Proceedings of the
47th IEEE Conference on Decision and Control
Cancun, Mexico, Dec. 9-11, 2008

A Sparsification Approach to Set Membership Identification of a Class of Affine Hybrid Systems

Necmiye Ozay* Mario Sznaier* Constantino Lagoa** Octavia Camps*.

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 optimizing a performance criteria (either minimum number of switches or minimum number of plants). Our main result shows that this problem can be reduced to a sparsification form, where the goal is to maximize sparsity of a given vector sequence. Although in principle this leads to an NP-hard problem, as we show in the paper, efficient convex relaxations can be obtained by exploiting recent results on sparse signal recovery. These results are illustrated using two non-trivial problems arising in computer vision applications: video-shot and dynamic texture segmentation.**

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 excellent tutorial paper [19] for a 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 [17] lead to nonconvex optimization problems, while optimization methods lead to generically NP–hard problems, either necessitating the use of relaxations [2] or restricted to small size problems [21].

Motivated by the computational complexity 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 optimizing a performance criteria (either minimum number of plants or minimum number of switches). The main result of the paper shows that this problem can be reduced to a sparsification form, where the goal is to minimize the number of non–zero elements of a given vector sequence. Although in principle this leads to an NP-hard problem, efficient convex relaxations can be obtained by exploiting recent results on sparse signal recovery based on ℓ_1-norm minimization [6], [22].

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

II. PRELIMINARIES

A. Notation and Definitions

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

$\mathbf{x}$	a vector in $\mathbb{R}^N$
$\|\mathbf{x}\|_p$	p-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 $\|\{\mathbf{x}\}\|_p \doteq \left(\sum_{i=1}^T \|\mathbf{x}(i)\|_p^p\right)^{1/p}$
$\|\{\mathbf{x}\}\|_0$	ℓ_o-quasinorm $\doteq$ number of non-zero vectors in the sequence (i.e. cardinality of the set $\{t \vert \mathbf{x}(t) \neq \mathbf{0}, t \in [1,T]\}$)

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 $\sigma_t \in \{1, \ldots, s\}$–the mode of the system– indicates which

*ECE Department, Northeastern University, Boston, MA 02115. {ozay.n,m.sznaier,camps}@neu.edu

**Department of Electrical Engineering, Penn State University, University Park, PA 16802. lagoa@engr.psu.edu

This work was supported in part by NSF grants ECS–0648054, IIS–0713003, and AFOSR grant FA9550–05–1–0437.

978-1-4244-3124-3/08/$25.00 ©2008 IEEE

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* [16], $\tau = \{I_i\}_{i=0}^{k}$, 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:

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

and the number of switches is equal to k.

An equivalent representation of (1) is:

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

where $\mathbf{r}(t) = [y(t-1), \ldots, y(t-n_a), u(t-1), \ldots, u(t-n_c), 1]^T$ is the regressor vector and $\mathbf{p}(\sigma_t) = [a_1(\sigma_t), \ldots, a_{n_a}(\sigma_t), c_1(\sigma_t), \ldots, c_{n_c}(\sigma_t), f(\sigma_t)]^T$ is the unknown coefficient vector at time t.

B. Background Results on Sparsification

In this section, we present some background results that will be used to recast the identification problem into a convex optimization form. We begin by recalling some results related to the problem of *sparse signal recovery* [6], [22]. This problem can be stated as: given some noisy linear measurements $\mathbf{y} = A\mathbf{x} + \eta$ of a discrete signal $\mathbf{x} \in \mathbb{R}^N$ where $A \in R^{m \times n}$, $m \ll n$ and a bound ϵ on the norm of the noise η are known, *find* the sparsest signal $\mathbf{x}^*$ consistent with the measurements. In terms of the ℓ_0 quasinorm, this problem can be recast into the following optimization form:

$$\min \|x\|_o \text{ subject to} : \|\mathbf{y} - A\mathbf{x}\| \leq \epsilon \tag{3}$$

It is well known that the problem above is at least generically NP–complete ([18], [1]). However, in the past few years a family of relaxations have been developed 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. Moreover, as shown in [22], under certain conditions the ℓ_1 minimization indeed yields the optimal ℓ_0 solution.

In this paper we will pursue a similar approach. However, we will work with sparsification problems in the space of vector valued finite[2] 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 necessitates 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 suggest 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 1: The convex envelope of the ℓ_0-norm of a vector valued sequence on $\|\{\mathbf{g}\}\|_\infty \leq 1$ is given by

$$\|\{\mathbf{g}\}\|_{0,env} \triangleq \sum_t \|\mathbf{g}(t)\|_\infty. \tag{4}$$

Proof: See the appendix. ■

III. PROBLEM STATEMENT

In this paper we consider the problem of identifying switched autoregressive exogenous (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 bound ϵ on the ℓ_p norm of the noise (i.e. $\|\eta\| \leq \epsilon$), find a hybrid affine model of the form (1) that is consistent with the a priori information and experimental data.

It is clear that this problem, though ensuring consistency, is not well-posed and has infinitely many solutions. For instance, one can always find a trivial hybrid 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. $n_a \leq n_y$ and $n_c \leq n_u$ 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. This criterion may also be useful when the piecewise constant mode signal σ_t is known to satisfy a dwell-time constraint (i.e. the time between any consecutive switches is bounded below by a dwell-time) or an average dwell-time constraint (i.e. the number of switches in any given interval is bounded above by its length normalized by an average dwell-time, plus a chatter bound)[3]. 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 bounds $\epsilon > \|\eta\|_p$, $n_u \geq n_c$ and $n_y \geq n_a$ on the ℓ_p norm of the noise and the order of the regressors, respectively, find a hybrid affine model of the form (1) that is consistent with the a priori

[1] Since it is not possible to deduce information for $t < \max(n_a, n_c)$ when the initial conditions are unknown, in the identification problem we take $t_0 = \max(n_a, n_c)$.

[2] Since the experimental data consists of only finite samples, we consider finite sequences. However, with appropriate modifications the discussions in this section can easily be extended to deal with infinite sequences.

[3] These are the discrete-time counterparts of some sets of mode signals defined in [12]. Detailed definitions of different sets of mode signals can be found therein.

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 [2], leads to the following identification problem:

Problem 3: **[Minimum Number of Submodels]** Given input/output data over the interval $[t_0, T]$, and bounds ϵ, n_y, n_u on the norm of the noise and 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. MAIN RESULTS

In this section we show that both, Problems 2 and 3, can be converted into an equivalent *sparsification* form where the objective is to maximize the number of zero elements of a suitably defined vector valued sequence. While in principle maximizing sparsity is a generically non-convex, hard to solve problem, recent developments in sparse signal recovery reveal that efficient, computationally tractable relaxations can be obtained by exploiting elements from convex analysis. To this effect, we start by defining a time varying parameter vector $\mathbf{p}(t) \in R^{n_y+n_u+1}$. Replacing $\mathbf{p}(\sigma_t)$ in (2) with $\mathbf{p}(t)$, allows for recasting the consistency problem into the following feasibility form:

$$\begin{array}{ll} \text{find} & \mathbf{p}(t) \\ \text{s.t} & y(t) - \mathbf{r}(t)^T\mathbf{p}(t) = \eta(t) \quad \forall t \\ & \|\{\eta\}\|_* \le \epsilon \end{array} \tag{5}$$

where $\|.\|_*$ denotes a suitable norm, specified according to the problem under consideration, and where ϵ is an upper bound on the noise level. Thus, restricting problems 2 and 3 to the feasible set of (5) guarantees consistency.

A. Identification with Minimum Number of Switches

In order to solve Problem 2, we consider the sequence of *first order differences* of the time varying parameters $\mathbf{p}(t)$, given by

$$\mathbf{g}(t) = \mathbf{p}(t) - \mathbf{p}(t+1) \tag{6}$$

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:

$$\begin{array}{ll} \min_{\mathbf{p}(t)} & \|\{\mathbf{p}(t) - \mathbf{p}(t-1)\}\|_0 \\ \text{s.t} & y(t) - \mathbf{r}(t)^T\mathbf{p}(t) = \eta(t) \quad \forall t \\ & \|\{\eta\}\|_* \le \epsilon \end{array} \tag{7}$$

From Lemma 1, it follows that a convex relaxation can be obtained replacing $\|.\|_o$ by $\|.\|_1$. A better heuristic can be obtained by adapting to this case the iterative weighted ℓ_1 heuristic proposed in ([9], [15], [3]). This requires solving, at each iteration, the following convex program:

$$\begin{array}{lll} \text{minimize}_{z,g,p} & \sum_t w_t^{(k)} z_t & \\ \text{subject to} & \|\mathbf{p}(t) - \mathbf{p}(t-1)\|_\infty \le z_t & \forall t \\ & y(t) - \mathbf{r}(t)^T\mathbf{p}(t) = \eta(t) & \forall t \\ & \|\{\eta\}\|_* \le \epsilon & \end{array} \tag{8}$$

where $w_t^{(k)} = (z_t^{(k)} + \delta)^{-1}$ and where $z_t^{(k)}$ denotes the optimal solution at the k^{th} iteration, with $z^{(0)} = [1, 1, \ldots, 1]^T$, and 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 $z_t^{(k)}$s, thus pushing these elements further towards zero.

1) A Greedy algorithm for the ℓ_∞ case: In the sequel we propose a computationally simpler alternative 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 [11] 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.

```
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 $\mathbf{p}$:
        $\mathcal{F} : \{ \ |y(t) - \mathbf{r}(t)^T\mathbf{p}| \le \epsilon \quad \forall t \in [\tau_k, i] \ \}$
    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 $k$
```

TABLE I

OPTIMAL GREEDY ALGORITHM FOR PROBLEM 2

Proposition 1: Let k^* denote the number of switches in an optimal solution to Problem 2 when the noise bound is given in terms of its ℓ_∞-norm. 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 $\tau_i' \ge \tau_j'^*$. Then, an induction step shows that, $\tau_i' \ge \tau_i'^*$ $\forall i \in \{0, \ldots, k^*\}$ implying $k \le k^*$.

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 $|y(t) - \mathbf{r}(t)^T\mathbf{p_j}| \le \epsilon$. 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: Algorithm (8) 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.

B. Identification with Minimum Number of Submodels

In this section, motivated by an idea used in [2], we present an iterative procedure for solving Problem 3. 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 $|y(t) - \mathbf{r}(t)^T\tilde{\mathbf{p}}| \leq \epsilon$ 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:

$$\begin{array}{ll} \min_{\mathbf{p}(t),\tilde{\mathbf{p}}} & \|\{\mathbf{p}(t) - \tilde{\mathbf{p}}\}\|_0 \\ \text{s.t} & |y(t) - \mathbf{r}(t)^T\mathbf{p}(t)| \leq \epsilon \quad \forall t \end{array} \tag{9}$$

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 up until all data points are clustered. The number of times (9) 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 [2] to re-estimate parameter values and reassign, if needed, data points, leads to the overall algorithm listed in Table II, where minimization of $\|.\|_0$ is (approximately) accomplished through the use of the weighted ℓ_1 norm minimization relaxation.

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   $\|\mathbf{p}(t) - \tilde{\mathbf{p}}\|_\infty \leq z_t$
                     $|y(t) - \mathbf{r}(t)^T\mathbf{p}(t)| \leq \epsilon$
                     $\forall t \in N_l$
        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 $\mathbf{z}^{(0)} = [1, 1, \ldots, 1]^T$; and $\delta$
        is the regularization constant.
    Let $i = 1$
    WHILE $i < l$
        Let $K_{il} = \{t \in N_i : |y(t) - \mathbf{r}(t)^T\tilde{\mathbf{p}}_l| \leq \epsilon\}$
        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 $K_l = \{t \in N_l : |y(t) - \mathbf{r}(t)^T\tilde{\mathbf{p}}_l| \leq \epsilon\}$
    Let $N_{l+1} = N_l \setminus K_l$
END WHILE
RETURN $s = l$ and $K_i, i = 1, \ldots, s$
```

TABLE II

ALGORITHM FOR PROBLEM 3

Remark 2: While consistent numerical experience shows that this algorithm works well in practice, counterexamples are available where it overestimates the number of systems. This is due to its greedy nature that 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. At this point the issues of existence of conditions under which the greedy algorithm is indeed optimal and bounds on its worst case performance are open research questions.

V. EXAMPLES

This section illustrates the proposed methods with some academic examples and compares their performance against the methods in [2] and [17].

In the first example, we considered input/output data generated by a hybrid system that switched among the following two ARX submodels:

$$y(t) = 0.2y(t-1) + 0.24y(t-2) + 2u(t-1);\ t \in [1, 25] \cup [51, 75]$$

$$y(t) = -1.4y(t-1) - 0.53y(t-2) + u(t-1);\ t \in [26, 50] \cup [76, 100]$$

with $\|\eta\|_\infty = 0.5$. The goal here was to identify a model that explained the experimental data record with the fewest possible number of switches. Figure 1 compares the performance of sparsification-based (both algorithm (7) and the greedy algorithm of Table I) against the algebraic method[4]. As shown there, the sparsification based methods correctly estimated the parameters and number of switches, while GPCA failed to do so (due to noise). Additional examples illustrating the use of sparsification to find the minimum number of switches are given in section VI.

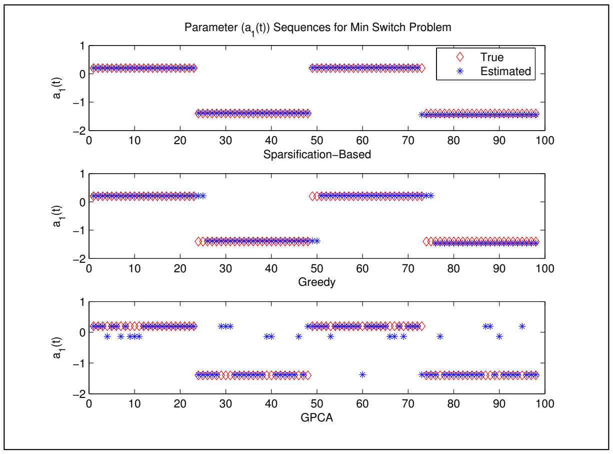


Fig. 1. True and estimated parameter sequences for parameter $a_1(\sigma_t)$ for Example 1.

The next two examples consider the problem of estimating the minimum number of systems. In the first case we used data generated by the SARX model:

$$y(t) = a_1(\sigma_t)y(t-1)+a_2(\sigma_t)y(t-2)+c_1(\sigma_t)u(t-1)+\eta(t) \quad (10)$$

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 $0.5 \le \|z\| \le 1$ with uniform random phase and magnitude, and $\eta(t)$ is an iid noise term uniformly distributed in $[-0.5, 0.5]$.

We randomly generated 100 SARX models of the form (10). For each model, we estimated the number of submodels by solving Problem 3 with our method and the bounded-error method; and by approximating the rank of an appropriate matrix obtained from data as proposed in [24] 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. Results on this experiment are summarized in Figure 2 and Table III.

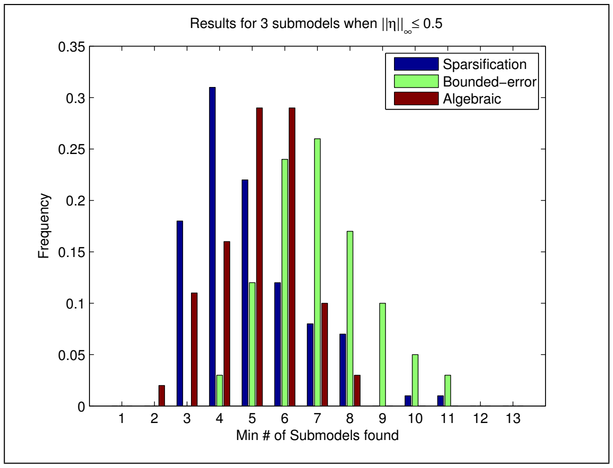


Fig. 2. Results of estimation of minimum number of submodels.

Absolute Error	Sparsification	Bounded-Error	Algebraic
Mean	1.93	4.07	2.18
Standard deviation	1.65	1.58	1.25

TABLE III

ERROR STATISTICS OF MINIMUM NUMBER OF SUBMODEL ESTIMATION WITH A NOISE LEVEL OF $\epsilon = 0.5$.

For the next example, we considered input/output data generated by an SARX system of the form (10) with the same mode signal σ_t and coefficients:

$$[a_1(1),\ a_2(1),\ c_1(1)] = [-1.6758,\ -0.8292,\ 1.8106]$$
$$[a_1(2),\ a_2(2),\ c_1(2)] = [-0.8402,\ -0.6770,\ 0.2150]$$
$$[a_1(3),\ a_2(3),\ c_1(3)] = [1.0854,\ -0.9501,\ 0.6941].$$

In this case we used two different criteria to assess the performance, for different noise levels, of the sparsification-based (Table II), bounded error, and algebraic algorithms, both in terms of quality of the segmentation and parameter identification error. Quality of the clustering was assessed using the Rand index [20] to compare the estimated mode signal $\hat{\sigma}_t$ against the true σ_t[5]. Quality of the parameter estimation was evaluated using the following error measure:

$$\Delta = \frac{\sqrt{\sum_{t=t_0}^{T} \|\mathbf{p}(\sigma_t) - \hat{\mathbf{p}}(\hat{\sigma}_t)\|_2^2}}{T - t_0 + 1} \quad (11)$$

For the noise level of $\epsilon = 0.05$, the sparsification, bounded error and algebraic methods found 4, 9 and 4 submodels, respectively. For the noise level of $\epsilon = 0.5$, the number of submodels found were 3, 9 and 4, respectively. The results of these experiments are summarized in Tables IV and V.

Noise Level	Sparsification	Bounded-error	Algebraic
$\epsilon = 0.05$	0.9681	0.9157	0.9212
$\epsilon = 0.5$	0.8436	0.7482	0.6849

TABLE IV

RAND INDICES THAT SHOW THE QUALITY OF MODE SIGNAL ESTIMATES.

[4] The bounded error based method was not used here since it does not attempt to minimize the number of switches.

[5] Recall that a Rand index of 1 corresponds to a perfect clustering.

Noise Level	Sparsification	Bounded-error	Algebraic
$\epsilon = 0.05$	0.0456	0.1355	0.0226
$\epsilon = 0.5$	0.0513	0.1757	0.2518

TABLE V

ERROR MEASURE Δ THAT SHOWS THE QUALITY OF PARAMETER ESTIMATES.

In these last two examples 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.

VI. 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 later, in a sequence of frames. Given the high dimensionality of the data, the starting point is to perform a principal component analysis (PCA) compression [7] to obtain low dimensional feature vectors $\mathbf{y}(t) \in \mathbb{R}^d$ representing each frame t. The next step is to assume, motivated by [23], [14], [4], [5], that each component $y_j(.)$ of the feature vector $\mathbf{y}(t)$ evolves independently, according to an unknown model of the form:

$$y_j(t) = \sum_{i=1}^{n_a} a_{i,j}(\sigma_t) y_j(t-1) + \eta(t), \; \|\eta(t)\|_2 \leq \epsilon \qquad (12)$$

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 $\|\mathbf{g}(.)\|_\infty$.

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. Figure 3 shows the ground truth and the segmentations obtained using the proposed method, GPCA [23], a histogram based method (bin to bin difference (B2B) with 256 bin histograms and window average thresholding [10]), and an MPEG-based method [25] for two sample sequences, *mountain.avi* and *family.avi*, available from `http://www.open-video.org`. Both the B2B and 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. As shown in Table VI, 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.

	Sparsification	MPEG	GPCA	B2B
mountain	0.9965	0.9816	0.9263	0.5690
family	0.9946	0.9480	0.8220	0.9078

TABLE VI

RAND INDICES FOR VIDEO-SHOT SEGMENTATION

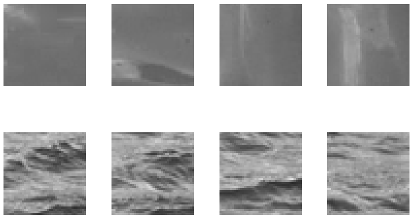
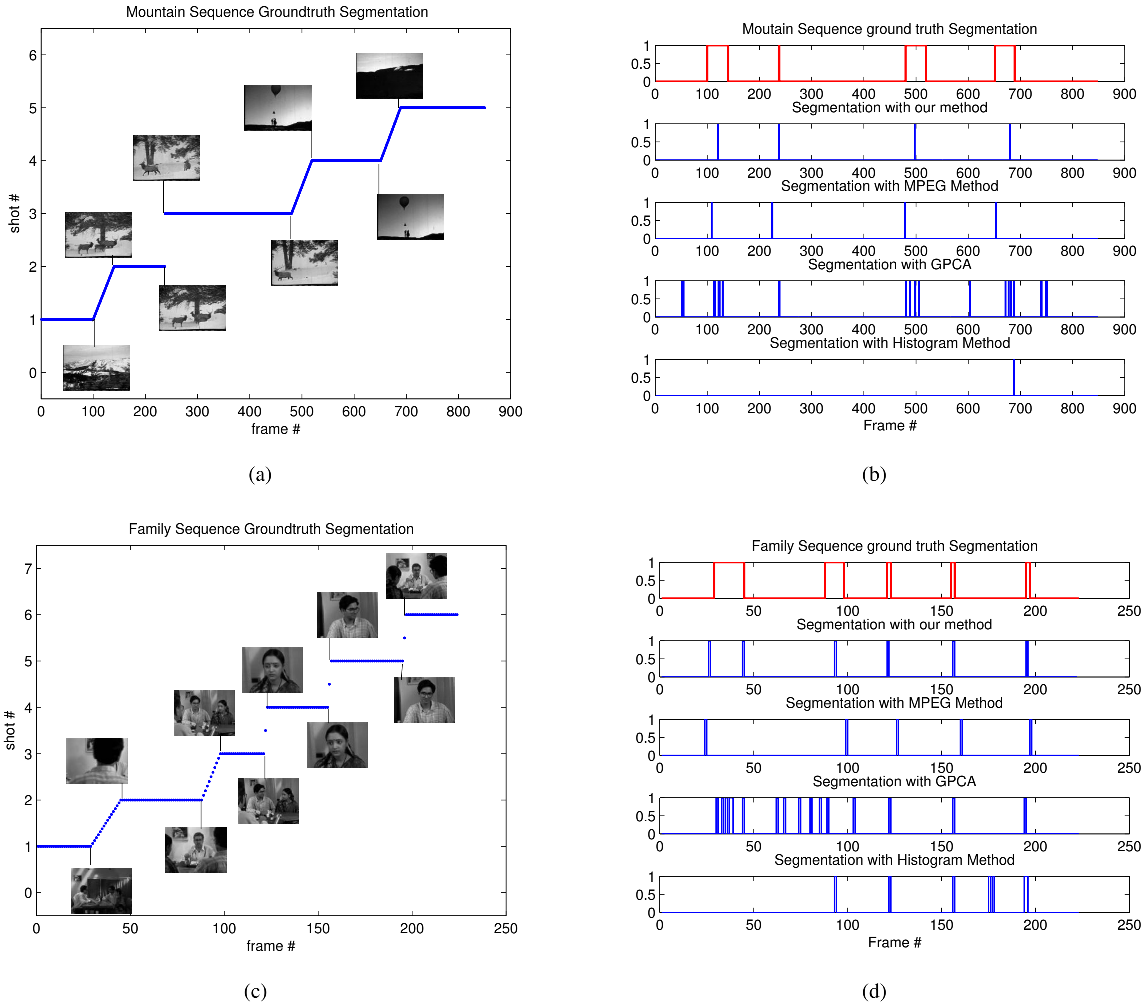


Fig. 4. Sample dynamic texture patches. Top: smoke, Bottom: river

Dynamic textures: Next, we consider two challenging sequences generated using the dynamic texture database `http://www.svcl.ucsd.edu/projects/motion-dytex/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 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. Sample frames from each sequence are shown in Figure 4. For these sequences both 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 both sequences. These results are summarized in Figure 5.

VII. CONCLUSIONS AND FUTURE WORK

In this paper we consider 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), the problem 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



(a)

(b)

(c)

(d)

Fig. 3. Video Segmentation Results. Left Column: Ground truth segmentation (jumps correspond to cuts and slanted lines correspond to gradual transitions). Right Column: Changes detected with different methods. Value 0 corresponds to frames within a segment and value 1 corresponds to the frames in transitions.

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 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.

REFERENCES

[1] Edoardo Amaldi and Viggo Kann. On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems. *Theoretical Computer Science*, 209(1–2):237–260, 1998.

[2] 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, 2005.

[3] E. J. Candes, M. Wakin, and S. Boyd. Enhancing sparsity by reweighted l1 minimization. Technical report, California Institute of Technology, 2007.

[4] A. B. Chan and N. Vasconcelos. Mixtures of dynamic textures. In *IEEE International Conference on Computer Vision*, volume 1, pages 641–647, 2005.

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

[6] D.L. Donoho, M. Elad, and V.N. Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Transactions on Information Theory*, 52(1):6–18, 2006.

[7] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley & Sons, Inc., second edition.

[8] M. Fazel, H. Hindi, and S. Boyd. A rank minimization heuristic with application to minimum order system approximation. In *American Control Conference*, 2001.

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

[10] U. Gargi, R. Kasturi, and S. H. Strayer. Performance characterization of video-shot-change detection methods. *IEEE Transactions on Circuits and Systems for Video Technology*, 10(1):1–13, 2000.

[11] A. Gionis and H. Mannila. Segmentation algorithms for time series and sequence data. In *SIAM International Conference on Data Mining*, 2005. Tutorial.

[12] J. P. Hespanha. Uniform stability of switched linear systems: Exten-

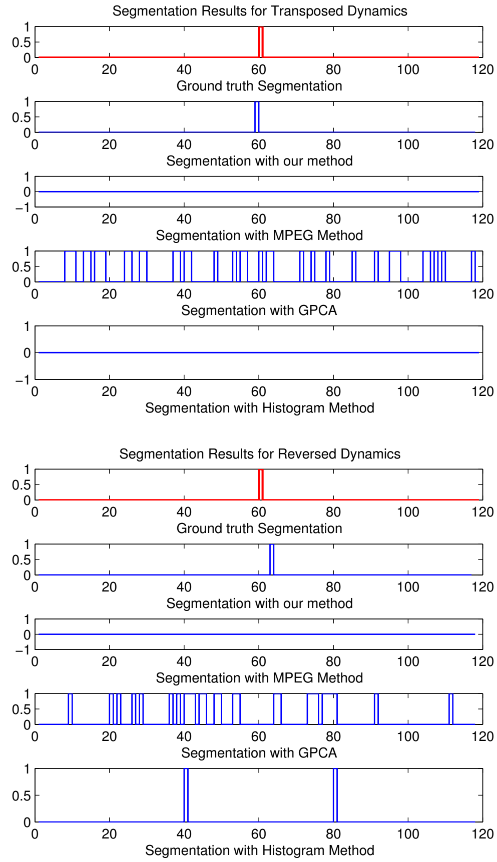


Fig. 5. Results for detecting change in dynamics only. Top: Smoke sequence concatenated with transposed dynamics. Bottom: River sequence concatenated with reversed dynamics.

sions of lasalle's invariance principle. *IEEE Transactions on Automatic Control*, 49(4):470–482, 2004.

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

[14] W. Hong, J. Wright, K. Huang, and Y. Ma. Multiscale hybrid linear models for lossy image representation. *IEEE Transactions on Image Processing*, 15(12):3655–3671, December 2006.

[15] M. Lobo, M. Fazel, and S. Boyd. Portfolio optimization with linear and fixed transaction costs. *Annals of Operations Research*, 152(1):376–394, 2007.

[16] J. Lygeros, K. H. Johansson, S. N. Simic, J. Zhang, and S. S. Sastry. Dynamical properties of hybrid automata. *IEEE Transactions on Automatic Control*, 48(1):2–17, 2003.

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

[18] B. K. Natarajan. Sparse approximate solutions to linear systems. *SIAM Journal on Computing*, 24(2):227–234, 1995.

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

[20] W.M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66:846–850, 1971.

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

[22] J.A. Tropp. Just relax: convex programming methods for identifying sparse signals in noise. *IEEE Transactions on Information Theory*, 52(3):1030–1051, 2006.

[23] R. Vidal, Y. Ma, and S. Sastry. Generalized principal component analysis (gpca). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(12):1945–1959, December 2005.

[24] R. Vidal, S. Soatto, and S. Sastry. An algebraic geometric approach to the identification of linear hybrid systems. In *IEEE Conference on Decision and Control*, pages 167–172, December 2003.

[25] Boon-Lock Y. and B. Liu. A unified approach to temporal segmentation of motion jpeg and mpeg compressed video. In *International Conference on Multimedia Computing and Systems*, pages 81–88, May 1995.

Appendix
Proof of Lemma 1

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

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

Under some technical conditions (see [13] 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 [8], 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 $\mathcal{C} \doteq \|x\|_\infty \le 1$ is:

$$\begin{array}{rcl} f^*(y) & = & \sup_{\|x\|_\infty \le 1} \{\langle x, y\rangle - f(x)\} \\ & = & \sum_{i\in\lambda} \|y(i)\|_1 - |\lambda| \end{array} \qquad (13)$$

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

$$\begin{array}{rcl} f^{**}(z) & = & \sup_{y\in\mathcal{S}} \{\langle y, z\rangle - f^*(y)\} \\ & = & \sup_{y\in\mathcal{S}} \{\sum_{i\in\lambda} y(i)^T z(i) \\ & & + \sum_{i\notin\lambda} y(i)^T z(i) - \sum_{i\in\lambda} \|y(i)\|_1 + |\lambda|\} \\ & = & \sup_{y\in\mathcal{S}} \{\sum_{i\in\lambda} y(i)^T \left[z(i) - \text{sign}(y(i))\right] \\ & & + \sum_{i\notin\lambda} y(i)^T z(i) + |\lambda|\} \end{array} \qquad (14)$$

Here we consider two cases:

1) If $\|z\|_\infty > 1$, it is possible to choose y such that the first term in (14) grows unboundedly and $f^{**}(z) \to \infty$. So the domain of f^{**} is $\|z\|_\infty \le 1$.

2) If $\|z\|_\infty \le 1$, the first term in the last line of (14) 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, $\sup_y \{\sum_{i\in\lambda} y(i)^T [z(i) - \text{sign}(y(i))] + |\lambda|\} = \sum_{i\in\lambda} [\|z(i)\|_\infty - 1] + |\lambda| = \sum_{i\in\lambda} \|z(i)\|_\infty$. Hence,

$$f^{**}(z) \;=\; \sum_{i=1}^{T} \|z(i)\|_\infty \,. \qquad (15)$$