a' - a)\|_1 \leq \frac{1}{1 - 2\xi_r^\circ(\varpi^N)} \big(\mathcal{J}(a') - \mathcal{J}(a) + 2\delta_r(a)\big) \tag{17}$$

with $\xi_r^\circ(\varpi^N)$ as in (13)*. Hence if X is full row rank then the left hand side constitutes a data-dependent norm on the error $a' - a$. If we let $\lambda = \inf_{\|\eta\|_1=1} \|X^\top \eta\|_1$, then $\|a' - a\|_1 \leq \frac{1}{\lambda(1 - 2\xi_r^\circ(\varpi^N))} (\mathcal{J}(a') - \mathcal{J}(a) + 2\delta_r(a))$.*

By interchanging the roles of A and A' in the inequality (15) one can obtain

$$\|\phi(A) - \phi(A')\|_1 \leq \frac{1}{1 - 2\xi_r(\varpi^N)} \big(\mathcal{J}(A) - \mathcal{J}(A') + 2\delta_r(A')\big)$$

Summing this with (15) then yields the following inequality

$$\|\phi(A') - \phi(A)\|_1 \leq \frac{1}{1 - 2\xi_r(\varpi^N)} \big(\delta_r(A') + \delta_r(A)\big). \tag{18}$$

Another immediate consequence of Lemma 2 can be stated as follows:

Lemma 3. *If $\xi_r(\varpi^N) < 1/2$ for some $r \in \{0,\dots,N\}$, then for all $\hat{A} \in \arg\min_A \mathcal{J}(A)$ and for all $A \in \mathbb{R}^{n\times s}$,*

$$\|\phi(A) - \phi(\hat{A})\|_1 \leq \frac{2}{1 - 2\xi_r(\varpi^N)} \delta_r(A). \tag{19}$$

[1] with the convention that $\mathcal{T}_r(v) = \emptyset$ for $r = 0$.

Moreover, if there exists a matrix $\tilde{A}$ such that $\|\phi(\tilde{A})\|_0 \leq r$ then

$$\begin{aligned}\arg\min_A \mathcal{J}(A) &= \left\{A \in \mathbb{R}^{n\times s} : \left\|\phi(A)\right\|_0 \leq r\right\} \\ &= \left\{A \in \mathbb{R}^{n\times s} : \phi(A) = \phi(\tilde{A})\right\}\end{aligned}$$

Proof. By Eq. (15), we have

$$\left\|\phi(A) - \phi(\hat{A})\right\|_1 \leq \frac{1}{1 - 2\xi_r(\varpi^N)}\left(\mathcal{J}(\hat{A}) - \mathcal{J}(A) + 2\delta_r(A)\right)$$

for all $A \in \mathbb{R}^{n\times s}$. Because $\mathcal{J}(\hat{A}) - \mathcal{J}(A) \leq 0$, this yields immediately (19). The second statement follows from the fact that if $\|\phi(\tilde{A})\|_0 \leq r$, then $\delta_r(\tilde{A}) = 0$. Therefore, replacing A with $\tilde{A}$ in (19) shows that $\phi(\tilde{A}) = \phi(\hat{A})$ and so, $\mathcal{J}(\tilde{A}) = \mathcal{J}(\hat{A})$. Hence such an $\tilde{A}$ is necessarily in $\arg\min_A \mathcal{J}(A)$. On the other hand, since $\phi(\tilde{A}) = \phi(\hat{A})$, any $\hat{A} \in \arg\min_A \mathcal{J}(A)$ satisfies $\|\phi(\hat{A})\|_0 \leq r$ hence concluding the proof. □

An interpretation of Lemma 3 is that if the data ϖ^N used to construct the map ϕ in (9) are generated by the switched system (1) and if the data is sufficiently informative in the sense that $\xi_r(\varpi^N) < 1/2$ for some r and the system parameter vectors are such that $\|\phi(A^\circ)\|_0 \leq r$ over the data, with A° denoting the true parameter matrix (see Eq. (1)), then $\mathrm{Set}(A^\circ) \in \Psi(\varpi^N)$. At this step, a question that needs to be discussed further is whether $\mathrm{Set}(A^\circ)$ may be the unique member of $\Psi(\varpi^N)$. For this purpose we need a property of uniform rank on the data X.

Definition 2 (An integer measure of genericity)**.** [1] 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,

$$\nu_n(X) = \min\left\{m : \forall\, \mathcal{S} \subset \mathbb{T} \text{ with } |\mathcal{S}| = m,\ \mathrm{rank}(X_{\mathcal{S}}) = n\right\}. \tag{20}$$

Here, $X_{\mathcal{S}}$ is a matrix formed with the columns of X indexed by $\mathcal{S}$.

This definition implies that any submatrix of $X \in \mathbb{R}^{n\times N}$ having at least $\nu_n(X)$ columns (with $n \leq \nu_n(X) \leq N$), has full row rank. The smaller $\nu_n(X)$, the more generic the regression data X are said to be. According to this rough criterion, the most generic data X achieve $\nu_n(X) = n$. This is typically the case when the regressors $\{x_t\}_{t\in\mathbb{T}}$ are in *general position* in $\mathbb{R}^n$. Under some minimality conditions [15] on the data-generating system (1), if the input signal $\{u_t\}$ is generated at random, then $\nu_n(X) = n$ with probability one.

Equipped with this notation and the definition of genericity index $\nu_n(X)$, we can now characterize uniqueness of the minimizer of $\mathcal{J}(A)$ based on the following lemma.

Lemma 4. *Consider a dataset ϖ^N of the form* (4) *and the notation $I_i(A)$ introduced at the beginning of Section III. Assume that there exists a matrix $\tilde{A} = [\tilde{a}_1 \quad \cdots \quad \tilde{a}_s] \in \mathbb{R}^{n\times s}$ with distinct columns $\tilde{a}_i$ such that*

$$\min_{i\in\mathbb{S}} \left|I_i(\tilde{A})\right| \geq s\nu_n(X) \tag{21}$$

on the data ϖ^N. Then the following holds:

$$\forall A \in \mathbb{R}^{n\times s},\ \phi(A) = \phi(\tilde{A}) \ \Rightarrow\ \mathrm{Set}(A) = \mathrm{Set}(\tilde{A}). \tag{22}$$

Proof. Let A be such that $\phi(A) = \phi(\tilde{A})$. Then for all $t \in \mathbb{T}$, $y_t - x_t^\top a_{\sigma_A(t)} = y_t - x_t^\top \tilde{a}_{\sigma_{\tilde{A}}(t)}$, which is equivalent to $x_t^\top(\tilde{a}_{\sigma_{\tilde{A}}(t)} - a_{\sigma_A(t)}) = 0$ for all $t \in \mathbb{T}$.
The next step of the proof is to show that for any $i \in \mathbb{S}$ there exists $j^\star \in \mathbb{S}$ such that $I_{ij^\star} \triangleq I_i(\tilde{A}) \cap I_{j^\star}(A)$ has a cardinality larger than or equal to $\nu_n(X)$. For this purpose we proceed by contradiction. Take an arbitrary $i \in \mathbb{S}$ and assume that $|I_{ij}| < \nu_n(X)\ \forall j \in \mathbb{S}$. Noting that

$$I_i(\tilde{A}) = I_i(\tilde{A}) \cap \mathbb{T} = I_i(\tilde{A}) \cap (\cup_{j=1}^s I_j(A)) = \cup_{j=1}^s I_{ij},$$

we obtain $|I_i(\tilde{A})| \leq \sum_{j=1}^s |I_{ij}| < s\nu_n(X)$. But this constitutes a contradiction to the assumption (21). In conclusion, for all $i \in \mathbb{S}$, there exists a $j^\star$ such that $|I_{ij^\star}| \geq \nu_n(X)$. Now we observe that for all $t \in I_{ij^\star}$, $x_t^\top(\tilde{a}_i - a_{j^\star}) = 0$ and so, $X_{I_{ij^\star}}^\top(\tilde{a}_i - a_{j^\star}) = 0$. But since $|I_{ij^\star}| \geq \nu_n(X)$, we have $\mathrm{rank}(X_{I_{ij^\star}}) = n$, which implies that $\tilde{a}_i = a_{j^\star}$. Since all columns of $\tilde{A}$ are distinct (no repetition), we conclude that $\tilde{A}$ and A have the same columns up to a permutation which is equivalent to saying that $\mathrm{Set}(\tilde{A}) = \mathrm{Set}(A)$. □

It is interesting to note that in the absence of noise in (1), having the true parameter matrix A° to obey (21) is a sufficient condition for exact recovery of that matrix from the data. What this means is that if $v_t = 0$ for all t and if all the subsystems have been sufficiently excited in the sense that condition (21) holds for $\tilde{A} = A^\circ$, then $\Psi(\varpi^N) = \{\mathrm{Set}(A^\circ)\}$.

The following theorem recapitulates the discussion of this section.

Theorem 1. *Consider the dataset ϖ^N in* (4)*, generated by the switched system* (1)*. Assume that:*

- *ϖ^N is informative enough in the sense that $\xi_r(\varpi^N) < 1/2$ for some $r \in \{0, \ldots, N\}$; let then*

$$r^*(\varpi^N) = \max\left\{r : \xi_r(\varpi^N) < 1/2\right\}.$$

- *There exists a matrix $\tilde{A} \in \mathbb{R}^{n\times s}$ satisfying the condition* (21) *and $\|\phi(\tilde{A})\|_0 \leq r^*(\varpi^N)$.*

Then the estimator Ψ defined in (6) *satisfies*

$$\Psi(\varpi^N) = \left\{\,\mathrm{Set}(\tilde{A})\right\}. \tag{23}$$

Proof. To begin with, note that for r^* defined as in the statement of the theorem, it holds that $\delta_{r^*}(\tilde{A}) = 0$ (see Eq. (11) for the definition of δ_r). Now, since the conditions of Lemma 3 are satisfied, we can apply it to infer that if $\hat{A} \in \arg\min_A \mathcal{J}(A)$, then $\phi(\hat{A}) = \phi(\tilde{A})$ so that $\mathcal{J}(\tilde{A}) = \min_A \mathcal{J}(A)$. Conversely, it is immediate to see that any $A' \in \mathbb{R}^{n\times s}$ which satisfies $\phi(A') = \phi(\tilde{A})$ lies necessarily in $\arg\min_A \mathcal{J}(A)$. Hence we can write

$$\arg\min_A \mathcal{J}(A) = \left\{A \in \mathbb{R}^{n\times s} : \phi(A) = \phi(\tilde{A})\right\}$$

Applying Lemma 4, we can then write

$$\arg\min_A \mathcal{J}(A) = \left\{A \in \mathbb{R}^{n\times s} : \mathrm{Set}(A) = \mathrm{Set}(\tilde{A})\right\}$$

and so, from (6) we see that $\Psi(\varpi^N) = \left\{\,\mathrm{Set}(\tilde{A})\right\}$. □

An interpretation of Theorem 1 is that if the data are sufficiently informative, then the set-valued estimator $\Psi(\varpi^N)$ returns only a singleton. We would of course like this singleton to coincide with the true set of parameter vectors $\{a_i^\circ\}_{i\in\mathbb{S}}$. For this to hold, it suffices that the true parameter matrix A° satisfies the second condition of the theorem. Note that such a condition is readily satisfied (with at least $r^* = 0$) when there is no noise in the data (i.e., $v_t = 0$ in (1) for all $t \in \mathbb{T}$) provided that each subsystem generates enough data. Moreover, by the second condition of the theorem, exact recovery of the true parameter matrix A° is still achievable by the estimator Ψ when $\{v_t\}$ is a *sparse noise* sequence containing at most r^* nonzero instances, regardless of the magnitude of these nonzero values. Hence, the larger r^* (i.e., the richer the regression data ϖ^N), the more outliers the least absolute deviation LSM estimator can handle. In contrast, the condition is unlikely to hold generally when *dense noise* is present in the data.

IV. Error bounds in the presence of noise

As mentioned above, we cannot hope for an exact recovery of the true parameter matrix A° by the estimator Ψ from data affected by a dense noise sequence $\{v_t\}$. We need instead to search for a possible bound on the estimation error in function of the magnitude of the noise and the richness properties of the data. Indeed (19) almost provides such a bound. The remaining question to be investigated is, under which conditions we can lower-bound $\|\phi(\hat{A}) - \phi(A^\circ)\|_1$ by a norm applying directly to $\hat{A} - A^\circ$.

A. A key step towards the obtention of an error bound

To begin with the analysis, we introduce some useful technical tools, the first of which is the class of $\mathcal{K}_\infty$ functions (see, e.g., [6]). This class of functions will be used to measure the increasing rate of the estimation error.

Definition 3 (class-$\mathcal{K}_\infty$ functions)**.** A function $\alpha : \mathbb{R}_+ \to \mathbb{R}_+$ is said to be of class-$\mathcal{K}_\infty$ if it is continuous, zero at zero, strictly increasing and satisfies $\lim_{s\to+\infty} \alpha(s) = +\infty$.

Using this definition we can state a technical lemma which will play an important role in the analysis.

Lemma 5 ([7])**.** *Let $f : \mathbb{R}^n \to \mathbb{R}_+$ be a positive continuous function satisfying the following properties:*

- *Positive definiteness: $f(x) = 0$ if and only if $x = 0$*
- *Relaxed homogeneity: There exists a $\mathcal{K}_\infty$ function q such that $f(x) \geq q(\frac{1}{\lambda})f(\lambda x)$ for all $\lambda > 0$.*

Then for any norm $\|\cdot\|$ on $\mathbb{R}^n$, there exists a constant $\alpha > 0$ such that $f(x) \geq \alpha q(\|x\|)$.

Our goal now is to derive a bound on a certain measure of the parametric estimation error between the true parameter matrix A° and the estimated ones $\hat{A} \in \arg\min_A \mathcal{J}(A)$. Recalling that $\mathcal{J}(A)$ is invariant under column permutation of the matrix A, for this metric to be pertinent, it needs to be specified in terms of distance between the sets $\mathrm{Set}(A^\circ)$ and $\mathrm{Set}(\hat{A})$. Hence we consider a metric d of the form $d(A, A') = \|A - A'_\pi\|$ where $\|\cdot\|$ is a norm on $\mathbb{R}^{n\times s}$ and $\pi : \mathbb{S} \to \mathbb{S}$ is a permutation depending on the matrices A and A'. Here, the notation A'_π is used to refer to the matrix obtained by permuting the columns of A as prescribed by π, $A'_\pi = [a'_{\pi(1)} \quad \cdots \quad a'_{\pi(s)}]$. The existence of a permutation π such that $d(A, A')$ is upper-bounded by $\|\phi(A) - \phi(A')\|_1$ will depend here on the partitions $\{I_i(A)\}_{i\in\mathbb{S}}$ and $\{I_i(A')\}_{i\in\mathbb{S}}$ achieved by A and A' respectively on the data set ϖ^N.

Definition 4. Consider the data set ϖ^N in (4), generated by the s-modes switched system (1). We say that two matrices $A \in \mathbb{R}^{n\times s}$ and $A' \in \mathbb{R}^{n\times s}$ are *comparable over the data set ω^N* if there exists a permutation $\pi : \mathbb{S} \to \mathbb{S}$ such that $\big|I_i(A) \cap I_{\pi(i)}(A')\big| \geq \nu_n(X)$ for all $i \in \mathbb{S}$.

Note, from Lemma 4 above, that any matrix $A \in \mathbb{R}^{n\times s}$ such that $\min_{i\in\mathbb{S}} |I_i(A)| \geq s\nu_n(X)$ is comparable to any other matrix A' with distinct columns satisfying $\phi(A) = \phi(A')$. In that case, it even holds that $A = A'_\pi$ for some permutation π on $\mathbb{S}$. We state hereafter a sufficient condition for comparability.

Lemma 6. *Consider a set ϖ^N of input-output data generated by system* (1) *as defined in* (4)*. Let $A \in \mathbb{R}^{n\times s}$ be a matrix obeying* (21)*. Then any matrix $A' \in \mathbb{R}^{n\times s}$ satisfying*

$$\begin{aligned} |I_i(A)| + |I_j(A)| & \\ \geq \max_{\ell\in\mathbb{S}} \big[\, \big|I_i(A) \cap I_\ell(A')\big| + \big|I_j(A) \cap I_\ell(A')\big| \,\big] & \\ + 2(s-1)\nu_n(X) \quad \forall (i,j) \in \mathbb{S}^2, i \neq j, & \end{aligned} \tag{24}$$

is comparable to A over ϖ^N in the sense of Definition 4.

Proof. See Appendix B. □

To illustrate the condition (24), consider the simple case where $|\mathbb{S}| = s = 2$. Then, under the assumption that A is subject to (21), A and A' are comparable over ϖ^N if $N \geq \max(|I_1(A')|, |I_2(A')|) + 2\nu_n(X)$. Noting that $\max(|I_1(A')|, |I_2(A')|) = N/2 + 1/2\big||I_1(A')| - |I_2(A')|\big|$ with the outer bars denoting the absolute value, (24) reduces to $N \geq 4\nu_n(X) + \big||I_1(A')| - |I_2(A')|\big|$. This relation identifies three factors which promote comparability: (i) the data X must be generic enough (i.e., $\nu_n(X)$ small); (ii) A' must partition the data into sets of balanced cardinalities; (iii) the number N of data must be large enough.

Theorem 2. *Consider the dataset ϖ^N in* (4)*, generated by the switched system* (1) *and assume that $\xi_r(\varpi^N) < 1/2$ for some $r \in \{0, \ldots, N\}$. Let $(A, A') \in \mathbb{R}^{n\times s} \times \mathbb{R}^{n\times s}$ be a pair of comparable matrices with respect to ω^N (as defined in Eq.* (4)*). Let π denote the associated permutation. Then for any norm $\|\cdot\|$ on $\mathbb{R}^{n\times s}$, there exists a strictly positive number D such that*

$$\|A'_\pi - A\| \leq \frac{1}{D\big(1 - 2\xi_r(\varpi^N)\big)} \big(\mathcal{J}(A') - \mathcal{J}(A) + 2\delta_r(A)\big). \tag{25}$$

Proof. We start by observing that all the conditions of Lemma 2 are satisfied. As a consequence, Eq. (15) holds. Departing from this equation, we just need to find an appropriate underestimate of $\|\phi(A) - \phi(A')\|_1$. To this end, note that

$$\begin{aligned} \big\|\phi(A) - \phi(A')\big\|_1 &= \sum_{t\in\mathbb{T}} \big|x_t^\top (a_{\sigma_A(t)} - a'_{\sigma_{A'}(t)})\big| \\ &= \sum_{(i,j)\in\mathbb{S}^2} \sum_{t\in I_i(A)\cap I_j(A')} \big|x_t^\top (a_i - a'_j)\big| \\ &\geq \sum_{i\in\mathbb{S}} \sum_{t\in I_i(A)\cap I_{\pi(i)}(A')} \big|x_t^\top \eta_i\big| \end{aligned}$$

where $\eta_i = a_i - a'_{\pi(i)}$ with $\pi : \mathbb{S} \to \mathbb{S}$ denoting the permutation defining the comparability of A and A' (see Definition 4). Recall that $\big|I_i(A) \cap I_{\pi(i)}(A')\big| \geq \nu_n(X)$, $i = 1, \ldots, s$. Let $g : \mathbb{R}^{n\times s} \to \mathbb{R}_+$ be the function defined by

$$g(\Lambda) = \inf_{\substack{\{J_i\}_{i\in\mathbb{S}} \\ |J_i|\geq\nu_n(X)}} \sum_{i\in\mathbb{S}} \big\|X_{J_i}^\top \eta_i\big\|_1 \tag{26}$$

where the infimum is taken over all s-tuples $(J_1, \ldots, J_s)$ of disjoint subsets of $\mathbb{T}$ with cardinality larger or equal to $\nu_n(X)$. Then by letting $\Lambda = A - A'_\pi$, it follows from the inequality above that

$$\big\|\phi(A) - \phi(A')\big\|_1 \geq g(\Lambda). \tag{27}$$

Since the infimum in (26) operates here on a finite set, it is reached by a certain $(J_1^\star, \ldots, J_s^\star)$. As a consequence g can be expressed by $g(\Lambda) = \sum_{i\in\mathbb{S}} \big\|X_{J_i^\star}^\top \eta_i\big\|_1$. The rest of the proof consists in showing that the function g satisfies the conditions of Lemma 5. Clearly, g is positive. If for some $E = [e_1 \quad \cdots \quad e_s] \in \mathbb{R}^{n\times s}$, $g(E) = 0$, then $X_{J_i^\star}^\top e_i = 0$ for all $i = 1, \ldots, s$. It follows, by the fact that $|J_i^\star| \geq \nu_n(X)$, that $e_i = 0$. Hence $E = 0$ and consequently, g is positive-definite. Moreover, g is continuous as a consequence of the ℓ_1 norm being continuous. Finally, g satisfies the relaxed homogeneity property with the $\mathcal{K}_\infty$ function q defined by $q(x) = x$. We can therefore apply Lemma 5 to conclude that $g(\Lambda) \geq D\,\|\Lambda\|$ with D being the strictly positive number defined by

$$D = \inf_{\|\Lambda\|=1} g(\Lambda). \tag{28}$$

This concludes the proof. □

The theorem establishes a bound on the metric $d(A, A')$ in case A and A' are comparable in the sense of Definition 4. For a given

r, it is interesting to note that the bound displayed in (25) is all the smaller as the data are more generic (i.e., $\xi_r(\varpi^N)$ defined in (12) is small for a relatively large r). We also note that if A and A' are not comparable as required in the statement of the theorem then, $\|A - A'_\pi\|$ can grow arbitrarily for any permutation π while $\|\phi(A) - \phi(A')\|_1$ remains small. To see this, take for example $s = 2$ and

$$A = \begin{bmatrix}\tilde{a}_1 & \tilde{a}_2\end{bmatrix}, \quad A' = \begin{bmatrix}\tilde{a}'_1 & \beta\tilde{a}'_2\end{bmatrix}$$

with the $\tilde{a}_i$ and $\tilde{a}'_i$ being unit ℓ_2-norm vectors and $\beta \in \mathbb{R}$. Then for a given dataset ϖ^N one can choose β sufficiently large such that $\sigma_{A'}(t) = 1$ for all $t \in \mathbb{T}$, i.e., $I_1(A') = \mathbb{T}$. For such values of β, A and A' are not comparable in the sense of Definition 4. We can see however that $\|\phi(A) - \phi(A')\|_1$ is independent of β while $\|A - A'_\pi\|$ will increase arbitrarily as β increases for any permutation π on $\mathbb{S} = \{1, 2\}$.

Remark 3. *Note that in the scope of Theorem 2, it is, in principle, possible to restrict the defining supremum of $\xi_r(\varpi^N)$ in* (12) *only to all pairs (A, A') of comparable matrices. The interest of such a slight reformulation is that it would produce a smaller value of $\xi_r(\varpi^N)$ and hence a potentially tighter bound in* (25)*.*

B. Estimation error bound for the switched system

An interesting situation is when (A, A') is taken in Theorem 2 to be equal to $(A^\circ, \hat{A})$ with $\hat{A} \in \arg\min_A \mathcal{J}(A)$. In this specific case, invoking the trick used to establish (19) yields the following statement.

Corollary 1. *Consider the data ϖ^N generated by system* (1) *and assume that $\xi_r(\varpi^N) < 1/2$ for some $r \geq 0$. Let $\hat{A} \in \arg\min_A \mathcal{J}(A)$. If $\hat{A}$ and the true parameter matrix A° are comparable in the sense of Definition 4 with $\pi : \mathbb{S} \to \mathbb{S}$ denoting the associated comparability permutation, then for any norm $\|\cdot\|$ on $\mathbb{R}^{n\times s}$, there exists a number $D > 0$ such that*

$$\|\hat{A}_\pi - A^\circ\| \leq \frac{2}{D\big(1 - 2\xi_r(\varpi^N)\big)}\delta_r(A^\circ). \tag{29}$$

Since r can be any integer in $\{0, \ldots, N\}$ such that $\xi_r(\varpi^N) < 1/2$, we can, at least formally, optimize the error bound over all such r's. Hence, whenever the comparability condition holds true, a better bound can, in principle, be obtained as

$$\big\|\hat{A}_\pi - A^\circ\big\| \leq \min_{r=0,\ldots,N}\Big\{\frac{2\delta_r(A^\circ)}{D\big(1 - 2\xi_r(\varpi^N)\big)} : \xi_r(\varpi^N) < \frac{1}{2}\Big\} \tag{30}$$

As already remarked, $\delta_r(A^\circ)$ measures how far $\phi(A^\circ)$ is from the set $\mathcal{S}_r$ of all r-sparse signals in $\mathbb{R}^N$. This is essentially a measure of the amount of noise $\{v_t\}$ in the system (1) which generates the data ϖ^N. More specifically, $\delta_r(A^\circ)$ equals the sum of the $N - r$ smallest elements in absolute value of the sequence $\{v_t^\circ\}_{t\in\mathbb{T}}$ defined by

$$v_t^\circ = v_t + x_t^\top(a^\circ_{\sigma(t)} - a^\circ_{\sigma_{A^\circ}(t)}) \tag{31}$$

with σ denoting the true switching signal from (1). From the definition of $\sigma_{A^\circ} \in \Sigma$ (see Eq. (7)), it is not hard to see that $|v_t^\circ| \leq |v_t|$ for all $t \in \mathbb{T}$ and so, $\delta_r(A^\circ) \leq \|\mathbf{v}\|_{1,r}$ with $\|\mathbf{v}\|_{1,r}$ denoting the sum, in absolute value, of the $N - r$ smallest entries of $\{v_t\}_{t\in\mathbb{T}}$. It follows that under the conditions of Corollary 1, $\|\hat{A}_\pi - A^\circ\| \leq 2/\big(D(1-2\xi_r(\varpi^N))\big)\|\mathbf{v}\|_{1,r}$. Hence, by considering the special case where r is taken equal to 0 (this is a reasonable choice e.g., when there is no outlier in the data), we get $\|\hat{A}_\pi - A^\circ\| \leq 2/D\,\|\mathbf{v}\|_1$ .Note that an underestimate $\hat{D}$ of the number D can be numerically found as suggested in Appendix E. Using $\hat{D}$ (instead of D) in the expression of the bound yields however a more pessimistic value of the bound. A question we ask now is, under which condition we may have $v_t^\circ = v_t$ from (31). Such a condition is given in the following proposition.

Proposition 1. *Consider the switched system* (1) *driven by the switching signal σ and the noise $\{v_t\}$. Then a necessary and sufficient condition for $\sigma_{A^\circ} = \sigma$ (irrespective of the values of σ and those of the noise) is*

$$|v_t| < \frac{1}{2}\min_{\substack{(i,j)\in\mathbb{S}^2\\ i\neq j}}\big|x_t^\top(a_i^\circ - a_j^\circ)\big| \;\forall t \in \mathbb{T}. \tag{32}$$

Proof. See Appendix C. □

The term on the right hand side of (32) can be interpreted as a measure of how distinguishable the subsystems are with respect to each other. Hence, what the proposition says is that if the noise level is below a certain threshold (which depends on the parametric distinguishability of the subsystems and on some genericity condition on the regressors), then the true switching signal coincides with σ_{A°. Finally, an interesting consequence of Proposition 1 is that, under condition (32), we obtain from (31) that $v_t^\circ = v_t$ for all $t \in \mathbb{T}$ with the consequence that $\delta_r(A^\circ)$ reduces to $\|\mathbf{v}\|_{1,r}$.

C. On the comparability of $\hat{A}$ and A°

According to Corollary 1, a sufficient condition for the estimation error induced by the estimator Ψ to be bounded as in (29), is that of comparability of $\hat{A}$ and A° over ϖ^N for all $\hat{A}$ such that $\mathrm{Set}(\hat{A}) \in \Psi(\varpi^N)$ (see Definition 4). Lemma 6 suggests that to favor the comparability of A° and $\hat{A}$, the data ϖ^N and the true parameter matrix A° should satisfy (21) and (24). Indeed these conditions impose, though in a non trivial way, some constraints on the distinguishability of the modes composing the switched system, the magnitude of the noise, the excitation signal $\{u_t\}$ and the switching signal σ.
Intuitively, if the level of the noise $\{v_t\}$ is low and if the constituent subsystems are distinguishable enough, then the true parameter matrix A° and its estimate $\hat{A}$ should be comparable. We formalize this as follows.

Lemma 7. *Assume that the input-output data ϖ^N* (4)*, generated by the s-mode switched system* (1) *is such that A° obeys $\min_{i\in\mathbb{S}}|I_i(A^\circ)| \geq sm$ with $m \geq \nu_n(X)$. Introduce the notation*

$$\gamma_m = \inf_{\substack{\|\eta\|_2=1\\ |I|\geq m}}\big\|X_I^\top\eta\big\|_1, \tag{33}$$

where the infimum is taken over all subsets I of $\mathbb{T}$ with cardinality at least m and over all $\eta \in \mathbb{R}^n$ with unit ℓ_2 norm.
If the subsystems of the switched system (1) *are parametrically distinguishable enough in the sense that*

$$\min_{i\neq j}\big\|a_i^\circ - a_j^\circ\big\|_2 > \frac{2\delta_r(A^\circ)}{\gamma_m\big(1 - 2\xi_r(\varpi^N)\big)} \tag{34}$$

for some $r \in \{0, \ldots, N\}$ such that $\xi_r(\varpi^N) < 1/2$, then A° and $\hat{A}$ are comparable over ϖ^N in the sense of Definition 4 for any $\hat{A} \in \arg\min_A \mathcal{J}(A)$.

Proof. See Appendix D. □

V. Conclusion

In this paper we have studied some properties of the least sum-of-minimums (LSM) absolute deviation estimator for switched system identification. Although this estimator is hard to implement numerically, it serves here as a reference estimator to analyze the degree of richness in the data for the identification scheme to be successful. In

particular, we have proposed a bound on the estimation error induced by this estimator. Interestingly, the expression of the proposed bound involves explicitly some informativity measures of the training data. The message of that expression in essence is that the richer the data, the smaller the estimation error. This opens a nice perspective for identification experiment design for switched systems. In effect, one can form an experiment design problem by searching for the input signal which optimizes the derived information-theoretic measures and thereby, the error bound delivered by the estimator. To further pave the avenue towards optimal experiment design, an intermediary step would, perhaps, be to complement the current analysis with one of the LSM estimator when used with the classical quadratic loss. Another important direction of research is to devise efficient numerical routines for estimating the informativity indices derived in this paper.

APPENDIX

A. Proof of Lemma 1

For the sake of notational simplicity we use $\mathcal{T}_r$ in place of $\mathcal{T}_r(v)$. Let $\mathcal{T}_r^c = \mathbb{T} \setminus \mathcal{T}_r$ be the complement of $\mathcal{T}_r$ in $\mathbb{T}$. Then

$$\begin{aligned}\left\|v' - v\right\|_1 &= \left\|(v'-v)_{\mathcal{T}_r}\right\|_1 + \left\|(v'-v)_{\mathcal{T}_r^c}\right\|_1 \\ &\leq \left\|(v'-v)_{\mathcal{T}_r}\right\|_1 + \left\|v'_{\mathcal{T}_r^c}\right\|_1 + \left\|v_{\mathcal{T}_r^c}\right\|_1 \\ &= \left\|(v'-v)_{\mathcal{T}_r}\right\|_1 + \left\|v'_{\mathcal{T}_r^c}\right\|_1 + \inf_{w\in\mathcal{S}_r}\|w-v\|_1\end{aligned}$$

The inequality is derived from the triangle inequality property of the ℓ_1 norm. The last equality relation relies on the fact that $\inf_{w\in\mathcal{S}_r}\|w-v\|_1 = \left\|v_{\mathcal{T}_r^c}\right\|_1$ (the sum of the $N-r$ smallest entries in absolute value of v). Considering the term $\left\|v'_{\mathcal{T}_r^c}\right\|_1$, we can write

$$\begin{aligned}\left\|v'_{\mathcal{T}_r^c}\right\|_1 &= \left\|v'\right\|_1 - \left\|v'_{\mathcal{T}_r}\right\|_1 \\ &= \|v_{\mathcal{T}_r}\|_1 - \left\|v'_{\mathcal{T}_r}\right\|_1 + \left\|v'\right\|_1 - (\|v\|_1 - \left\|v_{\mathcal{T}_r^c}\right\|_1) \\ &\leq \left\|(v'-v)_{\mathcal{T}_r}\right\|_1 + \left\|v'\right\|_1 - \|v\|_1 + \inf_{w\in\mathcal{S}_r}\|w-v\|_1\end{aligned}$$

Here, the second equality follows by adding and subtracting $\|v_{\mathcal{T}_r}\|_1$ while the last line is obtained by applying again the triangle inequality which gives $\left\|v_{\mathcal{T}_r}\right\|_1 - \left\|v'_{\mathcal{T}_r}\right\|_1 \leq \left\|(v'-v)_{\mathcal{T}_r}\right\|_1$. The result follows by combining the second inequality with the first one above. □

B. Proof of Lemma 6

By reasoning as in the proof of Lemma 4 thanks to the fact that A satisfies condition (21), we reach easily the conclusion that for all $i \in \mathbb{S}$, there exists $i^* \in \mathbb{S}$ such that $|I_i(A) \cap I_{i^*}(A')| \geq \nu_n(X)$. Let us define a map $\pi : \mathbb{S} \to \mathbb{S}$ by posing $\pi(i) = i^*$. We need to show that π can be selected to be a permutation under condition (24) of the lemma. For this purpose, we proceed by contradiction. Recall that π is a permutation here if and only if it is injective. And there is no injective map π that satisfies $\left|I_i(A) \cap I_{\pi(i)}(A')\right| \geq \nu_n(X)$ for all $i \in \mathbb{S}$, if and only if there is a pair (i,j), $i \neq j$, and an index $\ell \in \mathbb{S}$ such that

$$\begin{cases}\left|I_i(A) \cap I_\ell(A')\right| \geq \nu_n(X) \\ \left|I_j(A) \cap I_\ell(A')\right| \geq \nu_n(X)\end{cases} \tag{35a}$$

and $\forall k \neq \ell$,

$$\begin{cases}\left|I_i(A) \cap I_k(A')\right| < \nu_n(X) \\ \left|I_j(A) \cap I_k(A')\right| < \nu_n(X)\end{cases} \tag{35b}$$

Assume for contradiction that (35) holds. Then, because $\{I_r(A')\}_{r\in\mathbb{S}}$ forms a partition of $\mathbb{T}$, $|I_i(A)| = \sum_{r=1}^s |I_i(A) \cap I_r(A')| < (s-1)\nu_n(X) + |I_i(A) \cap I_\ell(A')|$. Similarly, we can write, $|I_j(A)| < (s-1)\nu_n(X) + |I_j(A) \cap I_\ell(A')|$. Hence $|I_i(A)| + |I_j(A)| < 2(s-1)\nu_n(X) + |I_i(A) \cap I_\ell(A')| + |I_j(A) \cap I_\ell(A')|$. This is in contradiction with (24). We therefore conclude on the existence of an injective map (and hence of a permutation) $\pi : \mathbb{S} \to \mathbb{S}$. □

C. Proof of Proposition 1

If (32) holds true, then for all $t \in \mathbb{T}$ and all $i \in \mathbb{S}$ with $i \neq \sigma(t)$,

$$\begin{aligned}\left|y_t - x_t^\top a^\circ_{\sigma(t)}\right| = |v_t| &< \frac{1}{2}\left|x_t^\top(a^\circ_{\sigma(t)} - a_i^\circ)\right| \\ &\leq \frac{1}{2}\left|y_t - x_t^\top a_i^\circ\right| + \frac{1}{2}\left|y_t - x_t^\top a^\circ_{\sigma(t)}\right|\end{aligned}$$

where the last inequality is derived from the triangle inequality property of $|\cdot|$. It follows that $\left|y_t - x_t^\top a^\circ_{\sigma(t)}\right| < \left|y_t - x_t^\top a_i^\circ\right|$ which implies that $\sigma_{A^\circ}(t) = \sigma(t)$ for all t. Conversely, if $\sigma_{A^\circ} = \sigma$, then for all $(j,t) \in \mathbb{S}\times\mathbb{T}$ such that $j \neq \sigma(t)$, we get immediately that $\left|v_t\right| < \left|y_t - x_t^\top a_j^\circ\right| = \left|x_t^\top(a^\circ_{\sigma(t)} - a_j^\circ) + v_t\right|$. Taking the square and dividing by $\left|x_t^\top(a^\circ_{\sigma(t)} - a_j^\circ)\right|$ gives $\left|x_t^\top(a^\circ_{\sigma(t)} - a_j^\circ)\right| > -2v_t s_j(t)$ with $s_j(t)$ denoting the sign of $x_t^\top(a^\circ_{\sigma(t)} - a_j^\circ)$. The last inequality holds for any possible values of σ if and only if $\left|x_t^\top(a_i^\circ - a_j^\circ)\right| > 2|v_t|$ for all $(i,j) \in \mathbb{S}^2$ with $i \neq j$. □

D. Proof of Lemma 7

To begin with, let us observe that by relying on Lemma 5, it can be shown that the number γ_m in (33) is well defined and satisfies $\gamma_m > 0$. By the same reasoning as in the proof of Lemma 4, we know that there exists a map $\pi : \mathbb{S} \to \mathbb{S}$ such $\left|I_i(A^\circ) \cap I_{\pi(i)}(\hat{A})\right| \geq m \geq \nu_n(X)$. We just need to establish that such a π is bijective under the conditions of the lemma, a property which is equivalent here just to injectivity of π. We proceed by contradiction. Suppose that π is not injective, that is, we can find $(i,j) \in \mathbb{S}^2$ with $i \neq j$ such that $\pi(i) = \pi(j)$. Let $J_i = I_i(A^\circ) \cap I_{\pi(i)}(\hat{A})$. By applying Lemma 3, we can write

$$\sum_{i\in\mathbb{S}} \left\|X_{J_i}^\top(a_i^\circ - \hat{a}_{\pi(i)})\right\|_1 \leq \left\|\phi(A^\circ) - \phi(\hat{A})\right\|_1 \leq d,$$

where we have posed $d = 2\delta_r(A^\circ)/(1 - 2\xi_r(\varpi^N))$ for conciseness. On the other hand, it follows from the definition (33) of γ_m that $\sum_{i\in\mathbb{S}} \left\|X_{J_i}^\top(a_i^\circ - \hat{a}_{\pi(i)})\right\|_1 \geq \gamma_m \sum_{i\in\mathbb{S}} \|a_i^\circ - \hat{a}_{\pi(i)}\|_2$. As a consequence, we can write $\sum_{i\in\mathbb{S}} \|a_i^\circ - \hat{a}_{\pi(i)}\|_2 \leq d/\gamma_m$. Hence, if $\pi(i) = \pi(j)$, then by virtue of the triangle inequality, $\|a_i^\circ - a_j^\circ\|_2 \leq \|a_i^\circ - \hat{a}_{\pi(i)}\|_2 + \|a_j^\circ - \hat{a}_{\pi(j)}\|_2 \leq d/\gamma_m$. This is in contradiction with the assumption (34). We therefore conclude that the claim of the lemma holds true. □

E. On the estimation of the number D in (28)

The following lemma provides a method for computing an underestimate of the parameter D in (28) for a particular choice of the norm involved in its definition though at the price of a combinatorial complexity.

Lemma 8. *Assume that the norm used for the definition of the number D in (28) is $\|\cdot\|_{2,\mathrm{col}}$ defined by $\|\Lambda\|_{2,\mathrm{col}} = \sum_{i=1}^s \|\eta_i\|_2$ for $\Lambda = \begin{bmatrix}\eta_1 & \cdots & \eta_s\end{bmatrix}$. Let $m = \nu_n(X)$. Then*

$$D \geq \gamma_m \geq \inf_{|I|=m} \lambda_{\min}^{1/2}(X_I X_I^\top), \tag{36}$$

where γ_m is the number defined in (33) and $\lambda_{\min}^{1/2}(\cdot)$ denotes the square root of the minimum eigenvalue. The infimum is taken over all subsets I of $\mathbb{T}$ with cardinality equal to m.

Proof. Recall from (26) and the proof of Theorem 2 the expression $g(\Lambda) = \sum_{i\in\mathbb{S}} \left\|X_{J_i^\star}^\top \eta_i\right\|_1$ of the function g, where the $J_i^\star$ are subsets

of $\mathbb{T}$ satisfying $|J_i^\star| \geq m = \nu_n(X)$. Then by substituting $\|\Lambda\|_{2,\text{col}}$ for the norm $\|\Lambda\|$ in Eq. (28), we have

$$\begin{aligned} D = \inf_{\|\Lambda\|_{2,\text{col}}=1} g(\Lambda) &= \inf_{\|\eta_1\|_2+\cdots+\|\eta_s\|_2=1} \sum_{i\in\mathbb{S}} \left\|X_{J_i^\star}^\top \eta_i\right\|_1 \\ &\geq \inf_{\|\eta_1\|_2+\cdots+\|\eta_s\|_2=1} \sum_{i\in\mathbb{S}} \gamma_m \|\eta_i\|_2 = \gamma_m \end{aligned}$$

The inequality follows as a consequence of the definition of γ_m by which $\left\|X_{J_i^\star}^\top \eta_i\right\|_1 \geq \gamma_m \|\eta_i\|_2$ since $|J_i^\star| \geq m$. Now, to prove the last inequality in (36), it suffices to notice that $\left\|X_I^\top \eta\right\|_1 \geq \left\|X_I^\top \eta\right\|_2$. As a result,

$$\begin{aligned} \gamma_m = \inf_{\substack{\|\eta\|_2=1 \\ |I|\geq m}} \left\|X_I^\top \eta\right\|_1 &\geq \inf_{\substack{\|\eta\|_2=1 \\ |I|\geq m}} \left\|X_I^\top \eta\right\|_2 \\ &= \inf_{|I|=m} \lambda_{\min}^{1/2}(X_I X_I^\top). \end{aligned}$$

□

Given $I \subset \mathbb{T}$, it is easy to obtain $\lambda_{\min}^{1/2}(X_I X_I^\top)$. Hence to obtain an (under)-estimate of D, we need to compute $\binom{N}{m}$ such values and take the minimum of them. Here the notation $\binom{N}{m}$ refers to the binomial coefficient. If we let $\hat{D} = \inf_{|I|=m} \lambda_{\min}^{1/2}(X_I X_I^\top)$, then it follows from (29) that $\|\hat{A}_\pi - A^\circ\| \leq \frac{2}{\hat{D}} \|\mathbf{v}\|_1$ in the particular case where r is taken equal to 0.

REFERENCES

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

[2] L. Bako. On a class of optimization-based robust estimators. *IEEE Transactions on Automatic Control*, 62:5990–5997, 2017.

[3] I. Daubechies, R. DeVore, M. Fornasier, and C. S. Güntürk. Iteratively reweighted least squares minimization for sparse recovery. *Communications on Pure and Applied Mathematics*, 63:1–38, 2010.

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

[5] A. Goudjil, M. Pouliquen, E. Pigeon, and O. Gehan. A real-time identification algorithm for switched linear systems with bounded noise. In *European Control Conference, Alborg, Denmark*, 2016.

[6] C. M. Kellett. A compendium of comparison function results. *Mathematics of Control, Signals, and Systems*, 26:339–374, 2014.

[7] A. Kircher, L. Bako, E. Blanco, and M. Benallouch. An optimization framework for resilient batch estimation in cyber-physical systems. Technical report, Ecole Centrale de Lyon (arxiv.org/abs/1906.01714), 2019.

[8] F. Lauer. Global optimization for low-dimensional switching linear regression and bounded-error estimation. *Automatica*, 89:73–82, 2018.

[9] F. Lauer and G. Bloch. *Hybrid System Identification: Theory and Algorithms for Learning Switching Models*. Springer International Publishing, 2019.

[10] D. Liberzon. *Switching in Systems and Control*. Birkhauser Boston Inc., 2003.

[11] J. Lunze and F. Lamnabhi-Lagarrigue (Eds). *Handbook of Hybrid Systems Control: Theory, Tools, Applications*. Cambridge University Press, 2009.

[12] N. Ozay, M. Sznaier, C. Lagoa, and O. Camps. A sparsification approach to set membership identification of switched affine systems. *IEEE Transactions on Automatic Control*, 57:634–648, 2012.

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

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

[15] M. Petreczky, L. Bako, S. Lecoeuche, and K. Motchon. Minimality and identifiability of discrete-time SARX systems. *To appear in International Journal of Robust and Nonlinear Control*, 2020.

[16] G. Pillonetto. A new kernel-based approach to hybrid system identification. *Automatica*, 70:21–31, 2016.

[17] Z. Sun. *Switched Linear Systems: Control and Design*. Springer-Verlag London, 2005.

[18] R. Vidal. Recursive identification of switched ARX systems. *Automatica*, 44:2274–2287, 2008.

[19] R. Vidal, S. Soatto, Y. Ma, and S. Sastry. An algebraic geometric approach to the identification of a class of linear hybrid systems. In *Conference on Decision and Control, Maui, Hawaii, USA*, 2003.