Set Membership Identification of Switched Linear Systems with Known Number of Subsystems ⋆

Necmiye Ozay [a], Constantino Lagoa [b], Mario Sznaier [c]

[a] *University of Michigan, Ann Arbor, MI, USA*

[b] *The Pennsylvania State University, University Park, PA, USA*

[c] *Northeastern University, Boston , MA, USA*

Abstract

This paper addresses the problem of robust identification of a class of discrete-time linear hybrid systems, switched linear models, in a set membership framework. Given a finite collection of noisy input/output data the objective is twofold: (i) establish whether this data was generated by a system that switches amongst an a-priori known number of subsystems, and (ii) in that case identify a suitable set of linear models along with a switching sequence that can explain the available experimental information. Our main result shows that these problems are equivalent to minimizing the rank of a matrix whose entries are affine in the optimization variables, subject to a convex constraint imposing that these variables are the moments of an (unknown) Borel measure with finite support. The use of well known (tight) convex relaxations of rank allows for further reducing the problem to a semidefinite optimization that can be efficiently solved. In the second part of the paper we extend these results to handle sensor failures that result in corrupted input/output measurements. Assuming that these failures are infrequent, we show that the problem can be recast into an optimization form where the objective is to simultaneously minimize the rank of a matrix and the number of nonzero rows of a second one. In both cases, appealing to well known convex relaxations of rank and sparsity leads to overall semidefinite optimization problems that can be efficiently solved. These results are illustrated with multiple examples showing substantially improved identification performance in the presence of noise and sensor faults.

Key words: System identification, hybrid systems, convex optimization.

1 Introduction and Motivation

In the past few years, considerable attention has been devoted to the problem of identifying hybrid systems from noisy experimental data, under different scenarios. The case where it is of interest to obtain a model that explains the observed data with the minimum possible number of switches (relevant for instance for fault detection applications) was recently addressed in [20], where it was shown that it can be reduced to a convex optimization problem. On the other hand, in several situations of practical interest, the number of subsystems s is known a-priori. For example, the system is known to switch amongst a given number of operating points (in the case of control systems), metabolic stages (in systems biology applications), or classes (computer vision and image processing). Thus, in these cases, the goal is to obtain a model that explains the data with s subsystems. Unfortunately it is known that, in the presence of unknown–but–bounded noise, this scenario leads to an NP–hard problem [21,8]. Several approaches have been proposed to address this difficulty [14,3,23,12]. While these are successful when dealing with relatively small noise levels or moderate size problems, performance deteriorates as the noise level or problem size increases. An alternative approach based on the use of convex optimization was proposed in [19,20,1]. This method tends to work well in practice, but it is not hard to construct academic counterexamples where it fails, due to its greedy nature.

Motivated by these difficulties, in this paper we propose a convex optimization–based approach to the problem of identifying hybrid systems from noisy input/output data and some minimal *a–priori* information (order and number of subsystems and bounds on the norm of the noise). Note that in this scenario, no information is available about the distribution of the noise, other than its support. Thus, in the spirit of set–membership identification, in the sequel the noise will

⋆ This paper was not presented at any IFAC meeting. Preliminary versions of this work appeared in [17,18]. Corresponding author N. Ozay. Tel. +1-(734)-936-0269.

Email addresses: necmiye@umich.edu (Necmiye Ozay), lagoa@engr.psu.edu (Constantino Lagoa), msznaier@coe.neu.edu (Mario Sznaier).

be treated as a deterministic, bounded sequence. The starting point is the algebraic procedure due to Vidal *et al.* [28], [14]. In the case of noiseless measurements, the (unknown) parameters of each subsystem are recovered from the null space of a matrix $\mathbf{V}(\mathbf{r})$ constructed from the input/output data $\mathbf{r}$ via a nonlinear embedding (the Veronese map). In the case of noisy data, the entries of this matrix depend polynomially on the unknown noise terms. Thus, finding a model in the consistency set (e.g. a model that interpolates the data within the noise level) is equivalent to finding an admissible noise sequence $\boldsymbol{\eta}$ that renders the matrix $\mathbf{V}(\mathbf{r})$ rank deficient, and a vector $\mathbf{c}$ in its null space. However, this is not trivial, given the polynomial dependence noted above. The main result of this paper shows that the problem of jointly finding $\boldsymbol{\eta}$ and $\mathbf{c}$ is equivalent to minimizing the rank of a matrix whose entries are affine in the optimization variables, subject to a convex constraint imposing that these variables are the moments of a suitable Borel measure. This result is achieved by using first an idea similar to that of [11] relating polynomial optimization and the problem of moments, to eliminate the polynomial dependence on the optimization variables, albeit at the price of introducing infinitely many constraints. The structure of the problem can then be exploited to decouple it into several *finite dimensional* smaller ones, each involving only the moments of a one–dimensional Borel measure. Combining these ideas with a convex relaxation, similar to the log-det heuristic of [7], that aims at dropping the rank of $\mathbf{V}$ by one and estimating a vector in its nullspace, allows for recasting the original problem into a semidefinite optimization form that can be solved efficiently.

In the second part of the paper, we generalize these results to handle potentially faulty measurements. This problem arises in many practical situations, typically involving data collected remotely and transmitted over a channel subject to outages. In principle, one can circumvent the difficulties associated with faulty measurements by restricting the identification algorithm to use only those measurements known to be reliable (note that data records with gaps pose no particular problems to the algebraic methods mentioned above). However, the main obstacle in pursuing such an approach is that neither the underlying dynamics, the switching sequence nor the location of the faulty measurements are a-priori known. To address this difficulty, in this paper we will exploit the fact that the failures under consideration are sparse. When combined with an algebraic geometric identification algorithm, this observation allows for recasting the identification with faulty measurements problem into an optimization form where the objective is to simultaneously minimize the rank of a matrix and the number of nonzero rows of a second one, both generated from the experimental information. Finally, the use of recently developed convex relaxations for problems of this type, leads to a computationally efficient algorithm, based on the solution of a semidefinite optimization problem. These results are illustrated with a non–trivial example: classifying human activities from video data transmitted over a noisy channel.

The rest of the paper is organized as follows. Section 2 presents some background results related to the problem of moments. In section 3, we formally state the problem under consideration and present a solution to this problem. In section 4, we address the issue of handling outliers. Section 5 illustrates the proposed method with several academic examples and a practical one. Finally, section 6 concludes the paper with some remarks.

2 PRELIMINARIES

For ease of reference, we summarize next the notation used in the paper and recall some results required to recast the identification problem into a convex optimization form.

2.1 Notation

By $\mathbf{x}$ and $\mathbf{M}$, we denote a vector in $\mathbb{R}^n$ and a matrix in $\mathbb{R}^{n\times m}$, respectively. $\|\mathbf{x}\|_\infty \doteq \sup_i |x_i|$ is the ∞-norm of a vector. When clear from context, we also use $\mathbf{x}$ to denote a sequence $\{\mathbf{x}\}_{t=0}^{T}$. $||\mathbf{M}||_{row,0}$ denotes the number of nonzero rows of the matrix $\mathbf{M}$. $\mathbf{I}$ denotes the identity matrix of appropriate dimensions. $\mathbf{M} \succeq \mathbf{N}$ indicates that the matrix $\mathbf{M}-\mathbf{N}$ is positive semidefinite. Given a subspace $W \subset R^n$, $W^\perp$ denotes its orthogonal complement. The Veronese map of degree s, denoted by ν_s, is a mapping from $\mathbb{R}^n$ to $\mathbb{R}^m$ with $m = \binom{s+n-1}{s}$, defined by $\nu_s([x_1,\ldots,x_n]^T) = [\ldots,\xi^s,\ldots]^T$ where $\xi^s \doteq x_1^{s_1}x_2^{s_2}\ldots x_n^{s_n}$, $\sum s_i = s$, e.g. all possible monomials of order s, in lexicographical order. For a given polynomial $p(\mathbf{x})$, its gradient is denoted by ∇p. Finally, in the sequel $\mathbb{P}^{[n]}(x)$ denotes the ring of multivariate polynomials in the indeterminate $\mathbf{x} \in R^n$. Finally, $\mathbb{P}_s^{[n]}$ denotes the subset of $\mathbb{P}^{[n]}$ formed by homogeneous polynomials of degree s.

2.2 Properties of Subspace Arrangements

In this section we briefly recall some results from algebraic geometric that are key in establishing the convergence properties of the proposed identification algorithm.

Definition 1 *The arrangement $\mathcal{A}(W)$ of a set of subspaces $W = \{W_i\}_{i=1}^{s} \subseteq R^n$ is defined as:*

$$\mathcal{A}(W) \doteq W_1 \cup W_2 \cup \ldots \cup W_s \tag{1}$$

Definition 2 *$I_s(\mathcal{A})$, the homogeneous component of degree s of the vanishing ideal of a subspace arrangement $\mathcal{A} \subseteq R^n$ is the set of all homogeneous multivariate polynomials of degree s in n variables that vanish on all points in $\mathcal{A}$, that is:*

$$I_s(\mathcal{A}) \doteq \left\{ p \in \mathbb{P}_s^{[n]} \colon p(\mathbf{z}) = 0 \ \forall \mathbf{z} \in \mathcal{A} \right\} \tag{2}$$

As we will show in the paper, convergence of any interpolatory algorithm for identifying a switched linear system with s subsystems is intimately related to the dimension of $I_s(\mathcal{A})$, where $\mathcal{A}$ is the subspace arrangement generated by the null spaces of each subsystem. Unfortunately, computing $dim[I_s(\mathcal{A})]$ for a generic arrangement is non-trivial (see for instance [15]). On the other hand, a simple expression exists for the practically relevant case of *transversal* arrangements, defined below. [1]

Definition 3 *Consider a subspace arrangement of the form* (1) *and, for each nonempty subset S of the index set $\{1, 2, .., s\}$ define its dimension d_S and co-dimension c_s as:*

$$d_S = dim(\cap_{i\in S} W_i),\ c_S = n - d_S \tag{3}$$

then, the arrangement $\mathcal{A}$ is said to be transversal if

$$c_S = \min\left(n, \sum_{i\in S} c_i\right) \text{ for all } S \subseteq \{1, 2, ..s\} \tag{4}$$

In this case, from the results in [15] it follows that if the co-dimension of each subspace in the arrangement is 1, then $dim(I_s) = 1$.

2.3 The Problem of Moments

Given a sequence of scalars $\{m_i\}_{i=1}^n$, the *problem of moments* is to determine whether there exists a representing Borel measure that has $\{m_i\}$ as its first n moments (see references [25], [10], [5] for a historical review and details of the problem). In particular, in the sequel we are interested in Borel measures whose support is included in bounded symmetric intervals of the real line. In this case, the following theorem provides necessary and sufficient conditions for the existence of such a measure.

Theorem 1 *Given a sequence $\{m_i : i = 1, 2, \ldots, n\}$, there exists a Borel measure $\mu(.)$ with support contained in $\mathcal{X} \doteq [-\epsilon, \epsilon]$ such that $\mu(\mathcal{X}) = 1$ and*

$$m_i = \mathcal{E}_\mu(x^i) = \int_{-\epsilon}^{\epsilon} x^i \mu(dx)$$

if and only if

- *when $n = 2k + 1$ (odd case), the following holds*

$$\epsilon \mathbf{M}(0, 2k) \succeq \mathbf{M}(1, 2k + 1) \tag{5}$$

$$\mathbf{M}(1, 2k + 1) \succeq -\epsilon \mathbf{M}(0, 2k) \tag{6}$$

- *when $n = 2k$ (even case), the following holds*

$$\mathbf{M}(0, 2k) \succeq 0 \tag{7}$$

$$\epsilon^2 \mathbf{M}(0, 2k - 2) \succeq \mathbf{M}(2, 2k) \tag{8}$$

where $\mathbf{M}(i, i + 2j)$ is the $(j + 1)$ by $(j + 1)$ Hankel matrix formed from the moments, that is:

$$\mathbf{M}(i, i + 2j) \doteq \begin{bmatrix} m_i & m_{i+1} & \ldots & m_{i+j} \\ m_{i+1} & \iddots & \iddots & m_{i+j+1} \\ \vdots & \iddots & \iddots & \vdots \\ m_{i+j} & \ldots & \ldots & m_{i+2j} \end{bmatrix}, \tag{9}$$

and where $m_0 = 1$.

Proof: Direct application of Theorem III.2.3 and Theorem III.2.4 in [10]. □

The problem of moments, especially its multivariate extensions, has been used in the optimization community to convert polynomial optimization problems into a hierarchy of convex semidefinite programming problems with increasing size ([11],[13]). In this paper, we apply similar ideas to feasibility problems involving a combination of polynomial and rank constraints. By exploiting the problem structure, we show that it suffices to use one dimensional Borel measures for which the moments can be precisely characterized by fixed sized linear matrix inequalities of the form given in Theorem 1.

3 Set Membership Identification of Hybrid Linear ARX models

In this paper we consider the problem of set membership identification of switched autoregressive exogenous (SARX) linear models of the form:

$$y_t = \sum_{j=1}^{n_a} a_j(\sigma_t) y_{t-j} + \sum_{j=1}^{n_c} c_j(\sigma_t) u_{t-j} + \eta_t \tag{10}$$

where $y_t \in \mathbb{R}$, $u_t \in \mathbb{R}$ and $\eta_t \in \mathbb{R}$, with $|\eta_t| \leq \epsilon\ \ \forall t$, denote outputs, inputs and noise respectively, $\sigma_t \in \{1, \ldots, s\}$ is the latent discrete state or mode of the system, and where a_j and c_j are unknown coefficients. Note that as stated, the problem above is ill defined, since, even in the absence of noise, a given input/output sequence $(\mathbf{u}, \mathbf{y})$ can be explained by an infinite number of models of the form (10), with different orders. To avoid this ambiguity, in the sequel we will work with minimal realizations, in the following sense. Let

$$\begin{array}{l} D_i(z) \doteq z^{\max(n_a,n_c)} - \sum_{j=1}^{n_a} a_j(i) z^{\max(n_a,n_c)-j} \\ N_i(z) \doteq \sum_{j=0}^{n_c} c_j(i) z^{\max(n_a,n_c)-j}. \end{array} \tag{11}$$

[1] Intuitively, a subspace arrangement is transversal if and only if all intersections are as small as possible [15].

Assumption 1 *For each subsystem, the polynomials* $\{D_i(z), N_i(z)\}$ *are coprime.*

For ease of reference, let

$$\mathbf{p}_i \doteq \begin{bmatrix} a_1(i) \ \dots \ a_{n_a}(i) \ c_1(i) \ \dots \ c_{n_c}(i) \end{bmatrix}$$
$$\mathbf{r}_t = \begin{bmatrix} y_t \ \dots \ y_{t-n_a} \ u_{t-1} \ \dots \ u_{t-n_c} \end{bmatrix}^T$$

and $\mathbf{p} \doteq \{\mathbf{p}_i\}_{i=1}^s$. Under Assumption 1, the problem of interest in this paper can be precisely defined as follows: Given experimental input/output data $\mathbf{r}_t$ over the interval $[t_0, T]$, and a priori information $\{s, n_a, n_c, \epsilon\}$, the number of submodels, their order and a bound on the ℓ_∞ norm of the noise, respectively, define the consistency set:

$$\mathcal{T}(\mathbf{r}) \doteq \{ \mathbf{p} \colon \text{ (10) holds for some sequences } \sigma_t \in [1 \dots s]^T \text{ and } \eta_t, |\eta_t| \leq \epsilon; \text{ for } t \in [t_0, T] \} \tag{12}$$

that is, the set of all switched linear models compatible with both the a-priori assumptions and the experimental data. In this context, the problem of interest here can formally be stated as follows:

Problem 1 (Consistency) *Determine whether or not* $\mathcal{T}(\mathbf{r}) = \emptyset$*, (that is whether or not the a-priori assumptions have been invalidated by the experimental data)*

Problem 2 (Identification) *If* $\mathcal{T}(\mathbf{r}) \neq \emptyset$*, then find a model* $\mathbf{p}^{id} \in \mathcal{T}(\mathbf{r})$

Remark 1 *Note that all elements in* $\mathcal{T}(\mathbf{r})$ *are indistinguishable based on the available information. Thus, interpolatory identification algorithms such as the one proposed in this paper simply select an arbitrary element of* $\mathcal{T}$*. Using information based complexity arguments [26], it can be shown that any such choice is, in terms of the worst case identification error, optimal within a factor of 2 [24]* [2]*.*

In the sequel, we show that Problem 1 above can be reduced to minimizing the rank of a matrix affine in the optimization variables, subject to finite–dimensional convex constraints. Further, if $\mathcal{T}(\mathbf{r}) \neq \emptyset$ then a solution to Problem 2 can be obtained from the solution to this rank minimization problem.

3.1 Algebraic formulation

In the noise free case (i.e., $\eta_t = 0 \ \forall t$), Problems 1 and 2 can be elegantly solved using an algebraic procedure, Generalized Principal Component Analysis (GPCA), proposed by Vidal *et al.* [28],[14]. Note that in this case an equivalent representation of (10) is:

$$\mathbf{b}(\sigma_t)^T \mathbf{r}_t = 0 \tag{13}$$

where $\mathbf{r}_t = [-y_t, y_{t-1}, \dots, y_{t-n_a}, u_{t-1}, \dots, u_{t-n_c}]^T$ and $\mathbf{b}(\sigma_t) = [1, a_1(\sigma_t), \dots, a_{n_a}(\sigma_t), c_1(\sigma_t), \dots, c_{n_c}(\sigma_t)]^T$, denote the regressor and (unknown) coefficients vectors at time t, respectively.

The idea behind the algebraic method is based on a polynomial constraint, the so-called *hybrid decoupling constraint*, that decouples the identification of model parameters from the identification of the discrete state and switching sequence. That is, (13) holds for some σ_t if and only if

$$p_s(\mathbf{r}_t) = \prod_{i=1}^{s} (\mathbf{b}_i^T \mathbf{r}_t) = \mathbf{c}_s^T \nu_s(\mathbf{r}_t) = 0 \tag{14}$$

holds for all t independent of which of the s submodels is active at time t. In the above equality, $\mathbf{b}_i \in \mathbb{R}^{n_a+n_c+1}$ is the parameter vector corresponding to the i^{th} submodel, $\mathbf{r}_t$ is the known regressor vector at time t, $\nu_s(.)$ is the Veronese map of degree s and $\mathbf{c}_s$ is a vector that collects unknown parameters from $\mathbf{b}_i$'s (see [27] for explicit definition). Collecting all data into a matrix form leads to [3]:

$$\mathbf{V}_s(\mathbf{r})\mathbf{c}_s \doteq \begin{bmatrix} \nu_s(\mathbf{r}_{t_0})^T \\ \vdots \\ \nu_s(\mathbf{r}_T)^T \end{bmatrix} \mathbf{c}_s = \mathbf{0}, \tag{15}$$

where $\mathbf{r}$, without the subscript, stands for the set of all regressor vectors for all $t \in [t_0, T]$. In the sequel, when it is clear from context, and with a slight abuse of notation, we will use $\mathbf{V}_s$ to denote the data matrix, $\mathbf{V}_s(\mathbf{r}) \in \mathbf{R}^{(T-t_0+1)\times\binom{s+n_a+n_c}{s}}$. Under suitable transversality assumptions (see Section 3.3 for details), $\mathcal{T}(\mathbf{r}) \neq \emptyset$ if and only if $\mathbf{V}_s$ is rank deficient. In that case, Problem 2 can be solved by first finding a vector $\mathbf{c}_s$ in the nullspace of $\mathbf{V}_s$ to find the coefficients of the multivariate polynomial $p_s(\mathbf{r})$. Then $\mathbf{b}_i$, the parameters of the models, can be computed from $\mathbf{c}_s$ via polynomial differentiation (see the Appendix).

In the presence of noise, the approach outlined above breaks down, since conditions (14) and (15) no longer hold. Indeed, the noisy equivalent of (14) is given by:

$$p_s(\mathbf{r}_t, \eta_t) = \prod_{i=1}^{s} (\mathbf{b}_i^T \tilde{\mathbf{r}}_t) = \mathbf{c}_s^T \nu_s(\tilde{\mathbf{r}}_t) = 0 \tag{16}$$

where $\tilde{\mathbf{r}}_t = [-y_t + \eta_t, y_{t-1}, \dots, y_{t-n_a}, u_{t-1}, \dots, u_{t-n_c}]^T$. Proceeding as in the noiseless case, a "noisy" data matrix

[2] In principle one could attempt to reduce this error by choosing the Chebyshev center of $\mathcal{T}$ as the identified system. However, finding this center is far from trivial, even for LTI systems. Hence, the usual practice is to select an arbitrary element of $\mathcal{T}$.

[3] Note that since the initial conditions are unknown, the regressor $\mathbf{r}_t$ is not available for $t < t_0 \doteq \max(n_a, n_c)$.

$\mathbf{V}_s(\mathbf{r},\boldsymbol{\eta}) \doteq \mathbf{V}_s(\tilde{\mathbf{r}})$ can be built. However, finding the coefficients of each subsystem entails now finding both an admissible noise sequence $\boldsymbol{\eta}^o$ and a vector $\mathbf{c}^o$ in the nullspace of $\mathbf{V}_s(\mathbf{r},\boldsymbol{\eta}^o)$ such that

$$\mathbf{V}_s(\mathbf{r},\boldsymbol{\eta}^o)\mathbf{c}^o = 0 \tag{17}$$

Since $\mathbf{V}_s$ is a polynomial function of the unknown noise terms η_t, this approach leads to a computationally very challenging nonlinear, nonconvex optimization problem. However, as we show next, by exploiting the method of moments, (17) can be recast into a constrained rank minimization form which in turn can be relaxed to an efficient convex optimization.

3.2 A moments based convex relaxation for robust identification of SARX models

Consider the following rank minimization problem:

$$\begin{array}{l}\text{minimize}_{\eta_t}\ \ \text{rank}\,\mathbf{V}_s(\mathbf{r},\boldsymbol{\eta}) \\ \text{subject to}\ \ \ |\eta_t| \le \epsilon,\ \ \forall t \in [t_0, T].\end{array} \tag{18}$$

Clearly, a necessary condition for consistency is the existence of rank deficient solutions to (18). Further, as shown in the sequel, the problem of searching for these solutions admits a computationally tractable relaxation.

Exploiting Theorem 1 and using the facts that (i) η_t and $\eta_{\bar{t}}$ are uncoupled (in the sense of having no functional dependency) for $t \neq \bar{t}$, and (ii) η_t only appears in the t^{th} row of $\mathbf{V}_s$, leads to the following moments optimization problem:

$$\begin{array}{l}\text{minimize}_{\mathbf{m}}\ \text{rank}\,\tilde{\mathbf{V}}_s(\mathbf{r},\mathbf{m}) \\ \text{subject to}\ \ (5)-(6)\ \ \forall \mathbf{m}^{(t)}\ \ \forall t \in [t_0,T]\ \text{if s is odd} \\ \qquad\qquad\ \ (7)-(8)\ \ \forall \mathbf{m}^{(t)}\ \ \forall t \in [t_0,T]\ \text{if s is even}\end{array} \tag{19}$$

where $\mathbf{m}^{(t)} = [m_1^{(t)}, \ldots, m_s^{(t)}]$ is the moment sequence corresponding to η_t, $\mathbf{m}$ is the collection of all $\mathbf{m}^{(t)}$ and $\tilde{\mathbf{V}}_s(\mathbf{r},\mathbf{m})$ is a matrix linear in the moments, obtained by replacing each k^{th} degree monomial η_t^k in $\mathbf{V}_s(\mathbf{r},\boldsymbol{\eta})$ with the corresponding k^{th} order moment $m_k^{(t)}$.

Example 1 *For instance when $s = 2$ and $(n_a, n_c) = (1, 1)$, then $\mathbf{r}_t = [-y_t, y_{t-1}, u_{t-1}]^T$. In this case the rows of $\mathbf{V}_s(\mathbf{r},\boldsymbol{\eta})$ and the corresponding rows of $\tilde{\mathbf{V}}_s(\mathbf{r},\mathbf{m})$ are given by:*

$$\nu_2(\mathbf{r}_t,\eta_t)^T = \begin{bmatrix} y_t^2 - 2y_t\eta_t + \eta_t^2 \\ -y_t y_{t-1} + y_{t-1}\eta_t \\ -y_t u_{t-1} + u_{t-1}\eta_t \\ y_{t-1}^2 \\ y_{t-1}u_{t-1} \\ u_{t-1}^2 \end{bmatrix}^T;$$

$$\mathcal{E}_\mu\left[\nu_2(\mathbf{r}_t,\eta_t)^T\right] = \begin{bmatrix} y_t^2 - 2y_t m_1^{(t)} + m_2^{(t)} \\ -y_t y_{t-1} + y_{t-1} m_1^{(t)} \\ -y_t u_{t-1} + u_{t-1} m_1^{(t)} \\ y_{t-1}^2 \\ y_{t-1}u_{t-1} \\ u_{t-1}^2 \end{bmatrix}^T. \tag{20}$$

Thus, $\tilde{\mathbf{V}}_s(\mathbf{r},\mathbf{m})$ is affine in the unknown moments.

Next, we present the main equivalence result of the paper:

Theorem 2 *Let $\boldsymbol{\eta}^*$ be an optimal solution to problem* (18) *and $\mathbf{m}^*$ be an optimal solution to* (19)*. Then, $\mathbf{V}_s(\mathbf{r},\boldsymbol{\eta}^*)$ is rank deficient if and only if $\tilde{\mathbf{V}}_s(\mathbf{r},\mathbf{m}^*)$ is rank deficient. Moreover, if $\mathbf{c}$ belongs to the nullspace of $\tilde{\mathbf{V}}_s(\mathbf{r},\mathbf{m}^*)$ then there exists a noise sequence $\boldsymbol{\eta}^{**}$ with $\|\boldsymbol{\eta}^{**}\|_\infty \le \epsilon$ such that $\mathbf{c}$ belongs to the nullspace of $\mathbf{V}_s(\mathbf{r},\boldsymbol{\eta}^{**})$.*

Proof: Assume that the minimum rank r_1 in (18) is achieved by some sequence η_t^* for all $t \in [t_0, T]$. Then $\tilde{\mathbf{V}}_s(\mathbf{r},\mathbf{m}^*)$ with $\mathbf{m}^{*(t)} = [\eta_t^*, (\eta_t^*)^2, \ldots, (\eta_t^*)^s]$ (i.e., all representative Borel measures have point support) has rank r_1 and $\mathbf{m}^{*(t)}$ satisfies the LMI constraints. Hence the minimum rank obtained by solving (19) is less than or equal to the minimum rank obtained by solving (18).

Consider now an optimal solution $\mathbf{m}^*$ of (19). Note that, from Theorem 1, this guarantees the existence of $T - t_0 + 1$ measures $\mu^{*(t)}$, each supported on $[-\epsilon, \epsilon]$. Let $\mathbf{c}$ be in the nullspace of $\tilde{\mathbf{V}}_s(\mathbf{r},\mathbf{m}^*)$ (i.e., $\tilde{\mathbf{V}}_s(\mathbf{r},\mathbf{m}^*)\mathbf{c} = \mathbf{0}$). Thus, for each row of $\mathbf{V}_s$, $\mathcal{E}_{\mu^{*(t)}}\left[\nu_s(\mathbf{r}_t,\eta_t)\right]\mathbf{c} = \mathcal{E}_{\mu^{*(t)}}\left[\nu_s(\mathbf{r}_t,\eta_t)\mathbf{c}\right] = 0$. By noting that $\nu_s(\mathbf{r}_t,\eta_t)\mathbf{c}$ is a polynomial function of η_t (hence continuous) and $\mu^{*(t)}$ is supported on $[-\epsilon,\epsilon]$, we can invoke the mean value theorem for integration to conclude that there exist $\eta_t^{**} \in [-\epsilon,\epsilon]$ for all t such that $\nu_s(\mathbf{r}_t,\eta_t^{**})\mathbf{c} = 0$.

Thus, whenever the nullspace of the solution of (18) is nontrivial, so is that of (19), which proves the theorem. □

An alternative way of obtaining an equivalent moment–based problem to (18) when $\mathbf{V}_s(\mathbf{r},\boldsymbol{\eta})$ is known to be rank

deficient by one is to define the equivalent polynomial objective function $\det[\mathbf{V}_s(\mathbf{r},\boldsymbol{\eta})^T\mathbf{V}_s(\mathbf{r},\boldsymbol{\eta})]$. However, in this case one would need higher order (possibly infinite number of) moments of a multidimensional Borel measure since it is not clear how to exploit the independence of noise terms while keeping the problem linear in moments of a one dimensional Borel measure. Although the rank objective is non-convex, it has the following advantages: (i) the equivalent moment based problem is finite dimensional (i.e., it requires only finite moment matrices); (ii) there are efficient convex relaxations for rank minimization; and (iii) extracting solutions requires only solving for the roots of a polynomial in one variable whereas for the case with multidimensional measures this is a non-trivial task. We elaborate on the last two points next.

Although rank minimization is an NP–Hard problem, efficient convex relaxations are available. In particular, good approximate solutions can be obtained by using a log–det heuristic [7] that relaxes rank minimization to a sequence of convex problems. Furthermore, since from a set membership point of view it suffices to find a rank deficient solution, we propose a modification of log–det heuristic that aims at dropping the rank by one. The algorithm, which is inspired by the adaptive step size defined for weighted ℓ_1 minimization in [4], is summarized next:

Algorithm 1: Drop Rank

$\mathbf{X} \in \mathbb{R}^{m\times n}$ and assuming wlog $m \leq n$, initialize:

$k=0$, $\epsilon_0 = \infty$, $\mathbf{W}_y^{(0)} = \mathbf{I}_{m\times m}$, $\mathbf{W}_z^{(0)} = \mathbf{I}_{n\times n}$

REPEAT

Solve

$$\min_{\mathbf{X}^{(k)},\mathbf{Y}^{(k)},\mathbf{Z}^{(k)}} \ \mathrm{Tr}\begin{bmatrix}\mathbf{W}_y^{(k)}\mathbf{Y}^{(k)} & 0\\ 0 & \mathbf{W}_z^{(k)}\mathbf{Z}^{(k)}\end{bmatrix}$$

$$\text{subject to} \quad \begin{bmatrix}\mathbf{Y}^{(k)} & \mathbf{X}^{(k)}\\ \mathbf{X}^{(k)^T} & \mathbf{Z}^{(k)}\end{bmatrix} \succeq 0$$

$$\mathbf{X}^{(k)} \in \mathcal{C}$$

Decompose $\mathbf{X}^{(k)} = \mathbf{U}\mathbf{D}\mathbf{V}^T$ using SVD.

Set $\epsilon_{k+1} = \min\{\epsilon_k, \mathbf{D}(m,m)\}$.

Set $\mathbf{W}_y^{(k+1)} = (\mathbf{Y}^{(k)} + \epsilon_{k+1}\mathbf{I})^{-1}$,

$\mathbf{W}_z^{(k+1)} = (\mathbf{Z}^{(k)} + \epsilon_{k+1}\mathbf{I})^{-1}$.

Set $k = k+1$.

UNTIL (a convergence criterion is reached)

RETURN $\mathbf{X}^{(k)}$

Above, for the sake of notational simplicity, we used $\mathbf{X} = \tilde{\mathbf{V}}_s(\mathbf{r},\mathbf{m})$; and $\mathbf{X}^{(k)} \in \mathcal{C}$ stands for convex constraints, that is, $\mathbf{m}^{(t)}$ lies on a convex set $\mathcal{C}$ defined by the LMIs in (19). In the first iteration this algorithm minimizes the nuclear norm, a well-known convex surrogate for rank, of $\tilde{\mathbf{V}}_s(\mathbf{r},\mathbf{m})$. Then, if it cannot drop the rank, it uses the weighting matrices, $\mathbf{W}_y$ and $\mathbf{W}_z$, to further decrease the small singular values towards zero. Explicit conditions under which the nuclear norm minimization is guaranteed to recover minimum rank solutions are given in [22], and convergence of this algorithm to a possible local minima when these fail follows from [7].

Assuming a rank deficient $\tilde{\mathbf{V}}_s(\mathbf{r},\mathbf{m})$ is found, a vector $\mathbf{c}$ in its nullspace can be obtained by simply performing a singular value decomposition. From Theorem 2, it follows that $\mathbf{c}$ is also in the nullspace of $\mathbf{V}_s(\mathbf{r},\boldsymbol{\eta})$ (i.e., $\mathbf{V}_s(\mathbf{r},\boldsymbol{\eta})\mathbf{c} = 0$). Hence, for each row, we have $\nu_s(\mathbf{r}_t,\eta_t^*)\mathbf{c} = 0$ which is a polynomial equation in one variable. One can solve for the noise values by finding the roots of this polynomial that lie in $[-\epsilon,\epsilon]$ (which are guaranteed to exist again by Theorem 2). The roots of a univariate polynomial can be easily computed, for instance, by computing the eigenvalues of the companion matrix of the polynomial - a method that is standard in most of the numerical computing packages. Once the noise values are estimated, the problem can be converted to the noise free case by plugging the noise estimates into $\mathbf{V}_s(\mathbf{r},\boldsymbol{\eta})$ and the system parameters can be computed using the procedure described in the Appendix.

Note that the proposed approach can be easily extended to accommodate additional *a priori* information about the noise. For instance, a switched system is likely to have different noise characterizations for different modes (e.g., each submodel has its own noise bound $\epsilon_1,\ldots,\epsilon_s$). This constraint can be written as $|\eta_t| \leq \epsilon_i$ if $\sigma_t = i$ and the hybrid decoupling constraint (16) can be modified by appropriately scaling the noise variables, that is

$$p_s(\mathbf{r}_t,\eta_t) = \prod_{i=1}^{s}(\mathbf{b}_i^T\tilde{\mathbf{r}}_t^{(i)}) = 0 \tag{21}$$

where $\tilde{\mathbf{r}}_t^{(i)} = [-y_t+\eta_t/\epsilon_i, y_{t-1},\ldots,y_{t-n_a},u_{t-1},\ldots,u_{t-n_c}]^T$. Thus, enforcing that $|\eta_t| \leq 1$ in the corresponding optimization problem (18) leads to submodel parameters and noise values that are consistent with this type of *a priori* information.

Remark 2 *When the number of the submodels s is unknown, it is possible to search for minimum number of submodels that explains the data. This can be accomplished with a simple iteration on s; starting with $s=1$ and increasing s up until a rank deficient solution to Problem* (19) *is found.*

Remark 3 *The results presented in this section can be directly applied for identifying multi-input single-output (MISO) systems. Under appropriate coprimeness/minimality conditions, it is possible to use these results for identification of multi-input multi-output (MIMO) SARX models. One extension of algebraic approach to MIMO case is presented in [2] by projecting the outputs $\mathbf{y}$ onto a single randomly chosen vector. As an alternative, which better utilizes the*

problem structure in moment-based reformulations, one can proceed by considering each component $[\mathbf{y}_t]_j$ of the output $\mathbf{y}_t$ of an MIMO system, as the output of a separate MISO system where the other components, $[\mathbf{y}_t]_i,\ i \neq j$, are treated as inputs. We omit the details due to limited space.

3.3 Identifiability and Convergence

Consider first the case of noiseless data and assume that the subspace arrangement $\mathcal{A} \doteq b_1^{\perp} \cup \dots b_s^{\perp}$ generated by the null spaces of the coefficient vectors of each subsystem is transversal [4]. Under these conditions, from the results in [15] it follows that $I_s(\mathcal{A})$, the subspace of homogeneous polynomials of degree s in n variables that vanishes on this arrangement, has dimension 1. Thus, if the input and switching sequences $(\mathbf{u}, \boldsymbol{\sigma})$ are such that the corresponding trajectories span sufficiently dense subsets of the each subspace in the arrangement, the null space of the corresponding data matrix, $\mathcal{N}(\mathbf{V}_s)$ has dimension one and hence, the consistency set $\mathcal{T}(\mathbf{r})$ is a singleton. As we show in Appendix B, the minimality assumption (Assumption 1) is a sufficient condition for the existence of pairs $(\mathbf{u}, \boldsymbol{\sigma})$ that render $dim(\mathcal{N}(\mathbf{V}_s)) = 1$.

Extending the ideas outlined above to the case of noisy measurements, requires introducing the following definition:

Definition 4 *Consider a transversal SARX system of the form* (10)*, given input and switching sequences $(\mathbf{u}, \boldsymbol{\sigma})$ and the corresponding regressor sequence $\mathbf{r}$ with elements $\mathbf{r}_t = \begin{bmatrix} y_t & \dots & y_{t-n_a} & u_{t-1} & \dots & u_{t-n_c} \end{bmatrix}^T$. The pair $(\mathbf{u}, \boldsymbol{\sigma})$ is said to be ϵ-robustly persistently exciting if the corresponding data matrix satisfies*

$$dim[\mathcal{N}(\mathbf{V}_s(\mathbf{r}, \boldsymbol{\eta})] \leq 1 \tag{22}$$

for all noise sequences such that $\|\eta_t\| \leq \epsilon$ for all $t \in [t_0, T]$.

Note that, for transversal systems, any persistently exciting pair $(\mathbf{u}, \boldsymbol{\sigma})$ is robustly persistently exciting for some ϵ small enough.

Clearly, under the transversality and robustly persistent excitation assumptions, a necessary condition for consistency is the existence of a noise sequence $\boldsymbol{\eta}^o$ such that the right null space of $\mathbf{V}_s(\mathbf{r}, \boldsymbol{\eta}^o)$ has dimension 1. As we show in the sequel, this is also sufficient, under suitable conditions on the noise level and the solution to the rank minimization problem (18). To this effect, we begin by introducing the following result:

Lemma 1 *Given m points $\mathbf{x}_i = \begin{bmatrix} x_{1,i} & \dots x_{n,i} \end{bmatrix} \in R^n$, denote by $\mathbf{V_s}(\mathbf{X}) \doteq \begin{bmatrix} \nu_s(\mathbf{x}_1) & \dots \nu_s(\mathbf{x}_m) \end{bmatrix}^T$ the data matrix obtained from the Veronese map of degree s. Assume $\mathbf{V_s}(\mathbf{X})$ is rank deficient and denote by p_s a monic (in lexicographical ordering) polynomial associated with its null space. Finally, given m numbers ρ_i, define the following sets:*

$$\mathcal{B} \doteq \left\{ \tilde{\mathbf{b}}_i \colon \begin{bmatrix} 1 & \tilde{\mathbf{b}}_i^T \end{bmatrix} = \frac{\nabla^T p_s}{\mathbf{e}_1^T \nabla p_s}\Big|_{\mathbf{x}_i},\ i = 1, \dots, m \right\}$$

$$\mathcal{B}_\rho \doteq \left\{ \ell_i \in R^n \colon \ell_i = \begin{bmatrix} 1 & (1+\rho_i)\tilde{\mathbf{b}}_i^T \end{bmatrix},\ i = 1, \dots, m \right\}$$

$$\mathcal{X}_\rho \doteq \Big\{ \mathbf{x} \in R^n \colon \mathbf{x} = \begin{bmatrix} (1+\rho_i)x_{1,i} & \dots x_{n,i} \end{bmatrix}, \quad i = 1, \dots, m \Big\} \tag{23}$$

If there exist m scalars $\rho_i \neq -1,\ i = 1, \dots, m$, such that $|\mathcal{B}_\rho| \leq s$ then there exists an arrangement $\mathcal{A}_s \subset R^n$ of s proper subspaces such that $\mathcal{X}_\rho \subset \mathcal{A}_s$.

Proof: For any given set $\{\rho_i\}$, define the arrangement $\mathcal{A}_\rho = \cup \ell_i^{\perp}$. Clearly [5] $|\mathcal{A}_\rho| = |\mathcal{B}_\rho|$. Since p_s is homogeneous, as all the monomials in the Veronese map have degree s, and $p_s(\mathbf{x}_i) = 0$, then $[\nabla^T p_s(\mathbf{x}_i)]\mathbf{x}_i = 0$. Therefore,

$$\ell_i \mathbf{x}_{i,\rho_i}^T = (1+\rho_i) \begin{bmatrix} 1 & \tilde{\mathbf{b}}_i^T \end{bmatrix} \mathbf{x}_i^T = 0 \tag{24}$$

where $\mathbf{x}_{i,\rho_i} \doteq \begin{bmatrix} (1+\rho_i)x_{1,i} & \dots x_{n,i} \end{bmatrix}$. Hence $\mathcal{X}_\rho \subset \mathcal{A}_\rho$. This together with $|\mathcal{A}_\rho| = |\mathcal{B}_\rho| \leq s$ implies $\mathcal{X}_\rho \subset \mathcal{A}_s$.

Next, we use the result above to establish necessary and sufficient conditions for consistency.

Theorem 3 *Consider T noisy measurements $\{y_t\}$ of the output of a transversal SARX system, excited by an ϵ robust pair $(\mathbf{u}, \boldsymbol{\sigma})$. Then the consistency set $\mathcal{T} \neq \emptyset$ if and only if* (18) *admits a solution $\boldsymbol{\eta}^o$ such that*

(i) $dim[\mathcal{N}(\mathbf{V}_s(\mathbf{r}, \boldsymbol{\eta}^o)] = 1$ and
(ii) there exist $T - t_0 + 1$ numbers $\{\rho_t\}$ such that $|\mathcal{B}_\rho| \leq s$ and

$$|\rho_t y_t + (1+\rho_t)\eta_t^o| \doteq |\tilde{\eta}_t| \leq \epsilon. \tag{25}$$

Proof: Under the transversality and robust persistency assumptions, if $\mathcal{T} \neq \emptyset$, then there exists at least one sequence $\boldsymbol{\eta}^o$ such that $\dim[\mathcal{N}(\mathbf{V}_s(\mathbf{r}, \boldsymbol{\eta}^o)] = 1$, where $\mathbf{r}$ is constructed according to the a priori information on model orders, and the corresponding polynomial can be factored as a product of linear forms. Thus, the conditions in the Theorem hold with $\rho_i \equiv 0$. On the other hand, if (18) admits a solution $\boldsymbol{\eta}^o$ such that condition (i) and (ii) hold, then from Lemma 1 it follows that the points $\tilde{\mathbf{r}}_t = [-y_t -$

[4] In the sequel, by a slight abuse of notation we will refer to these systems as transversal.

[5] By a slight abuse of notation, here we define cardinality of $\mathcal{A}_\rho$ as the number of distinct $n-1$ dimensional subspaces in the arrangement.

$\tilde{\eta}_t, y_{t-1}, \ldots, y_{t-n_a}, u_{t-1}, \ldots, u_{t-n_c}]^T$, with $|\tilde{\eta}_t| \leq \epsilon$, lie in an arrangement of s subspaces. □

Note that the existence of suitable ρ_t can be checked by performing $\mathcal{O}((T-t_o)^2)$ inner products, to find collinear $\tilde{\mathbf{b}}_i$, followed by Hadamard divisions to check whether (25) is satisfied. Further, a necessary condition for (25) to hold is that $|\rho_t| \leq \frac{2\epsilon}{|y_t+\eta_t^o|}$. Thus, intuitively, the result above states that if the rank minimization problem (18) leads to a polynomial p that is not factorable as a product of s linear varieties, then, for suitable small noise levels, the experimental data could not have been generated by points laying in an arrangement of s subspaces.

Finally, we consider the convergence properties of the proposed algorithm. Assume that persistence of excitation condition holds and that the uncorrupted experimental data $\mathbf{r}_t = [-y_t, y_{t-1}, \ldots, y_{t-n_a}, u_{t-1}, \ldots, u_{t-n_c}]^T$ is generated by s systems that correspond to a transversal subspace arrangement, so that $\dim \mathcal{N}(\mathbf{V}_s(\mathbf{r})) = 1$. Let c_o denote the (monic) vector in $\mathcal{N}(\mathbf{V}_s(\mathbf{r}))$, and, for a given noise level ϵ let $\eta_t^o(\epsilon)$ and $c(\epsilon)$ denote a solution to (18) and a monic vector in $\mathcal{N}(\mathbf{V}_s(\mathbf{r}, \boldsymbol{\eta}^o(\epsilon)))$, respectively. Since the entries of $\mathbf{V}_s(\mathbf{r}, \boldsymbol{\eta}^o(\epsilon))$ are polynomial functions of η_t^o, from a continuity argument, it follows that as the noise level $\epsilon \to 0$, then $\mathbf{c}(\epsilon) \to \mathbf{c}_o$. Hence the subsystems identified from the solution to (18) converge to the actual ones.

4 Handling sensor failures

In this section, we extend the results presented in the earlier section to the case where we allow for instantaneous failures in the measurement sensors at unknown times. These failures lead to corrupted input/output data, that if used in the identification process would result in substantial identification errors. This scenario is motivated by several application domains, including computer vision, where data is transmitted through channels subject to interference or outages. Exploiting the fact that these failures are infrequent, combined with ideas similar to the previous section, allows for recasting the problem into an optimization form where the objective is to simultaneously minimize the rank of a matrix and the number of nonzero rows of a second one.

In particular, we consider switched autoregressive exogenous (SARX) linear models of the form:

$$\begin{aligned} y_t &= \textstyle\sum_{i=1}^{n_a} a_i(\sigma_t) y_{t-i} + \sum_{i=1}^{n_c} c_i(\sigma_t) u_{t-i} + \eta_t \\ \tilde{y}_t &= \begin{cases} y_t \text{ if } f_t = 0 \\ \theta_t \text{ otherwise} \end{cases} \end{aligned} \tag{26}$$

where u, $\tilde{y}$ and η denote the input, output and process noise respectively; f_t is an unknown sparse binary sequence that represents the reliability of measurements (i.e., $f_t = 1$ for time instances when the measurement sensor fails), θ_t denotes the faulty measurements, and $\sigma_t \in \{1, \ldots, s\}$ is the discrete state.

If we assume, n_a and n_c are known exactly and equal for each submodel, and if each of the submodels are sufficiently excited, then it is possible to consider feasibility versions of (18) and (19):

$$\begin{aligned} &\text{find} && \boldsymbol{\eta} \\ &\text{subject to} && \operatorname{rank}[\mathbf{V}_s(\tilde{\mathbf{r}}, \boldsymbol{\eta})] \leq h \\ & && |\eta_t| \leq \epsilon, \ \ \forall t \in [t_0, T], \end{aligned} \tag{27}$$

and

$$\begin{aligned} &\text{find} && \mathbf{m} \\ &\text{subject to} && \operatorname{rank}[\tilde{\mathbf{V}}_s(\tilde{\mathbf{r}}, \mathbf{m})] \leq h \\ & && (5)-(6) \ \ \forall \mathbf{m}^{(t)}, \ \ \forall t \in [t_0, T] \text{ if s is odd} \\ & && (7)-(8) \ \ \forall \mathbf{m}^{(t)}, \ \ \forall t \in [t_0, T] \text{ if s is even} \end{aligned} \tag{28}$$

respectively, where $h =$ (number of rows of $\mathbf{V}$) $- 1$. In the presence of faulty measurements, (27) and (28) are typically no longer feasible, since the corrupted measurements do not satisfy (13), or equivalently, there is no vector $\mathbf{c} \neq 0$ such that $\tilde{\mathbf{V}}_s(\tilde{\mathbf{r}}_t, \mathbf{m}^{(t)})\mathbf{c} = 0$ for all t. The effect of the corrupted rows can be eliminated by introducing an error matrix $\mathbf{E}$ such that $\tilde{\mathbf{V}}_s(\tilde{\mathbf{r}}, \mathbf{m}) + \mathbf{E}$ is rank deficient. Further, under the assumption that faults are infrequent, $\mathbf{E}$ should be row sparse (that is, only a few rows, corresponding to the rows in $\tilde{\mathbf{V}}$ affected by the corrupted measurements, should be non-zero). From the reasoning above, it follows that faulty measurements can be accommodated by considering the following optimization problem:

$$\begin{aligned} &\text{minimize}_{\mathbf{m},\mathbf{E}} && ||\mathbf{E}||_{row,0} \\ &\text{subject to} && \operatorname{rank}[\tilde{\mathbf{V}}_s(\tilde{\mathbf{r}}, \mathbf{m}) + \mathbf{E}] \leq h \\ & && (5)-(6) \ \ \forall \mathbf{m}^{(t)}, \ \ \forall t \in [t_0, T] \ \text{ if s is odd} \\ & && (7)-(8) \ \ \forall \mathbf{m}^{(t)}, \ \ \forall t \in [t_0, T] \ \text{ if s is even.} \end{aligned} \tag{29}$$

Note that if there are k faults, $\mathbf{E}$ would have at most $k(n_a+1)$ nonzero rows, since a single faulty measurement affects n_a+1 rows of $\tilde{\mathbf{V}}$. Therefore, minimizing the nonzero rows of $\mathbf{E}$ amounts to minimizing the number of faulty measurements.

Theorem 4 *Let the optimum of* (29) *be* e *and let* $\{r_i\}_{i=1}^e$ *be the indices of nonzero rows of* $\mathbf{E}^*$ *in the optimal solution. If* $\mathbf{c}$ *is a vector in the null space of* $\tilde{\mathbf{V}}_s(\tilde{\mathbf{r}}, \mathbf{m}^*) + \mathbf{E}^*$*, then, for* $t \in [t_0, T] \backslash \{r_i\}_{i=1}^e$*, there exists a noise sequence* $\boldsymbol{\eta}^*$ *with* $\|\boldsymbol{\eta}^*\|_\infty \leq \epsilon$ *such that* $\mathbf{c}$ *belongs to the nullspace of* $\mathbf{V}_s(\tilde{\mathbf{r}}, \boldsymbol{\eta}^*)$.

Proof: Since $\mathbf{c}$ is in the nullspace of $\mathbf{N} \doteq \tilde{\mathbf{V}}_s(\tilde{\mathbf{r}}, \mathbf{m}^*) + \mathbf{E}^*$, it is also in the nullspace of the submatrix $\mathbf{N}'$ that is formed by eliminating the rows of $\mathbf{N}$ that correspond to $\{r_i\}_{i=1}^e$. Note that only $\tilde{\mathbf{V}}_s(\tilde{\mathbf{r}}, \mathbf{m}^*)$ contributes to $\mathbf{N}'$, since the corresponding rows in $\mathbf{E}^*$ are all zeros. Therefore, $\mathbf{N}'\mathbf{c} = 0$

together with the moment constraints in (29) implies that, for all $t \in [t_0, T] \backslash \{r_i\}_{i=1}^{e}$, there exists measures $\mu^{*(t)}$, each supported on $[-\epsilon, \epsilon]$ and having $\mathbf{m}^{(t)}$ as their moments. Hence $\mathcal{E}_{\mu^{*(t)}} \left[\nu_s(\tilde{\mathbf{r}}_t, \eta_t)^T\right] \mathbf{c} = 0$. Similar to the proof of Theorem 2, we can invoke the mean value theorem for integration to conclude that there exist $\eta_t^* \in [-\epsilon, \epsilon]$ for all $t \in [t_0, T] \backslash \{r_i\}_{i=1}^{e}$ such that $\nu_s(\tilde{\mathbf{r}}_t, \eta_t^*)^T \mathbf{c} = 0$. □

Intuitively, the result above states that, in the presence of outliers, it is possible to modify just a few rows of the Veronese map (precisely those corrupted by the outliers) in such a way that the modified matrix and the ideal, uncorrupted Veronese map have the same null space, for some admissible noise sequence $\boldsymbol{\eta}^*$. It follows that, if there are enough uncorrupted rows to completely characterize this null space, then it can be identified by solving Problem (29). Once this null space has been identified, the coefficients of each submodel can be recovered by proceeding as in [28,14]. Note that this approach is capable of handling faulty measurements of the input sequence as well, since these also result in corrupted rows in the embedded data matrix.

Remark 4 *It is worth emphasizing that in the proposed approach the matrix* $\mathbf{E}$ *enters linearly the constraints in* (29) *and its elements are decoupled from those of the moments sequence* $\mathbf{m}^{(t)}$*. For comparison, introducing a matrix* $\mathbf{E}$ *in* (18) *leads to polynomial constraints involving both the elements of* η_t *and* $\mathbf{E}$ *(originating from expanding the constraint rank*$[\mathbf{V}_s(\tilde{\mathbf{r}}, \boldsymbol{\eta}^*) + \mathbf{E}] \leq h$*), hence necessitating the consideration of the joint moments. Similarly, working directly with the sequence* θ_t*, unknowns* y_t *and* f_t *in* (10)*, requires considering the joint moments of the sequences* $\{\eta_t\}$*,* $\{\theta_t\}$*,* $\{y_t\}$ *and* $\{f_t\}$*, substantially increasing the computational complexity.*

4.1 Convex relaxations for rank and row sparsity

In principle, Problem (29) is generically NP-hard, due to both, the objective function and the rank constraint. However, as we show next, efficient convex relaxations can be obtained by combining recent results on rank minimization and block sparsification. The main idea is to first replace the objective function by

$$\text{rank}[\tilde{\mathbf{V}}_s(\tilde{\mathbf{r}}, \mathbf{m}) + \mathbf{E}] + \gamma \|\mathbf{E}\|_{row,0}$$

where γ is a suitably chosen regularization parameter. Next, note that enforcing row sparsity of $\mathbf{E}$ is equivalent to enforcing sparsity of a vector formed by norms of its rows. Following the arguments in [29,16], we relax $\|\mathbf{E}\|_{row,0}$ to $\sum_i \|\mathbf{E}_i\|_2$, where $\mathbf{E}_i$ denotes the i^{th} row of $\mathbf{E}$, and we relax rank to nuclear norm. Finally, using a semidefinite characterization of the nuclear norm [6], combined with a reweighted heuristic, leads to the following algorithm, based on solving a sequence of convex optimization problems:

Algorithm 2: SARX Identification with outliers

initialize:

$k = 0$, δ = small positive constant,

$W_y^{(0)} = I_{m \times m}$, $W_z^{(0)} = I_{n \times n}$, $w_i^{(0)} = \frac{1}{\sqrt{N_p}}$

REPEAT

Solve

$\min_{\mathbf{m}, Y^{(k)}, Z^{(k)}, \mathbf{E}^{(k)}} \text{Tr}\left[W_y^{(k)} Y^{(k)}\right]$

$+ \text{Tr}\left[W_z^{(k)} Z^{(k)}\right] + \gamma^{(k)} \sum_{i=1}^{N_p} w_i^{(k)} \|\mathbf{E}_i^{(k)}\|_2$

subject to

$$\begin{bmatrix} Y^{(k)} & \tilde{V}_s(\tilde{\mathbf{r}}, \mathbf{m}^{(k)}) + \mathbf{E}^{(k)} \\ [\tilde{V}_s(\tilde{\mathbf{r}}, \mathbf{m}^{(k)}) + \mathbf{E}^{(k)}]^T & Z^{(k)} \end{bmatrix} \geq 0$$

and (5)-(6) $\forall \mathbf{m}$ if s is odd, or (7)-(8) if s is even.

Let $\tilde{V}_s(\tilde{\mathbf{r}}, \mathbf{m}^{(k)}) + \mathbf{E}^{(k)} = UDV^T$.

Set $\epsilon = D(m, m)$.

Set $W_y^{(k+1)} = (Y^{(k)} + \epsilon I)^{-1}$.

Set $W_z^{(k+1)} = (Z^{(k)} + \epsilon I)^{-1}$.

Set $w_i^{(k+1)} = \frac{1}{\|\mathbf{E}_i^{(k)}\|_2 + \delta}$.

Set $\gamma^{(k+1)} = \gamma^{(0)} \frac{\|W_y^{(k)}\|_2 + \|W_z^{(k)}\|_2}{2\left\|[w_1^{(k)} \ \ldots \ w_{N_p}^{(k)}]\right\|_2}$.

Set $k = k + 1$.

UNTIL (a convergence criterion is reached)

RETURN $\tilde{V}_s(\tilde{\mathbf{r}}, \mathbf{m}^{(k)}), \mathbf{E}^{(k)}, \mathbf{m}^{(k)}$

where $\delta > 0$ is a regularization constant, and where the initial value of $\gamma^{(0)}$ is a tuning parameter that controls the number of measurements that are discarded. A small value of $\gamma^{(0)}$ will typically result in a large number of measurements labeled as faulty and dropped. On the other hand, a large value of $\gamma^{(0)}$ will result in fewer discarded measurements and hence may preclude finding a rank deficient solution. Once a suitable value of $\gamma^{(0)}$ is selected, the update rule for $\gamma^{(k)}$ in the algorithm above attempts to keep the relative weights of the rank minimization and row sparsity enforcing terms approximately constant throughout the optimization.

5 Illustrative Examples

In this section we use both numerical and practical examples to illustrate the effectiveness of the proposed methods.

5.1 Numerical Examples

5.1.1 SISO example

Consider a hybrid system that switches among the following three ARX subsystems

$$y_t = 0.2y_{t-1} + 0.24y_{t-2} + 2u_{t-1} + \eta_t \quad \text{(Submodel 1)}$$

$$y_t = -1.4y_{t-1} - 0.53y_{t-2} + u_{t-1} + \eta_t \quad \text{(Submodel 2)}$$

$$y_t = 1.7y_{t-1} - 0.72y_{t-2} + 0.5u_{t-1} + \eta_t \quad \text{(Submodel 3)}$$

modeled as

$$y_t = a_1(\sigma_t)y_{t-1} + a_2(\sigma_t)y_{t-2} + c_1(\sigma_t)u_{t-1} + \eta_t, \quad (30)$$

where $\sigma_t \in \{1, 2, 3\}$ depends on which model is active at time t. Experimental data was obtained by running simulations for $T = 96$ time steps where $\sigma_t = 1$ for $t = [1, 32]$, $\sigma_t = 2$ for $t = [33, 64]$ and $\sigma_t = 3$ for $t = [65, 96]$. We used random noise with $\|\boldsymbol{\eta}\|_\infty = 0.25$ and u_t were independent identically distributed samples from the standard normal distribution. For evaluating the results, the following maximum normalized parameter estimation error is defined:

$$\text{MNPEE} = \max_{i \in \{1,\ldots,s\}} \frac{\left\|\mathbf{b}_i - \hat{\mathbf{b}}_i\right\|_2}{\|\mathbf{b}_i\|_2} \quad (31)$$

where $\mathbf{b}_i$ and $\hat{\mathbf{b}}_i$ are, respectively, the true and the estimated parameters for the i^{th} submodel, and s is the number of submodels. As an additional performance measure, we consider fitting error defined as:

$$\text{FE} = \frac{1}{T - t_0 + 1} \sum_{t=t_0}^{T} \max(|\hat{y}_t - y_t| - \epsilon, 0) \quad (32)$$

where $\hat{y}_t$ is the result of simulating the system with identified dynamics and identified mode sequence. FE basically measures to what extend the a priori assumptions are violated. Ideally for the proposed method, one would expect FE to be zero but due to the relaxation in solving the rank minimization problem, which is not guaranteed to converge to a rank deficient solution, and due to numerical algorithms involved, there could be small deviations. Tables 1-2 summarize the results obtained by the proposed moments–based method together with results of the original algebraic method of [28] and product-of-errors framework of [12] with least squares cost function over 25 simulations per noise level, ϵ, with different realizations of noise and input. Although the method proposed in [12] seems to improve the results from the original algebraic method of [28], the proposed moments-based approach performs clearly superior to both in these experiments. Figure 1 shows, for a sample simulation run, the clustering of data into different submodels as well as the absolute value of noise given the identified model. The estimated noise levels are quite large for the method in [28] whereas they mostly satisfy the prior bound of $\epsilon = 0.25$ for the new method.

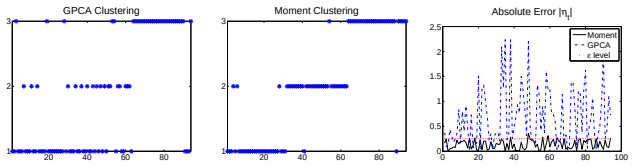

Fig. 1. Sample simulation results. Left: Clustering via GPCA. Middle: Clustering via moments–based method. Right: Comparison of noise estimates.

5.1.2 Example with faulty measurements

In this section, we illustrate the effectiveness of the extension proposed in Sec. 4 in the existence of both noise and faults in measurements with simulation examples. We considered a model of the form (26) that switches between (Submodel 1) and (Submodel 2) from the SISO example. We ran 25 experiments for $T = 100$ time steps where $\sigma_t = 1$ for $t = [1, 25] \cup [51, 75]$ and $\sigma_t = 2$ for $t = [26, 50] \cup [51, 100]$, and u_t were independent identically distributed samples from the standard normal distribution. The noise level was set to $\epsilon = 0.25$. Faults f_t were inserted at times $t = 10$, $t = 50$, and $t = 80$. We set $\gamma^{(0)} = 13$ for all experiments. Note that $\gamma^{(0)}$ can be adaptively optimized to improve the results but even with a single, fixed value, we see that the proposed method performs a lot better compared to methods that do not take faulty measurements into account. The results are summarized in Table 3. As can be seen from the table, if the data includes faulty measurements (i.e., outliers), ignoring them either by using a moments-based method or GPCA results in high errors in estimated parameters. We also ran the same tests with the product-of-errors framework with Hampel's and ϵ-insensitive loss functions as proposed in [12] to handle the case with outliers. Here we just report the results with ϵ-insensitive loss function which performed significantly better than Hampel's. However, still the overall performance of the moments-based method was better than that of [12].

5.2 Computer vision application: Activity analysis with sensor faults

Next, we illustrate the ability of the method to handle realistic scenarios by applying it to a non–trivial computer vision problem: human activity analysis. The goal here is to segment a video clip containing multiple activities into its constituent sub-activities and to find a model characterizing each of these, as a first step towards recognizing contextually abnormal situations. The data used in this particular example consists of 55 frames extracted from a video sequence of a person walking and bending in front of the camera. Three randomly chosen frames were corrupted with large amounts of noise to simulate instantaneous sensor failures (in this case interference in the wireless communication channel from the sensor to the base station). Figure 2 shows some sample frames from the sequence. Half way through the sequence the person bends down, then stands up and resumes walking. These frames were modeled as the

ϵ	MNPEE	Moments Based	GPCA [28]	PE [12]
0.05	mean±std	0.0053 ± 0.0033	0.1487 ± 0.2187	0.2620 ± 0.3827
	median±mad	0.0051 ± 0.0011	0.0406 ± 0.0313	0.0653 ± 0.0581
0.15	mean±std	0.0193 ± 0.0100	0.4532 ± 0.7759	0.2792 ± 0.2858
	median±mad	0.0167 ± 0.0029	0.2931 ± 0.2218	0.1580 ± 0.1173
0.25	mean±std	0.0633 ± 0.1005	4.4196 ± 18.1408	0.3565 ± 0.2902
	median±mad	0.0400 ± 0.0118	0.4103 ± 0.3045	0.2614 ± 0.2120

Table 1
Maximum normalized parameter estimation error (MNPEE) statistics (std: standard deviation, mad: median absolute deviation) over 25 simulations per noise level (ϵ) for the proposed method and the methods of [28] and [12].

ϵ	FE	Moments Based	GPCA [28]	PE [12]
0.05	mean±std	0.0017 ± 0.0066	0.1590 ± 0.2409	0.1798 ± 0.2355
	median±mad	0.0002 ± 0.0001	0.0320 ± 0.0288	0.0811 ± 0.070
0.15	mean±std	0.0056 ± 0.0140	0.3019 ± 0.2267	0.1737 ± 0.1681
	median±mad	0.0013 ± 0.0008	0.3021 ± 0.2229	0.1190 ± 0.1100
0.25	mean±std	0.0235 ± 0.0798	0.3990 ± 0.3463	0.1903 ± 0.1419
	median±mad	0.0048 ± 0.0028	0.3667 ± 0.2838	0.1842 ± 0.1102

Table 2
Fitting error (FE) statistics (std: standard deviation, mad: median absolute deviation) over 25 simulations per noise level (ϵ) for the proposed method and for the methods of [28] and [12].

MNPEE	Moments-based	PE [12]	GPCA [28]	Moments-based (no $\mathbf{E}$)
mean±std	0.0882 ± 0.0722	0.1446 ± 0.0967	0.6672 ± 1.0259	0.5852 ± 0.8605
median±mad	0.0603 ± 0.0254	0.1190 ± 0.0623	0.2485 ± 0.1449	0.2768 ± 0.2032

Table 3
Maximum normalized parameter estimation error (MNPEE) statistics) over 25 simulations for noise level ($\epsilon = 0.25$) and 3 faults for the proposed method with and without using the sparse matrix $\mathbf{E}$ introduced in Sec. 4 and for the algebraic geometric method of [28] and the product-of-errors method in [12] with ϵ-insensitive loss function.

output of an underlying switched affine system, with 2 submodels, each corresponding to a given activity. In particular, the horizontal [6] position of the center of mass was modeled as the output of a first order switched affine autoregressive system [7]:

$$x_t = a(\sigma_t)x_{t-1} + d(\sigma_t) + \eta_t \tag{33}$$

where $a(\sigma_t)$ and $d(\sigma_t)$ are unknown parameters. We set $\|\boldsymbol{\eta}\|_\infty = 3$, allowing ± 3 pixels noise.

For the measurements, we use a simple tracker based on background subtraction to estimate the location of the center of mass of the person in each frame. Sensor failures were captured with the following measurement equation:

$$\tilde{x}_t = \begin{cases} x_t \text{ if } f_t = 0 \\ \theta_t \text{ otherwise} \end{cases} \tag{34}$$

where f_t is an unknown (yet sparse) binary sequence with $f_t = 0$ for healthy conditions and $f_t = 1$ for times where the sensors fail, and where θ_t is random Gaussian noise with mean and standard deviation equal to those of x_t.

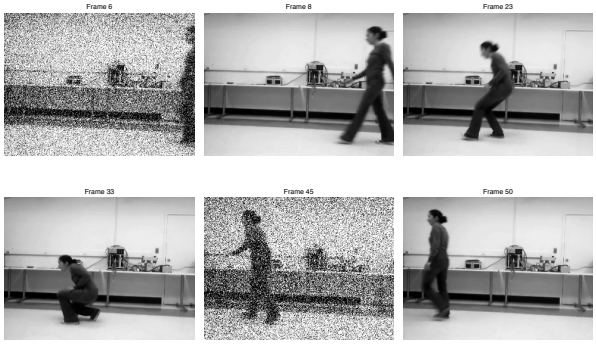

Fig. 2. Sample clean and corrupted frames from the video.

Figures 3 compares the results of applying the proposed identification method and GPCA in this scenario. As illustrated there, while the proposed method is able to identify the underlying subsystems, the switching sequence and the corrupted frames, GPCA fails to do so.

[6] It may seem more natural to use the vertical position. However, this would have resulted in 3 segments, corresponding to roughly no vertical motion, downward and upward motion, while there are only two different activities involved.

[7] Our approach trivially extends to affine systems where there is an offset term (d in this example) in addition to the linear model.

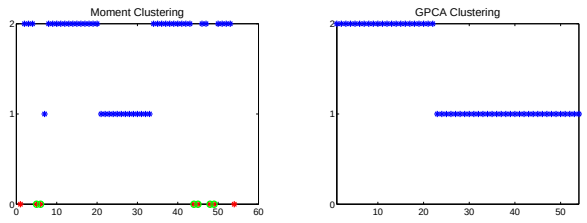

Fig. 3. Left:Activity segmentation via the proposed method. The outlying rows detected are denoted with red stars and the true outliers by green circles. Right: Activity segmentation via GPCA.

Remark 5 *Although the proposed method is computationally more expensive than GPCA (i.e., the former requires solving an SDP whereas the latter requires computing a singular value decomposition), it still leads to a tractable convex problem and performs significantly better in the presence of noise and sensor faults.*

6 Conclusions

This paper considered the problem of identifying switched linear systems. Its main contribution is two-fold: (i) A new formulation for set membership identification of switched linear systems from imperfect input/output measurements. The formulation is shown to be well-posed when certain identifiability and persistence of excitation conditions are satisfied. Moreover, the proposed algorithm is interpolatory, in the sense that the identified system is guaranteed to belong to the consistency set. (ii) A novel optimization procedure to effectively solve this problem. While the problem of identifying switched linear systems is known to be generically NP–hard, we showed that efficient convex relaxations can be obtained by recasting it into a moments optimization form. This procedure combines ideas from classical theory of moments, polynomial optimization, and rank/sparsity relaxations and it makes efficient use of problem structure. The effectiveness of the proposed algorithm and its robustness to noise and outliers was illustrated using extensive simulation examples and a non-trivial segmentation problems arising in computer vision.

Acknowledgements: The authors would like to thank Dr. F. Lauer for providing the code for the products of error framework [12] and the optimization package [9].

References

[1] L. Bako. Identification of switched linear systems via sparse optimization. *Automatica*, 47(4):668 – 677, 2011.

[2] L. Bako and R. Vidal. Algebraic identification of mimo sarx models. In M. Egerstedt and B. Mishra, editors, *Hybrid Systems: Computation and Control*, volume 4981 of *Lecture Notes in Computer Science*, pages 43–57. Springer Berlin / Heidelberg, 2008.

[3] A. Bemporad, A. Garulli, S. Paoletti, and A. Vicino. A bounded-error approach to piecewise affine system identification. *IEEE Trans. on Automatic Control*, 50(10):1567–1580, October 2005.

[4] E. J. Candes, M. Wakin, and S. Boyd. Enhancing sparsity by reweighted l1 minimization. *Journal of Fourier Analysis and Applications*, 14(5):877–905, December 2008.

[5] R.E. Curto and Fialkow L.A. Recursiveness, positivity, and truncated moment problems. *Houston J. Math*, 17:603–635, 1991.

[6] M. Fazel. *Matrix Rank Minimization with Applications*. PhD thesis, Stanford University, 2002.

[7] M. Fazel, H. Hindi, and S. Boyd. Log-det heuristic for matrix rank minimization with applications to hankel and euclidean distance matrices. In *American Control Conference*, June 2003.

[8] A. Garulli, S. Paoletti, and A. Vicino. A survey on switched and piecewise affine system identification. In *Proceedings of the IFAC Symposium on System Identification*, July 2012.

[9] W. Huyer and A. Neumaier. Global optimization by multilevel coordinate search. *Journal of Global Optimization*, 14:331–355, 1999. Software available at `http://www.mat.univie.ac.at/~neum/software/mcs/`.

[10] M. G. Krein and A. A. Nudelman. *The Markov Moment Problem and Extremal Problems*, volume 50 of *Translations of Mathematical Monographs*. American Mathematical Society, Providence, RI, 1977.

[11] J.B. Lasserre. Global optimization with polynomials and the problem of moments. *SIAM J. Optimization*, 11(3):796–817, 2001.

[12] F. Lauer, G. Bloch, and R. Vidal. A continuous optimization framework for hybrid system identification. *Automatica*, 47(3):608 – 613, 2011.

[13] M. Laurent. Sums of squares, moment matrices and optimization over polynomials. In M. Putinar and S. Sullivant, editors, *Emerging Applications of Algebraic Geometry*, pages 157–270. Springer, 2009.

[14] Y. Ma and R. Vidal. A closed form solution to the identification of hybrid arx models via the identification of algebraic varieties. In *Hybrid Systems Computation and Control*, pages 449–465, March 2005.

[15] Y. Ma, A. Yang, Derksen H., and Fossum R. Estimation of subspace arrangements with applications in modeling and segmenting mixed data. *SIAM Review*, 50:413–458, 2008.

[16] M. McCoy and J. Tropp. Two Proposals for Robust PCA using Semidefinite Programming. *Electronic Journal of Statistics*, 5:1123–1160, 2011.

[17] N. Ozay, C. Lagoa, and M. Sznaier. Robust identification of switched affine systems via moments-based convex optimization. In *Proc. 48^{th} IEEE Conf. Dec. Control*, pages 4686–4691, December 2009.

[18] N. Ozay and M. Sznaier. Hybrid system identification with faulty measurements and its application to activity analysis. In *Proc. 50^{th} IEEE Conf. Dec. Control and European Control Conf.*, Dec 2011.

[19] N. Ozay, M. Sznaier, C. Lagoa, and O. Camps. A sparsification approach to set membership identification of a class of affine hybrid systems. In *Proc. 47^{th} IEEE Conf. Dec. Control*, pages 123–130, December 2008.

[20] N. Ozay, M. Sznaier, C.M. Lagoa, and O.I. Camps. A sparsification approach to set membership identification of switched affine systems. *IEEE Trans. on Automatic Control*, 57(3):634 – 648, march 2012.

[21] S. Paoletti, A. Juloski, G. Ferrari-Trecate, and R. Vidal. Identification of hybrid systems: A tutorial. *European Journal of Control*, 13(2):242–260, 2007.

[22] B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Rev.*, 52:471–501, August 2010.

[23] J. Roll, A. Bemporad, and L. Ljung. Identification of piesewise affine systems via mixed-integer programming. *Automatica*, 40:37–50, 2004.

[24] R. Sánchez Peña and M. Sznaier. *Robust Systems Theory and Applications*. Wiley & Sons, Inc., 1998.

[25] J.A. Shohat and J.D. Tamarkin. *The Problem of Moments*. Math. Surveys I. American Mathematical Society, Providence, RI, 1943.

[26] J. F. Traub, G. W. Wasilkowski, and H Woźniakowski. *Information-Based Complexity*. Academic Press, 1988.

[27] R. Vidal, Y. Ma, and S. Sastry. Generalized principal component analysis (GPCA). *PAMI*, 27(12):1945–1959, Dec 2005.

[28] R. Vidal, S. Soatto, Y. Ma, and S. Sastry. An algebraic geometric approach to the identification of linear hybrid systems. In *Proc. 42nd IEEE Conf. Dec. Control*, pages 167–172, December 2003.

[29] H. Xu, C. Caramanis, and S. Sanghavi. Robust pca via outlier pursuit. *IEEE Trans. on Information Theory*, 58(5):3047 –3064, may 2012.

A Recovering the parameters of the model

Here we recall, for ease of reference, the polynomial differentiation based procedure proposed in [28] to recover the parameters of the model once $\mathbf{c_s}$ is computed. The derivative of $p_s(\mathbf{r})$ at a point $\mathbf{r}$ is given by

$$\nabla p_s(\mathbf{r}) = \frac{\delta p_s(\mathbf{r})}{\delta \mathbf{r}} = \sum_{i=1}^{s} \prod_{j \neq i} (\mathbf{b}_j^T \mathbf{r}) \mathbf{b}_i \qquad \text{(A.1)}$$

Since $\mathbf{b}_i^T \mathbf{r} = 0$ when $\mathbf{r}$ is generated by the i^{th} submodel (i.e., $\sigma_t(\mathbf{r}) = i$), it follows from (A.1) that the parameter vector is given by:

$$\mathbf{b}_i = \left. \frac{\nabla p_s(\mathbf{r})}{\mathbf{e}^T \nabla p_s(\mathbf{r})} \right|_{\sigma_t(\mathbf{r})=i} \qquad \text{(A.2)}$$

where $\mathbf{e}^T = [1, 0, \ldots, 0]$.

Since, in general, the association of data points with submodels $\sigma_t(\mathbf{r})$ is unknown, one can use the following heuristic function, suggested in [27], to choose one point from each submodel $\{\mathbf{r}_{t_i}\}_{i=1}^{s}$:

$$\mathbf{r}_{t_{i-1}} = \underset{\mathbf{r}_t : \nabla p_s(\mathbf{r}_t) \neq 0}{\operatorname{argmin}} \frac{\frac{|p_s(\mathbf{r}_t)|}{\|\nabla p_s(\mathbf{r}_t)\|} + \delta}{\left|(\mathbf{b}_i^T \mathbf{r}_t) \cdots (\mathbf{b}_s^T \mathbf{r}_t)\right| + \delta} \qquad \text{(A.3)}$$

where $\delta > 0$ is a small number to avoid division by zero.

Finally, given the parameter vectors $\{\mathbf{b}_i\}_{i=1}^{s}$, the mode signal can be computed as follows:

$$\sigma_t = \underset{i=1,\ldots,s}{\operatorname{argmin}} (\mathbf{b}_i^T \mathbf{r}_t)^2. \qquad \text{(A.4)}$$

B Existence of persistent excitations

Theorem 5 *Assumption 1 guarantees the existence of pairs $(\mathbf{u}, \boldsymbol{\sigma})$ such that the data matrix built from the corresponding trajectories satisfies dim$(\mathcal{N}(\mathbf{V_s})) = 1$*

Outline of the proof: Begin by noting that a minimal ARX system of the form

$$y_t = \sum_{j=1}^{n_a} a_j y_{t-j} + \sum_{j=1}^{n_c} c_j u_{t-j}. \qquad \text{(B.1)}$$

can be put as a series interconnection of the following two (controllable) systems

$$G_1 : v_t = \sum_{j=1}^{n_c} c_j u_{t-j+1} \text{ and } G_2 : y_t = \sum_{j=1}^{n_a} a_j y_{t-j} + v_{t-1}. \qquad \text{(B.2)}$$

Since (B.1) is minimal by assumption, it can be shown that the interconnection of G_1 and G_2 is controllable. Consider now an arbitrary target vector $\mathbf{r} \in R^{n_a+n_c+1}$ satisfying

$$[-1 \; a_1 \; \cdots \; a_{n_a} \; c_1 \; \cdots \; c_{n_c}]\mathbf{r} = 0. \qquad \text{(B.3)}$$

From controllability arguments it follows that, for any initial condition, there exists N and an input $\mathbf{u}$ such that $\mathbf{x}_{t_0+N-1} = \mathbf{r}(n_a + 1 : -1 : 2)$ and $\mathbf{z}_{t_0+N-1} = [\alpha \; \mathbf{r}(n_a + n_c + 1 : -1 : n_a + 3)^T]^T$, where α can be chosen arbitrarily. Setting $u_{t_0+N-1} = \mathbf{r}(n_a + 2)$ leads to a trajectory satisfying

$$[y_{t_0+N} \; \cdots \; y_{t_0+N-n_a} \; u_{t_0+N-1} \; \cdots \; u_{t_0+N-n_c}]^T = \mathbf{r}.$$

To complete the proof, start the system with $\sigma_t = 1$ and select enough points $\mathbf{r}^i$ satisfying $\mathbf{b}_1^T \mathbf{r}^i = 0$, $i = 1, 2, \ldots, N_1$ so that the following property is satisfied: Any homogenous polynomial $q(\mathbf{r})$ of degree less than or equal to s satisfying

$$q(\mathbf{r}_i) = 0 \text{ for all } i = 1, 2, \ldots, N_1$$

is of the form $q(\mathbf{r}) = q_{aux}(\mathbf{r})(\mathbf{b}_1^T \mathbf{r})$ for some polynomial $q_{aux}(\mathbf{r})$ of degree $s-1$. This can always be done if the points provide a sufficient cover of the hyperplane $\{\mathbf{r} : \mathbf{b}_1^T \mathbf{r} = 0\}$. Using the ideas outlined above it is always possible to find an input u_t such that the trajectory of the first subsystem goes through all $\mathbf{r}^i$, $i = 1, 2, \ldots, N_1$. Once the trajectory has passed through all of these points, the system is switched to the next subsystem and the procedure repeated.

Given that the points selected in each hyperplane are a sufficient cover (in the sense explained above) then there is only one homogeneous polynomial $p_s(\mathbf{r})$ of degree less than or equal to s that is zero in points of the obtained trajectory

$$p_s(\mathbf{r}) = \prod_{i=1}^{s} \mathbf{b}_i^T \mathbf{r}.$$

Hence, the data matrix $\mathbf{V}_s$ has a null space with dimension equal to one.