A Sparsification Approach to Set Membership Identification of Switched Affine Systems

Necmiye Ozay, *Member, IEEE*, Mario Sznaier, *Member, IEEE*, Constantino M. Lagoa, *Member, IEEE*, and Octavia I. Camps, *Member, IEEE*

***Abstract*—This paper addresses the problem of robust identification of a class of discrete-time affine hybrid systems, switched affine models, in a set membership framework. Given a finite collection of noisy input/output data and some minimal *a priori* information about the set of admissible plants, the objective is to identify a suitable set of affine models along with a switching sequence that can explain the available experimental information, while minimizing either the number of switches or subsystems. For the case where it is desired to minimize the number of switches, the key idea of the paper is to reduce this problem to a sparsification form, where the goal is to maximize sparsity of a suitably constructed vector sequence. Our main result shows that in the case of ℓ_∞ bounded noise, this sparsification problem can be exactly solved via convex optimization. In the general case where the noise is only known to belong to a convex set $\mathcal{N}$, the problem is generically NP-hard. However, as we show in the paper, efficient convex relaxations can be obtained by exploiting recent results on sparse signal recovery. Similarly, we present both a sparsification formulation and a convex relaxation for the (known to be NP hard) case where it is desired to minimize the number of subsystems. These results are illustrated using two non-trivial problems arising in computer vision applications: video-shot and dynamic texture segmentation.**

***Index Terms*—Hybrid systems, piecewise affine systems, set membership identification, sparse signal recovery.**

I. Introduction and Motivation

Hybrid systems –systems characterized by the interaction of both continuous and discrete dynamics– have been the subject of considerable attention during the past decade. These systems arise naturally in many different contexts, (e.g., biological systems, systems incorporating logical and continuous elements, manufacturing, etc,) and, in addition, can be used to approximate nonlinear dynamics. As a result of this research, an extensive body of results is now available addressing issues such as controllability/observability, stability analysis and control synthesis. However, applying these results requires using an explicit model of the system under consideration. While in some cases these models can be obtained from first principles, many practical applications require identifying the system from a combination of experimental data and some *a priori* information. This has prompted a substantial research effort devoted towards developing a framework for input/output identification of hybrid systems. As a result, several methods have been proposed addressing different aspects of the problem (see the tutorial paper [1] for an excellent summary of the main issues and recent developments in the field). While successful in many situations, a common feature of these methods is the computational complexity entailed in dealing with noisy measurements: in this case algebraic procedures [2] lead to nonconvex optimization problems, while optimization methods lead to generically NP–hard problems, either necessitating the use of relaxations [3] or restricted to small size problems [4]. Similarly, methods relying on probabilistic priors [5] also lead to combinatorial problems, once again requiring the use of relaxations in order to obtain computationally tractable algorithms. An alternative approach is provided by clustering based methods [6], [7]. Since these methods rely on identification performed on local clusters, in order to work well, they require both proximity of points corresponding to the same subsystem, and "fair sampling" of each cluster, which places constraints on the admissible input sequences. Further, the clustering step entails a non–convex minimization that can potentially get trapped in local minima.

Motivated by the difficulties noted above, in the first portion of this paper we propose a new approach to the problem of set membership identification of a class of hybrid systems: switched affine models. Specifically, given noisy input/output data and some minimal *a priori* information about the set of admissible plants, our goal is to identify a suitable set of affine models along with a switching sequence that can explain the available experimental information, while minimizing either the number of switches or the number of submodels. The key idea of the proposed solution is to reduce the identification problem to a sparsification form, where the objective is to minimize the number of non–zero elements of a suitable constructed vector sequence. The main result of the paper shows that in the case where it is desired to minimize the number of switches under ℓ_∞ bounded noise –a problem that arises in several practical applications including computer vision, medical image processing and fault detection– the resulting sparsification problem can

Manuscript received May 31, 2009; revised May 31, 2010; accepted July 17, 2011. Date of publication August 30, 2011; date of current version February 29, 2012. This work was supported in part by the NSF under Grants IIS–0713003, ECCS-0901433, AFOSR Grant FA9550–09–0253 and the Alert DHS Center of Excellence under Award Number 2008-ST-061-ED0001. Recommended by Associate Editor F. Dabbene.

N. Ozay is with the Department of Computing and Mathematical Sciences, California Institute of Technology, Pasadena, CA 91125 USA (e-mail: necmiye@caltech.edu; necmiye.ozay@gmail.com).

M. Sznaier is with the Department of Electrical and Computer Engineering, Northeastern University, Boston, MA 02115 USA. (e-mail: m.sznaier@neu.edu; msznaier@ece.neu.edu).

C. M. Lagoa is with the Department of Electrical Engineering, Pennsylvania State University, University Park, PA 16802 USA. (e-mail: lagoa@engr.psu.edu).

O. I. Camps is with the Department of Electrical and Computer Engineering, Northeastern University, Boston, MA 02115 USA (e-mail: camps@neu.edu; camps@ece.neu.edu).

Corresponding author: M. Sznaier.

Digital Object Identifier 10.1109/TAC.2011.2166295

be exactly solved via convex optimization, without the need to impose additional conditions, such as the Restricted Isometry Property [8], [9]. In addition, we show that, if the switches are identifiable from the available input/output data, then the proposed algorithm converges to the true switching sequence as the noise decreases towards zero. Here the difference with existing work on identifiability of hybrid systems [10] is the explicit focus on conditions for detecting switches. In the general case where the noise is only known to belong to a convex set $\mathcal{N}$, the associated sparsification problem is generically NP–hard. However, as we show in the paper, efficient convex relaxations can be obtained by exploiting recent results on sparse signal recovery based on ℓ_1-norm minimization [8], [9]. Finally, we also consider the case where it is of interest to minimize the number of plants. While this problem is also known to be NP–hard, we show that a convex relaxation based on sparsification works well in practice, typically outperforming existing methods.

In the second part of the paper we illustrate these results using two non-trivial problems arising in computer vision applications: segmentation of video sequences and of dynamic textures. As shown there, the proposed techniques outperform existing state-of-the-art algorithms.

II. Preliminaries

Notation and Definitions

For ease of reference, the notation used in the paper is summarized below:

$\mathbb{R}, \mathbb{Z}$	set of real numbers, integers
$\mathbf{x}$	a vector in $\mathbb{R}^N$
$\|\mathbf{x}\|_p \doteq (\sum_{i=1}^n \|x_i\|^p)^{1/p}$	p-norm in $\mathbb{R}^N$, $p \in [1, \infty)$
$\|\mathbf{x}\|_\infty \doteq \max_{1 \le i \le n} \|x_i\|$	∞-norm in $\mathbb{R}^N$
$\{\mathbf{x}(t)\}_{t=1}^T$, $\{\mathbf{x}\}$	a vector valued sequence of length T where each $\mathbf{x}(t) \in \mathbb{R}^N$
$\|\{\mathbf{x}\}\|_p$	ℓ_p norm of a vector valued sequence, $1 \le p < \infty$, $\|\{\mathbf{x}\}\|_p \doteq \left(\sum_{i=1}^T \|\mathbf{x}(i)\|_p^p\right)^{1/p}$, $\|\{\mathbf{x}\}\|_\infty \doteq \max_{1 \le i \le T} \|\mathbf{x}(i)\|_\infty$
$\|\{\mathbf{x}\}\|_0$	ℓ_0-quasinorm $\doteq$ number of non-zero vectors in the sequence (i.e., cardinality of the set $\{t \| \mathbf{x}(t) \neq \mathbf{0}, t \in [1, T]\}$)
$\mathbf{I}$	identity matrix of appropriate dimension

In this paper we will consider switched autoregressive exogenous (SARX) hybrid affine models of the form

$$y(t) = \sum_{i=1}^{n_a} a_i(\sigma_t) y(t-i) + \sum_{i=1}^{n_c} c_i(\sigma_t) u(t-i) + f(\sigma_t) + \eta(t) \quad (1)$$

where u, y and η denote the input, output and noise, respectively, and where $t \in [t_o, T]$. The discrete variable $\sigma_t \in \{1, \dots, s\}$ –the mode of the system– indicates which of the s *submodels* is active at time t. The time instants where the value of σ_t changes are called *discrete transitions* or *switches*. These switches partition the interval $[t_0, T]$ into a *discrete hybrid time set* [11], $\tau = \{I_i\}_{i=0}^k$, such that σ_t is constant within each subinterval $I_i = [\tau_i, \tau_i']$ and different in consecutive intervals. In the sequel we denote by τ_i and τ_i' the beginning and ending times of the i^{th} interval, respectively. Clearly, τ satisfies

- $\tau_0 = t_0$ and $\tau_k' = T$,
- $\tau_i \le \tau_i' = \tau_{i+1} - 1$,

and the number of switches is equal to k. An equivalent representation of (1) is

$$y(t) = \mathbf{p}(\sigma_t)^T \mathbf{r}(t) + \eta(t) \quad (2)$$

where $\mathbf{r}(t) = [y(t-1), \dots, y(t-n_a), u(t-1), \dots, u(t-n_c), 1]^T$ is the regressor vector and $\mathbf{p}(\sigma_t) = [a_1(\sigma_t), \dots, a_{n_a}(\sigma_t), c_1(\sigma_t), \dots, c_{n_c}(\sigma_t), f(\sigma_t)]^T$ is the unknown coefficient vector at time t. Note that if the initial conditions are unknown, it is not possible to identify $\mathbf{p}(\sigma_t)$ for $t < \max(n_a, n_c)$, even in the absence of noise. Thus, in the sequel we will take $t_0 = \max(n_a, n_c)$.

For notational simplicity, we begin by addressing first the case of SISO systems and extend our results to the MIMO case in Section IV-D.

III. Problem Statement

In this paper, we consider the problem of identifying SARX hybrid affine models from experimental measurements corrupted by noise. From a set-membership point of view, this problem can be formally stated as follows:

Problem 1: **[Consistency]** Given input/output data over the interval $[t_0, T]$, and *a priori* information consisting of a set membership noise description $\eta \in \mathcal{N}$, compact, convex, find a coefficient vector $\mathbf{p}(\sigma_t)$ and an admissible noise sequence $\eta(t) \in \mathcal{N}$ such that (2) holds for all $t \in [t_o, T]$, or prove that no such pair exists.

It is clear that this problem is not well-posed and has infinitely many solutions. For instance, one can always find a trivial piecewise affine model with $T - t_0 + 1$ submodels or one model with a large order that perfectly fits the data. This situation can be partially avoided by imposing upper bounds n_y and n_u on the order of each of the terms on the right hand side of (1), e.g., $n_a \le n_y$ and $n_c \le n_u$ for some known n_y, n_u. Still, even in this case the problem admits multiple solutions. More interesting problems can be posed by using the existing degrees of freedom to optimize suitable performance criteria.

One such criterion is to minimize the number of switches (i.e., minimum k), subject to consistency. Practical situations where this problem is relevant arise for instance in segmentation problems in computer vision and medical image processing, where it is desired to maximize the size of regions (roughly equivalent to minimizing the number of boundaries), and in fault-detection, in cases where it is desired to minimize the number of false alarms.

The formal statement of the identification problem with this criterion is as follows:

Problem 2: **[Minimum Number of Switches]** Given input/output data over the interval $[t_0, T]$, and *a priori* information consisting of a convex set membership noise description $\mathcal{N}$ and bounds $n_u \geq n_c$ and $n_y \geq n_a$ on the order of the regressors, find a hybrid affine model of the form (1) that is consistent with the *a priori* information and that can explain the experimental data with the minimum number of switches.

An alternative is to try to find the minimum number of submodels (i.e., minimum s) capable of explaining the data record. This criterion, used in [3], leads to the following identification problem:

Problem 3: **[Minimum Number of Submodels]** Given input/output data over the interval $[t_0, T]$, and *a priori* information consisting of a noise description $\eta \in \mathcal{N}$ and bounds $n_y,\ n_u$ on the regressor orders, find a hybrid affine model of the form (1) with minimum number of submodels that is consistent with the *a priori* information and experimental data.

IV. Identification With Minimum Number of Switches as a Sparsification Problem

In this section we address Problem 2 and show that it can be reduced to a sparsification form, where the objective is to maximize the number of zero elements of a suitably defined vector valued sequence. The starting point is to consider the sequence of *first order differences* of the time varying parameters $\mathbf{p}(t)$, given by

$$\mathbf{g}(t) = \mathbf{p}(t) - \mathbf{p}(t+1). \tag{3}$$

Clearly, since a non-zero element of this sequence corresponds to a *switch*, the sequence should be sparse having only k non-zero elements out of $T - t_0$. Thus, with this definition, Problem 2 is equivalent to the following (non–convex) sparsification problem:

$$\begin{aligned} \min_{\mathbf{p}(t)} \quad & \|\{\mathbf{p}(t) - \mathbf{p}(t-1)\}\|_0 \\ \text{s.t} \quad & y(t) - \mathbf{r}(t)^T\mathbf{p}(t) \in \mathcal{N} \quad \forall t. \end{aligned} \tag{4}$$

In the sequel, we consider two different situations depending on the characterization of the noise set $\mathcal{N}$: (i) The case where $\mathcal{N}$ is a ball in ℓ_∞, and (ii) the case where $\mathcal{N}$ is a general convex set. The main result of Section IV-A shows that, in the case of ℓ_∞ bounded noise, the sparsification problem (4) can be exactly solved via convex optimization, without the need to impose additional conditions. In the case of general noise descriptions, the problem is generically NP hard. However, as we show in Section IV-C, a convex relaxation can be obtained using Lemma 2 in the Appendix. In this case, exact recovery is no longer guaranteed, unless additional conditions are satisfied. However, extensive experiments show that the convex relaxation works well in practice.

A. A Greedy Algorithm for the ℓ_∞ Case

In this section we propose a computationally simple algorithm for solving Problem 2 in the case where the noise term is characterized in terms of its ℓ_∞ norm. This solution is motivated by existing results in time series clustering showing that a greedy sliding window algorithm [12] is optimal. As we show below, similar ideas can be applied to Problem 2, leading to an algorithm that entails solving a sequence of smaller linear programs in a greedy fashion.

TABLE I
Optimal Greedy Algorithm for Problem 2.

```
Greedy Algorithm
$k = 0$
$t_0 = \max(n_y, n_u)$
$\tau_k = t_0$
FOR $i = t_0 : T$
    Solve the following feasibility problem in $\mathbf{p}$:
        $\mathcal{F} : \{\ \left|y(t) - \mathbf{r}(t)^T\mathbf{p}\right| \leq \epsilon \quad \forall t \in [\tau_k, i]\ \}$
    IF $\mathcal{F}$ is infeasible
        Set $I_k = [\tau_k, i-1]$, $k = k+1$, and $\tau_k = i$
    END IF
END FOR
Set $I_k = [\tau_k, T]$ and $\tau = \{I_j\}_{j=0}^{k}$
RETURN $\tau$ and $k$
```

Theorem 1: Let k^* denote the number of switches in an optimal solution to Problem 2 (equivalently, to the sparsification problem (4)) when the noise is characterized in terms of an ℓ_∞ bound: $\|\{\eta\}\|_\infty \leq \epsilon$. Then the value k returned by the greedy algorithm outlined in Table I coincides with the optimal k^*.

Proof: Assume $\tau^* = \{I_i^*\}_{i=0}^{k^*}$ is the discrete hybrid time set corresponding to an optimal solution with k^* switches. Let $\tau = \{I_i\}_{i=0}^{k}$ and k be the pair of values returned by the greedy algorithm. In order to establish that the proposition is true, it is enough to show that if $\tau_i \in I_j^*$ then $\tau_i' \geq {\tau_j'}^*$. Then, an induction step shows that, $\tau_i' \geq {\tau_i'}^*\ \forall i \in \{0, \ldots, k^*\}$ implying $k \leq k^*$.

Since τ^* is optimal (hence feasible), $\mathbf{p}^*(t)$ is constant in each subinterval I^*. In particular, there exists $\mathbf{p_j}$ such that for all $t \in I_j^*$, $\mathbf{p}^*(t) = \mathbf{p_j}$ and $\left|y(t) - \mathbf{r}(t)^T\mathbf{p_j}\right| \leq \epsilon$. When $\tau_i \in I_j^*$, the same $\mathbf{p_j}$ is a feasible solution of $\mathcal{F}$ in the $({\tau_j'}^*)^{th}$ iteration of the greedy algorithm since $\tau_i \in I_j^*$ implies $[\tau_i, {\tau_j'}^*] \subseteq I_j^*$. Therefore, the algorithm will continue to the next iteration without entering the if condition within the for loop, which implies $\tau_i' \geq {\tau_j'}^*$.

Next, we show by induction that for all $i \leq k$, there exists $j \geq i$ such that $\tau_i' \geq {\tau_j'}^*$, hence $\tau_i' \geq {\tau_i'}^*$:

- For $i = 0$: $\tau_0 = \tau_0^* \in I_0^* \Rightarrow \tau_0' \geq {\tau_0'}^*$.
- For $i = m$: Assume $\exists j \geq m$ s.t. $\tau_m' \geq {\tau_j'}^*$.
- For $i = m+1$: From the previous line and properties of hybrid time sets, we have that $\tau_{m+1} = \tau_m' + 1 > \tau_m' \geq {\tau_j'}^* \Rightarrow \exists l > j$ (or equivalently $\exists l \geq j+1$) s.t. $\tau_{m+1} \in I_l^* \Rightarrow \tau_{m+1}' \geq {\tau_l'}^* \geq {\tau_{j+1}'}^*$. Since $j \geq m$ implies $j+1 \geq m+1$, this proves the induction hypothesis.

Using the fact that $T = \tau_k' = {\tau_{k^*}'}^*$ and the result of the induction particularly at $i = k$ leads to $\tau_k' \geq {\tau_k'}^* \Rightarrow {\tau_{k^*}'}^* \geq {\tau_k'}^* \Rightarrow k^* \geq k$.

Since by construction the result of the greedy algorithm is feasible for problem 2 and k^* is the minimum solution of the problem, $k^* \leq k$. Therefore, $k^* = k$. ■

Remark 1: By construction, the greedy algorithm pushes the end points of each interval forward in time as much as possible (i.e., τ_i' is as large as possible). Similarly, running the algorithm backwards (i.e., $i = T : t_0$) would push the start points of intervals (equivalently, end points of previous intervals) backward in time as much as possible. Therefore, running it once backwards and once forwards, it is possible to bracket the true locations of the switches. That is, τ_i, the actual time at which the actual switch occurs satisfies $\tau_i^{back} \leq \tau_i \leq \tau_i^{fwd}$, where τ_i^{fwd} and τ_i^{back} denote the i^{th} switching time obtained running the algorithm forward and backwards, respectively.

B. Identifiability of the Switches and Convergence of the Greedy Algorithm

In this section, we address the issue of identifiability of the switches from input/output data. We first present a necessary and sufficient condition under which the switches can be exactly identified in a noiseless setup. Later, we show that when these identifiability conditions hold, the greedy algorithm given in Table I finds the exact switching times for sufficiently small noise levels (i.e., as $\epsilon \rightarrow 0$).

Definition 1: Let $\tau = \{I_i\}_{i=0}^{k}$ be a hybrid time set corresponding to a particular trajectory of a switched linear ARX system. τ is said to be *causally identifiable* if whenever $\sigma_{t-1} \neq \sigma_t$, it is possible to detect the change in the value of σ_t as soon as $y(t)$ is observed.

Definition 2: Given the current regressor vector $\mathbf{r}(t)$, two submodels with parameter vectors $\mathbf{p_1}$ and $\mathbf{p_2}$ are *one-step indistinguishable* from $\mathbf{r}(t)$ if $\mathbf{r}(t)^T(\mathbf{p_1} - \mathbf{p_2}) = 0$.

Next, we present a necessary and sufficient condition for a switching sequence to be causally identifiable from input/output data. To this effect, we need to introduce first the following preliminary result, where, for notational simplicity, we defined $\mathbf{R}_{t_0,t_1} = [\mathbf{r}(t_0), \mathbf{r}(t_0+1), \ldots, \mathbf{r}(t_1)]$, $\mathbf{Y}_{t_0,t_1} = [y(t_0), y(t_0+1), \ldots, y(t_1)]^T$ and $\mathbf{N}_{t_0,t_1} = [\eta(t_0), \eta(t_0+1), \ldots, \eta(t_1)]^T$.

Lemma 1: If $\mathbf{r}(\tau_{i+1}) \in \text{range}\left(\mathbf{R}_{\tau_i,\tau_i'}\right)$ then there exists a constant γ_i such that $\mathbf{r}^T(\tau_{i+1})\left[\mathbf{p}_{i+1} - \mathbf{p}_i\right] = \gamma_i$ for all pairs $(\mathbf{p}_i, \mathbf{p}_{i+1})$ satisfying

$$\begin{aligned} \mathbf{Y}_{\tau_i,\tau_i'} &= \mathbf{R}^T_{\tau_i,\tau_i'}\mathbf{p}_i \\ y(\tau_{i+1}) &= \mathbf{r}^T(\tau_{i+1})\mathbf{p}_{i+1}. \end{aligned} \tag{5}$$

Proof: Since $\mathbf{r}(\tau_{i+1}) \in \text{range}\left(\mathbf{R}_{\tau_i,\tau_i'}\right)$ then $\mathbf{r}^T(\tau_{i+1}) = \mathbf{v}^T\mathbf{R}^T_{\tau_i,\tau_i'}$ for some $\mathbf{v} \neq 0$. Consider now two pairs $(\mathbf{p}_i, \mathbf{p}_{i+1})$ and $(\hat{\mathbf{p}}_i, \hat{\mathbf{p}}_{i+1})$ satisfying (5). Then

$$\begin{aligned} &\mathbf{r}^T(\tau_{i+1})\left[\mathbf{p}_{i+1} - \mathbf{p}_i\right] - \mathbf{r}^T(\tau_{i+1})\left[\hat{\mathbf{p}}_{i+1} - \hat{\mathbf{p}}_i\right] \\ &\quad = \mathbf{r}^T(\tau_{i+1})\left[\hat{\mathbf{p}}_i - \mathbf{p}_i\right] = \mathbf{v}^T\left[\mathbf{R}^T_{\tau_i,\tau_i'}\hat{\mathbf{p}}_i - \mathbf{R}^T_{\tau_i,\tau_i'}\mathbf{p}_i\right] = 0, \end{aligned}$$

where the last equality follows from the first equality in (5). ■

Theorem 2: In the noise free case, $\tau = \{I_i\}_{i=0}^{k}$ is causally identifiable from input/output data if and only if the following two conditions hold for all i:

$$\mathbf{r}^T(\tau_{i+1})\left[\mathbf{p}_{i+1} - \mathbf{p}_i\right] \neq 0 \tag{6}$$

$$\mathbf{r}(\tau_{i+1}) \in \text{range}\left(\mathbf{R}_{\tau_i,\tau_i'}\right). \tag{7}$$

Proof:

Necessity: Clearly (6) is necessary for the switch to be causally identifiable. To show that (7) is also necessary, assume that it fails. Then $\mathbf{r}^T(\tau_{i+1})\mathbf{R}^{\perp}_{\tau_i,\tau_i'} \doteq \mathbf{v}^T \neq 0$, where $\mathbf{R}^{\perp}_{\tau_i,\tau_i'}$ denotes a basis for the orthogonal complement of $\mathbf{R}_{\tau_i,\tau_i'}$. Define:

$$\mathbf{p} \doteq \mathbf{p}_i + \frac{y(\tau_{i+1}) - \mathbf{r}^T(\tau_{i+1})\mathbf{p}_i}{\|\mathbf{v}\|_2^2}\mathbf{R}^{\perp}_{\tau_i,\tau_i'}\mathbf{v}.$$

Simple algebra shows that $\mathbf{p}$ satisfies $\mathbf{Y}_{\tau_i,\tau_i'} = \mathbf{R}^T_{\tau_i,\tau_i'}\mathbf{p}$ and $y(\tau_{i+1}) = \mathbf{r}^T(\tau_{i+1})\mathbf{p}$. It follows that the model

$$y(t) = \mathbf{r}^T(t)\mathbf{p} \tag{8}$$

can explain all the data in the interval $[\tau_i, \tau_{i+1}]$, and thus the switch is not causally identifiable from the input/output data alone.

Sufficiency: Since $\mathbf{r}(\tau_{i+1}) \in \text{range}\left(\mathbf{R}_{\tau_i,\tau_i'}\right)$ and $\mathbf{r}^T(\tau_{i+1})\left[\mathbf{p}_{i+1} - \mathbf{p}_i\right] \neq 0$, it follows, from Lemma 1, that there does not exist a single $\mathbf{p}$ such that (8) holds for all $t \in [\tau_i, \tau_{i+1}]$. Hence the switch is causally identifiable from the input/output sequences $\{\mathbf{u}, y\}$. ■

Remark 2: The result above formalizes the intuition that a switch is causally identifiable if and only if the two modes involved are not one–step indistinguishable and no new modes of the present model have been excited at the last time step (condition (7)). In addition, it can be shown that conditions (6)–(7) are equivalent to

$$\text{rank}[\mathbf{Y}_{\tau_i,\tau_{i+1}}\ \mathbf{R}^T_{\tau_i,\tau_{i+1}}] > \text{rank}[\mathbf{R}^T_{\tau_i,\tau_{i+1}}]. \tag{9}$$

However, since rank is fragile to arbitrarily small perturbations, the former lead to a better approach for handling noise.

Theorem 3: If conditions (6)–(7) hold, then there exists a noise level ϵ_0 such that greedy algorithm correctly identifies the hybrid time set τ from the noisy trajectories, whenever the noise level ϵ is below ϵ_0.

Proof: In order to show that the greedy algorithm correctly identifies the hybrid time set τ, we need to show that

$$\left|y(\tau_{i+1}) - \mathbf{r}(\tau_{i+1})^T\mathbf{p}\right| > \epsilon \tag{10}$$

for all $\mathbf{p}$ such that

$$\mathbf{R}^T_{\tau_i,\tau_i'}\mathbf{p} = \mathbf{Y}_{\tau_i,\tau_i'} + \mathbf{N}_{\tau_i,\tau_i'} \tag{11}$$

or, equivalently $\left|\mathbf{r}(\tau_{i+1})^T\mathbf{p}_{i+1} - \mathbf{r}(\tau_{i+1})^T\mathbf{p}\right| \geq 2\epsilon$ for all $\mathbf{p}$ that satisfy (11). From condition (7), it follows that $\mathbf{r}(\tau_{i+1}) = \mathbf{R}_{\tau_i,\tau_i'}\boldsymbol{\lambda}$ for some $\boldsymbol{\lambda} \neq 0$, $\|\lambda\|_2$ finite, and $\|\mathbf{r}(\tau_{i+1})\|_2 \geq \underline{\sigma}_{R_i}\|\boldsymbol{\lambda}\|_2$, where $\underline{\sigma}_{R_i}$ denotes the smallest (non–zero) singular value of $\mathbf{R}_{\tau_i,\tau_i'}$. It follows that, for all $(\mathbf{p}, \hat{\mathbf{p}})$ that satisfy (11) we have

$$\mathbf{r}(\tau_{i+1})^T(\mathbf{p} - \hat{\mathbf{p}}) = \boldsymbol{\lambda}^T\mathbf{R}^T_{\tau_i,\tau_i'}(\mathbf{p} - \hat{\mathbf{p}}) = \boldsymbol{\lambda}^T\left(\mathbf{N}_{\tau_i,\tau_i'} - \hat{\mathbf{N}}_{\tau_i,\tau_i'}\right).$$

Hence

$$\begin{aligned} \left|\mathbf{r}(\tau_{i+1})^T(\mathbf{p} - \hat{\mathbf{p}})\right| &\leq \|\lambda\|_2\sqrt{(\tau_i - \tau_{i+1})}2\epsilon \\ &\leq 2\frac{\|\mathbf{r}(\tau_{i+1})\|_2\sqrt{(\tau_i - \tau_{i+1})}}{\underline{\sigma}_{R_i}}\epsilon \doteq b(\epsilon). \end{aligned} \tag{12}$$

From (12) and the fact that (6) implies that $\left|\mathbf{r}(\tau_{i+1})^T\mathbf{p}_{i+1} - \mathbf{r}(\tau_{i+1})^T\mathbf{p}_i\right| = \gamma_i > 0$, it follows that, if the noise level ϵ satisfies

$$\epsilon < \epsilon_o \doteq \min_i \frac{\underline{\sigma}_{R_i}\gamma_i}{2\left(\|\mathbf{r}(\tau_{i+1})\|_2\sqrt{(\tau_i - \tau_{i+1})} + \underline{\sigma}_{R_i}\right)} \tag{13}$$

then (10) holds for all τ_i and hence all switches will be correctly detected by the greedy algorithm. Note that, since we are working over finite horizons, $\epsilon_o > 0$. ■

It is important to note that in the case of ℓ_∞-norm bounded noise, our results do not explicitly depend on the level of sparsity of the sequence $\mathbf{p}(t) - \mathbf{p}(t-1)$. The greedy algorithm finds the solution that would explain the data with minimum number of switches even if the sequence $\mathbf{p}(t) - \mathbf{p}(t-1)$ is not too sparse. Moreover, the solution found by the greedy algorithm corresponds to the true switch sequence whenever the conditions given in Theorem 2 and 3 hold. On the other hand, if these conditions fail, no algorithm can causally identify the true switches.

The following examples illustrate some nontrivial facts about identifiability of the switches and provide further insight into the results of this section.

Example 1: This example illustrates the fact that dwell–time constraints are not necessary for identifiability of the switches. Consider three autonomous systems ($\sigma_t \in \{1,2,3\}$) of the form

$$y_t = a_1(\sigma_t)y_{t-1} + a_2(\sigma_t)y_{t-2} + a_3(\sigma_t)y_{t-3}$$

with

$$\begin{aligned}
\mathbf{p}_1 &= [a_1(1), a_2(1), a_3(1)] = [-3, 2, 1]\\
\mathbf{p}_2 &= [a_1(2), a_2(2), a_3(2)] = \left[\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right]\\
\mathbf{p}_3 &= [a_1(3), a_2(3), a_3(3)] = [2, -1, 1]\\
\text{and}\qquad \sigma_t &= \begin{cases} 1, & t \in [1,4]\\ 2, & t = 5\\ 3, & t = 6\end{cases}.
\end{aligned}$$

The trajectory corresponding to the initial conditions $y_0 = 0, y_{-1} = 7, y_{-2} = -12$, is given by $2, 1, 1, 1, 1, 2$. Thus, the rank condition (9) evaluated at $t = 6$ yields

$$\operatorname{rank}\begin{bmatrix}1 & 1 & 1 & 1\\ 2 & 1 & 1 & 1\end{bmatrix} = \operatorname{rank}\begin{bmatrix}1 & 1 & 1\\ 1 & 1 & 1\end{bmatrix} + 1$$

which implies that it is possible to detect the switch from $t = 5$ to $t = 6$ although the system remains in $\sigma_t = 2$ for only one time instant.

Example 2: The goal of this example is to illustrate that noiseless switch identifiability does not imply that mode switches are identifiable under arbitrarily small noise. To this effect consider a system with 2 submodels: the first corresponds to $\mathbf{p}_1 = [1/3 \quad 1/3 \quad 1/3]$, and is active for t = 1,2. The second corresponds to $\mathbf{p}_2 = [1 \quad -1 \quad 2]$ and is active for t = 3. The trajectory corresponding to the initial conditions $\mathbf{r}(1) = [1 \quad 1 \quad 1]$ and no external input is given by $y(1) = 1$, $y(2) = 1$ and $y(3) = 2$. In this case the associated matrices satisfy

$$\operatorname{rank}\begin{bmatrix}1 & 1 & 1 & 1\\ 1 & 1 & 1 & 1\\ 2 & 1 & 1 & 1\end{bmatrix} > \operatorname{rank}\begin{bmatrix}1 & 1 & 1\\ 1 & 1 & 1\\ 1 & 1 & 1\end{bmatrix}.$$

Hence the switch is causally identifiable. However, adding the noise sequence $\eta(1) = \epsilon$, $\eta(2) = 0$, $\eta(3) = 0$ leads to the trajectory: $y(1) = 1 + \epsilon$, $y(2) = 1 + \epsilon/3$, $\mathrm{y}(3) = 2 - 2(\epsilon/3)$. In this case, for any $\epsilon > 0$ the corresponding matrices satisfy

$$\operatorname{rank}\begin{bmatrix}1+\epsilon & 1 & 1 & 1\\ 1+\frac{\epsilon}{3} & 1+\epsilon & 1 & 1\\ 2-\frac{2\epsilon}{3} & 1+\frac{\epsilon}{3} & 1+\epsilon & 1\end{bmatrix} = \operatorname{rank}\begin{bmatrix}1 & 1 & 1\\ 1+\epsilon & 1 & 1\\ 1+\frac{\epsilon}{3} & 1+\epsilon & 1\end{bmatrix}.$$

Hence, the switch is not identifiable, regardless of how small ϵ is. This is due to the fact that in this case condition (7) fails since

$$\mathbf{r}(3) = [1+\epsilon/3 \quad 1+\epsilon \quad 1]^T \notin \operatorname{range}\left(\begin{bmatrix}1 & 1 & 1\\ 1+\epsilon & 1 & 1\end{bmatrix}^T\right).$$

C. The Case of General Convex Noise Descriptions

In the case of general noise descriptions $\eta \in \mathcal{N}$, all samples are coupled through the noise description. For instance, a noise description of the form $\|\eta\|_p \le \epsilon$, $p \ne \infty$, requires considering all elements of the noise sequence at once. Thus, batch algorithms that consider all available data must be used, as opposed to the greedy one used in the ℓ_∞ case. As we show next, in this case problem (4) can be relaxed to a convex optimization using the tools described in the Appendix. The starting point is to define the new variables $\mathbf{z}(0) = \mathbf{p}(\mathbf{0})$, $\mathbf{z}(t) = \mathbf{p}(t) - \mathbf{p}(t-1)$, $t \ge 1$. Further, without loss of generality, it can be assumed that $\mathbf{p}(0) \ne 0$, since $\mathbf{p}(0) = 0$ corresponds to the pathological case where the initial data consist purely of measurement noise. Under these conditions, minimizing $\|\{\mathbf{p}(t) - \mathbf{p}(t-1)\}\|_0$ and $\|\{\mathbf{z}(t)\}\|_0$ leads to the same optimal sequence. Moreover, if the noise description $\mathcal{N}$ is given in terms of a norm bound, then the constraints in (4) can be expressed as $\|\mathbf{A}\mathbf{z} + \mathbf{b}\|_* \le \epsilon$, where $\mathbf{A}$ is a matrix formed using the regressors $\mathbf{r}(t)$ and $\mathbf{b}$ is a vector formed by concatenating $y(t)$. Thus, problem (4) can be rewritten as

$$\begin{aligned}
\min_{\mathbf{z}(t)} \quad & \|\{\mathbf{z}(t)\}\|_0\\
\text{s.t} \quad & \|\mathbf{A}\mathbf{z} + \mathbf{b}\|_* \le \epsilon
\end{aligned} \tag{14}$$

which is in the standard form of a sparse recovery problem with structured sparsity, similar to those in [13]–[15]. Indeed proceeding as in there, conditions can be developed guaranteeing that minimizing an appropriate convex surrogate recovers the sparsest solution [16]. For cases where these conditions do not hold, from Lemma 2 in the Appendix, it follows that replacing $\|.\|_o$ by $\|.\|_1$ yields the tightest convex relaxation of the objective. Further, a better heuristic can be obtained by adapting to this case the iterative weighted ℓ_1 relaxation proposed in [17]–[19]. This requires solving, at each iteration, the following convex program:

$$\begin{aligned}
\text{minimize}_{v,z} \quad & \sum_t w_t^{(k)} v_t\\
\text{subject to} \quad & \|\mathbf{z}(t)\|_\infty \le v_t \quad \forall t\\
& \|\mathbf{A}\mathbf{z} + \mathbf{b}\|_* \le \epsilon
\end{aligned} \tag{15}$$

where $w_t^{(k)} = (v_t^{(k)} + \delta)^{-1}$, $v_t^{(k)}$ denotes the optimal solution at the k^{th} iteration, with $w^{(0)} = [1, 1, \ldots, 1]^T$, and where δ is a (small) regularization constant. In the first iteration, this method solves the standard ℓ_1-norm relaxation. Then at each subsequent iteration, it increases the weight $w_t^{(k)}$ associated with the small $v_t^{(k)}$s, thus pushing these elements further towards zero. Note in passing that, except in cases where the initial data consist purely of measurement noise, then $z(0) \neq 0$. Thus, only the sequence $\|z(t)\|_\infty,\ t = 1, N$ needs to be sparsified which allows for setting $w_0 = 0$ in (15).

Remark 3: Algorithm (15) requires solving m linear programs with $(n_y + n_u + 2) \times (T - t_0 + 1)$ variables and $2(n_y + n_u + 2) \times (T - t_0 + 1)$ inequality constraints, where m is the number of iterations required for convergence of the weighted ℓ_1-norm relaxation, typically around 5. On the other hand, the greedy algorithm requires solving $(T - t_0 + 1)$ linear programs with only $(n_y + n_u + 1)$ variables and at most $2(T - t_0 + 1)$ inequality constraints (the worst case scenario is when a single parameter value is feasible for the entire $[t_0, T]$ interval). Thus, in cases where both algorithms are applicable (e.g., when the noise is characterized in terms of its ℓ_∞ norm), the greedy algorithm is preferable from a computational complexity standpoint.

D. Extension To Multi-Input Multi-Output Models

It is straightforward to extend the sparsity based identification procedure with minimum number of switches criterion to multi–input multi–output (MIMO) models. Consider the MIMO switched ARX model with m_u inputs and m_y outputs

$$y(t)=\sum_{i=1}^{n_a} A_i(\sigma_t)y(t-i)+\sum_{i=1}^{n_c} C_i(\sigma_t)u(t-i) + f(\sigma_t)+\eta(t) \tag{16}$$

where $y \in \mathbb{R}^{m_y}$, $u \in \mathbb{R}^{m_u}$ are outputs and inputs, $A_i \in \mathbb{R}^{m_y \times m_y}$, $C_i \in \mathbb{R}^{m_y \times m_u}$ and $f \in \mathbb{R}^{m_y}$ are coefficient matrices, and $\eta \in \mathbb{R}^{m_y}$ is the noise, respectively. It is possible to solve for coefficient matrices in a similar manner as in (15). Thus, only the following modifications are required: (i) defining time varying coefficient matrices (i.e., $A_i(t)$, $C_i(t)$ and $f(t)$), (ii) forming $\mathbf{p}(t) \in \mathbb{R}^{m_y^2+m_y m_u+m_y}$ by stacking the elements of the coefficient matrices at time t into a column vector, and (iii) replacing the regressor equation in (15) with the multivariate regressor corresponding to(16).

E. Extension to Multidimensional Models

In this section we consider the identification of hybrid multidimensional systems (i.e., hybrid systems where the process dynamics depend on more than one indeterminate). In particular, we consider systems that are governed by *affine switched-coefficient difference equations* (ASCDEs). An n-dimensional ASCDE (a generalization of standard linear constant-coefficient difference equations) has the following form [20]:

$$\begin{aligned} y(t_1, \ldots, t_n) &= \sum_{(k_1,\ldots,k_n)\in\mathcal{R}_a} a_{k_1,\ldots,k_n}(\sigma_{t_1,\ldots,t_n}) y(t_1 - k_1, \ldots, t_n - k_n) \\ &+ \sum_{(k_1,\ldots,k_n)\in\mathcal{R}_c} c_{k_1,\ldots,k_n}(\sigma_{t_1,\ldots,t_n}) u(t_1 - k_1, \ldots, t_n - k_n) \\ &+ f(\sigma_{t_1,\ldots,t_n}) + \eta(t_1, \ldots, t_n) \end{aligned} \tag{17}$$

where y is the output; u is the input; η is noise; $\sigma_{t_1,\ldots,t_n} \in \{1, \ldots, s\}$ is the discrete mode signal as before. $\mathcal{R}_a, \mathcal{R}_c \subset \mathbb{Z}^n$ are coefficient support regions; and $a_{k_1,\ldots,k_n}(\sigma_{t_1,\ldots,t_n})$, $c_{k_1,\ldots,k_n}(\sigma_{t_1,\ldots,t_n})$ and $f_{k_1,\ldots,k_n}(\sigma_{t_1,\ldots,t_n})$ are the coefficients to be identified.

In particular, 3-D models of this form in noise-free setup were considered in [21] for spatiotemporal segmentation. Such a model can be useful in approximating the behavior of a wave traveling in an inhomogeneous space or images exhibiting regions with different textures.

In the sequel, as a shorthand notation, we denote the indeterminate in vector form, i.e., $\mathbf{t} = [t_1, \ldots, t_n] \in \mathbb{Z}^n$. Let $\mathcal{D} \subset \mathbb{Z}^n$ denote the domain over which experimental measurements are collected. Characterizing the interior of the domain as

$$\operatorname{int}(\mathcal{D}) = \{\mathbf{t} \in \mathcal{D} |\ \mathbf{t} - \mathbf{k} \in \mathcal{D}\ \forall \mathbf{k} \in \mathcal{R}_a\ \forall \mathbf{k} \in \mathcal{R}_c\} \tag{18}$$

allows for defining the set of neighboring indices as the following set of unordered pairs:

$$\mathcal{I} = \left\{ \{\mathbf{t}, \tilde{\mathbf{t}}\} \mid \left\|\mathbf{t} - \tilde{\mathbf{t}}\right\|_1 = 1, \text{ and } \mathbf{t}, \tilde{\mathbf{t}} \in \operatorname{int}(\mathcal{D}) \right\}. \tag{19}$$

In this context, a *switch* is defined between neighboring indices. That is, we say that there is a *switch* whenever $\sigma_{\mathbf{t}} \neq \sigma_{\tilde{\mathbf{t}}}$ for $\{\mathbf{t}, \tilde{\mathbf{t}}\} \in \mathcal{I}$ (analogous to the 1D case where switches are defined between t and $t+1$). A multidimensional hybrid "time" set is a partition $\{P_i\}$ of $\operatorname{int}(\mathcal{D})$ such that within each part (where we call the elements of partition as parts or segments) $\sigma_{\mathbf{t}}$ is constant and it is different between neighboring parts (P_i and P_j are called neighboring parts if there exists $\{\mathbf{t}, \tilde{\mathbf{t}}\} \in \mathcal{I}$ such that $\mathbf{t} \in P_i$ and $\tilde{\mathbf{t}} \in P_j$).

As in the 1D case, identification of a system of the form (17) is ill-posed since for example one can choose a partition where each part consists of a single $\mathbf{t}$. We are interested in finding a partition with minimum number of switches (this corresponds to minimizing the boundary of the segments in image segmentation problem). In order to minimize the number of switches, one should consider sparsifying the following difference sequence:

$$\mathbf{g}(i) = \mathbf{p}(\mathbf{t}) - \mathbf{p}(\tilde{\mathbf{t}}),\ \{\mathbf{t}, \tilde{\mathbf{t}}\} \in \mathcal{I} \tag{20}$$

where $\mathbf{p}(\mathbf{t}) = [a_{\mathbf{k}_1}(\sigma_{\mathbf{t}}), \ldots, a_{\mathbf{k}_{n_a}}(\sigma_{\mathbf{t}}), c_{\mathbf{k}_1}(\sigma_{\mathbf{t}}), \ldots, c_{\mathbf{k}_{n_c}}(\sigma_{\mathbf{t}}),\ f(\sigma_{\mathbf{t}})]$, and $i = 1, \ldots, |\mathcal{I}|$ is an index counting the elements of $\mathcal{I}$. Then, the identification problem can be written as

$$\begin{aligned} \min_{\mathbf{p}(\mathbf{t})} \quad & \|\{\mathbf{g}(i)\}\|_0 \\ \text{s.t} \quad & \text{Equation (17)} \quad \text{and } \eta \in \mathcal{N}. \end{aligned} \tag{21}$$

which can be solved, exactly as in the 1D case, using a weighted ℓ_1 norm relaxation.

V. Identification With Minimum Number of Submodels

In many cases of practical importance, it is of interest to find an SARX model that explains the data with the minimum number of submodels, rather than switches. For example, in some medical image processing problems the goal is to segment images in just two groups: healthy and diseased tissue. Similarly, in activity recognition applications, it is of interest to segment video clips into the minimum number of sub-activities, bearing in mind that frames corresponding to the same activity are not necessarily contiguous (e.g., a clip can consist of a person alternating between just two activities). Finally, the minimum number of submodels description provides an alternative for cases where the identifiability conditions of Theorems 2 or 3 fail, but the system is a-priori known to switch among a small number of submodels. The resulting optimization problem is related to partitioning a system of linear equations into the minimum number of feasible subsystems (MIN PFS problem [22]). Since it is well known that finding even an approximate solution to the MIN PFS problem is NP-Hard, [22] proposed a relaxation based on finding, at each step, a single vector that renders the maximum number of equations feasible (known as MAX FS problem). While this latter problem is still NP-Hard, it can be approximately solved using a thermal relaxation [22] (also adopted in [3]). In this section, motivated by the ideas in [3], [22], we provide an alternative solution to the MAX FS problem by recasting it into a sequence of sparsification problems. The main idea is to find one submodel at a time, along with the associated parameter vector $\tilde{\mathbf{p}}$, through the solution of a sparsification problem. This is accomplished by finding a parameter vector $\tilde{\mathbf{p}}$ that makes $\left[y(t)-\mathbf{r}(t)^T\tilde{\mathbf{p}}\right] \in \mathcal{N}$ feasible for as many time instants t as possible. Equivalently, defining $\tilde{\mathbf{g}}(t)=\mathbf{p}(t)-\tilde{\mathbf{p}}$, the goal is to maximize sparsity of $\tilde{\mathbf{g}}(t)$ leading to the following optimization problem:

$$\begin{aligned} \min_{\mathbf{p}(t),\tilde{\mathbf{p}}} \quad & \|\{\mathbf{p}(t)-\tilde{\mathbf{p}}\}\|_0 \\ \text{s.t} \quad & \left[y(t)-\mathbf{r}(t)^T\mathbf{p}(t)\right] \in \mathcal{N} \quad \forall t. \end{aligned} \tag{22}$$

Then, we can eliminate the time instants t for which $\tilde{\mathbf{g}}(t)$ is zero, and solve the same problem with the rest of the t's until all data points are clustered. The number of times (22) is solved gives an upper bound on the minimum number of submodels s. Combining this idea with a refinement step similar to the one proposed in [3] to re-estimate parameter values and reassign, if needed, data points, leads to the overall algorithm listed in Table II, where $\|.\|_0$ is (approximately) minimized by minimizing a weighted ℓ_1 norm surrogate.

Remark 4: Counterexamples are available where our algorithm overestimates the number of systems since MIN PFS and a sequence of MAX FS problems are not, in general, equivalent. Due to its greedy nature, MAX FS tends to assign as many points as possible to the parameters found earlier, possibly resulting in the later need to use additional parameter values in order to explain unassigned data points. Nevertheless, consistent numerical experience shows that MAX FS is a good alternative for MIN PFS. A geometric intuition as to under which conditions solving a sequence of MAX FS problems is indeed optimal for the MIN PFS problem can be found in [22]. It is worth mentioning that our alternative method for solving MAX FS, based on a weighted ℓ_1 minimization, works, in general, better than the thermal relaxation of [22] and [3]. This is illustrated in Section VI using a large number of random instances of Problem 3.

TABLE II
ALGORITHM FOR PROBLEM 3.

```
Algorithm for Minimum # of Submodels
$t_0 = \max(n_y, n_u)$
$N_1 = \{t_0, \ldots, T\}$
$l = 0$
WHILE $N_{l+1} \neq \emptyset$
  Let $l = l + 1$
  Find $\tilde{\mathbf{p}}_l$ by solving the re–weighted $\ell_1$ optimization:
    $\min_{z_t,\mathbf{p}(t),\tilde{\mathbf{p}}}$   $\sum_t w_t^{(k)} z_t$
    subject to   $\|\mathbf{p}(t)-\tilde{\mathbf{p}}\|_\infty \leq z_t$
                 $\left[y(t)-\mathbf{r}(t)^T\mathbf{p}(t)\right] \in \mathcal{N}$
                 $\forall t \in N_l$
    where $w_j^{(k)} = (z_j^{(k)} + \delta)^{-1}$ are weights with
    $z_j^{(k)}$ the arguments of the optimal solution in
    $k^{th}$ iteration and $\mathbf{w}^{(0)} = [1,1,\ldots,1]^T$; and $\delta$
    is the regularization constant.
  Let $i = 1$
  WHILE $i < l$
    Let $K_{il} = \left\{t \in N_i : \left[y(t)-\mathbf{r}(t)^T\mathbf{p}(t)\right] \in \mathcal{N}\right\}$
    IF $\#K_{il} > \#K_i$
      Let $\tilde{\mathbf{p}}_i = \tilde{\mathbf{p}}_l$ and $l = i$
    END IF
    Let i = i+1
  END WHILE
  Let $K_l = \left\{t \in N_l : \left[y(t)-\mathbf{r}(t)^T\mathbf{p}(t)\right] \in \mathcal{N}\right\}$
  Let $N_{l+1} = N_l \setminus K_l$
END WHILE
RETURN $s = l$ and $K_i, i = 1, \ldots, s$
```

VI. EXAMPLES

In this section, we provide simulation examples demonstrating the effectiveness of the proposed algorithms.

Example 3: In this example, we consider input/output data generated by a hybrid system that switches among two ARX submodels. For $t \in [1,25] \cup [51,75]$, the submodel

$$y(t) = 0.2y(t-1) + 0.24y(t-2) + 2u(t-1) + \eta(t)$$

is active; and for $t \in [26,50] \cup [76,100]$,

$$y(t) = -1.4y(t-1) - 0.53y(t-2) + u(t-1) + \eta(t)$$

is active with $\|\eta\|_\infty = 0.5$. The goal here is to identify a model that explained the experimental data record with the fewest possible number of switches. Fig. 1 compares the performance of sparsification-based (both the ℓ_1-based algorithm (4) and the greedy algorithm of Table I) against the algebraic method and the bounded error method. As shown there, the sparsification based methods correctly estimated the parameters and number of switches, while the other two failed to do so. The running times for ℓ_1-based, greedy, algebraic and bounded error methods are 5.9, 40.5, 1.1 and 17.9 seconds respectively. Additional examples illustrating the use of sparsification to find the minimum number of switches are given in Section VII.

For this example, ϵ_o in (13) was found to be 0.3136 which corresponds to 4.15% of the maximum absolute value of the output $y(t)$. Thus, in this case (13) does not hold, since the noise level $\epsilon = 0.5 \geq \epsilon_0$. Nevertheless, the greedy algorithm was able to correctly detect the switches. This is due to the fact that the analysis in Section IV-B is worst-case, in the sense that, for

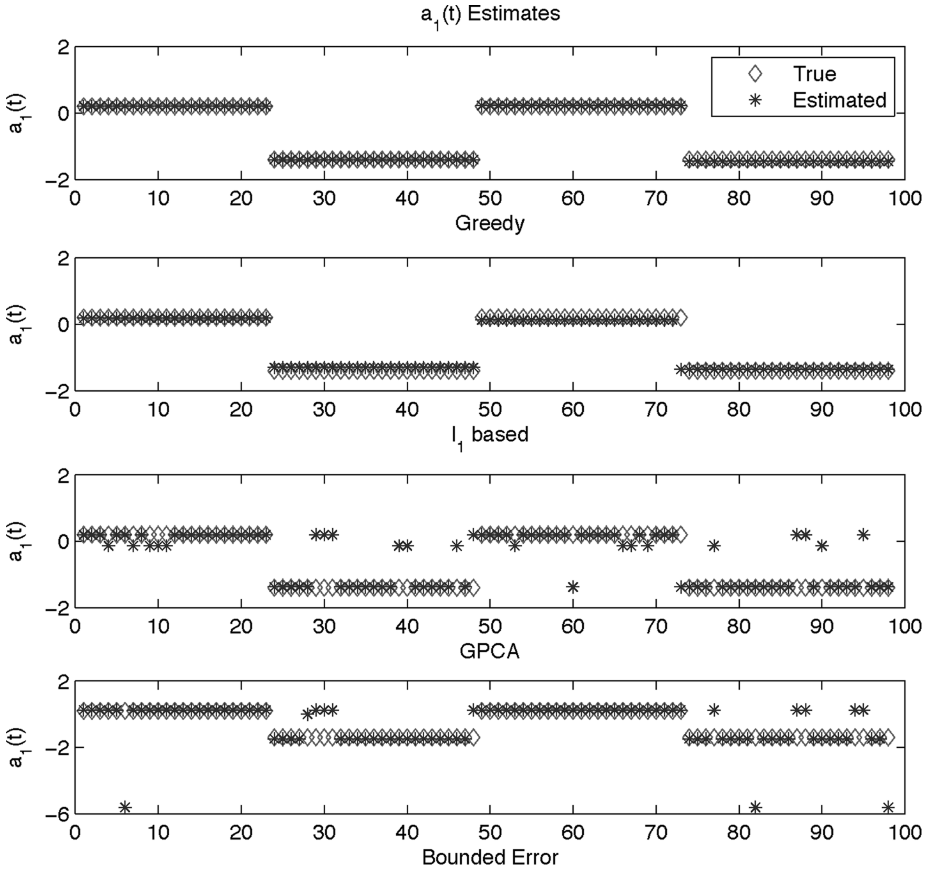


Fig. 1. True and estimated parameter sequences for parameter $a_1(\sigma_t)$ for Example 3.

all noise values below $\epsilon_o = 0.3136$, the greedy algorithm is guaranteed to find the correct switches. However, this analysis does not rule out the possibility of the algorithm finding out the correct switches for noise levels above ϵ_0, for noise sequences other than the worst–case one, as is the case here.

Example 4: This example considers the problem of estimating the minimum number of subsystems and investigates the effects of noise level on algorithm performance. The data used corresponded to the trajectories of 100 randomly generated SARX models of the form

$$y(t)=a_1(\sigma_t)y(t-1)+a_2(\sigma_t)y(t-2)+c_1(\sigma_t)u(t-1)+\eta(t) \tag{23}$$

with

$$\sigma_t = \begin{cases} 1, & t \in [1, 60] \\ 2, & t \in [61, 120] \\ 3, & t \in [121, 180] \end{cases}$$

where for all $i \in \{1, 2, 3\}$, $c_1(i)$ is a sample from a zero mean unit variance normal distribution, $a_1(i)$ and $a_2(i)$ are chosen such that the complex conjugate poles of the i^{th} submodel are distributed in $0.5 \leq ||z|| \leq 1$ with uniform random phase and magnitude, and $\eta(t)$ is an iid noise term uniformly distributed in $[-\epsilon, \epsilon]$. For each of these systems, the number of submodels was estimated by solving the minimum submodels problem with our method and the bounded-error method; and by approximating the rank of an appropriate matrix obtained from data as proposed in [23] for the algebraic method. The former two methods give upper bounds of true value $s = 3$, whereas the latter estimate depends on the threshold chosen to calculate the rank and could be lower than the true value. The same experiment was repeated for different noise levels. Results on these experiments are summarized in Fig. 2 and Table III.

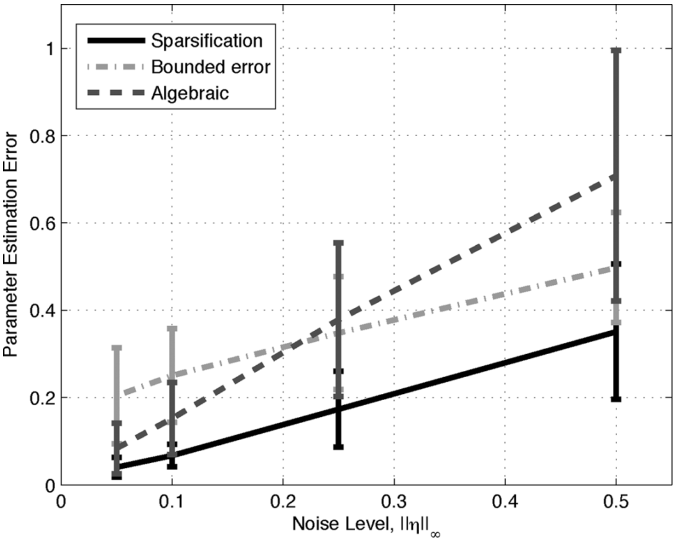
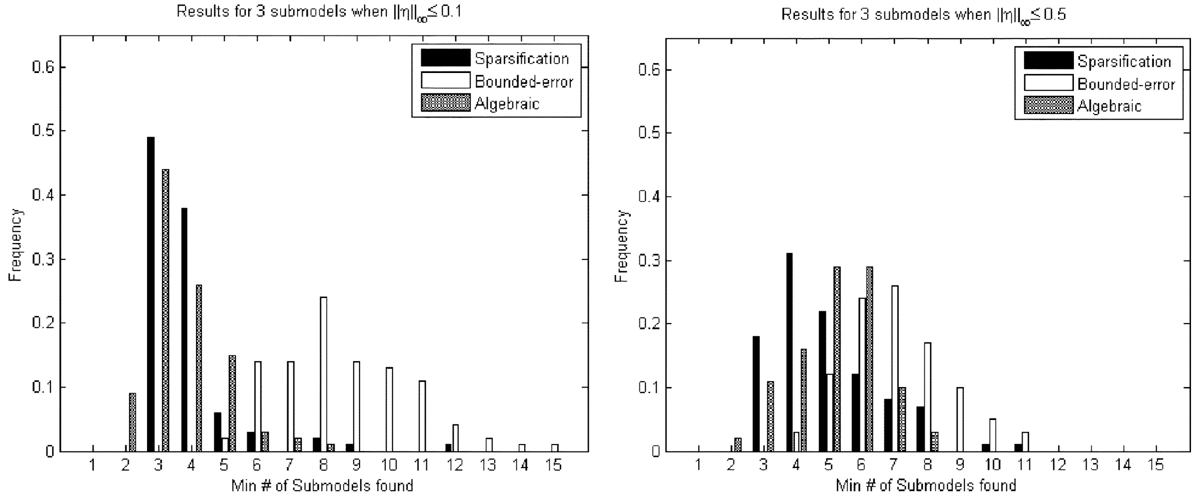


Fig. 2. Median of parameter estimation error Δ_n versus noise level ϵ. Error bars indicate the median absolute deviation.



(a) (b)

Fig. 3. Each histogram shows the frequency of estimated number of submodels for different noise levels. (a) $\epsilon = 0.15$, (b) $\epsilon = 0.5$. The true number of submodels is $s = 3$.

TABLE III
MINIMUM NUMBER OF SUBMODEL ESTIMATION ERROR STATISTICS FOR DIFFERENT NOISE LEVELS.

Noise Level ϵ	Absolute Error	Sparsification	Bounded-Error	Algebraic
0.1	Mean	0.84	5.64	0.87
	Standard deviation	1.36	2.03	1.02
	Median	1	5	1
	Median absolute deviation	1	1	1
0.5	Mean	1.93	4.07	2.18
	Standard deviation	1.65	1.58	1.25
	Median	2	4	2
	Median absolute deviation	1	1	1

TABLE IV
NORMALIZED PARAMETER IDENTIFICATION ERROR STATISTICS FOR THE MINIMUM NUMBER OF SUBMODELS PROBLEM WITH DIFFERENT NOISE LEVEL.

Noise Level ϵ	Δ_n	Sparsification	Bounded-Error	Algebraic
0.1	Mean	0.11	1.72	0.79
	Standard deviation	0.18	10.90	4.30
	Median	0.06	0.25	0.15
	Median absolute deviation	0.02	0.10	0.08
0.5	Mean	0.41	0.61	1.05
	Standard deviation	0.27	0.50	1.43
	Median	0.35	0.49	0.70
	Median absolute deviation	0.15	0.12	0.28

TABLE V
MEAN CPU TIMES OVER 100 RUNS IN EACH CASE (S:SECOND, MS:MILLISECOND).

Noise Level ϵ	Sparsification	Bounded-Error	Algebraic
0.1	205s	51s	18ms
0.5	76s	26s	39ms

Next we consider the parameter estimation accuracy for the same 100 random systems. To this end, the following normalized parameter identification error measure is defined:

$$\Delta_n = \frac{1}{T - t_0 + 1} \sum_{t=t_0}^{T} \frac{\|\mathbf{p}(\sigma_t) - \hat{\mathbf{p}}(\hat{\sigma}_t)\|_2}{\|\mathbf{p}(\sigma_t)\|_2}. \tag{24}$$

The parameter estimation results are summarized in Fig. 3 and Table IV. As shown there, the sparsification–based method outperformed both the bounded-error and algebraic procedures. While all methods proved considerably robust to noise in estimating the number of submodels, segmentation quality and parameter identification performance degraded significantly for the algebraic method as the noise level increased. On the other hand, sparsification was the most robust in terms of these performance criteria. The bounded-error method performed relatively poorly when estimating the number of submodels. Even though it clustered most of the data in the largest three submodels, it also generated superfluous submodels with parameter values far from the true values.

Finally, Table V summarizes the mean computation time for this set of simulations. As show there, algebraic method is the fastest. However, due to scalability issues that will be illustrated in Example 5, it applicability is restricted to relatively small data sets. Among bounded-error and sparsification, there is a trade-off between accuracy and computation time. It is also worth mentioning that it is not possible to solve problems of this size on the same machine using mixed integer programming (MIP) [4] since its complexity grows exponentially.

Example 5: This large scale example again considers the minimum number of systems problem and investigates the scalability of the different algorithms. The data was generated using a switched linear system with three submodels each having eight poles and four zeros. The mode signal was set to

$$\sigma_t = \begin{cases} 1, & t \in [1, 1000] \\ 2, & t \in [1001, 2000] \\ 3, & t \in [2001, 3000] \end{cases}.$$

For this example, the algebraic method failed due to insufficient memory since its complexity grows exponentially with the number of submodels. The running times for the sparsification method and bounded error methods were 74 minutes and 54 minutes, respectively.

Example 6: Textured Image Segmentation: The goal of this example is to illustrate the use of the proposed method to segment textured images. To this effect we combined two different textures to generate the two images shown in Fig. 4. In order to recast the segmentation problem into a hybrid system identification form, the grey-scale values of the pixels in each image were

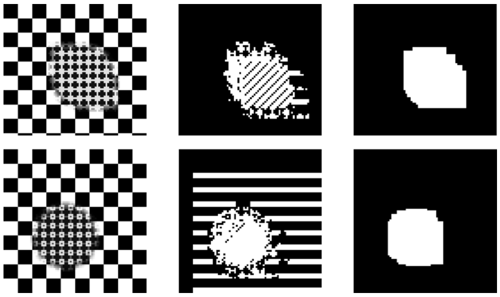


Fig. 4. Results for detecting switches (i.e., estimating $\hat{\sigma}_{i,j}$) in a texture image. Left: Original image. Middle: GPCA segmentation. Right: Segmentation via proposed method.

modeled using the following 2D autonomous linear switched-coefficient difference equation:

$$I(x,y) = \sum_{(k_x,k_y)\in\mathcal{R}_a} a_{k_x,k_y}(\sigma_{x,y}) I(x-k_x, y-k_y) + \eta(x,y) \tag{25}$$

where $I(x,y)$ denotes the intensity at pixel location (x,y) and the support region $\mathcal{R}_a$ was chosen according to the fundamental period of the textures. Fig. 4 shows the segmentation obtained when using our algorithm to minimize the number of switches, which in this case corresponds to minimizing the length of the boundaries between regions. As illustrated there, the proposed algorithm outperforms GPCA.

VII. Applications: Segmentation of Video Sequences

In this section we illustrate the application of the proposed identification algorithm to two non-trivial problems arising in computer vision: segmentation of video-shots and dynamic textures. Here the goal is to detect changes, e.g., scenes or activities in the former, texture in the latter, in a sequence of frames. Given the high dimensionality of the data, the starting point is to perform a principal component analysis (PCA) compression [24] to obtain low dimensional feature vectors $\mathbf{y}(t) \in \mathbb{R}^d$ representing each frame t. Specifically, each $N_x \times N_y$ size frame was represented by a vector $\mathbf{f}(t) \in \mathbb{R}^{N_x N_y}$ obtained by first converting it to gray scale and vectorizing. Next, the sample mean was found and used to construct the mean subtracted data matrix $\mathbf{F} = [\mathbf{f}(t_0) - \mathbf{m}, \ldots, \mathbf{f}(T) - \mathbf{m}]$. Finally, low dimensional representations $[\mathbf{y}(t_0), \ldots, \mathbf{y}(T)] = \mathbf{U}_{1:d}^T \mathbf{F}$ of the frames were obtained by performing a singular value decomposition $\mathbf{F} = \mathbf{UDV^T}$ followed by a projection of the data onto the subspace spanned by the first d columns of $\mathbf{U}$.

The next step is to assume, motivated by [25]–[28], that each component $y_j(.)$ of the feature vector $\mathbf{y}(t)$ evolves independently, according to an unknown multi-output model of the form described in Section IV-D

$$y_j(t) = \sum_{i=1}^{n_a} a_{i,j}(\sigma_t) y_j(t-1) + \eta_j(t), \ \|\eta(t)\|_2 \leq \epsilon. \tag{26}$$

Finally, defining $\mathbf{g}(t) = [\mathbf{p_1}(t) - \mathbf{p_1}(t+1), \ldots, \mathbf{p_d}(t) - \mathbf{p_d}(t+1)]$ allows to use the (minimum number of switches) sparsification-based approach to segment a given sequence according to the non-zero elements in the corresponding sequence $\|\mathbf{g}(.)\|_\infty$.

A. Video-Shot Segmentation

The goal here is to detect scene changes in video sequences. These changes can be categorized into two: (i) abrupt changes (cuts), and (ii) gradual transitions, e.g., various special effects that blend two consecutive scenes gradually. Fig. 5 shows the ground truth and the segmentations obtained using the proposed method (using 3^{rd} order models and $d = 3$) , GPCA [25], a histogram based method (bin to bin difference (B2B) with 256 bin histograms and window average thresholding [29]), and an MPEG-based method [30] for three sample sequences, *mountain.avi*, *family.avi* and *fisherman.mpg* available from http://www.open-video.org . A quantitative measure of the quality of a given segmentation $\mathcal{S}$ can be obtained using the *Rand* index [31], defined in this case as $RI = (N_{1,1} + N_{0,0})/(N_{1,1} + N_{0,0} + N_{1,0} + N_{0,1})$. Here $N_{1,1}$ denotes the number of pairs of points that belong to the same segment in both $\mathcal{S}$ and the ground truth (GT), $N_{0,0}$ is the number of pairs of points that belong to different segments in both $\mathcal{S}$ and GT, and $N_{1,0}$ ($N_{1,0}$) denotes the number of pairs of points that belong to the same segment in GT ($\mathcal{S}$) but were assigned to different segments in $\mathcal{S}$ (GT). Intuitively, this index measures the ratio of the number of agreements between the given segmentation $\mathcal{S}$ and the ground truth, to the number of agreements plus disagreements. Hence, $RI = 1$ indicates perfect clustering. A comparison of the performance of the four methods in terms of the Rand index is given in Table VI. Since the frames corresponding to gradual transitions do not belong to any cluster, these frames were excluded from the Rand index calculation. As an additional quantitative criterion, Table VII summarizes switch detection rates. As illustrated by these examples, the proposed method has slightly better performance than MPEG (the runner up), without the need to manually adjust seven parameters one of which, length of the transition, is very sensitive. B2B works well in finding cuts when there is a sudden change in color distribution as in the fisherman sequence, but fails otherwise. On the other hand, our method works well for different types and lengths of transitions. If the length of a gradual transition is compatible with the length of the segments (see, for instance, the first transition in family sequence), it might identify the transition as a separate segment since it is no longer possible to account for the dynamics of the transition within the noise level. It is also worth emphasizing that both, the B2B and the MPEG methods, rely on user adjustable parameters (two in the B2B case, seven for MPEG). In our experiments we adjusted these parameters, by trial and error, to get the best possible results. Hence the resulting comparisons against the proposed sparsification method correspond to best-case scenarios for both MPEG and B2B.

B. Dynamic Textures

Next, we consider three challenging sequences generated using the dynamic texture database http://www.svcl.ucsd.edu/projects/motiondytex/ synthdb/ . In the first one, we appended in time one patch from smoke to another patch from the same texture but transposed. Therefore, both sequences have the same photometric properties, but differ in the main motion direction: vertical in the first half and horizontal in the second half of the sequence. For the second example, we generated a sequence

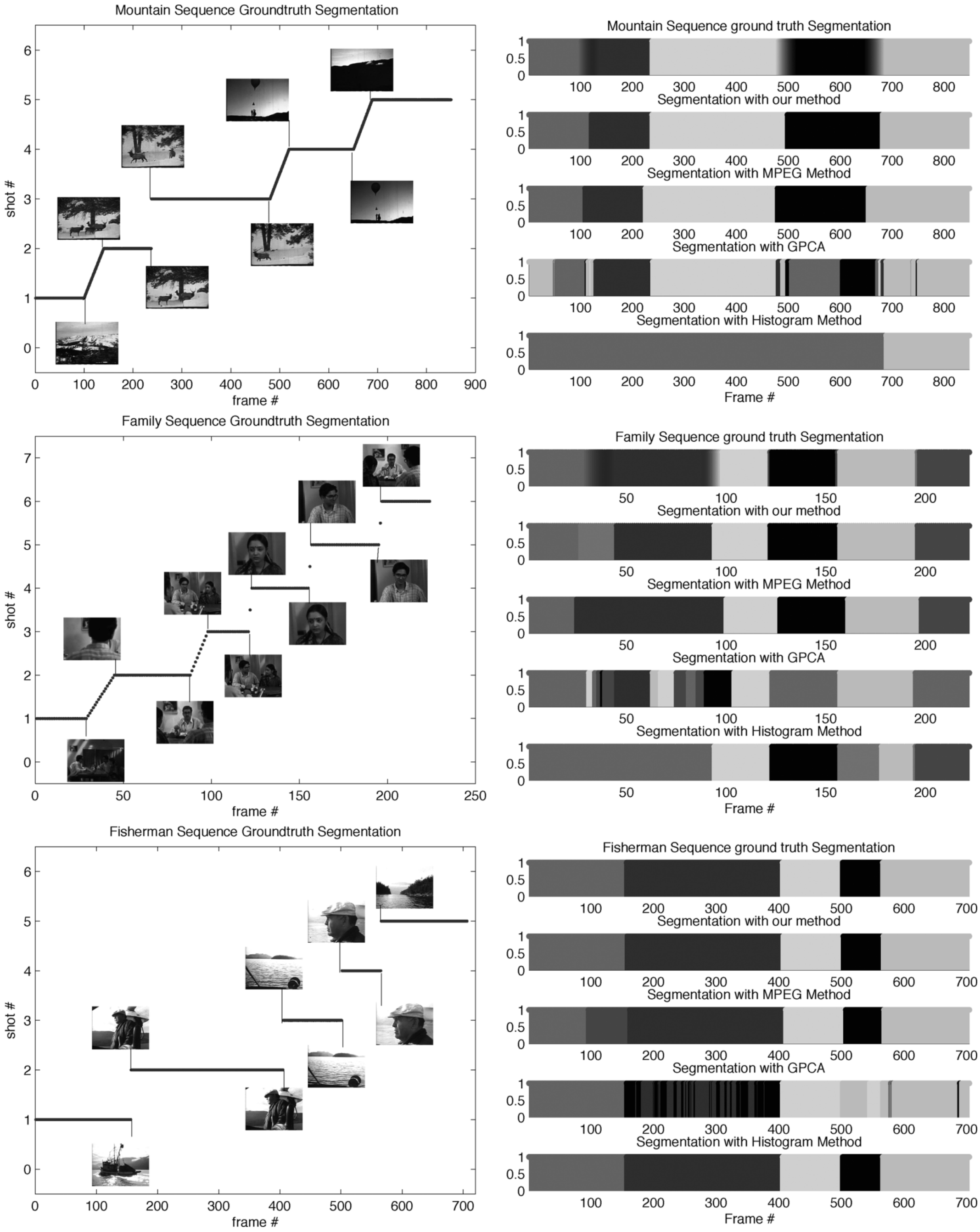


Fig. 5. Video Segmentation Results. Left Column: Ground truth segmentation (jumps correspond to cuts and slanted lines correspond to gradual transitions). Right Column: Changes and segments detected with different methods.

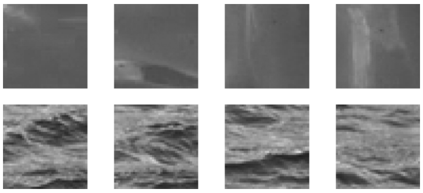


Fig. 6. Sample dynamic texture patches. Top: smoke, Bottom: river.

of river by sliding a window both in space and time (by going forward in time in the first half and by going backward in the second). Hence, the dynamics due to river flow are reversed.

TABLE VI
RAND INDICES FOR VIDEO-SHOT SEGMENTATION.

	Sparsification	MPEG	GPCA	B2B
mountain	0.9965	0.9816	0.9263	0.5690
family	0.9946	0.9480	0.8220	0.9078
fisherman	0.9955	0.9593	0.8966	1.0000

In the third example, we generated a sequence by using the river sequence with forward dynamics and subsampling the frames in the later part of the sequence. Hence, the river flow twice as fast in the second half of the clip. Sample frames from each sequence are shown in Fig. 6. For these sequences both

TABLE VII
SWITCH DETECTION RATES FOR VIDEO-SHOT SEGMENTATION. NS: TOTAL NUMBER OF SWITCHES IN THE SEQUENCE, TD: TRUE DETECTION, FA: FALSE ALARM, RD: REDUNDANT DETECTION (I.E., EXTRA SWITCHES FOUND ON GRADUAL TRANSITIONS).

		Sparsification			MPEG			GPCA			B2B		
	NS	TD	FA	RD	TD	FA	RD	TD	FA	RD	TD	FA	RD
mountain	4	4	0	0	4	0	0	4	8	12	1	0	0
family	5	5	0	1	5	0	0	5	6	5	4	2	1
fisherman	4	4	0	0	4	1	0	4	51	0	4	0	0

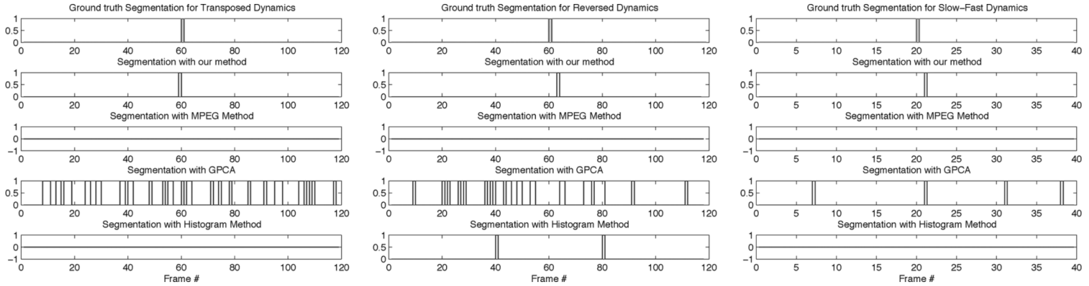


Fig. 7. Results for detecting change in dynamics only. Left: Smoke sequence concatenated with transposed dynamics. Center: River sequence concatenated with reversed dynamics. Right: River sequence with slow and fast dynamics.

histogram and MPEG methods failed to detect the cut since the only change is in the dynamics. On the other hand, the proposed method (using 5^{th} order models and $d = 3$) correctly segmented all sequences. These results are summarized in Fig. 7.

VIII. CONCLUSIONS AND FUTURE WORK

In this paper we considered the problem of identifying switched linear systems from input/output data and minimal *a priori* assumptions on the order of the subsystems and the magnitude of the noise. Our main result shows that, when an explanation with the minimum number of switches is sought (a problem relevant for instance in the context of segmentation), and ℓ_∞ bounded noise, the problem reduces to a convex optimization. In the case of general noise descriptions, the problem is no longer convex, but it can be recast into a sparsification form and efficiently solved using recently introduced relaxations. A similar idea can be also used when minimizing the number of systems. However, in this case, while usually working well in practice, the approach is suboptimal. The advantages of the proposed techniques over existing methods were illustrated using both academic examples and non-trivial segmentation problems arising in computer vision. As shown there, while most existing methods perform well in noiseless scenarios, sparsification–based techniques are substantially more robust to noise. Research currently under way seeks to address the issues of suboptimality of the approach for identifying the minimum number of systems consistent with the data, and to extend these approaches to classes of switched nonlinear systems, such as Hammerstein and Wiener. These problems are relevant to application domains such as computer vision where the high dimensionality of the data requires the use of, often non-linear, dimensionality reduction methods.

APPENDIX
BACKGROUND RESULTS ON SPARSIFICATION

In this appendix, we present the background results on the problem of *sparse signal recovery* [8], [9], [32] that motivate the approach pursued in the paper. This problem can be stated as: given some linear measurements $\mathbf{y} = A\mathbf{x}$ of a discrete signal $\mathbf{x} \in \mathbb{R}^n$ where $A \in \mathbb{R}^{m \times n}, \ m \ll n$, *find* the sparsest signal $\mathbf{x}^*$ consistent with the measurements. In terms of the ℓ_0 quasinorm (i.e., $\|\cdot\|_o$ satisfies all of the norm axioms except homogeneity since $\|c\mathbf{x}\|_o = \|\mathbf{x}\|_o$ for all non-zero scalars c), this problem can be recast into the following optimization form:

$$\min \|x\|_o \text{ subject to} : \mathbf{y} = A\mathbf{x}. \tag{27}$$

It is well known that the problem above is at least generically NP–complete ([33], [34]). Two fundamental questions in sparse signal recovery are: (i) the uniqueness of the sparse solution, (ii) existence of efficient algorithms for finding such a signal. In the past few years it has been shown that if the matrix A satisfies the so-called *restricted isometry property* (RIP), the solution is unique and can be recovered efficiently by several algorithms. These algorithms fall into two main categories: greedy algorithms (e.g., orthogonal matching pursuit [35]–[37]) and ℓ_1-based convex relaxation (also known as basis pursuit [8], [9], [32]). In this paper we follow the latter approach which is based on replacing $\|\mathbf{x}\|_o$ in the optimization above by $\|\mathbf{x}\|_1$. The idea behind this relaxation is the fact that the ℓ_1 norm is the *convex envelope* of the ℓ_0 norm, and thus, in a sense, minimizing the former yields the best convex relaxation to the (non-convex) problem of minimizing the latter. Morever, as shown in [8], [9], [32], this relaxation is stable and robust to noise. That is, even when only noisy linear measurements are available, if RIP holds for A, which is true with high probability for random matrices, recovery of the correct support of the original signal and approximating the true value within a factor of the noise are possible by solving

$$\min \|x\|_1 \text{ subject to} : \|\mathbf{y} - A\mathbf{x}\| \leq \epsilon \tag{28}$$

where ϵ is a bound on the norm of noise. This formulation arises naturally in many engineering applications such as magnetic resonance imaging, radar signal processing and image processing. Moreover, existence of efficient algorithms to solve this problem led to the compressed sensing framework which enabled speeding up signal acquisition considerably since the original sparse signal can be reconstructed using relatively few measurements. We refer the interested reader to the recent survey paper [38] for a comprehensive treatment of the subject.

The results above are not directly applicable to Problems (4) and (22) since these deal with sparsification problems in the space of vector valued finite sequences

$$\mathcal{S} = \left\{ \{\mathbf{g}(t)\}_{t=t_0}^{T} \mid \mathbf{g}(t) \in \mathbb{R}^m \right\}$$

rather than with vectors $\mathbf{x} \in \mathbb{R}^N$. This change requires extending the theory behind the ℓ_1-norm relaxation to the space $\mathcal{S}$. To this effect, begin by noting that the number of non-zero elements (i.e., vectors) in $\{\mathbf{g}\} \in \mathcal{S}$ (i.e., $\|\{\mathbf{g}\}\|_0$) is the same as in $\|\bar{\mathbf{g}}\|_0$ where $\bar{\mathbf{g}} = [\|\mathbf{g}(t_o)\|, \ldots, \|\mathbf{g}(T)\|]^T \in \mathbb{R}^{T-t_o+1}$. This suggests the use of $\|\bar{\mathbf{g}}\|_1 = \sum_t \|\mathbf{g}(t)\|$ as a convex objective function with an appropriate choice of the norm $\|\mathbf{g}(t)\|$. In particular, we will use $\|\mathbf{g}(t)\|_\infty$. The theoretical support for this intuitive choice is provided next.

Lemma 2: The convex envelope of the ℓ_0-norm of a vector valued sequence on $\|\{\mathbf{g}\}\|_\infty \leq 1$ is given by

$$\|\{\mathbf{g}\}\|_{0,env} \triangleq \sum_t \|\mathbf{g}(t)\|_\infty. \tag{29}$$

Proof: In order to prove the lemma, we need some preliminary results from convex analysis. For a function $f : \mathcal{C} \to \mathbb{R}$, where $\mathcal{C} \subseteq \mathbb{R}^n$, the conjugate $f^\star$ is defined as

$$f^\star(y) = \sup_{x \in \mathcal{C}} (\langle x, y \rangle - f(x)).$$

Under some technical conditions (see [39] Theorem 1.3.5), which are met here, the conjugate of the conjugate (i.e., $f^{\star\star}$) gives the convex envelope of the function f.

The proof proceeds now along the lines of that of the Theorem 1 in [40], by computing $\|x\|_o^{**}$, $x \in \mathcal{S}$. The isomorphism $\mathcal{I}$ from $\mathcal{S}$ to $\mathbb{R}^{m(T-t_o+1)}$, which simply stacks the elements of the sequence into a column vector, naturally induces an inner product on $\mathcal{S}$ as $\langle x, y \rangle = \langle \mathcal{I}(x), \mathcal{I}(y) \rangle = \sum_{t=1}^{T} x^T(t) y(t)$. For $f : \mathcal{S} \to \mathbb{R}$, $f(x) = \|x\|_0$, the conjugate function in $\mathcal{C} \doteq \|x\|_\infty \leq 1$ is

$$\begin{aligned} f^*(y) &= \sup_{\|x\|_\infty \leq 1} \{\langle x, y \rangle - f(x)\} \\ &= \sum_{i \in \lambda} \|y(i)\|_1 - |\lambda| \end{aligned} \tag{30}$$

where $\lambda = \{j : \|y(j)\|_1 > 1,\ j \in \{1, 2, \ldots, T\}\}$ is an index set and $|\lambda|$ is its cardinality.

$$\begin{aligned} & f^{**}(z) \\ &= \sup_{y \in \mathcal{S}} \{\langle y, z \rangle - f^*(y)\} \\ &= \sup_{y \in \mathcal{S}} \left\{ \sum_{i \in \lambda} y(i)^T z(i) + \sum_{i \notin \lambda} y(i)^T z(i) - \sum_{i \in \lambda} \|y(i)\|_1 + |\lambda| \right\} \\ &= \sup_{y \in \mathcal{S}} \left\{ \sum_{i \in \lambda} y(i)^T [z(i) - \operatorname{sign}(y(i))] + \sum_{i \notin \lambda} y(i)^T z(i) + |\lambda| \right\}. \end{aligned} \tag{31}$$

Here we consider two cases:

1) If $\|z\|_\infty > 1$, it is possible to choose y such that the first term in (31) grows unboundedly and $f^{**}(z) \to \infty$. So the domain of f^{**} is $\|z\|_\infty \leq 1$.
2) If $\|z\|_\infty \leq 1$, the first term in the last line of (31) is nonpositive. So to maximize the first term, $y(i)$ values should be chosen small in absolute value for $i \in \lambda$. Keeping in mind the bounds imposed on $y(i)$ values by λ, the maximum value of the second term is $\sum_{i \notin \lambda} \|z(i)\|_\infty$. Similarly,

$$\begin{aligned} & \sup_y \left\{ \sum_{i \in \lambda} y(i)^T [z(i) - \operatorname{sign}(y(i))] + |\lambda| \right\} \\ &\qquad = \sum_{i \in \lambda} [\|z(i)\|_\infty - 1] + |\lambda| = \sum_{i \in \lambda} \|z(i)\|_\infty . \end{aligned}$$

Hence,

$$f^{**}(z) = \sum_{i=1}^{T} \|z(i)\|_\infty . \tag{32}$$

■

A related line of results recently appeared in compressed sensing/sparse signal recovery community for structured sparsity (see for instance [13]–[15]).

References

[1] S. Paoletti, A. Juloski, G. Ferrari-Trecate, and R. Vidal, "Identification of hybrid systems: A tutorial," *Eur. J. Control*, vol. 13, no. 2, pp. 242–260, May 2007.

[2] Y. Ma and R. Vidal, "A closed form solution to the identification of hybrid ARX models via the identification of algebraic varieties," *Hybrid Syst. Comput. Control*, pp. 449–465, Mar. 2005.

[3] A. Bemporad, A. Garulli, S. Paoletti, and A. Vicino, "A bounded-error approach to piecewise affine system identification," *IEEE Trans. Autom. Control*, vol. 50, no. 10, pp. 1567–1580, Oct. 2005.

[4] J. Roll, A. Bemporad, and L. Ljung, "Identification of piecewise affine systems via mixed-integer programming," *Automatica*, vol. 40, pp. 37–50, 2004.

[5] A. Juloski, S. Wieland, and W. P. M. H. Heemels, "A bayesian approach to identification of hybrid systems," *IEEE Trans. Autom. Control*, vol. 50, no. 10, pp. 1520–1533, Oct. 2005.

[6] G. Ferrari-Trecate, M. Muselli, D. Liberati, and M. Morari, "A clustering technique for the identification of piecewise affine systems," *Automatica*, vol. 39, pp. 205–217, Feb. 2003.

[7] H. Nakada, K. Takaba, and T. Katayama, "Identification of piecewise affine systems based on statistical clustering technique," *Automatica*, vol. 41, no. 5, pp. 905–913, May 2005.

[8] D. Donoho, M. Elad, and V. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," *IEEE Transactions on Inf. Theory*, vol. 52, no. 1, pp. 6–18, Jan. 2006.

[9] J. Tropp, "Just relax: Convex programming methods for identifying sparse signals in noise," *IEEE Trans. Inf. Theory*, vol. 52, no. 3, pp. 1030–1051, Mar. 2006.

[10] R. Vidal, A. Chiuso, and S. Soatto, "Observability and identifiability of jump linear systems," in *Proc, 41st IEEE Conf. Decision and Control,* , Dec. 2002, vol. 4, pp. 3614–3619.

[11] J. Lygeros, K. H. Johansson, S. N. Simic, J. Zhang, and S. S. Sastry, "Dynamical properties of hybrid automata," *IEEE Trans. Autom. Control*, vol. 48, no. 1, pp. 2–17, Jan. 2003.

[12] A. Gionis and H. Mannila, "Segmentation algorithms for time series and sequence data," in *SIAM Int. Conf. Data Mining*, 2005, Tutorial.

[13] Y. Zhang, "On Theory of Compressive Sensing via ℓ_1-Minimization: Simple Derivations and Extensions," Tech. Rep. Rice University, 2008.

[14] Y. C. Eldar and M. Mishali, "Robust recovery of signals from a structured union of subspaces," *IEEE Trans. Inf. Theory*, vol. 55, no. 11, pp. 5302–5316, Nov. 2009.

[15] M. Stojnic, F. Parvaresh, and B. Hassibi, "On the reconstruction of block-sparse signals with an optimal number of measurements," *IEEE Trans. Signal Processing*, vol. 57, no. 8, pp. 3075–3085, Aug. 2009.

[16] N. Ozay, "Convex Relaxations for Robust Identification of Hybrid Models," Ph.D., Northeastern University, , 2010.

[17] M. Fazel, H. Hindi, and S. Boyd, "Log-det heuristic for matrix rank minimization with applications to hankel and euclidean distance matrices," in *Proc. American Control Conf.*, Jun. 2003.

[18] M. Lobo, M. Fazel, and S. Boyd, "Portfolio optimization with linear and fixed transaction costs," *Annals of Operations Research*, vol. 152, no. 1, pp. 376–394, Jul. 2007.

[19] E. J. Candes, M. Wakin, and S. Boyd, "Enhancing sparsity by reweighted l1 minimization," *Journal of Fourier Analysis and Applications*, vol. 14, no. 5, pp. 877–905, Dec. 2008.

[20] J. Woods*, Multidimensional Signal, Image and Video Processing and Coding*. : Academic Press, 2006.

[21] R. Vidal, "Identification of spatial-temporal switched ARX systems," in *Proc. 46th IEEE Conf. Decision and Control,*, Dec. 2007, pp. 4675–4680.

[22] E. Amaldi and M. Mattavelli, "The MIN PFS problem and piecewise linear model estimation," *Discrete Appl. Math.*, vol. 118, no. 1-2, pp. 115–143, Apr. 2002.

[23] 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 Conf. Decision and Control*, Dec. 2003, pp. 167–172.

[24] R. O. Duda, P. E. Hart, and D. G. Stork*, Pattern Classification*, 2nd ed. Hoboken: Wiley, 2000.

[25] R. Vidal, Y. Ma, and S. Sastry, "Generalized principal component analysis (gpca)," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 27, no. 12, pp. 1945–1959, Dec. 2005.

[26] W. Hong, J. Wright, K. Huang, and Y. Ma, "Multiscale hybrid linear models for lossy image representation," *IEEE Trans. Image Processing*, vol. 15, no. 12, pp. 3655–3671, Dec. 2006.

[27] A. B. Chan and N. Vasconcelos, "Mixtures of dynamic textures," in *Proc. IEEE Int. Conf. Computer Vision*, 2005, vol. 1, pp. 641–647.

[28] L. Cooper, J. Liu, and K. Huang, "Spatial segmentation of temporal texture using mixture linear models," in *Workshop on Dynamical Vision*, 2005, pp. 142–150.

[29] U. Gargi, R. Kasturi, and S. H. Strayer, "Performance characterization of video-shot-change detection methods," *IEEE Trans. Circuits Syst. for Video Technol.*, vol. 10, no. 1, pp. 1–13, Feb. 2000.

[30] Y. Boon-Lock and B. Liu, "A unified approach to temporal segmentation of motion jpeg and mpeg compressed video," in *Proc. Int. Conf. Multimedia Comput. and Syst.*, May 1995, pp. 81–88.

[31] W. Rand, "Objective criteria for the evaluation of clustering methods," *J. Amer. Statistical Assoc.*, vol. 66, pp. 846–850, 1971.

[32] E. J. Candès, J. K. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Commun. Pure and Appl. Math.*, vol. 59, no. 8, pp. 1207–1223, Aug. 2006.

[33] B. K. Natarajan, "Sparse approximate solutions to linear systems," *SIAM J. Computing*, vol. 24, no. 2, pp. 227–234, 1995.

[34] E. Amaldi and V. Kann, "On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems," *Theoretical Comp. Sci.*, vol. 209, no. 1–2, pp. 237–260, 1998.

[35] J. Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Trans. Inf. Theory*, vol. 50, no. 10, pp. 2231–2242, Oct. 2004.

[36] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Trans. Inf. Theory*, vol. 53, pp. 4655–4666, 2007.

[37] D. Needell and R. Vershynin, "Uniform uncertainty principle and signal recovery via regularized orthogonal matching pursuit," *Foundations of Comput. Math.*, vol. 9, no. 3, pp. 317–334, 2009.

[38] A. M. Bruckstein, D. L. Donoho, and M. Elad, "From sparse solutions of systems of equations to sparse modeling of signals and images," *SIAM Rev.*, vol. 51, no. 1, pp. 34–81, 2009.

[39] J.-B. Hiriart-Urruty and C. Lemarechal*, Convex Analysis and Minimization Algortihms II: Advanced Theory and Bundle Methods*. Berlin, Germany: Springer-Verlag, 1993, vol. 306, Grundlehrer der mathematischen Wissenschaften.

[40] M. Fazel, H. Hindi, and S. Boyd, "A rank minimization heuristic with application to minimum order system approximation," in *Proc. American Control Conf.*, Jun. 2001.

Necmiye Ozay (M'04) received the B.S. degree from Bogazici University, Istanbul, Turkey in 2004, the M.S. degree from the Pennsylvania State University, University Park, PA, in 2006 and the Ph.D. degree from Northeastern University, Boston, MA in 2010, all in electrical engineering.

Currently, she is a Postdoctoral Scholar at the California Institute of Technology, Pasadena, CA. In the summer of 2008, she was a Research Intern at GE Global Research, Niskayuna, NY. She has also held short-term visiting positions at Sabanci University, Istanbul in 2005 and Polytechnic University of Catalunya, Barcelona, Spain, in 2008. Her research interests lie at the broad interface of system identification and verification, convex optimization, control theory, and computer vision.

Dr. Ozay is the recipient of the 2008 IEEE Control Systems Society Conference on Decision and Control Best Student Paper Award and the 2009 IEEE Computer Society Biometrics Workshop Best Paper Honorable Mention Award. She is a member of SIAM.

Mario Sznaier (M'89) is currently the Dennis Picard Chaired Professor at the Electrical and Computer Engineering Department, Northeastern University, Boston MA. Prior to joining Northeastern University, Dr. Sznaier was a Professor of Electrical Engineering at the Pennsylvania State University, University Park, and also held visiting positions at the California Institute of Technology, Pasadena. His research interests include robust identification and control of hybrid systems, robust optimization, and dynamical vision.

Dr. Sznaier is currently serving as an associate editor for *Automatica* and as a member of the Board of Governors of the IEEE Control Systems Society. Additional recent service includes CSS Executive Director (2007–2011), Program Chair of the 2009 IFAC Symposium on Robust Control Design, and Program vice-chair of the 2008 IEEE Conference on Decision and Control. Dr. Sznaier was a plenary speaker at the 2009 and 2010 Interntional Conference on the Dynamics of Information Systems, and will deliver plenary lectures at the 2011 IFAC Symposium on Systems Identification and the 2011 Symposium on Robust Control Design. A list of publications and currently funded projects can be found at http://robustsystems.ece.neu.edu.

Constantino M. Lagoa (M'98) received the B.S. and M.Sc. degrees from the Instituto Superior Técnico, Technical University of Lisbon, Portugal in 1991 and 1994, respectively, and the Ph.D. degree from the University of Wisconsin at Madison in 1998.

He joined the Electrical Engineering Department of Pennsylvania State University, University Park, PA, in August 1998, where he currently holds the position of Professor. He has a wide range of research interests including robust control, controller design under risk specifications, system identification, control of computer networks and discrete event dynamical systems. He is the author or co-author of more than twenty journal papers and book chapters and more than forty conference publications.

Dr. Lagoa is currently Associate Editor of IEEE TRANSACTIONS ON CONTROL SYSTEMS TECHNOLOGY. He was a co-organizer of a CDC workshop on Distributional Robustness. He has also been a member of the Conference Editorial Board of the IEEE Control System Society since 2004. He is also a member of the IFAC Technical Committee on Robust Control and has served as a member of the international program committee of several conferences.

Octavia I. Camps (M'93) received a B.S. degree in computer science, a B.S. degree in electrical engineering from the Universidad de la Republica, Montevideo, Uruguay, an M.S. and Ph.D. degree in electrical engineering from the University of Washington, Seattle, in 1981, 1984, 1987, and 1992, respectively.

Since 2006, she has been a Professor in the Electrical and Computer Engineering Department at Northeastern University, Boston, MA. From 1991 to 2006 she was a faculty member of the Department of Electrical Engineering and the Department of Computer Science and Engineering at Pennsylvania State University, University Park, PA. In 2000, she was a visiting faculty at California Institute of Technology, Pasadena, and at the University of Southern California, Los Angeles. Her main research interests include robust computer vision, image processing, and machine learning.

Dr. Camps is a former associate editor of *Pattern Recognition, Machine Vision Applications and Image and Vision Computing*.