Efficient Identification of Error-in-Variables Switched Systems using a Riemannian Embedding

M. Sznaier, *Fellow, IEEE* X. Zhang *Member, IEEE* O. Camps, *Member, IEEE*

***Abstract*—This paper considers the problem of error in variables identification for switched affine models. Since it is well known that this problem is generically NP hard, several relaxations have been proposed in the literature. However, while these approaches work well for low dimensional systems with few subsystems, they scale poorly with both the number of subsystems and their memory. To address this difficulty, we propose a computationally efficient alternative, based on embedding the data in the manifold of positive semidefinite matrices, and using a manifold metric there to perform the identification. Our main result shows that, under dwell-time assumptions, the proposed algorithm is convergent, in the sense that it is guaranteed to identify the system for suitably low noise. In scenarios with larger noise levels, we provide experimental results showing that the proposed method outperforms existing ones. The paper concludes by illustrating these results with academic examples and a non-trivial application: action video segmentation.**

***Index Terms*—Error in Variables Identification, Switched Systems, Spectral Clustering.**

I. Introduction

Switched affine systems are important on their own, since they arise in the context of a wide range of application domains from fault-tolerant control to manufacturing, and as a "poor man's" model of non-linear phenomena. Given their importance, substantial research has been devoted to develop algorithms for stability analysis and controller synthesis for switched systems operating in different scenarios (see for instance [1]–[3] and references therein). However, in many practical scenarios, models of the system under consideration are not available and must be obtained from a combination of experimental data and a-priori information before these analysis and synthesis tools can be applied.

This work was supported in part by NSF under Grant CNS2038493 and CMMI2208182; in part by AFOSR under Grant FA9550-19-1-0005; and in part by ONR under Grant N00014-21-1-2431. M. Sznaier and O. Camps are with the ECE Department, Northeastern University, Boston, MA 02115, email {msznaier, camps}@coe.neu.edu. X. Zhang is with Microsoft, 3730 163rd Ave, Redmond, WA 98052, email zhangxk2009@gmail.com.

Identification of switched systems has been extensively studied in the past decade, mainly in the context of two different scenarios: (i) error-in-process models and (ii) error-in-variables models. The first case has been largely solved (see [4], [5] for a survey of earlier results), proceeding along three different approaches: optimization, algebraic, and clustering based methods. Earlier optimization based methods, [6], [7] recast the problem into an equivalent combinatorial optimization. Later approaches include sparse optimization [8]–[11], polynomial optimization [12]–[14], particle-swarm [15], difference of convex functions programming [16], and branch and bound [17] techniques. Notably, [18] established that, if the goal is to find a hybrid system that explains the observed data with the minimum number of switches and the noise is bounded in the ℓ^2 sense, then the problem can be solved in polynomial time. Algebraic based switched systems identification was first proposed in [19], showing that, in the case of noiseless data, the models can be recovered from a singular value decomposition of the embedded data matrix, followed by polynomial differentiation. While this method works well for low noise level, performance degrades quickly as this level increases. This issue was addressed in [20]–[22] which proposed to denoise the data using total least squares. The third class of methods, clustering based approaches, exploits tools from machine learning, for instance by first extracting relevant features and then resorting to methods such as k-means to estimate the discrete labels [23]–[29].

The case of error-in-variables models, where input/output measurements are corrupted by noise (and the related output estimation problem where only the outputs are affected by noise) is considerably less developed. Since in this case the problem is known to be NP hard, most existing methods are based upon convex relaxations of the original non-convex problem. Open issues are related to the computational complexity of the approaches and the quality of the identified models. Specifically, [30]–[32] presented an approach based on

recasting the problem into a rank constrained semi-definite program (SDP) using polynomial optimization arguments. Relaxing the rank constraints using the well known nuclear norm proxy for rank leads to a convex problem. This approach has been empirically shown to work well in a number of problems, but there are no theoretical convergence guarantees due to the rank relaxation. Further, computational complexity scales combinatorially both with the number of subsystems and their order, limiting the approach to systems consisting of relatively few low order subsystems.

This paper considers the error-in-variables (EiV) scenario. Our goal is to develop a method that addresses the computational complexity noted above, while, at the same time providing convergence guarantees, that is, showing that the proposed method will indeed recover the underlying system as the noise level approaches zero. The main result of the paper is a computationally efficient identification algorithm, based upon the idea of embedding the experimental data in the manifold of positive definite matrices and using a manifold metric to identify time intervals guaranteed to contain no switches. The key observation is the fact that the manifold distance between data points generated by the same subsystem is substantially smaller than the distance between points corresponding to different subsystems. Thus, switches can be detected by sharp increases in the manifold distance, and 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. Once the data is segmented, a model of each subsystem can be obtained by simply applying any EiV linear time invariant (LTI) systems identification technique to each cluster. The main theoretical result of the paper shows that, under minimum dwell time assumptions, if the noise level is below a threshold that depends on the subspace angle between subsystems, then this approach is guaranteed to produce the correct segmentation, and hence identify the correct model, provided that each cluster contains enough data to perform an LTI identification. Further, contrary to existing methods, the computational complexity of the proposed algorithm is mainly dominated by the number of switches, not the number or the order of the subsystems, and it scales linearly with the number of data points. In scenarios with higher noise levels, the theoretical convergence guarantees are lost, but extensive numerical experience shows that the proposed method consistently outperforms existing approaches in terms of computational burden, with comparable identification error.

These results are illustrated with two academic examples and a non-trivial practical one: activity segmentation from time traces of the position of a person's centroid. In all cases the proposed algorithm achieves performance comparable to the state of the art, while decreasing the computational time by at least one order of magnitude.

The paper is organized as follows: Section II provides the notation, background material on Riemannian metrics, and a statement of the problem. Section III presents the proposed solution, along with the supporting theory. Section IV illustrates these results with two academic and a practical example. Finally, Section V presents some concluding remarks. For ease of reading, all technical proofs are provided in the Appendix.

A preliminary version of this paper was presented at the 2018 CDC [33]. This version contains additional theoretical results regarding bounds on the noise level, complete proofs of all results and additional examples.

II. Preliminaries

In this section, we introduce the notation used in this paper, recall some needed background results on Riemannian metrics and normalized cuts, and formally introduce the problem under consideration.

A. Notation

$\mathbb{R}$	set of real numbers
$\mathcal{S}^n$	set of symmetric matrices in $\mathbb{R}^{n\times n}$
$\mathcal{S}^n_+(\mathcal{S}^n_{++})$	manifold of positive-semidefinite (-definite) matrices in $\mathcal{S}^n$
$\mathbf{x}$, $(\mathbf{M})$	a vector in $\mathbb{R}^n$ (matrix in $\mathbb{R}^{n\times m}$)
$\mathbf{M}^T$	transpose of matrix $\mathbf{M}$
$\det(\mathbf{M})$	determinant of a square matrix $\mathbf{M}$
$\mathcal{N}(\mathbf{M})$	null space of $\mathbf{M}$
$\sigma_{\max}(\mathbf{M})$	maximum singular value of $\mathbf{M}$
$\sigma_{\min}(\mathbf{M})$	smallest non-zero singular value of $\mathbf{M}$
$\sigma_{\mathrm{r}}(\mathbf{M})$	r^{th} singular value of $\mathbf{M}$
$\|\mathbf{M}\|_2$	2 norm of $\mathbf{M}$, $\|\mathbf{M}\|_2 = \sigma_{\max}(\mathbf{M})$
$\|\mathbf{M}\|_*$	nuclear norm of $\mathbf{M}$, $\|\mathbf{M}\|_* = \sum \sigma_i(\mathbf{M})$
$\|\mathbf{M}\|_F$	Frobenius norm of $\mathbf{M}$, $\|\mathbf{M}\|_F^2 = \sum M_{ij}^2$
$\mathbf{x}_{i:j}$	a sequence of scalars $\{x_i, x_{i+1}, \cdots, x_j\}$
$\mathbf{H}^r_{\mathbf{x}}$	Hankel matrix with r rows associated with a scalar sequence $\mathbf{x}_{i:j}$

$$\mathbf{H}^r_{\mathbf{x}} \doteq \begin{bmatrix} x_i & x_{i+1} & \cdots & x_{j-r+1} \\ x_{i+1} & x_{i+2} & \cdots & x_{j-r+2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{i+r-1} & x_{i+r} & \cdots & x_j \end{bmatrix}$$

$\mathbf{G}$	Gram matrix associated with a given Hankel matrix: $\mathbf{G} = \mathbf{H}\mathbf{H}^T$
$\hat{\mathbf{G}}(\epsilon)$	normalized, regularized Gram matrix: $\hat{\mathbf{G}}(\epsilon) = \frac{\mathbf{G}}{\|\mathbf{G}\|_*} + \epsilon \mathbf{I}_r$

B. The Jensen-Bregman Log-Det divergence

The key idea behind the approach proposed in this paper is to embed data in $\mathcal{S}_{++}^n$, the manifold of positive definite matrices, and use a suitable manifold distance to compare systems and detect switches. The intrinsic metric in $\mathcal{S}_{++}^n$, induced by the geodesic length along the manifold curvature, is the Affine Invariant Riemannian Metric (AIRM) [34], [35], defined as:

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

The main disadvantage of this metric is its high computational cost. To circumvent this difficulty [36] introduced a computationally effective surrogate, the Jensen-Bregman Log-Det Divergence (JBLD) given by:

$$J_{ld}(\mathbf{X}, \mathbf{Y}) \doteq \log\det \left(\frac{\mathbf{X}+\mathbf{Y}}{2} \right) - \frac{1}{2} \log\det (\mathbf{X}\mathbf{Y}) \quad (1)$$

Attractive properties of the JBLD include the facts that (i) its square root is a geometry aware metric in $\mathcal{S}_{++}^n$ [37], [38], in the sense that any manifold curve has the same length under the JBLD and AIRM distances, up to a $\sqrt{2}$ factor, and (ii) its low computational burden, compared against the AIRM.

A potential difficulty is that (1) is only well defined for matrices in $\mathcal{S}_{++}^n$ while this paper requires comparing positive semi-definite matrices. The following result and its corollaries extending the JBLD to $\mathcal{S}_{+}^n$, provide the theoretical justification for the proposed method.

Theorem 1. *Given $\mathbf{X}, \mathbf{Y}, \mathbf{\Delta_x}, \mathbf{\Delta_Y} \in \mathcal{S}_+^n$, assume that rank$(\frac{\mathbf{X}+\mathbf{Y}}{2}) = r$ and $\max\{\sigma_{max}(\mathbf{\Delta_x}), \sigma_{max}(\mathbf{\Delta_Y})\} \leq \bar{\delta}$, $\min\{\sigma_{min}(\mathbf{\Delta_x}), \sigma_{min}(\mathbf{\Delta_Y})\} \geq \underline{\delta}$. Then*

$$\begin{aligned}
&n \log \frac{\underline{\delta}}{\bar{\delta}} + (\bar{r} - r) \log \underline{\delta} + r \log(\sigma_{min}(\frac{\mathbf{X}+\mathbf{Y}}{2}) + \underline{\delta}) \\
&- \frac{r_x}{2} \log \frac{(\|\mathbf{X}\|_* + r_x \bar{\delta})}{r_x} - \frac{r_y}{2} \log \frac{(\|\mathbf{Y}\|_* + r_y \bar{\delta})}{r_y} \\
&\leq J_{ld}(\mathbf{X} + \mathbf{\Delta_x}, \mathbf{Y} + \mathbf{\Delta_Y}) \leq \\
&n \log \frac{\bar{\delta}}{\underline{\delta}} + (\bar{r} - r) \log \bar{\delta} + r \log(\frac{0.5(\|\mathbf{X}\|_* + \|\mathbf{Y}\|_*) + r\bar{\delta}}{r}) \\
&- \frac{r_x}{2} \log[\sigma_{min}(\mathbf{X}) + \underline{\delta})] - \frac{r_y}{2} \log[\sigma_{min}(\mathbf{Y}) + \underline{\delta})]
\end{aligned} \quad (2)$$

where:

$$\begin{aligned}
r_x &= \text{rank}(\mathbf{X}),\ r_y = \text{rank}(\mathbf{Y}),\ \text{and} \\
\bar{r} &\doteq \tfrac{r_x + r_y}{2}
\end{aligned}$$

Corollary 1.

$$\lim_{\epsilon \to 0} J_{ld}(\mathbf{X} + \epsilon\mathbf{I}, \mathbf{Y} + \epsilon\mathbf{I}) \neq \infty \iff \mathcal{N}(\mathbf{X}) = \mathcal{N}(\mathbf{Y})$$

Corollary 2. *Consider two rank r matrices $\mathbf{X}, \mathbf{Y} \in \mathcal{S}_+^n$ with $\|\mathbf{X}\|_* = \|\mathbf{Y}\|_* = 1$ and such that $\mathcal{N}(\mathbf{X}) = \mathcal{N}(\mathbf{Y})$. Then*

$$J_{ld}(\mathbf{X} + \epsilon\mathbf{I}, \mathbf{Y} + \epsilon\mathbf{I}) < r\left[\log(1 + r\epsilon) - \log r(\underline{\sigma} + \epsilon)\right]$$

for any $\underline{\sigma} \leq \min\{\sigma_r(\mathbf{X}), \sigma_r(\mathbf{Y})\}$.

The following result provides a bound on the smallest non-zero singular value of $\frac{\mathbf{X}+\mathbf{Y}}{2}$. It will be used in Section III-A to partition the data record into segments generated by a single subsystem.

Theorem 2. *Given $\mathbf{X}, \mathbf{Y} \in \mathcal{S}_+^n$ with rank$(\mathbf{X}) =$ rank$(\mathbf{Y}) = n - 1$ and $\min\{\sigma_{n-1}(\mathbf{X}), \sigma_{n-1}(\mathbf{Y})\} \geq \underline{\sigma}$, let $\mathbf{n_X}$ and $\mathbf{n_Y}$ denote the corresponding (normalized) null vectors. Then $\sigma_{min}(\frac{\mathbf{X}+\mathbf{Y}}{2}) \geq \underline{\sigma}(1 - |\mathbf{n}_\mathbf{X}^T \mathbf{n_Y}|) \doteq \sigma^*$.*

Corollary 3. *Consider two matrices $\mathbf{X}, \mathbf{Y} \in \mathcal{S}_+^n$, $n \geq 2$, with $\|\mathbf{X}\|_* = \|\mathbf{Y}\|_* = 1$, with rank$(\mathbf{X}) = n - 1$ and rank$(\mathbf{Y}) \geq n - 1$ and such that $\mathcal{N}(\mathbf{X}) \neq \mathcal{N}(\mathbf{Y})$. Then*

$$\begin{aligned}
J_{ld}(\mathbf{X} + \epsilon\mathbf{I}, \mathbf{Y} + \epsilon\mathbf{I}) > n \log(\sigma^* + \epsilon) - \frac{\log \epsilon}{2} \\
- (n-1) \log \left[\frac{1 + (n-1)\epsilon}{n-1} \right]
\end{aligned}$$

C. Spectral Clustering

Clustering algorithms seek to group a set of data points into clusters according to a given similarity measure. Of particular interest to this paper are spectral clustering techniques that solve the problem by recasting it into a graph cut form and exploiting properties of the eigen-decomposition of the associated Laplacian matrix. Specifically, in this context the data is represented using a similarity graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{W})$ where each node $V_i \in \mathcal{V}$ corresponds to a data point, $\mathcal{E}$ is the set of edges connecting these nodes, and each element W_{ij} of the symmetric weighting matrix $\mathbf{W} \in \mathbb{R}^{n \times n}$ measures the similarity between V_i and V_j, with $W_{ij} = 0$ if there is no edge connecting V_i and V_j, and $W_{ii} = 1$. The corresponding degree and Laplacian matrix are given by:

$$\begin{aligned}
&\mathbf{D} = \text{diag}\{d_1, \ldots, d_n\} \text{ where } d_i = \sum_j W_{ij} \\
&\mathbf{L} = \mathbf{D} - \mathbf{W}
\end{aligned} \quad (3)$$

It can be shown [39] that $\mathbf{L} \in \mathcal{S}_+^n$, and always has an eigenvalue at zero. Moreover, the multiplicity of the zero eigenvalue equals the number of connected components in the graph and the corresponding eigenspace is spanned by the indicator vectors of those components.

In cases where small perturbations (due for instance to noise) render a disconnected graph connected, then the number of close-to-zero eigenvalues indicates the number of components (see for instance [39]), with the corresponding eigenvectors characterizing each of the clusters. In particular, the following two results will be useful to provide quantitative results on the size of the perturbations that can be tolerated while still recovering the correct clustering.

Lemma 1. *Consider a connected graph $G = \{\mathcal{V}, \mathcal{E}, \mathbf{W}\}$ with n vertices. Assume that the non-zero elements of $\mathbf{W}$ are bounded below by some $\underline{w} > 0$. Then, the second smallest eigenvalue of the associated graph Laplacian $\mathbf{L}$ satisfies $\lambda_{n-1}(\mathbf{L}) \geq \frac{4\underline{w}}{n \cdot diam(G)}$, where $diam(G)$ denotes the diameter of the graph (e.g. the greatest distance between any pair of vertices).*

Lemma 2. *Consider the graph Laplacian $\mathbf{L}$ corresponding to a graph with n_s connected components and a perturbation $\tilde{\mathbf{L}}$. Let $\mathbf{V}_1$ and $\tilde{\mathbf{V}}_1$ denote the unitary matrices whose columns are the eigenvectors associated with the smallest n_s eigenvalues of $\mathbf{L}$ and $\tilde{\mathbf{L}}$. Then, there exist a unitary matrix $\mathbf{R}$ such that $\|\mathbf{V}_1 - \tilde{\mathbf{V}}_1\mathbf{R}\|_2 \leq 2\frac{\|\mathbf{L}-\tilde{\mathbf{L}}\|_2}{\lambda_{n-n_s-1}(\mathbf{L})}$,where λ_{n-n_s-1} denotes the smallest non-zero eigenvalue of $\mathbf{L}$.*

D. Problem Statement

The goal of this paper is to develop a computationally efficient algorithm for identifying Error-in-Variables Switched ARX models (EIV-SARX) from experimental input/output data and some *a-priori* information. Specifically, we consider switched autoregressive exogenous (SARX) systems of the form:

$$\tilde{y}_t = \sum_{k=1}^{n_a} a_k(s_t)\tilde{y}_{t-k} + \sum_{k=1}^{n_b} b_k(s_t)\tilde{u}_{t-k},\ n_a \geq n_b \quad (4)$$

consisting of n_s subsystems, each defined by the vector of model coefficients

$$\mathbf{m}_i \doteq \begin{bmatrix} -1\ a_1(i) \dots a_{n_a}(i) & b_1(i) \dots b_{n_b}(i) \end{bmatrix}\ 1 \leq i \leq n_s$$

In the sequel, we make the following assumptions:

- **A.1** *Dwell time.* Once the system (4) switches to a given subsystem at some time T_s, it does not switch again in the interval $[T_s+1, T_s+T_{dwell}-1]$, with $T_{dwell} \geq 3n_a + 2n_b + 1$.
- **A.2** *Distinguishability.* There exists some θ_{min} such that $\frac{\mathbf{m}_k^T \mathbf{m}_j}{\|\mathbf{m}_k\|\|\mathbf{m}_j\|} \leq \cos(\theta_{min}) < 1$ for all $k \neq j$.

Assumption A.1 allows for obtaining a computationally tractable algorithm, by enabling the use of manifold distances to determine whether two given data segments can be considered behaviors of the same underlying dynamics. While this dwell time constraint may restrict the applicability of the method, it arises naturally in many practical scenarios such as biological systems and manufacturing. In these cases, the proposed method is able to exploit this additional dwell time information to substantially reduce the computational burden. Assumption A.2 is needed to guarantee that the angle between the subspaces spanned by the different subsystems is large enough so that these systems can be unequivocally identified from noisy data.

Under these assumptions we are interested in solving:

Problem 1. *Given: (a)* a priori *information consisting of system orders (n_a, n_b), dwell time T_{dwell}, and input and measurement noise variances $\sigma_\eta^2, \sigma_\nu^2$; and (b) N_p noise corrupted input/output experimental data points $\{u_t = \tilde{u}_t + \nu_t, y_t = \tilde{y}_t + \eta_t\}_{t=1}^{N_p}$}; *find the minimum number of subsystems n_s, a switching sequence $s_t \in [1, n_s]$ and n_s vectors $\begin{bmatrix} a_{id,1}(i) \dots a_{id,n_a}(i) & b_{id,1}(i) \dots b_{id,n_b}(i) \end{bmatrix}$ such that*

$$\begin{aligned} y_t - \eta_t = &\sum_{k=1}^{n_a} a_{id,k}(s_t)(y_{t-k} - \eta_{t-k}) \\ &+ \sum_{k=1}^{n_b} b_{id,k}(s_t)(u_{t-k} - \nu_{t-k}) \end{aligned} \quad (5)$$

for some zero mean white noise sequences η_t, ν_t with variances $\sigma_\eta^2, \sigma_\nu^2$.

Note that the problem above becomes ill posed if the input is not informative enough, for instance if there is a pole/zero cancellation between a subsystem and the input. Thus, to avoid this scenario in the sequel we make the following additional assumption:

A.3: *Data informativity.* There exists some $\underline{\sigma}$ such that, for any time interval $[k, k+h-1]$ of length $h \geq 2n_a + n_b + 1$ where a single system is active, $\sigma_{n_a+n_b}(\hat{\mathbf{G}}_k) \geq \underline{\sigma} > 0$, where:

$$\begin{aligned} \mathbf{G}_k &\doteq \begin{bmatrix} \mathbf{H}^{n_a+1}_{\tilde{y}_{k:k+h-1}} \\ \mathbf{H}^{n_b}_{\tilde{u}_{k_u:k+h-2}} \end{bmatrix} \begin{bmatrix} \mathbf{H}^{n_a+1}_{\tilde{y}_{k:k+h-1}} \\ \mathbf{H}^{n_b}_{\tilde{u}_{k_u:k+h-2}} \end{bmatrix}^T \\ \hat{\mathbf{G}}_k &\doteq \frac{\mathbf{G}_k}{\|\mathbf{G}_k\|_*} \end{aligned} \quad (6)$$

Here $k_u = k + n_a - n_b$ and $\mathbf{H}^{n_b}_{\tilde{u}}, \mathbf{H}^{n_a+1}_{\tilde{y}}$ denote the Hankel matrices with $h - n_a$ columns and n_b, n_a+1 rows respectively, corresponding to $\tilde{u}_{k_u:k+h-2}$ and $\tilde{y}_{k:k+h-1}$, the input/output sequences to (4).

III. EIV-SARX IDENTIFICATION VIA JBLD BASED SPECTRAL CLUSTERING

The main idea underlying the proposed method is to partition the data into short segments guaranteed

not to contain switches and then group these segments into subsets that have been generated by the same LTI system. The later step is accomplished by recasting the problem into a spectral clustering form, where the distance between segments is computed using a manifold distance between the dynamics associated with these segments. Finally, the parameters of the subsystems are recovered by performing a LTI systems identification step on each cluster. A high level outline of the proposed approach is provided in Algorithms 1 and 2 .

Algorithm 1 JBLD based switched system identification

Inputs: input sequence $\mathbf{u} \in \mathbb{R}^{N_p}$, output sequence $\mathbf{y} \in \mathbb{R}^{N_p}$, system orders n_a and n_b, window size h and regularization parameter ϵ.

Step 1: Data Segmentation. Use Algorithm 2 to partition the input and output sequences into segments such that each segment is generated by a single LTI system. The s^{th} segment of input and output are denoted $\mathbf{u}_s$ and $\mathbf{y}_s$, respectively.

Step 2: Spectral Clustering.

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

$\mathbf{H}^{n_b}_{\mathbf{u}_s} \leftarrow$ Hankelize $\mathbf{u}_s$ with n_b rows

$\mathbf{H}^{n_a+1}_{\mathbf{y}_s} \leftarrow$ Hankelize $\mathbf{y}_s$ with n_a+1 rows

$\mathbf{G}_s \leftarrow \begin{bmatrix} \mathbf{H}^{n_a+1}_{\mathbf{y}_s} \\ \mathbf{H}^{n_b}_{\mathbf{u}_s} \end{bmatrix} \begin{bmatrix} \mathbf{H}^{n_a+1}_{\mathbf{y}_s} \\ \mathbf{H}^{n_b}_{\mathbf{u}_s} \end{bmatrix}^T$

$\hat{\mathbf{G}}_s(\epsilon) \leftarrow \frac{\mathbf{G}_s}{\|\mathbf{G}_s\|_*} + \epsilon \mathbf{I}_{n_a+n_b+1}$

end for

Compute the similarity matrix $\mathbf{W}_{ij} = e^{-0.5 J_{ld}(\hat{\mathbf{G}}_i(\epsilon), \hat{\mathbf{G}}_j(\epsilon))}$

Cluster labels $\mathbf{z} \leftarrow$ spectral cluster on $\mathbf{W}$

Step 3: Subsystem Identification.

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

Perform a Systems Identification on cluster z_k

end for

Outputs: $\mathbf{a}_k, \mathbf{b}_k$ (parameters of each subsystem)

Next we provide the theoretical justification for the proposed algorithm, by analyzing the properties of the data segmentation and clustering steps.

A. Step 1: Data Segmentation and Switch Detection

The goal of this step, implemented in Algorithm 2, is to segment the data into portions where only one subsystem is active. Towards this goal, the algorithm builds two sequences $\{T_i^-\}$ and $\{T_i^+\}$, such that actual switch instants T_i satisfy $T_i^- \leq T_i \leq T_i^+$. Thus, the interval $[T_i^+ + 1, T_{i+1}^- - 1]$ contains no switches (equivalently, the data y_k, $k \in [T_i^+ + 1, T_{i+1}^- - 1]$ must have been generated by a single subsystem). The algorithm uses a sliding window of size $T_{dwell} - n_a - n_b \geq h \geq 2n_a + n_b + 1$

Algorithm 2 JBLD based data segmentation

Inputs: input sequence $\mathbf{u} \in \mathbb{R}^{N_p}$, output sequence $\mathbf{y} \in \mathbb{R}^{N_p}$, system orders n_a and n_b, sliding window size h, threshold and regularization parameters τ, ϵ.

Step 1: Segmentation. Use a sliding window to partition the input and output sequences into $N_p - h + 1$ segments of length h. The i^{th} input and output segments are denoted by $\mathbf{u}_{i:i+h-1}$ and $\mathbf{y}_{i:i+h-1}$.

Step 2: Build Gram matrix.

for $i = 1$ to $N_p - h + 1$ **do**

$\mathbf{H}^{n_b}_{\mathbf{u},i} \leftarrow$ Hankelize $\mathbf{u}_{i+n_a-n_b:i+h-2}$ with n_b rows.

$\mathbf{H}^{n_a+1}_{\mathbf{y},i} \leftarrow$ Hankelize $\mathbf{y}_{i:i+h-1}$ with n_a+1 rows

$\mathbf{G}_i \leftarrow \begin{bmatrix} \mathbf{H}^{n_a+1}_{\mathbf{y},i} \\ \mathbf{H}^{n_b}_{\mathbf{u},i} \end{bmatrix} \begin{bmatrix} \mathbf{H}^{n_a+1}_{\mathbf{y},i} \\ \mathbf{H}^{n_b}_{\mathbf{u},i} \end{bmatrix}^T$

$\hat{\mathbf{G}}_i(\epsilon) \leftarrow \frac{\mathbf{G}_i}{\|\mathbf{G}_i\|_*} + \epsilon \mathbf{I}_{n_a+n_b+1}$

end for

Step 3: Search system switches.

$\mathbf{T}^+ \leftarrow \emptyset, \mathbf{T}^- \leftarrow \emptyset, T^- \leftarrow 0, T^+ \leftarrow 1$

$j^+ \leftarrow 1$ ▷ forward search index

while $j^+ < N_p - h + 1$ **do**

$j^+ \leftarrow j^+ + 1$

if $J_{ld}(\hat{\mathbf{G}}_{T^+}(\epsilon), \hat{\mathbf{G}}_{j^+}(\epsilon)) > \tau$ **then**

$T^+ \leftarrow j^+ + h - 1$, $\mathbf{T}^+ \leftarrow \begin{bmatrix} \mathbf{T}^+ & T^+ \end{bmatrix}$

$j_{\text{high}} \leftarrow T^+ - n_a - 1$, $j_{\text{low}} \leftarrow T^- + T_{\text{dwell}} - n_a$

$j^- \leftarrow j_{\text{high}}$ ▷ backward search index

while $j^- \geq j_{\text{low}}$ **do**

if $J_{ld}(\hat{\mathbf{G}}_{T^+}(\epsilon), \hat{\mathbf{G}}_{j^-}(\epsilon)) < \tau$ **then**

$T^- \leftarrow j^- + n_a$

else

$\mathbf{T}^- \leftarrow \begin{bmatrix} \mathbf{T}^- & T^- \end{bmatrix}$

break

$j^- \leftarrow j^- - 1$

end if

end while

$j^+ \leftarrow T^+$

end if

end while

Outputs: $\mathbf{T}^+$, $\mathbf{T}^-$

to find switches by detecting sharp increases in the JBLD distance between Gram matrices corresponding to adjacent segments. Specifically, given a regularization parameter ϵ, let $\hat{\mathbf{G}}_k(\epsilon)$ denote the (normalized, regularized) Gram matrix built from the data in the interval $[k, k+h-1]$, where h is the chosen window length:

$$\hat{\mathbf{G}}_k(\epsilon) = \hat{\mathbf{G}}_k + \epsilon \mathbf{I}_r \tag{7}$$

where $r \doteq n_a + n_b + 1$ and $\hat{\mathbf{G}}_k$ is defined in (6). Next, consider an increasing sequence $\{i\}$ and define $T_0^- = 0$, $T_0^+ = 1$ and:

$$
\begin{aligned}
j_i^+ &= \underset{j\geq T_{i-1}^+ +1}{argmin} \left\{ j : J_{ld}(\hat{\mathbf{G}}_{T_{i-1}^+}(\epsilon), \hat{\mathbf{G}}_j(\epsilon)) \geq \tau \right\} \\
T_i^+ &= j_i^+ + h - 1 \\
j_{high} &= T_i^+ - n_a - 1;\ j_{low} = T_{i-1}^- + T_{dwell} - n_a \\
j_i^- &= \underset{j_{high}\geq j\geq j_{low}}{argmin} \left\{ j : J_{ld}(\hat{\mathbf{G}}_j(\epsilon), \hat{\mathbf{G}}_{T_i^+}(\epsilon)) < \tau \right\} \\
T_i^- &= j_i^- + n_a
\end{aligned}
\tag{8}
$$

where τ denotes a suitable threshold to be determined later. 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 with $n_a = n_b = 1$ and dwell time $T_{dwell} = 3n_a + 2n_b + 1 = 6$:

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

$$y_{k+1} = 2u_k \qquad \text{(system2)}$$

and the following input/output sequences:
u(1:11)={0,-1,2,0,-1,1,2,0,-1,1,0},
y(1:12)={1,1,0,2,2,1,2,4,0,-2,2,0}.
The corresponding Hankel matrices, $\mathbf{H}_i^3 = \begin{bmatrix}(\mathbf{H}_{y,i}^2)^T & (\mathbf{H}_{u,i}^1)^T\end{bmatrix}^T$ with $h = 2n_a + n_b + 1 = 4$ are

$$
\mathbf{H}_1 = \begin{bmatrix}1 & 1 & 0\\ 1 & 0 & 2\\ 0 & -1 & 2\end{bmatrix}, \mathbf{H}_2 = \begin{bmatrix}1 & 0 & 2\\ 0 & 2 & 2\\ -1 & 2 & 0\end{bmatrix}, \mathbf{H}_3 = \begin{bmatrix}0 & 2 & 2\\ 2 & 2 & 1\\ 2 & 0 & -1\end{bmatrix},
$$
$$
\mathbf{H}_4 = \begin{bmatrix}2 & 2 & 1\\ 2 & 1 & 2\\ 0 & -1 & 1\end{bmatrix}, \mathbf{H}_5 = \begin{bmatrix}2 & 1 & 2\\ 1 & 2 & 4\\ -1 & 1 & 2\end{bmatrix}, \mathbf{H}_6 = \begin{bmatrix}1 & 2 & 4\\ 2 & 4 & 0\\ 1 & 2 & 0\end{bmatrix},
$$
$$
\mathbf{H}_7 = \begin{bmatrix}2 & 4 & 0\\ 4 & 0 & -2\\ 2 & 0 & -1\end{bmatrix}, \mathbf{H}_8 = \begin{bmatrix}4 & 0 & -2\\ 0 & -2 & 2\\ 0 & -1 & 1\end{bmatrix}, \mathbf{H}_9 = \begin{bmatrix}0 & -2 & 2\\ -2 & 2 & 0\\ -1 & 1 & 0\end{bmatrix}
$$

Here the points in [1,6] were generated by system 1 and those in [9,12] were generated by system 2. Note that since y_7 and y_8 satisfy both models, they can be assigned to either one. Applying the procedure outlined above[1] to the corresponding Gram matrices yields $j_1^+ = 6$, $T_1^+ = 9$, $j_1^- = 6$, $T_1^- = 7$. This correctly indicates that the earliest possible switch happened at $T = 7$, and the latest possible one at $T = 9$. Further, due to the dwell time constraints, a second switch cannot happen until $T_2 = T_1^- + T_{dwell} = 13$. Hence, the points in the intervals $[1, 6]$ and $[9, 12]$ each belong to a single class. Since this guarantees that $\mathbf{H}_9$ contains points from a single system, it is rank deficient and thus j_1^- in (8) is well defined. Note in passing that the ambiguity in detecting the switches arises from the fact that the subsystems in this example are one-step indistinguishable [8].

[1]Since we are using noiseless data, we can take $\epsilon = 0$ and $\tau = \infty$

Theorem 3. *Suppose that assumptions A.1–A.3 hold, and define $\tau_{lb}(\epsilon)$ and $\tau_{ub}(\epsilon)$ as follows*

$$
\begin{aligned}
\tau_{lb}(\epsilon) \doteq & (n_a + n_b)\left[\log(1 + (n_a + n_b)\epsilon) - \right. \\
& \left. \log\left((n_a + n_b)(\underline{\sigma} + \epsilon)\right)\right] \\
\tau_{ub}(\epsilon) \doteq & (n_a + n_b + 1)\log(\sigma^* + \epsilon) - \log\epsilon \\
& - (n_a + n_b)\log\frac{1 + (n_a + n_b)\epsilon}{n_a + n_b}
\end{aligned}
\tag{9}
$$

where $\sigma^* \doteq \underline{\sigma}[1 - \cos(\theta_{min})]$. Then, there exist ϵ and τ such that $\tau_{lb}(\epsilon) < \tau < \tau_{ub}(\epsilon)$. Further, the sequences generated using (8) *with these values satisfy $T_i^- \leq T_i \leq T_i^+$, where T_i denotes the actual switching instants.*

B. Step 2: Spectral Clustering

After $\mathcal{I} \doteq \cup_k [T_k^+, T_{k+1}^- - 1]$, the set of all intervals guaranteed to contain no switches, has been obtained, the next step is to group these intervals into clusters, each generated by a single subsystem. Note that in principle, in the noiseless case, this can be accomplished by simply selecting ϵ and τ as in Theorem 3 and, proceeding pairwise, grouping together all segments (i, j) where $J_{ld}(\hat{\mathbf{G}}_i(\epsilon), \hat{\mathbf{G}}_j(\epsilon)) \leq \tau$. However, with an eye towards handling noisy data, here we will pursue an alternative approach, based on spectral clustering. Specifically, we will consider a graph where the nodes consist of the intervals $[T_k^+, T_{k+1}^- - 1]$ and where the edge connecting two nodes has associated a weight $\mathbf{W}_{ij}$:

$$\mathbf{W}_{ij} = e^{-0.5 J_{ld}(\hat{\mathbf{G}}_i(\epsilon), \hat{\mathbf{G}}_j(\epsilon))} \tag{10}$$

As shown next, if ϵ is suitably chosen, the graph Laplacian corresponding to this matrix $\mathbf{W}$ has exactly n_s eigenvalues close to zero (in a sense to be precisely defined next) and the corresponding eigenvectors are the indicators of the clusters generated by each of the n_s subsystems.

Theorem 4. *If ϵ is selected such that $\tau_{ub}(\epsilon) - \tau_{lb}(\epsilon) > 2\log\frac{n_o^2(n_o-1)}{2}$, where $n_o \doteq \frac{N_p}{3n_a+2n_b+1}$, the graph Laplacian corresponding to $\mathbf{W}$ defined in* (10) *has exactly n_s eigenvalues $\lambda_i \leq 2(n_o - 1)e^{-0.5\tau_{ub}(\epsilon)}$, where n_s is the minimum number of subsystems required to explain the observed data.*

C. Handling Noise

When the measured data is corrupted by noise, it is not immediate whether the manifold distance can be used to separate data originating from different systems, since the corresponding Gramians $\mathbf{G}_i$ are generically full rank. As we show next, Theorem 3 still holds for noisy data, provided that the input and measurement noises are white, with suitable low variance.

Theorem 5. *Assume that A1-A3 hold and the input and output noises are white, with zero mean and variance $\sigma_\nu^2, \sigma_\eta^2$ and uncorrelated with the input/output sequences $\{u, y\}$. Then, there exist some σ_{ub} function only of the problem data such that if $\max\{\sigma_\nu, \sigma_\eta\} \leq \sigma_{ub}$, Algorithm 1 recovers, with high probability, the correct data segmentation.*

D. Step 3: Subsystem Identification

After the switch detection and spectral clustering steps, each cluster contains data segments generated from a single subsystem. Thus, at this point, any LTI system identification technique that handles EIV scenarios can be used to recover the parameters that characterize each subsystem, that is, finding a parameter vector $\boldsymbol{\theta_i}$ and associated noise matrices satisfying:

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

where

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

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

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

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

$$\begin{aligned} \boldsymbol{\theta}_i &= \begin{bmatrix} a_{i,1} & \cdots & a_{i,n_a} & b_{i,1} & \cdots & b_{i,n_b} \end{bmatrix}^T \\ \mathbf{b}_{t_{ij}} &= \begin{bmatrix} y_{t_{ij}^s} & y_{t_{ij}^s+1} & \cdots & y_{t_{ij}^e} \end{bmatrix}^T \\ \mathbf{f}_{t_{ij}} &= \begin{bmatrix} \eta_{t_{ij}^s} & \eta_{t_{ij}^s+1} & \cdots & \eta_{t_{ij}^e} \end{bmatrix}^T \end{aligned}$$

where i is the index of a subsystem, j indexes the disconnected segments generated by this subsystem, t_{ij}^s and t_{ij}^e denoted the starting and ending times of the segment ij, $\boldsymbol{\theta}_i$ is the identified model of subsystem i, and $\mathbf{E}_{ij}, \mathbf{F}_{ij}$ are the (structured) noise terms. In this paper we solved (11) using a simple regularized least squares approach [40], since consistent numerical experience shows that it is substantially faster than competing methods, with comparable errors.

E. Step 4: Labeling ambiguous data points

While Step 3 above generates the solution to Problem 1, in many applications (e.g anomaly detection), it is of interest to assign labels to all data points, including the ambiguous ones not utilized for the identification. This can be accomplished by searching each interval $[T_k^-, T_k^+]$ for the location that minimizes the simulation error (note that under the dwell time constraints each of these intervals is known to contain a single switch). 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:

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

where

$$\boldsymbol{\phi}_j = [y_{j-1} \ \cdots \ y_{j-n_a} \ u_{j-1} \ \cdots \ u_{j-n_b}]^T$$

Once the estimate T_k^{est} of the switching time T_k is obtained, the data points in $[T_k^-, T_k^{\text{est}} - 1]$ are labeled s_1, and those in $[T_k^{\text{est}}, T_k^+]$ are labeled as s_2.

IV. Experiments

In this section we illustrate the advantages of the proposed method with two academic and one practical examples. Applying Algorithms 1 and 2 requires selecting the parameters h, ϵ and τ. In general, h should be as small as possible, to produce the smallest $[T^-, T^+]$ intervals that contain the switches, leading to more accurate estimates of the model. Thus, for the noiseless case $h = 2n_a + n_b + 1$. However, for noisy data, using larger values of h leads to Gram matrices less sensitive to noise. So in these scenarios, it is beneficial to use $h > 2n_a + n_b + 1$, with higher h values corresponding to higher noise levels. In principle, determining the range of values for ϵ in Theorem 3 so that $\tau_{lb}(\epsilon) < \tau_{ub}(\epsilon)$, requires knowledge of the parameter $\underline{\sigma}$, which essentially quantifies the informativity of the data and the observability of the system. An estimate of $\underline{\sigma}$ can be obtained by considering the data in $[1, T_{\text{dwell}} - 1]$, since this data has been generated by a single system. In addition, for

many practical examples (such as the activity recognition one discussed below), data-sets with sample clips of the activity may be available and can be used to estimate a lower bound of $\underline{\sigma}$ over the data-set. Note that since $\underline{\sigma}$ appears inside a log, the bounds (9) are relatively robust to errors in its estimation. Given $\{\sigma_\nu^2, \sigma_\eta^2\}$, the regularization parameter ϵ can be obtained from the proof of Theorem 5. However, consistent numerical experience shows that the proposed algorithm is largely insensitive to the value of ϵ, since most of the regularization effect is provided directly by the noise. Once the range of values for τ has been estimated from Theorem 3, its final value can be chosen using cross-validation.

A. Academic Example 1

In this example we consider a system composed of three second-order subsystems with $n_a = 2$ and $n_b = 2$. The corresponding models are given by $\tilde{y}_t = \boldsymbol{\theta}_{s_t} \begin{bmatrix} \tilde{y}_{t-1} & \tilde{y}_{t-2} & \tilde{u}_{t-1} & \tilde{u}_{t-2} \end{bmatrix}^T$, with

$$\boldsymbol{\theta}_1 = \begin{bmatrix} -0.1 & 0.42 & -0.55 & 0.08 \end{bmatrix} \quad \text{(Subsystem1)}$$
$$\boldsymbol{\theta}_2 = \begin{bmatrix} 1.55 & -0.58 & -2.10 & 0.96 \end{bmatrix} \quad \text{(Subsystem2)}$$
$$\boldsymbol{\theta}_3 = \begin{bmatrix} 1 & -0.24 & -0.65 & 0.30 \end{bmatrix} \quad \text{(Subsystem3)}$$

We randomly generated the switch locations and thus the vector $\mathbf{s}$, discarding those realizations that did not satisfy the dwell time constraint $T_{dwell} \geq 30$. The system was excited by a Gaussian random input $\mathbf{u}$ corrupted by zero-mean Gaussian noise $\boldsymbol{\nu}$ with standard deviation 1. Without loss of generality, initial conditions were set to $\tilde{y}_1 = 5, \tilde{y}_2 = 5$, and in each time $t \geq 3$, we used the model parameter $\boldsymbol{\theta}_{s_t}$ to generate the corresponding output $\tilde{y}_t, \quad t \in [3, 600]$. Finally, these values were corrupted with Gaussian noise with standard deviation η. Table II shows the results of experiments with η taking values 0.1, 0.15, 0.2, and 0.3.

1) Switch detection: First, we evaluate the performance of the switch detection module in terms of recall, precision and F_1 score [41] defined as:

$$\begin{aligned} \text{Recall} &= \tfrac{TP}{TP+FN}, \text{ Precision} = \tfrac{TP}{TP+FP} \text{ and} \\ F_1 &= \tfrac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \end{aligned}$$

Here TP stands for the total number of true positive detections (e.g. correctly detected switches); FP stands for the total number of false positives (e.g. switches detected where there are none), and FN stands for the total number of false negatives (e.g. switches that have not been detected). Ideally, the goal is to have Precision, Recall and F1 close to 1.

A comparison of the performance of different approaches over 100 random experiments is shown in Table I. Here the first 300 input/output data points (out of a total of 600) were used for identification and the entire sequence for validation. For this problem, we used $h = 10 > 2n_a + n_b + 1$ and estimated an initial value of τ in the range $[1.6,\ 3.6]$ using Theorem 3 for different noise levels. Finally, we used cross-validation to set $\tau = 2$. For benchmarking purposes we applied several existing methods to this example, using the default values provided by the authors, leading to the results shown in Table I. As illustrated there, the proposed method is both the fastest and the best performer in terms of precision and F_1 scores, and the second best in terms of recall.

2) SARX-EIV system identification: Next, we applied the proposed method to identify the parameters of each subsystem. To compare against existing methods, performance was evaluated using five criteria: success rate, parameter error, validation error, fitting accuracy, and running time. Following [25] we evaluated the parameter estimation error using a Normalized Mean Square Error (NMSE), defined as

$$\text{NMSE} = \frac{1}{n_s} \sum_{i=1}^{n_s} \frac{\|\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i\|_2^2}{\|\boldsymbol{\theta}_i\|_2^2} \tag{12}$$

To compute the validation error, we generated a simulated output $\hat{\mathbf{y}}$ using the estimated parameters and ground truth subsystem labels and compared the results against the sequence $\tilde{\mathbf{y}}$ generated by the ground truth parameters and switching sequence using the following criterion

$$\text{VE} = \frac{\|\hat{\mathbf{y}} - \tilde{\mathbf{y}}\|_2}{n - n_a} \tag{13}$$

where n is the length of $\tilde{\mathbf{y}}$. Following [25], statistics of the metrics were computed only on successful experiments, defined as those where the validation error satisfied VE $\leq 10^{\frac{\|\tilde{\mathbf{y}}\|_2}{n}}$. Finally, the fitting of the simulated sequence against $\mathbf{y}$, the given data, was evaluated using the FIT score defined as:

$$\text{FIT} = 1 - \frac{\|\hat{\mathbf{y}} - \mathbf{y}\|_2}{\|\mathbf{y} - \text{mean}(\mathbf{y})\|_2} \tag{14}$$

As shown in Table II, the proposed algorithm achieves the smallest parameter errors and best fit to the data, with the second best running time. In this example, the fastest algorithm (k-LinReg, [25]) is up to two orders of magnitude faster, but at the price of substantial increase in the validation and parameter fit errors (up to a fourth-fold increase). It is also worth noting that this algorithm failed to identify a system in at least 25% of the runs (that is, the algorithm identified a system with validation error below an acceptable threshold only in 75% of the experiments). This is due to the fact that it was designed to handle the case of error–in–model (EiM), as opposed

TABLE I
ACADEMIC EXAMPLE 1: $\sigma_\eta = 0.1$, SWITCH DETECTION, RUN 100 TIMES RANDOMLY.

Methods	Recall	Precision	F_1 score	Time
Min # of Switches [8]	32.67%	46.09%	0.3824	492.9s
Sum Of Norms [9]	66.53%	42.37%	0.5177	202.6s
DpSwitch [18]	55.25%	79.71%	0.6526	78.9s
Proposed	84.16%	84.33%	0.8424	11.7s

TABLE II
ACADEMIC EXAMPLE 1, SYSTEM IDENTIFICATION AS A FUNCTION OF NOISE OVER 100 RANDOM RUNS. NMSE STANDS FOR NORMALIZED MEAN SQUARE ERROR OF PARAMETERS; VE STANDS FOR VALIDATION ERROR.

Methods	noise std. dev.	Succ (%)	NMSE (10^{-1})	VE (10^{-2})	FIT (%)	Time (s)
Lauer11 [42]	0.1	91	6.1 ± 10.7	14.4 ± 18.7	86.3 ± 7.6	137
	0.15	89	7.9 ± 13.4	17.7 ± 18.6	82.5 ± 9.2	147
	0.2	88	7.6 ± 11.5	23.6 ± 26.9	80.2 ± 9.7	141
	0.3	79	6.5 ± 6.6	30.0 ± 28.5	76.0 ± 9.9	142
SON-EM [27]	0.1	99	1.8 ± 4.9	2.9 ± 3.6	92 ± 7.8	168
	0.15	99	2.4 ± 4.7	3.9 ± 4.0	89.4 ± 7.7	190
	0.2	100	3.8 ± 6.3	5.2 ± 5.1	86.0 ± 9.0	183
	0.3	100	5.9 ± 7.5	7.4 ± 7.0	79.8 ± 10.0	211
k-LinReg [25]	0.1	75	5.6 ± 14.8	14.0 ± 33.4	88.6 ± 12.2	0.3
	0.15	75	5.0 ± 12.0	13.4 ± 26.5	86.3 ± 9.3	0.3
	0.2	71	6.2 ± 12.1	14.9 ± 22.3	85.0 ± 7.2	0.4
	0.3	71	5.1 ± 8.0	19.0 ± 26.0	80.8 ± 7.2	0.4
Proposed	0.1	100	1.8 ± 2.1	1.9 ± 1.8	93.4 ± 2.1	16.0
	0.15	100	2.5 ± 3.0	3.5 ± 4.0	89.8 ± 4.4	16.1
	0.2	100	3.0 ± 3.1	5.0 ± 5.2	86.8 ± 5.0	16.3
	0.3	100	6.1 ± 8.0	8.2 ± 7.7	79.9 ± 8.7	16.7

to error-in-variables, and thus, the bounds in [25] no longer apply. Further, as shown there, even for the EiM case, for a given number of points, the probability of failure is proportional to $e^{2^{(0.5n_s)}}$, so performance is expected to degrade as n_s increases.

B. High Order Example

In this example we consider a system composed of $n_s = 10$ subsystems, each having $n_a = n_b = 10$. We randomly generated 5 sets of such systems, subject to the constraint that the cosine of the angle between subsystems should be less than 0.65, to guarantee well separated systems. We excited the system with a pseudo-random binary sequence of length 8000 generated using Matlab's command `idinput`, and created 20 random switches, with a dwell time $T_{\text{dwell}} \geq 80$. The first 4000 input/output pairs were used for identification and the entire sequence for validation. In this case we used $h = 3n_a + 2n_b + 1$. An initial value of τ was estimated to be in the range $[3.4\ \ 8.1]$ using Theorem 3 and the information on the noise variance, and the actual value used, $\tau = 5$, was fine-tuned by cross-validation, leading to the results shown in Table III. As shown there, the proposed approach achieves performance comparable to SON-EM, but it is approximately 30 times faster. In this example, k-LinReg performed very poorly, probably due to the large number of subsystems, and [42], failed in all instances, after running for more than 8270 seconds.

C. Action Segmentation Example

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) walking from right to left, (ii) squatting and standing up, and (iii) resume walking to the left. Sample frames from this video are shown in Figure 1. The data used here consists of the y coordinate of the centroid of the subject in each frame, obtained using background subtraction. In this case, we modeled the trajectory of the centroid as a no-input switched system. The parameters we used are $n_a = 3$, $n_b = 0$, $h = 10$, $\epsilon = 10^{-8}$ and $\tau = 2$. The segmentation obtained using the proposed method is shown in Figure 2, and a comparison of the proposed method against existing techniques is given in Table IV. As shown there the proposed method achieved the highest label identity accuracy, 99.2%, with a modest computational burden. As before, k-LinReg was the fastest method, but at the price of a 60% decrease in label accuracy, yielding a result only 10% better than a random choice of labels.

TABLE III

HIGH-ORDER ACADEMIC EXAMPLE, $n_a = 10, n_b = 10, n_s = 10$, SYSTEM IDENTIFICATION RUN OVER 5 RANDOM SYSTEMS. NMSE STANDS FOR NORMALIZED MEAN SQUARE ERROR OF PARAMETERS; VE STANDS FOR VALIDATION ERROR.

Methods	noise std. dev.	Succ (%)	NMSE (10^{-1})	VE (10^{-2})	FIT (%)	Time (s)
k-LinReg [25]	0.05	0	–	–	–	134
	0.1	40	14.03 ± 1.2	1.58 ± 0.2	55.0 ± 3.9	169.1
SON-EM [27]	0.05	100	4.13 ± 1.9	0.99 ± 0.29	86.93 ± 13.9	379.7
	0.1	100	4.02 ± 0.76	0.98 ± 0.29	85.88 ± 3.0	381.3
Proposed	0.05	100	4.56 ± 0.58	0.96 ± 0.28	88.59 ± 4.47	13.7
	0.1	100	4.65 ± 0.50	0.96 ± 0.26	85.04 ± 4.47	12.9

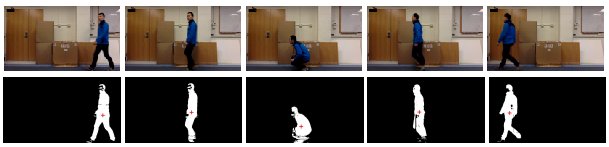

Fig. 1. Top: Frames from a video of a subject walking and squatting. Bottom: foreground blobs and the center of mass of the subject.

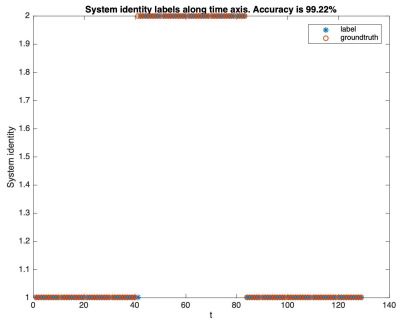

Fig. 2. Action segment labels obtained using the proposed method.

TABLE IV

ACTION SEGMENTATION EXAMPLE: COMPARISON OF THE PROPOSED APPROACH AGAINST OTHER METHODS. LABEL ACC STANDS FOR LABEL ACCURACY.

Methods	Label acc (%)	FIT (%)	Time(s)
Lauer11 [42]	51.9	86.4	1.42
k-LinReg [25]	57.4	86.0	0.01
MinSubmodels [8]	65.1	75.3	2.71
SON-EM [27]	89.9	90.6	3.6
Proposed	99.2	90.6	0.2

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 data in the positive definite manifold using regularized Gram matrices, 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 below a given number related to the subspace angle between the subspaces spanned by each subsystem. 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 both academic and practical examples, the proposed algorithm is computationally efficient and outperforms most existing techniques in terms of the identification error and computation time. An exception is the k-LinReg algorithm that, for small n_s, runs close to an order of magnitude faster than the proposed algorithm, but at the price of a similar increment in the identification error, pointing out to the existence of a severe computation time versus identification error trade-off. Moreover, while the proposed algorithm is guaranteed, under suitable conditions on the noise, to yield the correct data segmentation and converge to the actual system, no such guarantees exist for k-LinReg in the case where both the output and input measurements are corrupted by noise. Current research seeks to remove the dwell time constraint by exploiting semi-algebraic optimization tools.

REFERENCES

[1] J. Lunze and F. lamnabhi lagarrigue. *Handbook of Hybrid Systems Control: Theory, Tools, Applications*. Cambridge University Press, 2009.

[2] Daniel Liberzon. *Switching in systems and control*. Springer Science & Business Media, 2012.

[3] Mirko Fiacchini and Marc Jungers. Necessary and sufficient condition for stabilizability of discrete-time linear switched systems: A set-theory approach. *Automatica*, 50(1):75–83, 2014.

[4] Simone Paoletti, Aleksandar Lj Juloski, Giancarlo Ferrari-Trecate, and René Vidal. Identification of hybrid systems a tutorial. *European journal of control*, 13(2-3):242–260, 2007.

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

[6] Jacob Roll, Alberto Bemporad, and Lennart Ljung. Identification of piecewise affine systems via mixed-integer programming. *Automatica*, 40(1):37–50, 2004.

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

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

[9] Henrik Ohlsson, Lennart Ljung, and Stephen Boyd. Segmentation of arx-models using sum-of-norms regularization. *Automatica*, 46(6):1107–1111, 2010.

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

[11] Fabien Lauer Van Luong Le and Gérard Bloch. Selective l1 minimization for sparse recovery. *IEEE Transactions on Automatic Control*, 59(11):3008–3013, 2014.

[12] Necmiye Ozay, Constantino Lagoa, and Mario Sznaier. Robust identification of switched affine systems via moments-based convex optimization. In *Decision and Control, 2009 held jointly with the 2009 28th Chinese Control Conference. CDC/CCC 2009. Proceedings of the 48th IEEE Conference on*, pages 4686–4691. IEEE, 2009.

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

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

[15] Ichiro Maruta, Toshiharu Sugie, and Tae-Hyoung Kim. Identification of multiple mode models via distributed particle swarm optimization. *IFAC Proceedings Volumes*, 44(1):7743–7748, 2011.

[16] Tao Pham Dinh, Hoai Minh Le, Hoai An Le Thi, and Fabien Lauer. A difference of convex functions algorithm for switched linear regression. *IEEE Transactions on Automatic Control*, 59(8):2277–2282, 2014.

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

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

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

[20] Sohail Nazari, Qing Zhao, and Biao Huang. An improved algebraic geometric solution to the identification of switched arx models with noise. In *American Control Conference (ACC), 2011*, pages 1230–1235. IEEE, 2011.

[21] Sohail Nazari, Qing Zhao, and Biao Huang. Matrix-wise approach for identification of multi-mode switched arx models with noise. In *American Control Conference (ACC), 2012*, pages 3402–3407. IEEE, 2012.

[22] Sohail Nazari, Bahador Rashidi, Qing Zhao, and Biao Huang. An iterative algebraic geometric approach for identification of switched arx models with noise. *Asian Journal of Control*, 18(5):1655–1667, 2016.

[23] Giancarlo Ferrari-Trecate, Marco Muselli, Diego Liberati, and Manfred Morari. A clustering technique for the identification of piecewise affine systems. *Automatica*, 39(2):205–217, 2003.

[24] Laurent 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.

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

[26] Abdelhak Goudjil, Mathieu Pouliquen, Eric Pigeon, and Olivier Gehan. A real-time identification algorithm for switched linear systems with bounded noise. In *Control Conference (ECC), 2016 European*, pages 2626–2631. IEEE, 2016.

[27] András Hartmann, João M Lemos, Rafael S Costa, João Xavier, and Susana Vinga. Identification of switched arx models via convex optimization and expectation maximization. *Journal of Process Control*, 28:9–16, 2015.

[28] Mohammad Gorji Sefidmazgi, Mina Moradi Kordmahalleh, Abdollah Homaifar, and Ali Karimoddini. Switched linear system identification based on bounded-switching clustering. In *American Control Conference (ACC), 2015*, pages 1806–1811. IEEE, 2015.

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

[30] Chao Feng, Constantino M Lagoa, and Mario Sznaier. Hybrid system identification via sparse polynomial optimization. In *American Control Conference (ACC), 2010*, pages 160–165. IEEE, 2010.

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

[32] Yongfang Cheng, Yin Wang, Mario Sznaier, and Octavia Camps. Subspace clustering with priors via sparse quadratically constrained quadratic programming. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5204–5212, 2016.

[33] X. Zhang, M. Sznaier, and O. Camps. Efficient identification of error-in variables switched systems based on riemannian distance-like functions. In 57^{th} *IEEE Conference on Decision and Control (CDC)*, pages 3006–3011, 2018.

[34] Xavier Pennec, Pierre Fillard, and Nicholas Ayache. A riemannian framework for tensor computing. *International Journal of Computer Vision*, 66(1):41–66, 2006.

[35] Rajendra Bhatia. *Positive definite matrices*. Princeton University Press, 2009.

[36] Arun Cherian, Suvrit Sra, Adrish Banerjee, and Nikolaos Papanikolopoulos. Jensen-bregman logdet divergence with application to efficient similarity search for covariance matrices. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(9):2161–2174, 2013.

[37] S. Sra. A new metric on the manifold of kernel matrices with application to matrix geometric means. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, pages 144–152, 2012.

[38] Mehrtash T Harandi, Mathieu Salzmann, and Richard Hartley. From manifold to manifold: geometry-aware dimensionality reduction for spd matrices. In *Computer Vision–ECCV 2014*, pages 17–32. Springer, 2014.

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

[40] M. Sznaier. Control oriented learning in the era of big data. *IEEE Control Systems Letters*, 5(6):1855–1867, 2021.

[41] C. J. van Rijsbergen. *Information Retrieval*. London, GB; Boston, MA: Butterworth, 2nd edition, 1979.

[42] Fabien Lauer, Gérard Bloch, and René Vidal. A continuous optimization framework for hybrid system identification. *Automatica*, 47(3):608–613, 2011.

[43] Roger A. Horn, Noah H. Rhee, and So Wasin. Eigenvalue inequalities and equalities. *Linear Algebra and its Applications*, 270(1):29 – 44, 1998.

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

[45] C Davis and W. M. Kahan. The rotation of eigenvectors by a perturbation. *SIAM J. Numer. Anal.*, 7(1):1–46, 1970.
[46] Radosaw Adamczak, Alexander E. Litvak, Alain Pajor, and Nicole Tomczak-Jaegermann. Sharp bounds on the rate of convergence of the empirical covariance matrix. *Comptes Rendus Mathematique*, 349(3):195–200, 2011.

APPENDIX

Proof of Theorem 1

In order to prove the Theorem and its Corollaries, we need the following preliminary result.

Lemma 3. *For any $\mathbf{X} \in \mathcal{S}_+^n$ with rank$(\mathbf{X}) = r$, and $\Delta \in \mathcal{S}_+^n$, the following inequality holds*

$$|\mathbf{X}+\Delta| \leq \left(\frac{\|\mathbf{X}\|_* + r\sigma_{max}(\Delta)}{r}\right)^r \sigma_{max}(\Delta)^{n-r} \quad (15)$$

Proof.

$$\begin{aligned}|\mathbf{X}+\Delta| &= \prod_{i=1}^{r}\lambda_i(\mathbf{X}+\Delta)\prod_{i=r+1}^{n}\lambda_i(\mathbf{X}+\Delta)\\ &\leq \left(\frac{\sum_{i=1}^{r}(\sigma_i(\mathbf{X})+\sigma_{max}(\mathbf{\Delta}))}{r}\right)^r \sigma_{max}(\mathbf{\Delta})^{n-r}\\ &= \left(\frac{\|\mathbf{X}\|_* + r\sigma_{max}(\mathbf{\Delta})}{r}\right)^r \sigma_{max}(\mathbf{\Delta})^{n-r}\end{aligned}$$

where we used Weyl's and the geometric-arithmetic mean inequalities. □

Proof of Theorem 1, lower bound: Let $\mathbf{\Delta} \doteq \frac{\mathbf{\Delta_X}+\mathbf{\Delta_Y}}{2}$. Then

$$\begin{aligned}&\left|\frac{\mathbf{X}+\mathbf{Y}}{2}+\mathbf{\Delta}\right| =\\ &\prod_{i=1}^{r}\lambda_i(\frac{\mathbf{X}+\mathbf{Y}}{2}+\mathbf{\Delta})\prod_{i=r+1}^{n}\lambda_i(\frac{\mathbf{X}+\mathbf{Y}}{2}+\mathbf{\Delta})\\ &\geq [\sigma_{\min}(\frac{\mathbf{X}+\mathbf{Y}}{2})+\underline{\delta}]^r\underline{\delta}^{n-r}\end{aligned}$$

$$|\mathbf{X}+\mathbf{\Delta_X}| \leq \left(\frac{\|\mathbf{X}\|_* + r_x\bar{\delta}}{r_x}\right)^{r_x}\bar{\delta}^{n-r_x}$$

$$|\mathbf{Y}+\mathbf{\Delta_Y}| \leq \left(\frac{\|\mathbf{Y}\|_* + r_y\bar{\delta}}{r_y}\right)^{r_y}\bar{\delta}^{n-r_y}$$

Thus

$$\begin{aligned}&J_{\text{ld}}(\mathbf{X}+\mathbf{\Delta_X},\mathbf{Y}+\mathbf{\Delta_Y}) \geq (n-r)\log(\underline{\delta})+\\ &r\log(\sigma_{\min}(\frac{\mathbf{X}+\mathbf{Y}}{2}+\underline{\delta}) - (n-\bar{r})\log\bar{\delta}\\ &-\frac{r_x}{2}\log\frac{(\|\mathbf{X}\|_*+r_x\bar{\delta})}{r_x} - \frac{r_y}{2}\log\frac{(\|\mathbf{Y}\|_*+r_y\bar{\delta})}{r_y}\\ &\geq n\log\frac{\underline{\delta}}{\bar{\delta}} + (\bar{r}-r)\log\underline{\delta} + r\log[\sigma_{\min}(\frac{\mathbf{X}+\mathbf{Y}}{2})+\underline{\delta}]\\ &-\frac{r_x}{2}\log\frac{(\|\mathbf{X}\|_*+r_x\bar{\delta})}{r_x} - \frac{r_y}{2}\log\frac{(\|\mathbf{Y}\|_*+r_y\bar{\delta})}{r_y}\end{aligned}$$

Proof of the upper bound:

$$\begin{aligned}&\left|\frac{\mathbf{X}+\mathbf{Y}}{2}+\mathbf{\Delta}\right| \leq \bar{\delta}^{n-r}\left(\frac{0.5(\|\mathbf{X}\|_*+\|\mathbf{Y}\|_*)+r\bar{\delta}}{r}\right)^r\\ &|\mathbf{X}+\mathbf{\Delta_X}| \geq \underline{\delta}^{n-r_x}(\sigma_{\min}(\mathbf{X})+\underline{\delta})^{r_x}\\ &|\mathbf{Y}+\mathbf{\Delta_Y}| \geq \underline{\delta}^{n-r_y}(\sigma_{\min}(\mathbf{Y})+\underline{\delta})^{r_y}\end{aligned}$$

Hence

$$\begin{aligned}&J_{\text{ld}}(\mathbf{X}+\mathbf{\Delta_X},\mathbf{Y}+\mathbf{\Delta_Y}) \leq (n-r)\log\bar{\delta} - (n-\bar{r})\log\underline{\delta}\\ &+r\log(\frac{0.5(\|\mathbf{X}\|_*+\|\mathbf{Y}\|_*)+r\bar{\delta}}{r})\\ &-\frac{r_x}{2}\log[\sigma_{\min}(\mathbf{X})+\underline{\delta})] - \frac{r_y}{2}\log[\sigma_{\min}(\mathbf{Y})+\underline{\delta}] \leq\\ &n\log\frac{\bar{\delta}}{\underline{\delta}} + (\bar{r}-r)\log\bar{\delta} + r\log(\frac{0.5(\|\mathbf{X}\|_*+\|\mathbf{Y}\|_*)+r\bar{\delta}}{r})\\ &-\frac{r_x}{2}\log[\sigma_{\min}(\mathbf{X})+\underline{\delta}] - \frac{r_y}{2}\log[\sigma_{\min}(\mathbf{Y})+\underline{\delta}]\end{aligned}$$

□

*P*roof of Corollary 1: Follows from Theorem 1 by setting $\mathbf{\Delta} = \epsilon\mathbf{I}$ and noting that, due to term $(\bar{r}-r)\log\underline{\delta}$ in (2),

$$\begin{aligned}&\lim_{\epsilon\to 0} J_{\text{ld}}(\mathbf{X}+\epsilon\mathbf{I},\mathbf{Y}+\epsilon\mathbf{I}) \neq \infty \iff \bar{r}=r \iff\\ &\text{rank}(\mathbf{X}+\mathbf{Y}) = \frac{\text{rank}(\mathbf{X})+\text{rank}(\mathbf{Y})}{2} \iff\\ &\dim(\mathcal{N}(\mathbf{X})\cap\mathcal{N}(\mathbf{Y}) = \frac{\dim(\mathcal{N}(\mathbf{X}))+\dim(\mathcal{N}(\mathbf{Y}))}{2}\\ &\iff \mathcal{N}(\mathbf{X}) = \mathcal{N}(\mathbf{Y})\end{aligned}$$

□

*P*roof of Corollary 2: Follows from Theorem 1 by setting $\mathbf{\Delta_X} = \mathbf{\Delta_Y} = \epsilon\mathbf{I}$ and noting that in this case rank$(\mathbf{X}+\mathbf{Y}) = r$ and

$$\frac{r}{2}\log[\sigma_{\min}(\mathbf{X})+\epsilon] + \frac{r}{2}\log[\sigma_{\min}(\mathbf{Y})+\epsilon] \geq r\log(\underline{\sigma}+\epsilon)$$

□

Proof of Theorem 2: Consider the svd of $\mathbf{X}$ and $\mathbf{Y}$

$$\mathbf{X} = \begin{bmatrix}\mathbf{U}_x & \mathbf{n}_x\end{bmatrix}\begin{bmatrix}\mathbf{\Sigma}_x & 0\\ 0 & 0\end{bmatrix}\begin{bmatrix}\mathbf{U}_x^T\\ \mathbf{n}_x^T\end{bmatrix}$$

$$\mathbf{Y} = \begin{bmatrix}\mathbf{U}_y & \mathbf{n}_y\end{bmatrix}\begin{bmatrix}\mathbf{\Sigma}_y & 0\\ 0 & 0\end{bmatrix}\begin{bmatrix}\mathbf{U}_y^T\\ \mathbf{n}_y^T\end{bmatrix}$$

Let $\sigma_x = \sigma_{n-1}(\mathbf{X})$ and $\sigma_y = \sigma_{n-1}(\mathbf{Y})$. From Weyl's inequality [43] we have that

$$\begin{aligned}\sigma_{\min}(\mathbf{X}+\mathbf{Y}) = \sigma_n(\mathbf{X}+\mathbf{Y}) &\geq \sigma_n(\sigma_x\mathbf{U}_x\mathbf{U}_x^T + \sigma_y\mathbf{U}_y\mathbf{U}_y^T)\\ &= \sigma_n[(\sigma_x+\sigma_y)\mathbf{I} - \sigma_x\mathbf{n}_x\mathbf{n}_x^T - \sigma_y\mathbf{n}_y\mathbf{n}_y^T] \geq\\ &(\sigma_x+\sigma_y) - \sigma_{\max}(\sigma_x\mathbf{n}_x\mathbf{n}_x^T + \sigma_y\mathbf{n}_y\mathbf{n}_y^T)\end{aligned} \quad (16)$$

Next, note that the rank 2 matrix $\mathbf{M} \doteq \sigma_x \mathbf{n}_x \mathbf{n}_x^T + \sigma_y \mathbf{n}_y \mathbf{n}_y^T$ can be factored as $\mathbf{M} = \mathbf{L}\mathbf{L}^T$ with $\mathbf{L} \doteq \begin{bmatrix} \sqrt{\sigma_x}\mathbf{n}_x & \sqrt{\sigma_y}\mathbf{n}_y \end{bmatrix}$. Thus,

$$\begin{aligned} &\sigma_{\max}(\mathbf{M}) = \sigma_{\max}(\mathbf{L}\mathbf{L}^T) = \sigma_{\max}(\mathbf{L}^T\mathbf{L}) \\ &= \sigma_{\max}(\begin{bmatrix} \sigma_x & \sqrt{\sigma_x\sigma_y}\mathbf{n}_x^T\mathbf{n}_y \\ \sqrt{\sigma_x\sigma_y}\mathbf{n}_x^T\mathbf{n}_y & \sigma_y \end{bmatrix}) \\ &= \frac{\sigma_x+\sigma_y}{2} + \frac{1}{2}\sqrt{(\sigma_x-\sigma_y)^2 + 4\sigma_x\sigma_y(\mathbf{n}_x^T\mathbf{n}_y)^2} \end{aligned} \tag{17}$$

Combining (16) and (17) yields:

$$\sigma_{\min}(\mathbf{X}+\mathbf{Y}) \geq \frac{\sigma_x+\sigma_y}{2} - \frac{1}{2}\sqrt{(\sigma_x-\sigma_y)^2 + 4\sigma_x\sigma_y(\mathbf{n}_x^T\mathbf{n}_y)^2}$$

Finally, noting the the expression above is an increasing function of σ_x, σ_y and setting $\sigma_x = \sigma_y = \underline{\sigma}$ leads to

$$\sigma_{\min}(\mathbf{X}+\mathbf{Y}) \geq \underline{\sigma} - \underline{\sigma}\sqrt{(\mathbf{n}_x^T\mathbf{n}_y)^2} = \underline{\sigma}(1 - |\mathbf{n}_x^T\mathbf{n}_y|)$$

□

*P*roof of Corollary 3. Set $\mathbf{\Delta_X} = \mathbf{\Delta_Y} = \epsilon\mathbf{I}$. By hypothesis, $r_x = n-1$, $r_y \geq n-1$. Hence $\bar{r} \leq n - \frac{1}{2}$ and $r = n$ (Since $\mathcal{N}(\mathbf{X}) \neq \mathcal{N}(\mathbf{Y})$). From Theorem 1 we have that:

$$\begin{aligned} &J_{\text{ld}}(\mathbf{X}+\epsilon\mathbf{I}, \mathbf{Y}+\epsilon\mathbf{I}) \geq +(\bar{r}-r)\log\epsilon + \\ &n\log(\sigma_{\min}(\frac{\mathbf{X}+\mathbf{Y}}{2}) + \epsilon) - \frac{n-1}{2}\log\frac{1+(n-1)\epsilon}{n-1} \\ &- \frac{r_y}{2}\log\frac{(1+r_y\epsilon)}{r_y} \geq -\frac{\log\epsilon}{2} + n\log(\sigma^* + \epsilon) \\ &- (n-1)\log\frac{1+(n-1)\epsilon}{n-1} \end{aligned} \tag{18}$$

where we used the fact that $f(r) \doteq -\frac{r}{2}\log\frac{(1+r\epsilon)}{r}$ is increasing in $r \geq 1$ if $\epsilon < 0.76$. □

Proof of Lemma 1: Let $\mathbf{e}$ denote an eigenvector of $\mathbf{L}$ associated with $\lambda_n(\mathbf{L}) = 0$. Then

$$\begin{aligned} \lambda_{n-1}(\mathbf{L}) &= \min_{\substack{\|\mathbf{x}\|=1 \\ \mathbf{x}^T\mathbf{e}=0}} \mathbf{x}^T\mathbf{L}\mathbf{x} = \min_{\substack{\|\mathbf{x}\|=1 \\ \mathbf{x}^T\mathbf{e}=0}} \frac{1}{2}\sum_{i,j} W_{ij}(x_i - x_j)^2 \\ &\geq \underline{w} \min_{\substack{\|\mathbf{x}\|=1 \\ \mathbf{x}^T\mathbf{e}=0}} \sum_{(i,j)\in\mathcal{E}} (x_i - x_j)^2 \geq \underline{w}\frac{4}{n\cdot\text{diam}(G)} \end{aligned}$$

where the last inequality follows from Theorem 4.2 in [44] □

Proof of Theorem 3: The first statement follows from the fact that, since $\sigma^* > 0$, $\lim_{\epsilon\to 0}\tau_{ub}(\epsilon) = \infty$, while $\tau_{lb}(\epsilon)$ remains finite. To prove the second, consider the trajectories that start at $t = 1$. Due to the dwell time constraint, all the data in $[1, h]$ is generated by a single system. Hence $\hat{\mathbf{G}}_1$ is rank deficient. From Corollary 3 and Theorem 2, if $J_{ld}(\hat{\mathbf{G}}_1(\epsilon), \hat{\mathbf{G}}_j(\epsilon)) \leq \tau < \tau_{ub}(\epsilon)$, the matrices $\hat{\mathbf{G}}_1, \hat{\mathbf{G}}_j$ share the same null space. Hence, the data in the interval $[1, j_1^+ + h - 2]$ can be explained by a single model of orders (n_a, n_b). At the same time, since $J_{ld}(\hat{\mathbf{G}}_1, \hat{\mathbf{G}}_{j_1^+}) \geq \tau > \tau_{lb}(\epsilon)$, from Corollary 2 it follows that the data in the interval $[1, j_1^+ + h - 1]$ cannot be explained by a single system of order n_a. Hence, a switch must have taken place no later than $T_1^+ \doteq j_1^+ + h - 1$, and the earliest such a switch could have happened is $T_1^+ - n_a - n_b$. Thus, assuming a dwell time $T_{dwell} \geq 3n_a + 2n_b + 1$, the data in the interval $[T_1^+, T_1^+ + h - 1]$ must have been generated by the same system, and hence the matrix $\mathbf{G}_{T_1^+}$ contains data from a single system. Moreover, no switches could have taken place in $[T_1^+ - n_a, T_1^+]$. Working now backwards from $T_1^+ - n_a - 1$, consider the first column of the matrix $\mathbf{H}_{j_1^-}$. Since $\mathbf{H}_{j_1^-}$ and $\mathbf{H}_{j_1^- - 1}$ have different null spaces, it follows that there is no single system that could have generated the data in the interval $[j_1^- - 1, T_1^+ + h - 1]$, hence, taking into account the previous switch, the earliest that a switch could have happened is $\max\{j_1^- + n_a, T_{i-1}^- + T_{dwell}\}$[2]. The proof is completed by induction, starting now from T_i^+ and repeating the reasoning above. □

Note that the dwell time constraint guarantees that

$$\begin{aligned} T_i^- - T_{i-1}^+ &\geq T_{i-1}^- + T_{dwell} - T_{i-1}^+ \\ &\geq T_{i-1}^+ - (n_a+n_b) + T_{dwell} - T_{i-1}^+ \\ &= T_{dwell} - (n_a+n_b) \geq 2n_a + n_b + 1 \end{aligned}$$

so even in the worst case, when all unreliable data points have been discarded, the remaining intervals are long enough to form the matrices $\mathbf{G}_i$ and perform the clustering step.

Proof of Lemma 2

Start by noting that

$$\begin{aligned} &\|\mathbf{V}_1 - \tilde{\mathbf{V}}_1\tilde{\mathbf{V}}_1^T\mathbf{V}_1\|_2 = \|\mathbf{V}_1 - \tilde{\mathbf{V}}_1 + \tilde{\mathbf{V}}_1(\mathbf{I} - \tilde{\mathbf{V}}_1^T\mathbf{V}_1)\|_2 \\ &\geq \|\mathbf{V}_1 - \tilde{\mathbf{V}}_1\|_2 - \|\tilde{\mathbf{V}}_1(\mathbf{I} - \tilde{\mathbf{V}}_1^T\mathbf{V}_1)\|_2 \\ &\geq \|\mathbf{V}_1 - \tilde{\mathbf{V}}_1\|_2 - \|\tilde{\mathbf{V}}_1\|_2\|(\mathbf{I} - \tilde{\mathbf{V}}_1^T\mathbf{V}_1)\|_2 \end{aligned}$$

Hence, for any unitary $\mathbf{R}$

$$\begin{aligned} &\|\mathbf{V}_1 - \tilde{\mathbf{V}}_1\mathbf{R}\|_2 \leq \|\mathbf{V}_1 - \tilde{\mathbf{V}}_1\tilde{\mathbf{V}}_1^T\mathbf{V}_1\|_2 + \\ &\|\mathbf{I} - \mathbf{R}^T\tilde{\mathbf{V}}_1^T\mathbf{V}_1\|_2 \end{aligned} \tag{19}$$

To bound the second term, let $\tilde{\mathbf{V}}_1^T\mathbf{V}_1 \doteq \mathbf{U}\mathbf{S}\mathbf{V}^T$ and take $\mathbf{R} = \mathbf{U}\mathbf{V}^T$. Then

$$\begin{aligned} &\|\mathbf{I} - \mathbf{R}^T\tilde{\mathbf{V}}_1^T\mathbf{V}_1\|_2 = \|\mathbf{I} - \mathbf{S}\|_2 = \max_i(1 - \cos\theta_i) \\ &\leq \max_i \sin\theta_i = \|\sin\mathbf{\Theta}(\mathbf{V}_1, \tilde{\mathbf{V}}_1\mathbf{R})\|_2 \end{aligned} \tag{20}$$

[2]Since if for instance a switch happened at $t_s \doteq j_1^- + n_a - 1$, $\mathbf{r}_1$, the first column of the Hankel matrix $\mathbf{H}_{j_1^- - 1}$ satisfies $\mathbf{m}^T\mathbf{r}_1 = 0$. where $\mathbf{m}^T\mathbf{H}_{j_1^-} = 0$, and hence $\mathbf{H}_{j_1^-}$ and $\mathbf{H}_{j_1^- - 1}$ would have the same null space, contradicting the definition of j^-.

where θ_i are the principal angles between the subspaces spanned by $\mathbf{V}_1$ and $\tilde{\mathbf{V}}_1$. Explicitly computing $\|\mathbf{V}_1 - \tilde{\mathbf{V}}_1\mathbf{R}(\tilde{\mathbf{V}}_1\mathbf{R})^T\mathbf{V}_1\|_2$ yields

$$\begin{aligned}&\|\mathbf{V}_1 - \tilde{\mathbf{V}}_1\mathbf{R}(\tilde{\mathbf{V}}_1\mathbf{R})^T\mathbf{V}_1\|_2^2 = \\ &\max_i \sigma_i\{(\mathbf{V}_1 - \tilde{\mathbf{V}}_1\tilde{\mathbf{V}}_1^T\mathbf{V}_1)^T(\mathbf{V}_1 - \tilde{\mathbf{V}}_1\tilde{\mathbf{V}}_1^T\mathbf{V}_1)\} = \\ &\max_i \sigma_i\{\mathbf{I} - (\mathbf{V}_1^T\tilde{\mathbf{V}}_1)(\mathbf{V}_1^T\tilde{\mathbf{V}}_1)^T\} \\ &= \|\sin\mathbf{\Theta}(\mathbf{V}_1, \tilde{\mathbf{V}}_1\mathbf{R})\|_2^2\end{aligned} \tag{21}$$

Finally, from Davis-Kahan Theorem [45] we have that

$$\|\sin\mathbf{\Theta}(\mathbf{V}_1, \tilde{\mathbf{V}}_1\mathbf{R})\|_2 \leq \frac{\|\mathbf{L} - \tilde{\mathbf{L}}\|_2}{\lambda_{n-n_s-1}(\mathbf{L})} \tag{22}$$

Combining (19)- (22) yields the desired bound. □

Proof of Theorem 4: Start by considering an ideal graph where $\mathbf{W}_{i,j}^{ideal} = 0$ if the data in the $[T_i^+, T_{i+1}^- - 1]$ and $[T_j^+, T_{j+1}^- - 1]$ was generated by different subsystems and let $\mathcal{E}^{ideal}$ denote its edge set. Since this graph is disconnected, with n_s connected components, it follows that its associated Laplacian $\mathbf{L}^{ideal}$ has exactly n_s eigenvalues at zero. Further, from Lemma 1 and Corollary 2, it follows that, for a given ϵ, the smallest non-zero eigenvalue of $\mathbf{L}^{ideal}$ satisfies

$$\lambda_{n_w-n_s} \geq e^{-0.5\tau_{lb}(\epsilon)}\frac{4}{n_w^2}$$

The actual Laplacian obtained using (10) can be considered as a result of adding a Hermitian perturbation $\mathbf{\Delta}$ to $\mathbf{L}^{ideal}$, $\mathbf{L} = \mathbf{L}^{ideal} + \mathbf{\Delta}$ where

$$\begin{aligned}\mathbf{\Delta}_{i,i} &= \sum_{(i,j)\notin\mathcal{E}^{ideal}} e^{-0.5J_{ld}(\hat{\mathbf{G}}_i(\epsilon),\hat{\mathbf{G}}_j(\epsilon))} \\ \mathbf{\Delta}_{i,j} &= \begin{cases} 0 & \text{if } (i,j)\in\mathcal{E}^{ideal} \\ e^{-0.5J_{ld}(\hat{\mathbf{G}}_i(\epsilon),\hat{\mathbf{G}}_j(\epsilon))} & \text{otherwise}\end{cases}\end{aligned}$$

Note that by construction the smallest eigenvalue of $\mathbf{\Delta}$ is zero, since this matrix is also a graph Laplacian. Since $\mathbf{\Delta}_{i,j} \leq e^{-0.5\tau_{ub}(\epsilon)}$, a Gershgorin disk's argument shows that the largest eigenvalue of $\mathbf{\Delta}$ satisfies

$$\lambda_1(\Delta) \leq 2(n_w - 1)e^{-0.5\tau_{ub}(\epsilon)} \tag{23}$$

Hence, from Weyl's inequality we have that

$$\begin{aligned}&\lambda_{n_w-i}(\mathbf{L}) \leq 2(n_w-1)e^{-0.5\tau_{ub}(\epsilon)}, i = 0,\ldots,n_s-1 \\ &\lambda_{n_w-n_s}(\mathbf{L}) \geq \lambda_{n_w-n_s}(\mathbf{L}^{ideal}) \geq e^{-0.5\tau_{lb}(\epsilon)}\frac{4}{n_w^2}\end{aligned} \tag{24}$$

where we have used the facts that Δ is a graph Laplacian (and hence its smallest singular value is 0). Finally note that, due to Assumption A.1, $n_w \leq \frac{N_p}{3n_a+2n_b+1} \doteq n_o$, where N_p denotes the total number of data points. It follows that if ϵ is selected such that

$$\tau_{ub}(\epsilon) - \tau_{lb}(\epsilon) > 2\log\frac{n_o^2(n_o-1)}{2} \tag{25}$$

then n_s is given by the number of eigenvalues of $\mathbf{L}$ smaller than $\lambda(\epsilon) \doteq 2(n_o-1)e^{-0.5\tau_{ub}(\epsilon)}$. Finally, from Lemma 2 combined with (24) there exists a unitary $\mathbf{R}$ such that the eigenvectors corresponding to the n_s smallest eigenvalues of $\mathbf{L}$ and $\mathbf{L}^{ideal}$ satisfy:

$$\|\mathbf{V}_{ideal} - \mathbf{V}\mathbf{R}\|_2 \leq n_w^2(n_w-1)e^{-\frac{\tau_{ub}(\epsilon)-\tau_{lb}(\epsilon)}{2}} \tag{26}$$

It follows that if ϵ is selected such that (25) holds, then $|\mathbf{V}_{ij,ideal} - \mathbf{V}\mathbf{R}_{ij}| < 0.5$ and thus the rows of $\mathbf{V}$ can be rearranged to cluster the data (elements of the j^{th} cluster correspond to those indexes where $V_{i,j} > 0.5$). □

In order to prove Theorem 5 we need the following concentration of measure results:

Lemma 4. *[46]. Let $\{\mathbf{x}_i\}_{i=1}^N$, $\mathbf{x}_i \in \mathbb{R}^n$, be a sequence of sub-Gaussian i.i.d random vectors with zero mean and covariance $\mathbf{P}$. Then, with probability $p \geq 1 - e^{-c\sqrt{n}}$*

$$\|\frac{1}{N}\sum_{i=1}^N \mathbf{x}_i\mathbf{x}_i^T - \mathbf{P}\|_2 \leq C\sqrt{\frac{n}{N}}\|\mathbf{P}\|_2 \tag{27}$$

where c and C are universal constants.

Lemma 5. *Consider two white Gaussian sequences of length N_w, $\{\eta_i\}_{i=1}^{N_w}$ and $\{\nu_i\}_{i=1}^{N_w}$ with variances σ_η^2 and σ_ν^2 and define $\mathbf{\Sigma} \doteq \begin{bmatrix}\mathbf{H}_\eta^{n_a+1} \\ \mathbf{H}_\nu^{n_a}\end{bmatrix}\begin{bmatrix}\mathbf{H}_\eta^{n_a+1} \\ \mathbf{H}_\nu^{n_a}\end{bmatrix}^T$. Then, the following inequality holds with probability $p \geq 1 - e^{-c\sqrt{2n_a+1}}$:*

$$\|\mathbf{\Sigma} - h\begin{bmatrix}\sigma_\eta^2\mathbf{I} & \mathbf{0} \\ \mathbf{0} & \sigma_\nu^2\mathbf{I}\end{bmatrix}\|_2 \leq C(n_a+1)\sqrt{2h}\max\{\sigma_\eta^2, \sigma_\nu^2\} \tag{28}$$

where $h = N_w - n_a$.

Proof. Let $\boldsymbol{\mu}_i \doteq \begin{bmatrix}\eta_{i:i+n_a}^T \\ \nu_{i:i+n_a-1}^T\end{bmatrix}$. Then, $\mathbf{\Sigma} = \sum_{i=1}^h \boldsymbol{\mu}_i\boldsymbol{\mu}_i^T$. The vectors $\boldsymbol{\mu}_i$ are identically distributed but not independent, due to the Hankel structure, and thus Lemma 4 cannot be applied to the entire sequence. On the other hand, $\boldsymbol{\mu}_i$ and $\boldsymbol{\mu}_j$ are uncorrelated if $|j-i| > n_a$. Hence

$$\begin{aligned}&\|\mathbf{\Sigma} - h\begin{bmatrix}\sigma_\eta^2\mathbf{I} & \mathbf{0} \\ \mathbf{0} & \sigma_\nu^2\mathbf{I}\end{bmatrix}\|_2 \\ &\leq \sum_{j=1}^{n_a+1}\|\sum_{k=0}^{\frac{h}{n_a+1}-1}\boldsymbol{\eta}_{j+k(n_a+1)}\boldsymbol{\eta}_{j+k(n_a+1)}^T \\ &-\frac{h}{n_a+1}\begin{bmatrix}\sigma_\eta^2\mathbf{I} & \mathbf{0} \\ \mathbf{0} & \sigma_\nu^2\mathbf{I}\end{bmatrix}\| \leq \\ &C(n_a+1)\sqrt{2h}\|\begin{bmatrix}\sigma_\eta^2\mathbf{I} & \mathbf{0} \\ \mathbf{0} & \sigma_\nu^2\mathbf{I}\end{bmatrix}\|_2\end{aligned}$$

where the last inequality follows from Lemma 4. □

Proof of Theorem 5: For simplicity we assume $n_a = n_b$ but the proof holds in general, with minimal modifications. Consider the model (5) and the associated Gramian matrix built from the input/output sequences in an interval of length h:

$$\begin{aligned}\mathbf{G} &= \begin{bmatrix}\mathbf{H}_y \\ \mathbf{H}_u\end{bmatrix}\begin{bmatrix}\mathbf{H}_y^T & \mathbf{H}_u^T\end{bmatrix} = \\ &\begin{bmatrix}\mathbf{H}_{\tilde{y}} + \mathbf{H}_\eta \\ \mathbf{H}_{\tilde{u}} + \mathbf{H}_\nu\end{bmatrix}\begin{bmatrix}\mathbf{H}_{\tilde{y}}^T + \mathbf{H}_\eta^T & \mathbf{H}_{\tilde{u}}^T + \mathbf{H}_\nu^T\end{bmatrix} \\ &= \mathbf{G}_{ideal} + \mathbf{M} + \mathbf{\Sigma}\end{aligned}$$

where we defined

$$\begin{aligned}\mathbf{G}_{ideal} &\doteq \begin{bmatrix}\mathbf{H}_{\tilde{y}} \\ \mathbf{H}_{\tilde{u}}\end{bmatrix}\begin{bmatrix}\mathbf{H}_{\tilde{y}}^T & \mathbf{H}_{\tilde{u}}^T\end{bmatrix} \\ \mathbf{M} &\doteq \begin{bmatrix}\mathbf{H}_\eta \\ \mathbf{H}_\nu\end{bmatrix}\begin{bmatrix}\mathbf{H}_{\tilde{y}}^T & \mathbf{H}_{\tilde{u}}^T\end{bmatrix} + \begin{bmatrix}\mathbf{H}_{\tilde{y}} \\ \mathbf{H}_{\tilde{u}}\end{bmatrix}\begin{bmatrix}\mathbf{H}_\eta^T & \mathbf{H}_\nu^T\end{bmatrix} \\ \mathbf{\Sigma} &\doteq \begin{bmatrix}\mathbf{H}_\eta \\ \mathbf{H}_\nu\end{bmatrix}\begin{bmatrix}\mathbf{H}_\eta^T & \mathbf{H}_\nu^T\end{bmatrix}\end{aligned}$$

Let $\hat{\mathbf{G}} \doteq \mathbf{G}_{ideal} + \mathbf{M}$, and note that, for any $\mathbf{v} \in \mathcal{N}(\mathbf{G}_{ideal})$ we have that $\mathbf{v}^T\mathbf{M}\mathbf{v} = 0$. Since, from Assumption A3, $\sigma_{n_a+n_b}(\mathbf{G}_{ideal}) \geq \underline{\sigma}$, it follows that if $\sigma_{max}(\mathbf{M}) \leq \underline{\sigma}$, then $\hat{\mathbf{G}} \succeq 0$ and $\mathcal{N}(\mathbf{G}_{ideal}) = \mathcal{N}(\hat{\mathbf{G}})$. From Lemma 5 we have that $\mathbf{\Sigma} = h\begin{bmatrix}\sigma_\eta^2\mathbf{I} & \mathbf{0} \\ \mathbf{0} & \sigma_\nu^2\mathbf{I}\end{bmatrix} + \mathbf{\Delta}$, where, with high probability, $\|\mathbf{\Delta}\|_* \leq (2n_a+1)\|\mathbf{\Delta}\|_2 \leq 2C(n_a+1)^2\sqrt{2h}\max\{\sigma_\eta^2, \sigma_\nu^2\}$. Thus, with high probability

$$\frac{|\hat{\mathbf{G}}\|_*}{\|\mathbf{G}\|_*} \leq 1 + \frac{\|\Sigma\|_*}{\|\mathbf{G}\|_*} \leq + \frac{C_{noise}\max\{\sigma_\eta^2, \sigma_\nu^2\}}{\|\mathbf{G}\|_*} \text{ where} \tag{29}$$

$$C_{noise} \doteq 2(n_a+1)\left[h + C(n_a+1)\sqrt{2h}\right]$$

Let $\epsilon_{\max} \doteq \max\{\epsilon : \tau_{lb}(\epsilon) < \tau_{ub}(\epsilon)\}$ and $\sigma_{ub}^2 = \frac{\|\mathbf{G}\|_*\epsilon_{\max}}{C_{noise}}$. The proof follows now by applying Theorem 1 with $\mathbf{X} = \frac{\hat{\mathbf{G}}}{\|\mathbf{G}\|_*}$ and $\mathbf{\Delta}_\mathbf{x} = \frac{\mathbf{\Sigma}}{\|\mathbf{G}\|_*}$ and noting that the effect of measurement noise is equivalent to having a perturbation term $\|\Delta_\mathbf{x}\|_* \leq \frac{C_{noise}}{\|\mathbf{G}\|_*}$. Thus Algorithm 1 still produces, with high probabilty, the correct segmentation provided that the noise is small enough so that $\max\{\sigma_\eta^2, \sigma_\nu^2\} \leq \sigma_{ub}^2$. □

Mario Sznaier is currently the Dennis Picard Chaired Professor at the Electrical and Computer Engineering Department, Northeastern University, Boston. Prior to joining Northeastern University, Dr. Sznaier was a Professor of Electrical Engineering at the Pennsylvania State University and also held visiting positions at the California Institute of Technology. His research interest include robust identification and control of hybrid systems, robust optimization, and dynamical vision. Dr. Sznaier is currently serving as an associate editor for the journal Automatica. Additional recent service includes chair of the CSS Technical Committee on Computational Aspects of Control System Desig (2012-2016), General co-Chair of the 2016 MSC, Program Chair of the 2017 CDC, CSS Executive Director (2007-2011) and member of the CSS Board of Governors (2006-2014). He is a distinguished member of the IEEE Control Systems Society and an IEEE Fellow.

Xikang Zhang is currently a software engineer at Microsoft. Prior to joining Microsoft, he received his Ph.D. degree in electrical engineering from Northeastern University, Boston. Prior to that, he received his M.E. in Communication and Information System from China University of Geosciences (Beijing) and a B.E. degree in Information Engineering from Zhejiang University. His main research interests include computer vision, control and machine learning. His main research focus is activity recognition using dynamic system model. He is also interested in the problems of tracking, person re-identification, and switch system identification.

Octavia Camps Octavia Camps received a B.S. degree in computer science and a B.S. degree in electrical engineering from the Universidad de la Republica (Uruguay), and a M.S. and a Ph.D. degree in electrical engineering from the University of Washington. Since 2006, she is a Professor in the Electrical and Computer Engineering Department at Northeastern University. From 1991 to 2006 she was a faculty of Electrical Engineering and of Computer Science and Engineering at The Pennsylvania State University. Prof. Camps was a visiting researcher at the Computer Science Department at Boston University during Spring 2013 and in 2000, she was a visiting faculty at the California Institute of Technology and at the University of Southern California. She is an associate editor of Computer Vision and Image Understanding (CVIU) and IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI). Her main research interests include dynamics-based computer vision and machine learning.