2019 IEEE 58th Conference on Decision and Control (CDC)
Palais des Congrès et des Expositions Nice Acropolis
Nice, France, December 11-13, 2019

Efficient Identification of Error-in-Variables Switched Systems via a Sum-of-Squares Polynomial Based Subspace Clustering Method

B. Ozbay O. Camps M. Sznaier

Abstract— **This paper addresses the problem of identification of error in variables switched linear models from experimental input/output data. This problem is known to be generically NP hard and thus computationally expensive to solve. To address this difficulty, several relaxations have been proposed in the past few years. While solvable in polynomial time these (convex) relaxations tend to scale poorly with the number of points and number/order of the subsystems, effectively limiting their applicability to scenarios with relatively small number of data points. To address this difficulty, in this paper we propose an efficient method that only requires performing (number of subsystems) singular value decompositions of matrices whose size is independent of the number of points. The underlying idea is to obtain a sum-of-squares polynomial approximation of the support of each subsystem one-at-a-time, and use these polynomials to segment the data into sets, each generated by a single subsystem. As shown in the paper, exploiting ideas from Christoffel's functions allows for finding these polynomial approximations simply by performing SVDs. The parameters of each subsystem can then be identified from the segmented data using existing error-in-variables (EIV) techniques.**

I. INTRODUCTION

The problem of identifying switched affine systems from experimental data is ubiquitous in several domain applications such as electronic circuits [1], biological systems [2], computer vision [3], [4] and automated machines [5], [6], to name just a few. In the so-called error in the process model case, the problem has been thoroughly studied in the past few years (see e.g. [5], [7]–[14] and references therein). Recently, it has been shown in [15] that, when the goal is to find a switching model that interpolates the data within a given noise level with the minimum number of switches, then the problem can be solved in polynomial time.

On the other hand, many scenarios require fitting the data with a minimum (or known) number of subsystems. Examples of these situations include not only control applications (fault tolerant control and anomaly detection), but also, among others, computer vision and machine learning (activity recognition, subspace clustering of dynamic data). Unfortunately, in Error in Variables (EIV) cases where the measured data is corrupted by noise, the minimum number of subsystems scenario leads to a very challenging NP hard problem. This difficulty has motivated the search for tractable relaxations, leading to several approaches, which can be roughly divided into optimization-based methods [13], [14], [16]–[18], algebraic methods [19]–[21], and recursive methods [22]–[25]. While these methods have been shown to be effective, most scale poorly with the number of data points. Thus, their application is restricted to moderate number of data points and models with relatively few low order subsystems (typically, these methods scale polynomially or worse with the number of points, number of subsystems and subsystem order). Methods designed explicitly to handle large data sets include [14], [18], [26]. However, [18] needs information about the moments of the noise distribution (something that is not always available) and relies on the solution to a (simple) optimization problem. [14] introduced an efficient branch and bound approach, but the worst case complexity of the method is still exponential in the dimension of the data and number of modes. Finally, [26] relies on a spectral clustering step that can be computationally expensive in situations where the number of clusters is large.

To address these difficulties, motivated by an earlier algebraic approach proposed in [19] for the noiseless case and its stochastic reformulation in [27], in this paper we propose a one-at-a-time algebraic method for efficient identification of switched systems with a given number of subsystems. The advantages of the proposed method are:

1) It can handle error-in-variables (EIV) scenarios
2) It only requires performing (number of subsystems) singular value decompositions of a matrix whose dimension is independent of the number of data points
3) It scales linearly with the number of data points

The basic idea behind this method is to find, for each subsystem, a sum-of-squares polynomial that approximates its support, and use this polynomial to segment the data. Once the data is segmented, the parameters of each subsystem can be obtained using existing EIV methods for LTI systems. Our main result, motivated by [27], [28], shows that these polynomials can be found by performing a singular value decomposition of the empirical moments matrix.

The rest of the paper is organized as follows. Section II presents the notation used in the paper and some background material related to the Generalized Principal Component Analysis (GPCA) method and Christoffel functions. In Section III-A, we formally state the error-in-variables switched auto-regressive exogenous (EIV-SARX) problem. Section III-B, presents the proposed subspace clustering based approach and discusses its connection to algebraic methods and iterative application of [27]. Section IV illustrates the effectiveness of the proposed method compared to existing approaches. Finally, Section V concludes the paper with some remarks and possible directions for future research.

This work was supported in part by NSF grants IIS–1318145 and ECCS–1404163; AFOSR grant FA9550-19-1-0005; and the Alert DHS Center of Excellence under Award Number 2013-ST-061-ED0001.

The authors are with the ECE Department, Northeastern University, Boston, MA 02115. emails: ozbay.b@husky.neu.edu, {msznaier,camps}@coe.neu.edu

978-1-7281-1398-2/19/$31.00 ©2019 IEEE

II. PRELIMINARIES

A. Notation

$(\mathbf{A})_i$	i^{th} row of $\mathbf{A}$.
$\mathbf{A} \succeq 0$	matrix $\mathbf{A}$ is positive semidefinite.
$\mathscr{P}^n_{d,h}$	subspace of n^{th} degree homogeneous multivariate polynomials in d variables.
$s_{n,d} \doteq \binom{n+d-1}{d}$	number of monomials of degree n in d variables.
$\mathbf{v}_n(\mathbf{x})$	Veronese map of degree n: $\mathbf{v}_n \begin{pmatrix} x_1 & \dots & x_d \end{pmatrix} \doteq \begin{bmatrix} x_1^n & x_1^{n-1}x_2 & \dots & x_d^n \end{bmatrix}^T$
$\mathscr{E}_\mu(x)$	Expected value of x with respect to the probability density function μ.
$\mathbf{H}_x^{r,c}$	Hankel matrix with c columns and r rows associated with a vector sequence $\mathbf{x}$, with elements $(\mathbf{H}_x)_{i,j} = \mathbf{x}_{i+j-1}$

B. Moment Matrices and Sum-of-Squares (SoS) Polynomials

Given a probability measure μ supported on $\mathbb{R}^d$, its associated moments sequence is given by

$$m_\alpha = \mathscr{E}_\mu(\mathbf{x}^\alpha) = \int_{\mathbb{R}^d} \mathbf{x}^\alpha d\mu \tag{1}$$

where $\mathbf{x} \doteq \begin{bmatrix} x_1 & x_2 & \dots & x_d \end{bmatrix}^T$, α is shorthand for the multi-index $\begin{bmatrix} \alpha_1 & \alpha_2 & \dots & \alpha_d \end{bmatrix}$ and $\mathbf{x}^\alpha \doteq x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_d^{\alpha_d}$. In the sequel we will denote by $\mathbf{M}(\mathbf{m})$ the moments matrix with elements $\mathbf{M}(\alpha,\beta) = m_{\alpha+\beta}$ (where we have assumed that the sequence $\mathbf{m}$ has been ordered in a graded reverse lexicographic order) and by $\mathbf{M}_n$ its truncated version containing moments of order up to $2n$. For instance, in the case of 2 variables, x_1 and x_2, there are $\binom{4}{2} = 6$ monomials of degree up to $2n = 2$ and the corresponding moment matrix is

$$\mathbf{M}_1 = \begin{bmatrix} 1 & m_{(1,0)} & m_{(0,1)} \\ m_{(1,0)} & m_{(2,0)} & m_{(1,1)} \\ m_{(0,1)} & m_{(1,1)} & m_{(0,2)} \end{bmatrix} \tag{2}$$

A homogeneous polynomial $P(\mathbf{x}) = \sum_\alpha p_\alpha \mathbf{x}^\alpha \in \mathscr{P}^n_{d,h}$ can be written in terms of the coefficients vector $\mathbf{c}^T$ as $P(\mathbf{x}) = \mathbf{c}^T \mathbf{v}_n(\mathbf{x})$. Using this notation, the expected value (w.r.t μ) of the product of two polynomials P_1, P_2 is given by

$$\langle P_1(.), P_2(.) \rangle_\mu = \int_{\mathbb{R}^d} \mathbf{c}_1^T \mathbf{v}_n(\mathbf{x}) \mathbf{v}_n^T(\mathbf{x}) \mathbf{c}_2 d\mu = \mathbf{c}_1^T \mathbf{L}_n \mathbf{c}_2 \tag{3}$$

where we have defined

$$\mathbf{L}_n = \int_{\mathbb{R}^d} \mathbf{v}_n(\mathbf{x}) \mathbf{v}_n^T(\mathbf{x}) d\mu \tag{4}$$

(note that $\mathbf{L}_n$ is the bottom right submatrix of $\mathbf{M}_n$, composed of moments of order $2n$). Since $\mathbf{L}_n \succeq 0$, it follows that (3) defines an inner product (induced by the measure μ) in $\mathscr{P}^n_{d,h}$. In the sequel, with a slight abuse of notation, we will refer sometimes to $\mathbf{L}_n$ as the truncated moment matrix.

C. Christoffel Polynomials

In this section we recall some results showing the relationship between a sum of squares polynomial that approximates the support set of a given distribution μ and the empirical moment matrix formed using points sampled from this distribution. Given a set of samples $\mathbf{x}_i$, $i = 1, \dots, N$ drawn from a distribution μ let $\mathbf{L}_n \doteq \frac{1}{N} \sum \mathbf{v}_n(\mathbf{x}_i) \mathbf{v}_n^T(\mathbf{x}_i)$ denote the (truncated) empirical moment matrix of order n, with corresponding singular vectors and values $\mathbf{u}_i, \sigma_i$. From (3) it can be easily shown that the polynomials having coefficient vectors $\mathbf{c}_i \doteq \frac{1}{\sqrt{\sigma_i}} \mathbf{u}_i$. form an orthonormal basis, with respect to μ, of $\mathscr{P}^n_{d,h}$. Further, [29] these orthonormal polynomials define a reproducing Kernel:

$$K_n(\mathbf{x},\mathbf{y}) \doteq \sum_{i=1}^{s_{n,d}} (\mathbf{c}_i^T \mathbf{v}_n(\mathbf{x}))(\mathbf{c}_i^T \mathbf{v}_n(\mathbf{y})) \tag{5}$$

Define now the SoS polynomial

$$Q_n(\mathbf{x}) \doteq K_n(\mathbf{x},\mathbf{x}) = \sum_{i=1}^{s_{n,d}} (\mathbf{c}_i^T \mathbf{v}_n(\mathbf{x}))^2 = \mathbf{v}_n^T(\mathbf{x}) \mathbf{L}_n^{-1} \mathbf{v}_n(\mathbf{x}) \tag{6}$$

The function $Q_n^{-1}(\mathbf{x})$is known as the Christoffel function. The following result, taken from [28] and [29] relates this function to the probability measure μ[1]

$$Q_n^{-1}(\xi) = \min_{P \in \mathscr{P}^n_{d,h}} \int_{\mathbb{R}^p} P^2(\mathbf{x}) d\mu \text{ s. t. } P(\xi) = 1 \tag{7}$$

Further, it can be shown [27] that an explicit expression for the coefficients of the minimizing polynomial is given by:

$$\mathbf{c}^*(\xi) = \frac{1}{\sum_{i=1}^{s_{n,d}} (\frac{1}{\sqrt{\sigma_i}} \mathbf{u}_i^T \mathbf{v}_n(\xi))^2} \sum_{i=1}^{s_{n,d}} \frac{1}{\sigma_i} \mathbf{u}_i^T \mathbf{v}_n(\xi) \mathbf{u}_i \tag{8}$$

Note in passing that finding $\mathbf{c}^*$ involves a single SVD of a matrix $\mathbf{L}_n$ whose size is independent of the number of data points. In the sequel, we will refer to the polynomial $P(\mathbf{x},\xi) = \mathbf{v}^T(\mathbf{x}) \mathbf{c}^*(\xi)$ as the "Christoffel support polynomial at ξ", since it provides (locally) an approximation to the support of μ. To see this, note that if $\mu(\xi) \approx 0$, the polynomial P can be selected so that $P(\xi) \approx 0$ at points where μ is not small, leading to a small objective in (7). On the other hand, if $\mu(\xi) \not\approx 0$, since by continuity, any polynomial satisfying $P(\xi) = 1$ will have values close to 1 in a neighborhood of ξ, the corresponding values of Q_n^{-1} will be large (e.g. $Q_n(\xi)$ is small). As noted in [28], this observation, combined with the fact that $\mathscr{E}_\mu(Q_n) = \binom{n+d-1}{d}$ and Markov's inequality, allows for using $Q_n(\mathbf{x})$ to approximate the overall level sets of the distribution, in the sense that points where $Q_n(\mathbf{x}) \gg \binom{n+d-1}{d}$ have low probability of belonging to the distribution. These results will play a key role in segmenting the given data into subspaces, one subspace at a time.

D. Subspace Arrangements

In this section we briefly recall some definitions connecting the problem of subspace clustering with the properties of certain polynomials. These results form the basis of algebraic methods such as GPCA and its variants [30]–[32].

[1] Note that [28] considers polynomials of degree up to $2n$ and the associated moment matrix $\mathbf{M}_n$ while here we consider only homogeneous polynomials and the submatrix $\mathbf{L}_n$. This choice reflects the fact that in this paper we are interested in distributions supported on subspace arrangements, rather than general ones.

Definition 2.1: The arrangement $\mathscr{A}(S)$ of a set of subspaces $S=\{S_i\}_{i=1}^n \subseteq R^n$ is defined as:

$$\mathscr{A}(S) \doteq S_1 \cup S_2 \cup \ldots \cup S_n \tag{9}$$

Definition 2.2: The vanishing ideal $I(\mathscr{A})$ of a subspace arrangement $\mathscr{A} \subseteq R^d$ is the set of all multivariate polynomials in d variables that vanish on all points in $\mathscr{A}$, that is:

$$I(\mathscr{A}) \doteq \left\{ P \in \mathscr{P}^d \colon P(\mathbf{z}) = 0 \, \forall \mathbf{z} \in \mathscr{A} \right\} \tag{10}$$

The subset $I_n(\mathscr{A}) \subseteq I(\mathscr{A})$ formed by homogeneous polynomials of degree n is known as the homogeneous component of degree n of $I(\mathscr{A})$.

Definition 2.3: Given a set I of polynomials, $\mathscr{Z}(I)$, the zero set of I is the set of all common roots, that is

$$\mathscr{Z}(I) \doteq \{\mathbf{x} \in \mathbb{R}^d \colon P(\mathbf{x}) = 0 \text{ for all } P \in I\} \tag{11}$$

The following result shows that the arrangement $\mathscr{A}$ is completely characterized by its associated homogeneous ideal:

Lemma 2.4 (Lemma 2.8, [30]): The subspace arrangement $\mathscr{A}$ is the zero set of $I_n(\mathscr{A})$, e.g. $\mathscr{A} = \mathscr{Z}\left[I_n(\mathscr{A})\right]$

These results form the basis of algebraic based approaches, which, rather than directly estimating the parameters of S_i, seek to estimate first I_n from the data. Under mild conditions, I_n has dimension 1 (e.g. it is a principal ideal) and the parameters of each subspace can then be estimated from its generator, for instance via polynomial differentiation.

III. Christoffel Function Based Identification of EIV SARX systems

Next we state the problem of interest and propose a solution that scales linearly with the number of data points.

A. Problem Statement

Consider an error-in-variables switched auto-regressive exogenous (SARX) linear model

$$\hat{y}_t = \sum_{i=1}^{n_a} a_i(\gamma_t)\hat{y}_{t-i} + \sum_{i=0}^{n_b} b_i(\gamma_t)\hat{u}_{t-i} \tag{12}$$

$$y_t = \hat{y}_t + \zeta_t \quad , \quad u_t = \hat{u}_t + \eta_t$$

where $\hat{y}_t$ and $\hat{u}_t$ denote the actual output/input signals respectively corrupted by additive bounded noise $\|\zeta_t\|_\infty \leq \varepsilon_\zeta$ and $\|\eta_t\|_\infty \leq \varepsilon_\eta$ where ε_ζ, ε_η are bounds on the output and input noise respectively and γ_t is the mode variable indicating which subsystem is active at time t. Our goal is to identify the parameters $\{a_{k=1}^{n_a}(j), b_{k=1}^{n_b}(j)\}$ that characterize each of the subsystems in (12) from the input/output experimental data $(u_k,\, y_k)$ and the a-priori information $\{s, n_a, n_b\}$, where s is number of the subsystems, n_a and n_b are the input and output order of the subsystems.

B. Algebraic Reformulation With a Stochastic Perspective

Consider a trajectory of (12) corresponding to a given input and switching sequences, and, for ease of notation, define

$$\begin{aligned} \mathbf{r}_t &= [-y_t, y_{t-1}, \ldots, y_{t-n_a}, u_{t-1}, \ldots, u_{t-n_b}]^T \\ \mathbf{b}(\gamma_t) &= [1, a_1(\gamma_t), \ldots, a_{n_a}(\gamma_t), b_1(\gamma_t), \ldots, b_{n_b}(\gamma_t)]^T \end{aligned} \tag{13}$$

Note that $\mathbf{b}^T(\gamma_t)\mathbf{r}_t = 0$ holds for all time instants where $\gamma_t = \gamma$. Thus, the corresponding regressors $\mathbf{r}_{t,\gamma}$ live in a subspace normal to $\mathbf{b}(\gamma)$. We will use this fact to recast the systems identification problem into a (noisy) subspace clustering form. The proposed method is based on the observation that, in the noise free case, the vanishing ideal of the arrangement of subspaces defined by the collection of vectors $\mathbf{b}_i, i = 1, \ldots, s$ is generated by the polynomial [19]

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

where $\mathbf{b}_i \in R^{n_a+n_b+1}$ is the vector corresponding to parameters of the i^{th} subsystem, $\mathbf{v}_s(.)$ denotes the Veronese map of degree s, and where the entries of the vector $\mathbf{c}_s$ are only functions of the entries of the vectors $\mathbf{b}_i$. Evaluating this polynomial at each data point and collecting the results in a matrix leads to

$$\mathbf{V}_s\mathbf{c}_s \doteq [\mathbf{v}_s(\mathbf{r}_{t_0}) \quad \cdots \quad \mathbf{v}_s(\mathbf{r}_T)]^T \; \mathbf{c}_s = 0 \tag{15}$$

In the noise free case, the identification problem can be solved by using the GPCA algorithm proposed in [19], [32], simply by finding a vector $\mathbf{c}_s$ in the null space of $\mathbf{V}_s{}^2$ and then recovering the parameters of each subsystem via polynomial differentiation. Unfortunately, as shown for instance in [13], this approach is fragile, and even small amount of noise can lead to large errors in the identified parameters. Next, we indicate how to circumvent this difficulty by exploiting the properties of the Christoffel functions.

Consider an arrangement of subspaces $\mathscr{A}(S) \doteq S_1 \cup S_2 \cup \ldots \cup S_s$, $S_i \subset \mathbb{R}^d$, where the normal to each subspace is $\mathbf{b}_i$ and let μ denote a probability measure supported in this arrangement. Given a point $\mathbf{x}_o \notin \mathscr{A}(S)$, define the following polynomial optimization problem

$$P^*_{\mathbf{x}_o}(\mathbf{x}) = \left\{ \underset{P \in \mathscr{P}^s_{d,h}}{argmin} \int_\mu P^2(\xi) d\mu \text{ subject to } P(\mathbf{x}_o) = 1 \right\} \tag{16}$$

that is, $P^*_{\mathbf{x}_o}(.)$ is the minimum variance (w.r.t μ) homogeneous polynomial of degree s, whose value at $\mathbf{x}_o$ is fixed. Note that this is precisely the Christoffel support polynomial at $\mathbf{x}_o$ defined in Section II-C. If the support of μ is the arrangement $\mathscr{A}(S)$, then it is easy to see that a solution to the problem above is given by

$$P^*_{\mathbf{x}_o}(\mathbf{x}) = \frac{\prod_{i=1}^{s}(\mathbf{b}_i^T \mathbf{x})}{\prod_{i=1}^{s}(\mathbf{b}_i^T \mathbf{x}_o)} \tag{17}$$

Further, if the arrangement is transversal [30], then this solution is unique, since there exists only one homogeneous polynomial of degree s that vanishes on $\mathscr{A}$ and has a fixed value at a given point $\mathbf{x}_o \notin \mathscr{A}$. In the systems identification scenario of interest to this paper, the measure μ is unknown. Rather, the experimental data consists of points $\mathbf{x}_i, i = 1, \ldots, N_p$ drawn from these subspaces and corrupted by noise. In this case, an approximation to the problem above can be obtained by replacing the probability measure μ by

[2] Under mild conditions this vector is unique, since the vanishing ideal of the arrangement is a principal ideal [30].

the empirical distribution μ_{emp}, and an approximation to (16) using the empirical distribution is given by:

$$c*_{\mathbf{x}_o} = \underset{\mathbf{c}}{argmin} \frac{1}{N_p}\mathbf{c}^T\mathbf{L}_{s,emp}\mathbf{c} \quad \text{subject to } \mathbf{c}^T\mathbf{v}_s(\mathbf{x}_o) = 1 \tag{18}$$

where

$$\mathbf{L}_{s,emp} \doteq \frac{1}{N_p}\sum_{i=1}^{N_p}\mathbf{v}_s(\mathbf{x}_i)\mathbf{v}_s^T(\mathbf{x}_i) = \frac{1}{N_p}\mathbf{V}_s^T\mathbf{V}_s \tag{19}$$

and $\mathbf{V}_s$ is a matrix having as its k^{th} row $(\mathbf{V}_s)_k = \mathbf{v}_s^T(\mathbf{x}_k)$. The solution to the problem above is given by (8), using the singular vectors/values of $\mathbf{L}_{s,emp}$ instead of $\mathbf{L}_s$.

Remark 1: Intuitively, the Christoffel support polynomial $P_{\mathbf{x}_o}(\mathbf{x}) = \mathbf{c}^{*\,T}_{\mathbf{x}_o}\mathbf{v}_s(\mathbf{x})$ provides an approximation to the support set of $\mathscr{A}$ in the sense that it is expected to be close to zero for points in the subspace arrangement, and, by continuity, close to one for points outside the arrangement and close to $\mathbf{x}_o^*$. We will exploit this property in the next section.

C. Proposed Algorithm

The proposed algorithm is based on iteratively using the Christoffel support polynomial with coefficients given by (8) using $\mathbf{L}_{s,emp}$ instead of $\mathbf{L}_s$ to find points generated by a single subsystem. A high level description is given in Algorithm 1.

Step 1: Finding reliable data for the first s-1 subspaces

As indicated above, we propose to find points generated by each subsystem proceeding in a "one-at-a-time" fashion. Let $\mathscr{X}_{a,k}$ denote the set of available points at the k^{th} stage of the algorithm (that is points that have not yet been assigned to a given subsystem). The goal here is to find a set of points $\mathscr{X}_k \subseteq \mathscr{X}_{a,k}$ that, with high probability, were generated by the same subsystem, or equivalently, belong to the same subspace. Motivated by the discussion above, we propose to accomplish this by selecting a point $\mathbf{x}_{o,k} \in \mathscr{X}_{a,k}$, computing the corresponding Christoffel support polynomial $P_{\mathbf{x}_{o,k}}(\mathbf{x})$ of degree $s-k$ (since at this stage only $s-k+1$ subspaces remain for which data has not been assigned) and defining

$$\mathscr{X}_k \doteq \{\mathbf{x} \in \mathscr{X}_{a,k} : P^2_{\mathbf{x}_{o,k}}(\mathbf{x}) \geq t\} \tag{20}$$

where t is a given threshold. Implementation of this idea requires addressing the issues of how to select the point $\mathbf{x}_{o,k}$ and the threshold t.

Selecting $\mathbf{x}_{o,k}$: At the k^{th} stage of the algorithm, the remaining points belong to an arrangement $\mathscr{A}_k$ consisting of $s-k+1$ subspaces. For a given point $\mathbf{x}_o$, consider the Christoffel support polynomial $P_{\mathbf{x}_{o,n_k}}(.)$ of degree $n_k = s-k$. From (7) and Markov's inequality (chapter 8 in [33]) it follows that, for any point $\mathbf{x}$ drawn from an arrangement formed by n_k of the remaining subspaces

$$\mathbf{Prob}_\mu\left[P^2_{\mathbf{x}_{o,n_k}}(\mathbf{x}) > t\right] \leq \frac{1}{tQ_{n_k}(\mathbf{x}_o)} \tag{21}$$

The inequality above suggests selecting $\mathbf{x}_o^* = argmin_{\mathbf{x}\in\mathscr{X}_{a,k}}Q_{n_k}(\mathbf{x}_o)$, since this choice maximizes the number of points where $P^2_{\mathbf{x}^*_{o,n_k}}(.)$ is large. By construction degree$[P_{\mathbf{x}^*_o,n_k}(.)] = n_k$, and thus $P_{\mathbf{x}^*_{o,n_k}}(.)$ can't be identically zero in $\mathscr{A}_k - \{\mathbf{x}_o^*\}$. Further, since $P_{\mathbf{x}^*_o,n_k}(\mathbf{x}_o^*) = 1$, by continuity $P_{\mathbf{x}^*_o,n_k}(.)$ will be close to 1 on points close to $\mathbf{x}_o^*$. Thus, intuitively, most of the points where $P_{\mathbf{x}^*_o,n_k}(.)$ is large belong to the same subspace as $\mathbf{x}_o^*$. Once $\mathbf{x}_o^*$ is selected, t can be chosen from (21) to obtain a suitable upper bound $\frac{1}{tQ_{n_k}(\mathbf{x}_o^*)}$ on the probability of miss-classification. However, extensive experimental results show that this bound is typically conservative and better results are obtained selecting t using Otsu's algorithm [34].

Algorithm 1 Subspace identification using SOS polynomial

```
Inputs: $\mathbf{X_y} \leftarrow \mathbf{H}^{n_a+1,N-n_a-1}_{\mathbf{y}_t}$ Hankelized output data, $\mathbf{X_u} \leftarrow \mathbf{H}^{n_b,N-n_a-1}_{\mathbf{u}_t}$ Hankelized input data where $N$ is the horizon length, $s \leftarrow$ number of subsystems, $k \leftarrow 1$
$\mathscr{X}_a \leftarrow \begin{bmatrix}\mathbf{X_y}\\ \mathbf{X_u}\end{bmatrix}$
for $k := 1$ to $s$ do
    if $k < s-1$ then ▷ Find a group of points outside of the union of $s-k$ of the remaining $s-k+1$ subspaces
        Step 1: Finding reliable data for the first $s-1$ subspaces
        Selecting $\mathbf{x}_{o,k}$:Find the best "anchor" point $\mathbf{x}_o^*$ using the SoS Polynomial $Q_{k-1}(\mathbf{x})$ representing a union of $s-k$ of the $s-k+1$ available clusters
        Compute $P_{\mathbf{x}^*_o,k}(\mathbf{x})$, the Christoffel support polynomial, for the union of $s-k$ clusters treating the point $\mathbf{x}_o^*$ as an outlier.
        Compute an optimal threshold $t$ and assign points where $P^2_{\mathbf{x}_{o,k}}(\mathbf{x}) \geq t$ to the set $\mathscr{X}_k$. This set approximates a subset of $S_{o,k}$ the subspace that contains $\mathbf{x}^*_{o,k}$.
    else if $k = s-1$ then                ▷ Last subspace case
        Step 2:Handling the last subspace
        Generate a set containing reliable data drawn from the first $s-1$ subspaces: $\mathscr{X}_{rel} \doteq \cup_{i=1}^{s-1}\mathscr{X}_{i,rel}$
        Compute and threshold $| P^2_{s-1,\mathscr{X}_{rel}}(\mathbf{x}) |$ for $\mathbf{x} \in \mathscr{X}_a$ to find a set $\mathscr{X}_s \subset S_s$, the last subspace
    end if
    Finding a reliable set $\mathscr{X}_{k,rel} \subset \mathscr{X}_k$ via outlier rejection on $\mathscr{X}_k$
    Update available data: $\mathscr{X}_a \leftarrow \mathscr{X}_a \setminus \mathscr{X}_{k,rel}$
end for
Step 3: Labeling the entire data set ▷ using the SoS polynomial computed with the sets $\mathscr{X}_{j,rel}$
for $j := 1$ to $s$ do
    Compute $Q_{1,\mathscr{X}_{j,rel}}(\mathbf{x})$ for each subspace
end for
Assign each point $\mathbf{x}$ to the cluster $j$ which gives the smallest normalized $(Q_{1,\mathscr{X}_{j,rel}}(\mathbf{x})/norm(Q_{1,\mathscr{X}_{j,rel}(\mathbf{x})})$
```

Step 2: Handling the last subspace

After the $(s-1)^{th}$ iteration of the algorithm, since each iteration is not guaranteed to remove all points in a given subspace, the set $\mathscr{X}_{a,s-1}$ can potentially include points from each subspace in the arrangement. As we show next, points in the last subspace S_s can also be extracted from $\mathscr{X}_{a,s-1}$

by finding a suitable Christoffel support polynomial. To this effect, consider the set $\mathscr{X}_{rel} \doteq \cup_{i=1}^{s-1} \mathscr{X}_{i,rel}$ and the associated moments submatrix $\mathbf{L}_{rel} \doteq \frac{1}{|\mathscr{X}_{rel}|} \sum_{\mathbf{x}_i \in \mathscr{X}_{rel}} \mathbf{v}(\mathbf{x}_i)\mathbf{v}^T(\mathbf{x}_i)$. As discussed in section II-C, the Christoffel function $Q(\mathbf{x}) \doteq \mathbf{v}^T(\mathbf{x})\mathbf{L}_{rel}^{-1}\mathbf{v}(\mathbf{x})$ should be (with high probability) smaller for points that belong to the arrangement $\cup_{i=1}^{s-1} S_i$ than for those outside it. Thus, if the switching sequence visited all subsystems (and hence the experimental data comes from s subspaces), it follows (from Markov's inequality) that the point

$$\mathbf{x}_s^* = \underset{\mathbf{x} \in \mathscr{X}_{a,s-1}}{argmax} [\mathbf{v}^T(\mathbf{x})\mathbf{L}_{rel}^{-1}\mathbf{v}(\mathbf{x})]$$

is, with high probability, an outlier to the set $\cup_{i=1}^{s-1} S_i$, or, equivalently, an inlier to the last subspace S_s. Thus, a set of points in this subspace can be obtained by computing $P_{\mathbf{x}_s^*, s-1}$, the Christoffel support polynomial of $\mathscr{X}_{rel}$ with respect to $\mathbf{x}_s^*$ and setting

$$\mathscr{X}_s \doteq \{\mathbf{x} \in \mathscr{X}_{a,s-1} : P_{\mathbf{x}_s^*, s-1} \geq t\} \tag{22}$$

Finding a reliable set $\mathscr{X}_{k,rel}$: At the end of each iteration, after the subset $\mathscr{X}_k$ of candidate points in the k^{th} subspace has been found, a "reliable" subset of these, $\mathscr{X}_{k,rel} \subset \mathscr{X}_k$ can be found using the outlier rejection method proposed in [27]. These points are then removed from the pool of available data, e.g. $\mathscr{X}_{a,k} \setminus \mathscr{X}_k \to \mathscr{X}_{a,k+1}$, $k \to k+1$ and the process is repeated until $k = s$.

Step 3: Labeling the entire data set

The sets $\mathscr{X}_{i,rel}$ can be used as "seeds" to reliably assign points to subsystems as follows. Under the assumption that the switching sequence and inputs are sufficiently rich to generate enough data in each set $\mathscr{X}_{i,rel}$, the corresponding moment matrices $\mathbf{L}_{i,rel}$ provide a good characterization of the support of each subspace. That is, given a point $\mathbf{x}_i$ the Christoffel function $\mathbf{v}^T(\mathbf{x}_i)\mathbf{L}_{j,rel}^{-1}\mathbf{v}(\mathbf{x}_i)$ is (with high probability) small if $\mathbf{x}_i \in S_j$ and large otherwise. Thus, points can be assigned to subspaces by simply selecting the one corresponding to the smallest Q, that is, computing

$$j^* = \underset{j \in [1,s]}{argmin} [\mathbf{v}^T(\mathbf{x}_i)\mathbf{L}_{j,rel}^{-1}\mathbf{v}(\mathbf{x}_i)]$$

and assigning $\mathbf{x}_i$ to S_{j^*}.

Step 4: Finding the parameters of each subsystem

Once the data has been segmented into subspaces S_i, the identification process is completed by computing the parameters of each subsystem. In principle, in relatively low noise scenarios, the data matrix $\mathbf{X}_i$ whose columns are the coordinates of the points in S_i should be close to rank $n_a + n_b$. Hence the parameters $\mathbf{b}_i$ are given by the null vector of $\mathbf{X}_i\mathbf{X}_i^T$, normalized so that $b_i(1) = 1$.

It is worth emphasizing that the proposed method requires performing singular value decompositions of the moment matrices $\mathbf{L}_i$ whose size depends only on the dimension of the data (system order) and number of subsystems, together with point evaluations of the polynomial $P^2_{\mathbf{x}_{o,k}}(\mathbf{x})$ for all data points. Hence computational time grows only linearly with the number of data points, but combinatorially with the number of subsystems and their order.

IV. Illustrative Example

In this section we illustrate the advantages of the proposed approach using the following system, used in [10]:

$$\begin{aligned} \hat{y}_t &= a_1(\gamma_t)\hat{y}_{t-1} + a_2(\gamma_t)\hat{y}_{t-2} + b_1(\gamma_t)\hat{u}_{t-1} \\ y_t &= \hat{y}_t + \zeta_t \quad , \quad u_t = \hat{u}_t + \eta_t \end{aligned} \tag{23}$$

and switches between

$$\begin{aligned} \hat{y}_t &= 0.2\hat{y}_{t-1} + 0.24\hat{y}_{t-2} + 2\hat{u}_{t-1} (\text{Subsystem } 1) \\ \hat{y}_t &= -1.4\hat{y}_{t-1} + -0.53\hat{y}_{t-2} + 1\hat{u}_{t-1} (\text{Subsystem } 2) \end{aligned} \tag{24}$$

We considered a scenario with 7 switches, occurring at time instances $t = \frac{k \times N}{8}$ where $k = 1, 2 \ldots, 7$ such that $\gamma_t = 1$ *for odd* k and $\gamma_t = 2$ *for even* k. Data was generated from 20 random runs for various combinations of noise levels $\varepsilon \in \{0.05, 0.15, 0.25\}$ and horizon length $N \in [96, 192, 396]$. The results of applying several different approaches to this data (manifold embedding [26], algebraic methods [18], [19]) are summarized in Table I. There, the column PE shows the parameter estimation error defined as,

$$\text{PE} = \max \|\rho - \hat{\rho}\|_2 \tag{25}$$

where ρ and $\hat{\rho}$ are the normalized true and estimated parameters of the subsystems. As shown in the table, the proposed method is at least one order of magnitude faster than competing methods (except GPCA) while achieving a comparable (or better) identification error, and the effect gets more pronounced as the number of points increases. Regarding GPCA, it runs an order of magnitude faster than the proposed SoS based method, but as expected, the identification error becomes far worse (an order of magnitude) as the noise level increases.

V. CONCLUSIONS

The problem of EIV identification of switched ARX models (and the related problem of subspace clustering) arises in many domains ranging from control to machine learning and computer vision. For instance, in the context of control, solving this problem is a prerequisite to design controllers for switched systems in cases where models of the system are a-priori unavailable. Unfortunately, this problem is known to be NP-hard, which has prompted a large research effort seeking to develop computationally tractable relaxations. As a result, a number of techniques are currently available that have been shown to work well in practice (and, under certain conditions exact). Nevertheless, virtually all of these techniques scale polynomially with the number of data points and their dimension[3], which limits their applicability to moderately large data sets (typically no more than a few thousands). To circumvent this difficulty, in this paper we proposed a method that scales linearly with the number of data points and only requires performing order of s singular value decompositions of matrices whose size is independent of the number of data points. This is accomplished by recasting the problem as that of finding an

[3] An exception is [26], that scales linearly with the number of points but polynomially with the number of switches.

TABLE I

NUMERICAL EXAMPLES: COMPARISON OF PROPOSED METHOD WITH OTHER METHODS

N	ε	GPCA [19]		Hojjatinia et. al. [18] - unknown parameters		JBLD-Based [26]		SOS-Based (Proposed)	
		PE	time (secs)	PE	time (secs)	PE	time (secs)	PE	time (secs)
96	0.05	0.0539	0.0027	0.0571	6.5201	0.0099	0.0384	0.0527	0.0503
	0.15	0.0603	0.0005	0.3053	6.4554	0.0496	0.0306	0.0628	0.0056
	0.25	0.1764	0.0005	0.1224	6.3717	0.2090	0.0393	0.0731	0.0056
192	0.05	0.0577	0.0005	0.0247	6.3217	0.0061	0.0995	0.0079	0.0051
	0.15	0.1576	0.0006	0.1018	6.2882	0.0262	0.1572	0.0182	0.0064
	0.25	0.2342	0.0006	0.1398	6.3510	0.0692	0.1273	0.0334	0.0054
396	0.05	0.0109	0.0006	0.0106	6.3165	0.0270	0.4213	0.0935	0.0057
	0.15	0.1626	0.0005	0.0786	6.3753	0.0850	0.4129	0.0123	0.0056
	0.25	0.1274	0.0005	0.1022	6.5183	0.1124	0.4337	0.0223	0.0053

SoS polynomial that approximates the support of each subspace, one at a time. As shown in the paper this polynomial can be directly constructed from the aforementioned SVDs. The effectiveness of the proposed approach was illustrated with an example used in the literature to compared SARX identification methods, where it was shown to outperform existing methods, specially as the number of data points increases. Note that as stated, the proposed method scales badly with the order and number of subsystems, since the size of the corresponding matrix depends combinatorially on s and n_a. On going research seeks to address this problem by using tools from randomized linear algebra.

REFERENCES

[1] G. Ferrari-Trecate, M. Muselli, D. Liberati, and M. Morari. A clustering technique for the identification of piecewise affine systems. *Automatica*, 39(2):205–217, 2003.

[2] M. Musters, D. Lindenaar, A. Juloski, and N. Van Riel. Hybrid identification of nonlinear biochemical processes. *IFAC Proceedings Volumes*, 39(1):350–355, 2006.

[3] R. Vidal, S. Soatto, and A. Chiuso. Applications of hybrid system identification in computer vision. In *2007 European Control Conference (ECC)*, pages 4853–4860.

[4] R. Vidal and Y. Ma. A unified algebraic approach to 2-d and 3-d motion segmentation and estimation. *Journal of Mathematical Imaging and Vision*, 25(3):403–421, 2006.

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

[6] A Lj Juloski, WPMH Heemels, and G Ferrari-Trecate. Data-based hybrid modelling of the component placement process in pick-and-place machines. *Control Engineering Practice*, 12(10):1241–1252, 2004.

[7] J. Roll, A. Bemporad, and L. Ljung. Identification of piecewise affine systems via mixed-integer programming. *Automatica*, 40(1):37–50, 2004.

[8] N. Ozay, C. Lagoa, and M. Sznaier. Robust identification of switched affine systems via moments-based convex optimization. In *Proc. of the 48h IEEE Conf. on Dec. and Ctrl. (CDC)*, pages 4686–4691, 2009.

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

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

[11] D. Piga and R. Tóth. An sdp approach for 0-minimization: Application to arx model segmentation. *Automatica*, 49(12):3646–3653, 2013.

[12] T. P. Dinh, H. M. Le, H. A. Le Thi, and F. Lauer. A difference of convex functions algorithm for switched linear regression. *IEEE Trans. Aut. Ctrl.*, 59(8):2277–2282, 2014.

[13] N. Ozay, C. Lagoa, and M. Sznaier. Set membership identification of switched linear systems with known number of subsystems. *Automatica*, 51:180–191, 2015.

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

[15] N. Ozay. An exact and efficient algorithm for segmentation of arx models. In *2016 American Control Conference (ACC)*, pages 38–41.

[16] C. Feng, C. Lagoa, and M. Sznaier. Hybrid system identification via sparse polynomial optimization. In *Proceedings of the 2010 ACC*, pages 160–165.

[17] Y. Cheng, Y. Wang, and M. Sznaier. A convex optimization approach to semi-supervised identification of switched arx systems. In *53rd IEEE Conference on Decision and Control*, pages 2573–2578, 2014.

[18] S. Hojjatinia, C. M Lagoa, and F. Dabbene. Identification of switched arx systems from large noisy data sets. *arXiv preprint arXiv:1804.07411*, 2018.

[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 *42nd Conf. on Dec. and Control*, volume 1, pages 167–172, 2003.

[20] S. Nazari, Q. Zhao, and B. Huang. An improved algebraic geometric solution to the identification of switched arx models with noise. In *Proceedings of the 2011 ACC*, pages 1230–1235.

[21] S. Nazari, B. Rashidi, Q. Zhao, and B. Huang. An iterative algebraic geometric approach for identification of switched arx models with noise. *Asian J. of Control*, 18(5):1655–1667, 2016.

[22] L. Bako, K. Boukharouba, E. Duviella, and S. Lecoeuche. A recursive identification algorithm for switched linear/affine models. *Nonlinear Analysis: Hybrid Systems*, 5(2):242–253, 2011.

[23] A. Goudjil, M. Pouliquen, E. Pigeon, and O. Gehan. A real-time identification algorithm for switched linear systems with bounded noise. In *2016 European Control Conf. (ECC)*, pages 2626–2631.

[24] R. Vidal. Recursive identification of switched arx systems. *Automatica*, 44(9):2274–2287, 2008.

[25] O. Ayad, M. Sayed-Mouchweh, and booktitle=Int. Conf. on Fuzzy Syst. pages=1–7 year=2010 Billaudel, P. Switched hybrid dynamic systems identification based on pattern recognition approach.

[26] X. Zhang, M. Sznaier, and O. Camps. Efficient identification of error-in variables switched systems based on riemannian distance-like functions. In *2018 IEEE Conf. on Dec. and Control*, pages 3006–3011.

[27] M. Sznaier and O. Camps. Sos-rsc: A sum-of-squares polynomial approach to robustifying subspace clustering algorithms. In *IEEE CVPR*, pages 8033–8041, 2018.

[28] Edouard Pauwels and Jean B Lasserre. Sorting out typicality with the inverse moment matrix sos polynomial. In *Advances in Neural Information Processing Systems*, pages 190–198, 2016.

[29] Y. Xu. On orthogonal polynomials in several variables. *Special functions, q-series and related topics, The Fields Institute for Research in Mathematical Sciences, Communications Series*, 14:247–270, 1997.

[30] Y. Ma, A. Y Yang, H. Derksen, and R. Fossum. Estimation of subspace arrangements with applications in modeling and segmenting mixed data. *SIAM review*, 50(3):413–458, 2008.

[31] N. Ozay, M. Sznaier, C. Lagoa, and O. Camps. Gpca with denoising: A moments-based convex approach. In *2010 CVPR*, pages 3209–3216.

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

[33] R. Tempo, G. Calafiore, and F. Dabbene. *Randomized algorithms for analysis and control of uncertain systems: with applications*. Springer Science & Business Media, 2012.

[34] N. Otsu. A threshold selection method from gray-level histograms. *IEEE Trans. on systems, man, and cybernetics*, 9(1):62–66, 1979.