2018 IEEE Conference on Decision and Control (CDC)
Miami Beach, FL, USA, Dec. 17-19, 2018

Efficient Identification of Error-inVariables Switched Systems Based on Riemannian Distance-Like Functions

X. Zhang M. Sznaier O. Camps
Electrical and Comp. Engineering Department,
Northeastern University,
Boston, MA 02115.

Abstract— **This paper considers the problem of identifying error in variables switched affine models from experimental input/output data. Since this problem is generically NP hard, several relaxations have been proposed in the past. While these relaxations work well for low dimensional systems with few subsystems, they scale poorly with both the number of subsystems and their memory. As an alternative, in this paper we present a computationally efficient method based on embedding the data in the manifold of positive semidefinite matrices, and using a manifold metric to detect switches and identify subsystems. The main result of the paper shows that, under dwell-time assumptions, the method is guaranteed to identify the system, for suitably low noise scenarios. In the case of larger noise levels, consistent numerical experience shows that the proposed method outperforms existing ones. These results are illustrated with a non-trivial practical example: action segmentation.**

I. Introduction

Switched systems are pervasive across applications domains ranging from systems biology to manufacturing. Thus, a large research effort has been devoted in the past decade to develop controllers for such systems. Successful application of this control design framework to practical problems requires the ability to identify switched models from experimental data. Identification of switched systems in the case of error in the process models has been extensively studied in the past decades [29], [11]. The related methods can be generally divided into three categories: optimization based methods, algebraic methods, and clustering based methods. In optimization based methods, [32] and [3] recast the problem into a combinatorial optimization. In the case of bounded noise, [17] proposed to use a branch and bound approach to efficiently solve the optimization problem. Additional approaches include sparse optimization [28], [1], polynomial optimization [26], [27], [31], particle-swarm [19] and difference of convex functions programming [8]. Very recently, it was shown in [25] that indeed, this problem can be solved in polynomial time, provided that the goal is to find a hybrid system that explains the observed data with the minimum number of switches, rather than the minimum number of subsystems. This method is very efficient for ℓ^2 bounded noise but performance degrades substantially in the case of ℓ^∞ bounds.

This work was supported in part by NSF grants CNS–1646121, CMMI–1638234, IIS–1814631 and ECCS–1808381; AFOSR grant FA9550-15-1-0392; and the Alert DHS Center of Excellence under Award Number 2013-ST-061-ED0001. email {zhangxk,msznaier,camps}@ece.neu.edu

Regarding algebraic methods, to the best of our knowledge, [36] was the first to apply these techniques to switched systems identification. While the approach works well for noiseless data, performance degrades quickly with the noise level. [23], [24], [22] addressed this issue by using methods such as total least squares to denoise the data. Finally, clustering based approaches (see for instance [10], [2], [16], [12], [14], [34], [15] and references therein) borrow tools from pattern recognition and machine learning.

On the other hand, far fewer methods exist to handle the case of error-in-variables models, where each of the measurements is assumed to be corrupted by noise. In this case, the problem is known to be NP hard and most existing methods are based upon convex relaxations of the original non-convex problem. In particular, [9], [5], [6] proposed a polynomial optimization based approach, which in turn is relaxed to a sequence of convex semi-definite programs. While in principle this sequence of approximations is guaranteed to converge to the actual system, in practice, computational complexity prevents considering higher elements of the sequence, necessitating adding rank constraints to the algorithm. Thus, while empirically shown to work, theoretical convergence guarantees are lost. Further, even when considering low order relaxations, computational complexity scales combinatorially both with the number of subsystems and their order.

To avoid this difficulty, in this paper we propose a computationally efficient alternative, based upon the idea of embedding the data in the manifold of positive definite matrices and using a manifold distance-like function to compare the dynamics underlying short segments of the data record. In this context, switches in the system are characterized by sharp increases in this distance. Further, segments where the same subsystem is active can be identified by finding clusters where this distance is small, a problem that can be efficiently solved by recasting it into a graph cut form. Our theoretical results show that, under minimum dwell time assumptions, in the case of suitably low noise, this approach is guaranteed to identify the minimum number of subsystems that explains the observed data. In scenarios with higher noise levels these theoretical guarantees are lost, but numerical experience shows that the proposed method consistently outperforms existing approaches in terms of computational burden and identification error. These results are illustrated with a non-trivial practical example: activity segmentation from time traces of the position of a person's

978-1-5386-1395-5/18/$31.00 ©2018 IEEE

centroid. While the proposed technique is less general than those in [9], [5], [6], since it assumes dwell time constraints, it is able to exploit these constraints to substantially lower the computational burden, as opposed to polynomial optimization based approaches. Thus, it substantially outperforms these approaches in many practical scenarios (e.g. biological systems) where dwell time constraints arise naturally.

The paper is organized as follows: Section II contains the notation, some required background on Riemannian metrics, and formally states the problem under consideration. Section III presents the proposed solution, along with the supporting theory. Section IV illustrates these results with a practical example. Finally, Section V presents some concluding remarks and points out to some open questions.

II. Preliminaries

In this section we introduce the notation used in the paper and recall, for ease of reference, some key results.

A. Notation

$\mathbb{R}$	set of real numbers
$\mathcal{S}^n$	set of symmetric matrices in $\mathbb{R}^{n\times n}$
$\mathcal{S}^n_+(\mathcal{S}^n_{++})$	cone of positive-semidefinite (-definite) matrices in $\mathcal{S}^n$
$\mathbf{M}^T$	transpose of matrix $\mathbf{M}$
$\lvert\mathbf{M}\rvert$	determinant of $\mathbf{M}$
$\mathcal{N}(\mathbf{M})$	null space of $\mathbf{M}$
$\lVert\mathbf{M}\rVert_*$	nuclear norm of $\mathbf{M}$
$\lVert\mathbf{M}\rVert_F$	Frobenius norm of $\mathbf{M}$
$\sigma_{\mathrm{r}}(\mathbf{M})$	r^{th} singular value of $\mathbf{M}$
$\mathbf{H}^{r,q}_{\mathbf{x}}$	Hankel matrix associated with a vector sequence $\mathbf{x}_{1:n}$ where $n = r+q-1$ and

$$\mathbf{H}^{r,q}_{\mathbf{x}} \doteq \begin{bmatrix} x_0 & x_1 & \cdots & x_{q-1} \\ x_1 & x_2 & \cdots & x_q \\ \vdots & \vdots & \ddots & \vdots \\ x_{r-1} & x_r & \cdots & x_n \end{bmatrix}$$

$\mathbf{G}$	Gram matrix associated with a given Hankel matrix: $\mathbf{G} = \mathbf{H}\mathbf{H}^T$
$\mathbf{I}_n$	Identity matrix, $\mathbf{I}_n \in \mathbb{R}^{n\times n}$
$J_{ld}(\mathbf{X},\mathbf{Y})$	Jensen-Bregman LogDet Divergence:

$$J_{ld}(\mathbf{X},\mathbf{Y}) = \log\left|\frac{\mathbf{X}+\mathbf{Y}}{2}\right| - \frac{1}{2}\log|\mathbf{X}\mathbf{Y}|,\ \mathbf{X},\mathbf{Y}\in\mathcal{S}^n_{++}$$

B. Background results on the Jensen-Bregman LogDet divergence

In this paper we will embed data in the manifold of positive definite matrices and use distances there to detect switches and compare systems. To this effect, we need a metric that takes into account the non-flat geometry of $\mathcal{S}^n_{++}$. The intrinsic metric in $\mathcal{S}^n_{++}$ is the so-called Affine Invariant Riemannian Metric (AIRM) [30], [4], defined by

$$J_R(\mathbf{X},\mathbf{Y}) \doteq \|\log\left(\mathbf{X}^{-\frac{1}{2}}\mathbf{Y}\mathbf{X}^{-\frac{1}{2}}\right)\|_F$$

While this metric is induced by the geodesic length along the manifold curvature, its main disadvantage is its high computational cost. A recently introduced, computationally effective surrogate is the Jensen-Bregman LogDet Divergence (JBLD) [7]:

$$J_{ld}(\mathbf{X},\mathbf{Y}) \triangleq \log\left|\frac{\mathbf{X}+\mathbf{Y}}{2}\right| - \frac{1}{2}\log|\mathbf{X}\mathbf{Y}| \tag{1}$$

As shown in [13] the JBLD is *geometry aware* and is closely related to AIRM (Theorem 1 in [13]).

Note that (1) is only well defined for matrices in $\mathcal{S}^n_{++}$. However, this paper requires comparing positive semidefinite matrices, for which the JBLD is not defined. The following result extending the JBLD to $\mathcal{S}^n_+$, provides the theoretical justification for the proposed identification method.

Theorem 1 ([38]): Given $\mathbf{X},\mathbf{Y}\in\mathcal{S}^n_+$, define the regularized matrices $\mathbf{X}(\sigma) = \mathbf{X}+\sigma\mathbf{I},\ \mathbf{Y}(\sigma) = \mathbf{Y}+\sigma\mathbf{I}$, where $\sigma > 0$. Then

$$\lim_{\sigma\to 0} J_{ld}(\mathbf{X}(\sigma),\mathbf{Y}(\sigma)) \neq \infty \iff \mathcal{N}(\mathbf{X}) = \mathcal{N}(\mathbf{Y}) \tag{2}$$

C. Problem Statement

In this paper we are interested in identifying Error-in-Variables Switched ARX models (EIV-SARX) from experimental input/output data and some minimal a-priori information on the system to be identified and the statistics of the noise. Specifically, we are interested in solving:

Problem 1: Given experimental input/output data $\{(u_t,y_t)_{t=t_0}^T\}$ and bounds σ_η,σ_ν on the covariance of the noise, find a set of coefficients $\{a_{k=1}^{n_a}(i), b_{k=1}^{n_a}(i)\}$ so that the EIV-SARX model

$$\begin{aligned} \hat{y}_t &= \sum_{k=1}^{n_a} a_k(s_t)\hat{y}_{t-k} + \sum_{k=1}^{n_a} b_k(s_t)\hat{u}_{t-k} \\ y_t &= \hat{y}_t + \eta_t,\ u_t = \hat{u}_t + \nu_t \end{aligned} \tag{3}$$

explains the observed data. Here $\hat{y},\hat{u}$ represent the actual input/output variables, corrupted by additive noise η_k,ν_k, y_t,u_t denote their measured values, and s_t is the mode variable indicating which subsystem is active at time t.

Assumption A1: Minimality To avoid ambiguities, we will assume that each LTI subsystem has a minimal representation of the form above. Equivalently, if the input/output pairs (y_{t_i},u_{t_i}) are explained by an n_a^{th} order ARX model, there does not exist another ARX model with $n_a' < n_a$ that explains the same subset of data.

Remark 1: Note that the problem above is ill posed, even under Assumption 1, since the experimental data can always be explained by fitting a model to each data point. Thus, to regularize the problem we will seek to explain the data with the minimum possible number of subsystems. This scenario arises often in practical applications such as fault tolerant control, fault detection and isolation, image segmentation, and activity recognition.

III. EIV-SARX IDENTIFICATION VIA JBLD BASED SPECTRAL CLUSTERING

In this section we present the proposed method, summarized in Algorithm 1, and provide its theoretical justification. The main idea is to partition the experimental data into subsets, each of which are known to have been generated by an LTI system of order at most n_a, and then compute a manifold distance between the dynamics underlying each segment. Once these distances are obtained, a graph cut step is performed to assign segments to clusters, each corresponding to a single subsystem. Finally, the parameters of the subsystems are recovered by simply performing a LTI systems identification step on each cluster. A high level outline of this algorithm is provided in Algorithm 1.

Algorithm 1 Riemannian distance-like function based switched system identification

Inputs: input sequence $\mathbf{u}_{1:n}$, output sequence $\mathbf{y}_{1:n}$, sliding window size h, Hankel row size $r = 2n_a + 1$.

Step 1: Data Segmentation. Use (8) to partition the input and output sequences into segments of length h, each generated by a single LTI system. The ith segment of input and output are denoted $\mathbf{u}_{c_i:c_i+h-1}$ and $\mathbf{y}_{c_i:c_i+h-1}$, respectively.

for i = 1 to # of segments **do**

$\mathbf{H}_{\mathbf{u},i} \leftarrow$ Hankelize $\mathbf{u}_{c_i:c_i+h-2}$

$\mathbf{H}_{\mathbf{y},i} \leftarrow$ Hankelize $\mathbf{y}_{c_i:c_i+h-1}$

$\mathbf{G}_i \leftarrow \begin{bmatrix}\mathbf{H}_{\mathbf{y},i}\\ \mathbf{H}_{\mathbf{u},i}\end{bmatrix}\begin{bmatrix}\mathbf{H}_{\mathbf{y},i}^T & \mathbf{H}_{\mathbf{u},i}^T\end{bmatrix}$

$\hat{\mathbf{G}}_i(\sigma) \leftarrow \frac{\mathbf{G}_i}{\|\mathbf{G}_i\|_*} + \sigma\mathbf{I}_r$

end for

Step 2: Spectral Clustering.

Compute $\mathbf{D}$, where $\mathbf{D}_{ij} = J_{ld}(\hat{\mathbf{G}}_i(\sigma), \hat{\mathbf{G}}_j(\sigma))$

Compute the similarity matrix $\mathbf{W}$, where $\mathbf{W}_{ij} = e^{-\frac{D_{ij}^2}{2}}$

Cluster labels $\mathbf{z} \leftarrow$ normalized cuts on $\mathbf{W}$

Step 3: Subsystem Identification.

for i = 1 to # of clusters **do**

Perform a LTI SysId step on cluster z_i

end for

Outputs: $\mathbf{a}_i, \mathbf{b}_i$

The next result provides the theoretical underpinnings of the proposed algorithm:

Theorem 2: Under the following assumptions:

A.2 Noiseless data[1].

A.3 Dwell time: once the system (3) switches to a given subsystem, it remains there for $T \geq 4n_a + 1$ sampling instants.

A.4 Persistence of the excitation. For any window j of length $h \geq 3n_a + 1$ where the i^{th} system is active, then $\sigma_r(\hat{\mathbf{G}}_j^i) \geq \underline{\sigma}_i > 0$, where, for a rank r matrix, $\sigma_r(.)$ denotes the smallest non-zero singular value, and $\hat{\mathbf{G}}_j^i \doteq \frac{\mathbf{G}_j^i}{\|\mathbf{G}_j^i\|_*}$.

[1]The more general case with noisy data is considered in Theorem 3.

Then, there exists some $\sigma > 0$ such that Algorithm 1 exactly recovers the underlying system.

Establishing this result, requires analyzing first the properties of the data segmentation and clustering steps.

A. Step 1: Switch Detection and Data Segmentation

The goal of this step is to segment the data into portions where only one subsystem is active. To this effect, we will build two sequences $\{T_i^-\}$ and $\{T_i^+\}$, such that actual switch instants T_i satisfy $T_i^- \leq T_i \leq T_i^+$. Hence, the data y_k at time instants $k \in [T_i^+, T_{i+1}^- - 1]$ is known to have been generated by a single subsystem. The algorithm (outlined in (8)), uses a sliding window of size $h \geq 3n_a + 1$ to detect switches by comparing the corresponding Gram matrices. Specifically, given a time instant k and a regularization parameter σ, let $\hat{\mathbf{G}}_k(\sigma)$ denote the (normalized, regularized) Gram matrix built from the data in the segment $[k, k + h - 1]$, where $h \geq 3n_a + 1$ is the chosen window length:

$$\hat{\mathbf{G}}_k(\sigma) = \frac{\mathbf{G}_k}{\|\mathbf{G}_k\|_*} + \sigma\mathbf{I}_r \tag{4}$$

where $r \doteq 2n_a + 1$,

$$\mathbf{G}_k = \begin{bmatrix}\mathbf{H}_{\mathbf{y},k}\\ \mathbf{H}_{\mathbf{u},k}\end{bmatrix}\begin{bmatrix}\mathbf{H}_{\mathbf{y},k}\\ \mathbf{H}_{\mathbf{u},k}\end{bmatrix}^T \tag{5}$$

and

$$\mathbf{H}_{\mathbf{y},k} = \begin{bmatrix} y_k & y_{k+1} & \cdots & y_{k-n_a+h-1} \\ y_{k+1} & y_{k+2} & \cdots & y_{k-n_a+h} \\ \vdots & \vdots & \ddots & \vdots \\ y_{k+n_a} & y_{k+n_a+1} & \cdots & y_{k+h-1} \end{bmatrix} \tag{6}$$

$$\mathbf{H}_{\mathbf{u},k} = \begin{bmatrix} u_k & \cdots & u_{k-n_a+h-1} \\ \vdots & \ddots & \vdots \\ u_{k+n_a-1} & \cdots & u_{k+h-2} \end{bmatrix} \tag{7}$$

Given an increasing sequence $\{i\}$, define j_i by:

$$\begin{aligned} j_i^+ &= \underset{j \geq i+1}{argmin} \left\{ j : J_{ld}(\hat{\mathbf{G}}_i(\sigma), \hat{\mathbf{G}}_j(\sigma)) \geq \tau \right\} \\ T_i^+ &= j_i^+ + h - 1 \\ j_i^- &= \underset{j \leq T_i^+}{argmin} \left\{ j : J_{ld}(\hat{\mathbf{G}}_j(\sigma), \hat{\mathbf{G}}_{T_i^+}(\sigma)) < \tau \right\} \\ T_i^- &= j_i^- + n_a \end{aligned} \tag{8}$$

where τ denotes a suitable threshold. As we show next, in the case of noiseless data, the sequences T_i^-, T_i^+ bracket the actual switching sequence. Before establishing a formal proof of this result, below we illustrate the intuition behind it with a simple example. Consider the following two models:

$$y_{k+1} = 0.5y_k + 0.5u_k \qquad \text{(system1)}$$

$$y_{k+1} = u_k \qquad \text{(system2)}$$

and the following input/output sequences:
u(1:8)={1,4,0,1,2,3,0,4}, y(1:9)={-1,0,2,1,1,2,3,0,4}.

The corresponding Hankel matrices, with $h = 3n_a + 1 = 4$ are

$$\mathbf{H}_1 = \begin{bmatrix} -1 & 0 & 2 \\ 0 & 2 & 1 \\ 1 & 4 & 0 \end{bmatrix}, \mathbf{H}_2 = \begin{bmatrix} 0 & 2 & 1 \\ 2 & 1 & 1 \\ 4 & 0 & 1 \end{bmatrix}, \mathbf{H}_3 = \begin{bmatrix} 2 & 1 & 1 \\ 1 & 1 & 2 \\ 0 & 1 & 2 \end{bmatrix}$$

$$\mathbf{H}_4 = \begin{bmatrix} 1 & 1 & 2 \\ 1 & 2 & 3 \\ 1 & 2 & 3 \end{bmatrix}, \mathbf{H}_5 = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 3 & 0 \\ 2 & 3 & 0 \end{bmatrix}, \mathbf{H}_6 = \begin{bmatrix} 2 & 3 & 0 \\ 3 & 0 & 4 \\ 3 & 0 & 4 \end{bmatrix}$$

Here the points 1-4 were generated with the model $y_{k+1} = 0.5y_k + 0.5u_k$ and the points 6-9 were generated by the model $y_{k+1} = u_k$. Note that the point y_5 satisfies both models, and hence it could be assigned to either class. In terms of the corresponding Gram matrices, it is easily seen that the procedure above yields $j_1^+ = 3$, $T_1^+ = 6$, $j_1^- = 4$, $T_1^- = 5$. This correctly indicates that the earliest possible switch happened at $T = 5$, and the latest possible one at $T = 6$. Thus, the points in the intervals $[1, 4]$ and $[6, 9]$ each belong to a single class.

Lemma 1: Under Assumptions A.2–A.4, there exists some $\sigma > 0$ small enough such that the sequences generated using (8) satisfy $T_i^- \leq T_i \leq T_i^+$, where T_i denotes the actual switching instants.

Proof: Omitted for space reasons, can be obtained by contacting the authors. ■

B. Step 2: Spectral Clustering

Once $\mathcal{I} \doteq \cup_k [T_k^+, T_{k+1}^- - 1]$, the set of all intervals known to contain data generated by a single subsystem, has been found proceeding as outlined in the previous section, the next step is to identify the minimum number of subsystems of order $n_r \leq n_a$ that could have generated the observed data. We propose to solve this problem by recasting it into a spectral clustering form. Specifically, consider all possible sliding windows w_i of length h contained in some interval $[T_k^+, T_{k+1}^- - 1]$ and a graph where each node represents one such window, and where the edge connecting two nodes has associated a weight $\mathbf{W}_{ij}$ given by:

$$\mathbf{W}_{ij} = \begin{cases} 1 \text{ if } w_i, w_j \subseteq [T_k^+, T_{k+1}^- - 1] \text{ for some } k \\ \exp(-\frac{\mathbf{D}_{ij}^2}{2}) \text{ otherwise} \end{cases} \quad (9)$$

where $\mathbf{D}_{ij}$ is the manifold distance between the Gram matrices corresponding to the windows (w_i, w_j). Note that $\mathbf{W}$ measures the similarity between the underlying systems: if the data in w_i and w_j was generated by different subsystems, as $\sigma \to 0$, D_{ij} becomes large and hence $\mathbf{W}_{ij} \to 0$. Thus, intuitively, the minimum number of subsystems required to explain the data can be obtained by performing a graph cut that seeks to minimize the total cost of cut (that is the sum of the distances on the edges that cross the cut).

While there are several algorithms available for performing graph cuts, in this paper we will use the normalized cuts algorithm proposed in [35], since it has been proven to be effective and robust. Given a undirected graph with n nodes, the degree of a node is defined as the sum of the weights of all edges that are connected to it. Thus, we can define a degree matrix $\mathbf{\Lambda}^{n,n} = \text{diag}\{\lambda_1, \lambda_2, \cdots, \lambda_n\}$, where λ_i is the degree of the ith node. The normalized graph Laplacian matrix is then defined as:

$$\mathbf{L} = \mathbf{I}_n - \mathbf{\Lambda}^{-\frac{1}{2}} \mathbf{W} \mathbf{\Lambda}^{-\frac{1}{2}} \quad (10)$$

where $\mathbf{W}$ is the similarity matrix. As shown in [21], [20], $\mathbf{L} \in \mathcal{S}_+^n$, and its smallest eigenvalue is always zero. Moreover, the multiplicity of the eigenvalue 0 equals the number of connected components in the graph, and the corresponding eigenspace is spanned by the indicator vectors of those components (see for instance [37]).

Lemma 2: Given n_w windows of length h, each containing noiseless data generated by a single subsystem, the spectral clustering step described above generates the minimum number of subsystems that interpolate the observed data.

Proof: Omitted for space reasons, can be obtained by contacting the authors. ■

Proof of Theorem 2: Follows from combining Lemmas 1 and 2. ■

C. Handling Noisy Data

Consider now the case where the measured data is corrupted by noise. In this case, the matrices $\mathbf{G}$ are generically full rank and the result above no longer holds. In principle, such a case can be handled by "denoising" the data, that is, looking for matrices $\mathbf{E}_i \in \mathcal{S}_+^r$, $\|\mathbf{E}_i\| \leq \sigma_\eta$ such that $\text{rank}(\mathbf{G}_i + \mathbf{E}_i) \leq 2n_a$. While it can be shown that for σ_η small enough, Algorithm 1 applied to the denoised data is still guaranteed to recover the correct system, the entailed computational complexity is not trivial. Rank minimization problems are generically NP-hard and the standard convex relaxation where a re-weighted nuclear norm is used as a surrogate for rank requires solving a sequence of semidefinite programs whose complexity scales at least as n_a^6.

As an alternative, as we show next, under suitable assumptions on the noise, Theorem 2 still holds.

Theorem 3: Assume that the noise is white, with zero mean and covariance σ_{noise} and uncorrelated with the input/output sequences $\{u, y\}$. Then, if σ_{noise} is small enough, Algorithm 1 recovers the correct data segmentation.

Proof: Omitted for space reasons, can be obtained by contacting the authors ■

Note that the results above only guarantee perfect recovery for small enough noise level. Nevertheless, consistent numerical experience shows that the JBLD metric separates well data segments corresponding to different subsystems, even for moderately large noise levels.

D. Step 3: Subsystem Identification

After Step 2, each cluster should contain data segments generated from a single subsystem. Hence, at this point any LTI system identification technique that handles EIV scenarios can be used to recover the parameters that characterize each subsystem. In this paper we will use a Structured Total Least Norm (STLN) based approach [33], since consistent numerical experience shows that it is substantially faster than competing methods (e.g. matlab's *ssest* command), with

comparable errors. Briefly, in this context the goal is to find a parameter vector $\boldsymbol{\theta}$ that satisfies:

$$\begin{bmatrix} \mathbf{Y}_{i1} - \mathbf{E}_{i1} & \mathbf{U}_{i1} - \mathbf{F}_{i1} \\ \mathbf{Y}_{i2} - \mathbf{E}_{i2} & \mathbf{U}_{i2} - \mathbf{F}_{i2} \\ \vdots & \vdots \\ \mathbf{Y}_{im} - \mathbf{E}_{im} & \mathbf{U}_{im} - \mathbf{F}_{im} \end{bmatrix} \boldsymbol{\theta}_i = \begin{bmatrix} \mathbf{b}_{i1} - \mathbf{f}_{i1} \\ \mathbf{b}_{i2} - \mathbf{f}_{i2} \\ \vdots \\ \mathbf{b}_{im} - \mathbf{f}_{im} \end{bmatrix} \tag{11}$$

where

$$\mathbf{Y}_{ij} = \begin{bmatrix} y_{t_{ij}^s-1} & y_{t_{ij}^s-2} & \cdots & y_{t_{ij}^s-n_a} \\ y_{t_{ij}^s} & y_{t_{ij}^s-1} & \cdots & y_{t_{ij}^s-n_a+1} \\ \vdots & \vdots & \ddots & \vdots \\ y_{t_{ij}^e-1} & y_{t_{ij}^e-2} & \cdots & y_{t_{ij}^e-n_a} \end{bmatrix} \tag{12}$$

$$\mathbf{E}_t = \begin{bmatrix} \eta_{t_{ij}^s-1} & \eta_{t_{ij}^s-2} & \cdots & \eta_{t_{ij}^s-n_a} \\ \eta_{t_{ij}^s} & \eta_{t_{ij}^s-1} & \cdots & \eta_{t_{ij}^s-n_a+1} \\ \vdots & \vdots & \ddots & \vdots \\ \eta_{t_{ij}^e-1} & \eta_{t_{ij}^e-2} & \cdots & \eta_{t_{ij}^e-n_a} \end{bmatrix} \tag{13}$$

$$\mathbf{U}_t = \begin{bmatrix} u_{t_{ij}^s-1} & u_{t_{ij}^s-2} & \cdots & u_{t_{ij}^s-n_a} \\ u_{t_{ij}^s} & u_{t_{ij}^s-1} & \cdots & u_{t_{ij}^s-n_a+1} \\ \vdots & \vdots & \ddots & \vdots \\ u_{t_{ij}^e-1} & u_{t_{ij}^e-2} & \cdots & u_{t_{ij}^e-n_a} \end{bmatrix} \tag{14}$$

$$\mathbf{F}_t = \begin{bmatrix} \nu_{t_{ij}^s-1} & \nu_{t_{ij}^s-2} & \cdots & \nu_{t_{ij}^s-n_a} \\ \nu_{t_{ij}^s} & \nu_{t_{ij}^s-1} & \cdots & \nu_{t_{ij}^s-n_a+1} \\ \vdots & \vdots & \ddots & \vdots \\ \nu_{t_{ij}^e-1} & \nu_{t_{ij}^e-2} & \cdots & \nu_{t_{ij}^e-n_a} \end{bmatrix} \tag{15}$$

$$\boldsymbol{\theta}_i = \begin{bmatrix} a_{i,1} & \cdots & a_{i,n_a} & b_{i,1} & \cdots & b_{i,n_a} \end{bmatrix}^T \tag{16}$$

$$\mathbf{b}_{t_{ij}^s} = \begin{bmatrix} y_{t_{ij}^s} & y_{t_{ij}^s+1} & \cdots & y_{t_{ij}^e} \end{bmatrix}^T \tag{17}$$

$$\mathbf{f}_{t_{ij}^s} = \begin{bmatrix} \eta_{t_{ij}^s} & \eta_{t_{ij}^s+1} & \cdots & \eta_{t_{ij}^e} \end{bmatrix}^T \tag{18}$$

Here i denotes the index of the subsystem, j is the index of disconnected segments generated by the same subsystem, t_{ij}^s and t_{ij}^e denote i the starting and ending switch time of a segment, $\boldsymbol{\theta}_i$ denotes the model of subsystem i, and $\mathbf{E}_{ij}, \mathbf{F}_{ij}$ are the structured error terms. While this problem is non-convex, consistent numerical experience shows that the STLN approach in [33], although only guaranteed to provide a local optimum, works well in practice.

E. Step 4: Labeling ambiguous data points

While Step 3 above generates the solution to Problem 1, many applications, such as anomaly detection, require, in addition to identifying the parameters of each subsystem, assigning labels to each point and estimating the actual switching instant. This can be accomplished by searching each interval $[T_k^-, T_k^+]$, known to contain a switch, for the location that minimizes the identification error. Specifically, assume that the data before T_k^- was generated by the subsystem s_1 with parameters $\boldsymbol{\theta}_{s_1}$, and the one after T_k^+ by s_2 with parameters $\boldsymbol{\theta}_{s_2}$. The best estimate of the actual switch location is given by

$$i_{\text{min}} = \underset{i \in [T_k^-, T_k^+ - 1]}{argmin} \sum_{j=T_k^-}^{i-1} (\boldsymbol{\theta}_{s_1}^T \boldsymbol{\phi}_j - \mathbf{y}_j)^2 + \sum_{j=i}^{T_k^+} (\boldsymbol{\theta}_{s_2}^T \boldsymbol{\phi}_j - \mathbf{y}_j)^2 \tag{19}$$

where

$$\boldsymbol{\phi}_j = \begin{bmatrix} y_{j-1} & \cdots & y_{j-n_a} & u_{j-1} & \cdots & u_{j-n_a} \end{bmatrix}^T \tag{20}$$

Then, data points in $[T_k^-, i_{\text{min}} - 1]$ are labeled s_1, and those in $[i_{\text{min}}, T_k^+]$ are labeled as s_2.

IV. A PRACTICAL APPLICATION: ACTION SEGMENTATION

In this section, we applied the proposed method to real data from a computer vision action segmentation problem. We recorded a video in our lab with the following sequence of actions. (i) The subject began with walking from the right to the left; (ii) halfway, he squatted and stood up, and (iii) then he continued walking to the left. Sample frames from this video are shown in Figure 1. We used background subtraction to get the blob of the subject. Then, we computed the center of the blob in each frame, and used the y coordinate of its trajectory as the data. Since there is no input available here, we model the trajectory as impulse response. The segmentation result using the proposed method is shown in Figure 2. A comparison of the proposed method against existing techniques is given in Table I. As shown there the proposed method achieved the highest identity accuracy 95.35% with a modest computational burden.

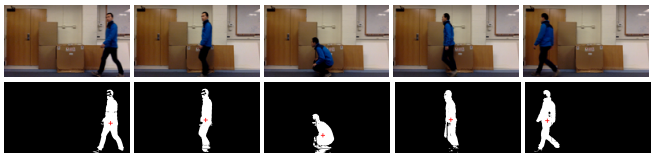

Fig. 1. Top figure: Sample frames from a walking and squatting sequence. Bottom figure: foreground blobs and the center of mass of the subject.

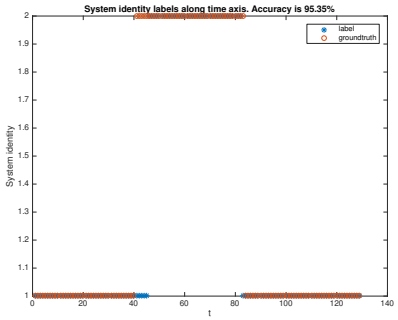

Fig. 2. Action segment labels obtained using the proposed method.

V. CONCLUSIONS

Despite its practical relevance, identification of Error-In-Variables SARX models is far from solved. In this paper we propose an approach based upon firstly embedding the

TABLE I

ACTION SEGMENTATION EXAMPLE: COMPARISON OF THE PROPOSED APPROACH AGAINST OTHER METHODS.

Methods	Label accuracy	Running time
Continuous [18]	51.94%	1.42s
kLinReg [16]	57.36%	0.01s
MinSubmodels [28]	59.69%	1.85s
SON-EM [14]	89.92%	3.58s
Proposed	95.35%	0.29s

data in the PSD manifold and then segmenting it there using graph cuts, where the weights of the edges are given by the manifold distance between segments. Once the data is segmented, the parameters of each subsystem can be extracted by any EIV LTI systems identification method. Theoretical results are provided showing that this approach is guaranteed to identify time intervals where a single system is active, and to correctly cluster all segments corresponding to the same underlying dynamics, provided that the noise level is suitably low. Further, in cases where the number of subsystems is a-priori unknown, it can be estimated from the eigenvalues of the Laplacian of the associated graph. While for higher noise levels these theoretical guarantees no longer hold, consistent numerical experience shows that the method works well, even for moderately large noise. As illustrated with a practical example, the proposed algorithm is computationally efficient and leads to better segmentations than existing ones. Current research seeks to extend these results to classes of non-linear systems.

REFERENCES

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

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

[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, 2005.

[4] R. Bhatia. *Positive definite matrices*. Princeton University Press, 2009.

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

[6] Y. Cheng, Y. Wang, M. Sznaier, and O. Camps. Subspace clustering with priors via sparse quadratically constrained quadratic programming. In *Proc. of the IEEE CVPR*, pages 5204–5212, 2016.

[7] A. Cherian, S. Sra, A. Banerjee, and N. Papanikolopoulos. Jensen-bregman logdet divergence with application to efficient similarity search for covariance matrices. *IEEE Trans. on Pattern Analysis and Machine Intelligence,* 35(9):2161–2174, 2013.

[8] 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. on Automatic Control*, 59(8):2277–2282, 2014.

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

[10] 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.

[11] A. Garulli, S. Paoletti, and A. Vicino. A survey on switched and piecewise affine system identification. *IFAC Proceedings Volumes*, 45(16):344–355, 2012.

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

[13] M. T. Harandi, M. Salzmann, and R. Hartley. From manifold to manifold: geometry-aware dimensionality reduction for spd matrices. In *Proc. 2014 ECCV*, pages 17–32. Springer.

[14] A. Hartmann, J. M. Lemos, R. S. Costa, J. Xavier, and S. Vinga. Identification of switched arx models via convex optimization and expectation maximization. *J. Process Control*, 28:9–16, 2015.

[15] N. Hure and M. Vašak. Clustering-based identification of mimo piecewise affine systems. In *Process Control (PC), 2017 21st International Conference on*, pages 404–409. IEEE, 2017.

[16] F. Lauer. Estimating the probability of success of a simple algorithm for switched linear regression. *Nonlinear Analysis: Hybrid Systems*, 8:31–47, 2013.

[17] F. Lauer. Global optimization for low-dimensional switching linear regression and bounded-error estimation. *arXiv:1707.05533*, 2017.

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

[19] I. Maruta, T. Sugie, and T.-H. Kim. Identification of multiple mode models via distributed particle swarm optimization. *IFAC Proceedings Volumes*, 44(1):7743–7748, 2011.

[20] B. Mohar. Some applications of laplace eigenvalues of graphs. In *Graph symmetry*, pages 225–275. Springer, 1997.

[21] B. Mohar, Y Alavi, G Chartrand, and OR Oellermann. The laplacian spectrum of graphs. *Graph theory, combinatorics, and applications*, 2(871-898):12, 1991.

[22] 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.

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

[24] S. Nazari, Q. Zhao, and B. Huang. Matrix-wise approach for identification of multi-mode switched arx models with noise. In *Proc 2012 ACC*, pages 3402–3407.

[25] N. Ozay. An exact and efficient algorithm for segmentation of arx models. In *Proc. 2016 ACC*, pages 38–41.

[26] N. Ozay, C. Lagoa, and M. Sznaier. Robust identification of switched affine systems via moments-based convex optimization. In *Proc. 48th IEEE CDC*, pages 4686–4691, 2009.

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

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

[29] S. Paoletti, A. Juloski, G. Ferrari-Trecate, and R. Vidal. Identification of hybrid systems a tutorial. *European journal of control*, 13(2-3):242–260, 2007.

[30] X. Pennec, P. Fillard, and N. Ayache. A riemannian framework for tensor computing. *Int. J. of Computer Vision*, 66(1):41–66, 2006.

[31] D. Piga and R. Tóth. An sdp approach for l0-minimization: Application to arx model segmentation. *Automatica*, 49(12):3646–3653, 2013.

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

[33] J Ben Rosen, H. Park, and J. Glick. Total least norm formulation and solution for structured problems. *SIAM J.on Matrix Analysis and Applications*, 17(1):110–126, 1996.

[34] M. Gorji Sefidmazgi, M. Moradi Kordmahalleh, A. Homaifar, and A. Karimoddini. Switched linear system identification based on bounded-switching clustering. In Proc. 2015 ACC, pages 1806–1811.

[35] J. Shi and Ji. Malik. Normalized cuts and image segmentation. *IEEE Trans. PAMI*, 22(8):888–905, 2000.

[36] R. Vidal, S. Soatto, Y. Ma, and S. Sastry. An algebraic geometric approach to the identification of a class of linear hybrid systems. In *Proc. 42nd IEEE CDC*, pages 167–172. 2003.

[37] U. Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

[38] X. Zhang, Y. Wang, M. Gou, M. Sznaier, and O. Camps. Efficient temporal sequence comparison and classification using gram matrix embeddings on a riemannian manifold. In *Computer Vision and Pattern Recognition (CVPR)*, 2016.