Surviving the Upcoming Data Deluge: A Systems and Control Perspective

Mario Sznaier Octavia Camps Necmiye Ozay Constantino Lagoa

Abstract— **Arguably, one of the biggest challenges facing the systems and control community stems from the exponential growth in data collection capabilities, made possible by the development of low cost, ultra low power sensors. These developments have rendered feasible a spectrum of new control applications, ranging from zero emission buildings to reconfigurable, self aware environments, that can profoundly impact society. However, realizing this potential, requires endowing controllers with the ability to timely extract actionable information from the very large data streams generated by the sensors, a goal that challenges the capabilities of existing techniques. The goal of this paper is to show the key role that dynamics can play in accomplishing this task. This is accomplished by establishing a connection, largely unexplored until recently, between the problems of information extraction, manifold embedding and identification of switched systems, and showing that this connection allows for recasting the problem of decision making in "data deluged" scenarios into a tractable convex optimization form.**

I. INTRODUCTION

Recent advances in the past few years, such as the development of inexpensive, energy harvesting sensors, combined with a similar growth in actuation capabilities, rendered feasible a spectrum of new control applications, ranging from zero-emissions buildings to smart grids and managed aquifers to achieve long term sustainable use of scarce resources. Arguably, a major road-block in achieving this vision stems from the curse of dimensionality. Successful operation in these scenarios, requires controllers endowed with the ability to timely extract actionable information from the very large data streams generated by the ubiquitous sensors. However, existing techniques are ill-equipped to deal with this data deluge.

Non-linear dimensionality reduction, e.g. finding low dimensional parsimonious representations of high dimensional correlated data, is a well studied problem in machine learning, where a large number of methods has been proposed. These include, among others, Locally Linear Embeddings (LLE) [31], semi-definite embedding (SDE) [35], Global coordination of local linear models (GCM) [29] and Dynamic Global Coordinate Model (DGCM) [18]. While these methods have proved very efficient in handling static data, most do not exploit dynamical information, encapsulated in the temporal ordering of the data, and thus may fail to capture the underlying temporal dynamics. This issue becomes particularly relevant when these dynamics are the key factor that allows for early detection and classification of anomalies.

The goal of this tutorial paper is to illustrate the key role that dynamics can play in timely extracting and exploiting actionable information that is very sparsely encoded in high dimensional data streams, and to seek a rapprochement between techniques used in the systems and control community and those used in machine learning. The main idea is to treat time series as the output of an underlying switched dynamical system, typically represented by a difference inclusion characterized by a relatively small set of parameters, with jumps indicating the occurrence of events. The key observation, illustrated in Figure 1, is the fact that higher degrees of spatio-temporal correlations in the data lead to lower complexity models, allowing for recasting the problem of information extraction into a sparsification form, which in turn can be reduced to a convex semidefinite optimization problem by exploiting recent results in semi-algebraic optimization. As illustrated in the sequel, embedding the problem of actionable information extraction in the conceptual world of dynamical systems leads to scalable, computationally tractable algorithms, compatible with real time operation in fast changing scenarios, where critical decisions must be made based on information that is very sparsely encoded in very large data streams. As an example, in this framework anomaly detection reduces to simply computing the null

M. Sznaier and O. Camps are with the ECE Dept., Northeastern University, Boston, MA, 02115, email: {msznaier,camps}@coe.neu.edu. N. Ozay is with the EECS Dept., University of Michigan, Ann Arbor, MI 48109, email: necmiye@umich.edu. C. Lagoa is with the EE Dept., Penn State University, University Park, PA, 16802, email: lagoa@engr.psu.edu. This work was supported in part by NSF grants ECCS–1201973, IIS–1318145, CNS–1329422 and ECCS–1404163; AFOSR grant FA9550-12-1-0271, DARPA grant N66001-14-1-4045 and the Alert DHS Center of Excellence under Award Number 2008-ST-061-ED0001.

Sparse Signal Recovery:	**Sparse Information Recovery:**
Strong prior	**Strong prior**
Signal has a sparse representation: $f(t) = \sum_i c_i \psi(t)$ with only a few $c_i \neq 0$.	Actionable information is generated by a bounded complexity switched dynamical system
Signal recovery	**Information recovery**
Sparsify the coefficients:	Sparsify the dynamics:
$\min \lVert [c_1, \ldots, c_n] \rVert_o$	$\mathbf{min}_y \mathbf{rank} \mathbf{V}(\mathbf{y})$
subject to $f(t) = y(t)$.	where $\mathbf{V}(\mathbf{y})$ depends polynomially in the data
Relax to Linear Programming	**Relax to Semidefinite Programming through the use of sparse polynomial optimization methods**

Fig. 1. Sparse dynamical information recovery versus sparse signal recovery.

space of suitably constructed matrices.

The paper is organized as follows. Section II introduces the notation used throughout the paper and some key background results. Section III establishes a connection between dimensionality reduction and Wiener systems identification, and illustrates its application to the problem of activity classification. Section IV shows that several problems arising in the context of information extraction (e.g. segmentation, change detection, identification of contextually abnormal time series) can be recast as identification/model (in)validation of switched affine models and discusses computationally tractable algorithms to solve these problems. The effectiveness of these approaches is illustrated with several practical examples. Section V briefly discusses the use of efficient first order algorithms to address computational complexity and scaling issues. Finally, in Section VI we provide some concluding remarks and point out to open research directions.

II. PRELIMINARIES

For ease of reference, in this section we summarize the notation used in this paper and recall some results on sparse polynomial optimization that play a key role in establishing the main results of this paper.

A. Notation

$\mathbb{R}$, $\mathbb{N}$	set of real number and non-negative integers
$\mathbf{x}, \mathbf{M}$	a vector in $\mathbb{R}$ (matrix in $\mathbb{R}^{m \times n}$)
$\lVert\mathbf{x}\rVert_{w,1}$	weighted ℓ^1 norm: $\lVert\mathbf{x}\rVert_{w,1} \doteq \sum \lvert w_i x_i \rvert$
$\lVert\mathbf{x}\rVert_0$	ℓ^0 quasi-norm, number of non-zero elements in $\mathbf{x}$
$\mathbf{I}$	Identity matrix
$\mathbf{M} \succeq \mathbf{N}$	the matrix $\mathbf{M} - \mathbf{N}$ is positive semidefinite
$\lVert\mathbf{A}\rVert_*$	Nuclear norm: $\lVert\mathbf{A}\rVert_* \doteq \sum \mathrm{svd}(\mathbf{A})$
$\mathbf{H}_y^{m,n}$	Hankel matrix associated with a vector sequence $\mathbf{y}(.)$:

$$\mathbf{H}_y^{m,n} \doteq \begin{bmatrix} \mathbf{y}_0 & \mathbf{y}_1 \cdots & \mathbf{y}_m \\ \mathbf{y}_1 & \mathbf{y}_2 \cdots & \mathbf{y}_m \\ \vdots & \vdots \ddots & \vdots \\ \mathbf{y}_n & \mathbf{y}_{n+1} \cdots & \mathbf{y}_{m+n-1} \end{bmatrix}$$

In the sequel the indexes m, n may be omitted when clear from the context.

B. Moments Based Polynomial Optimization

In this paper, we will reduce the information extraction problem to a polynomial optimization over a semialgebraic set, that is, a problem of the form:

$$p_K^* := \min_{\mathbf{x} \in K} p(\mathbf{x}) \doteq \sum_\alpha p_\alpha \mathbf{x}^\alpha \qquad \text{(P1)}$$

where $\mathbf{x}^\alpha = x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_n^{\alpha_n}$ and $K \subset \mathbb{R}^n$ is a compact semi-algebraic set defined by a collection of polynomial inequalities of the form $g_k(\mathbf{x}) \doteq \sum_\beta g_{k,\beta} \mathbf{x}^\beta \geq 0$, $k = 1, \cdots, d$. In general, problem (P1) is non-convex, and hence hard to solve. Instead, we consider a related problem:

$$\tilde{p}_K^* := \min_{\mu \in \mathcal{P}(K)} \int p(\mathbf{x}) \mu(dx) := \min_{\mu \in \mathcal{P}(K)} \mathbf{E}_\mu \left[p(\mathbf{x}) \right] \qquad \text{(P2)}$$

where $\mathcal{P}(K)$ is the space of positive Borel measures on K with $\int_K d\mu = 1$ and $\mathbf{E}_\mu$ denotes expectation with respect to μ. Although (P2) is an infinite dimensional problem, it is, in contrast to (P1), convex. As shown in [14], Problems (P1) and (P2) are equivalent, in the sense that

- $\tilde{p}_K^* = p_K^*$.
- For every optimal solution μ^* of (P2), $p(x) = p_K^*$, μ^*–almost everywhere.

As briefly summarized next, a finite dimensional sequence of approximations to problem (P2) can be obtained exploiting results from the theory of moments.

Given a sequence of scalars $\{m_\alpha\}$, indexed by a multi-index $\alpha \in \mathbb{N}^n$, the *K-moment problem* is to determine whether there exists a probability measure μ supported on K that has $\{m_\alpha\}$ as its α^{th} moments. That is:

$$m_\alpha = \mathbf{E}_\mu(\mathbf{x}^\alpha) \doteq \int_K \mathbf{x}^\alpha \mu(dx) \tag{1}$$

As shown in [14], existence of such a measure is equivalent to positive semidefiniteness of the (infinite) moment $\mathbf{M}(\mathbf{m})$ and localization $\mathbf{L}(g_k\mathbf{m})$ matrices. Truncated versions of these matrices are given by:

$$\begin{aligned}
&\mathbf{M}_N(\mathbf{m})(i,j) = m_{\alpha^{(i)}+\alpha^{(j)}},\, \forall i,\, j = 1, \cdots, S_N \\
&\mathbf{L}_N(g_k\mathbf{m})(i,j) = \sum_\beta g_{k,\beta^l} m_{\beta^{(l)}+\alpha^{(i)}+\alpha^{(j)}}, \\
&\forall i,\, j = 1, \cdots, S_{N-\lfloor \frac{\text{degree}(g_k)}{2} \rfloor}
\end{aligned} \tag{2}$$

where $S_N = \begin{pmatrix} N+n \\ n \end{pmatrix}$(e.g. the number of moments in $\mathbb{R}^n$ up to order N) and the moments have been arranged according to grevlex ordering of the corresponding monomials so that $\mathbf{0} = \boldsymbol{\alpha}^{(1)} < \ldots < \boldsymbol{\alpha}^{(S_N)}$.

It follows (see [14], [15] for more details) that problem (P1) can be reduced to a sequence of Linear Matrix Inequalities (LMI) optimization problems in the moments of the unknown Borel measure of the form

$$\begin{array}{rll}
p_N^* = & \min_{\mathbf{m}} & \sum_\alpha p_{\boldsymbol{\alpha}} m_{\boldsymbol{\alpha}} \\
& \text{s.t.} & \mathbf{M}_N(\mathbf{m}) \succeq 0, \\
& & \mathbf{L}_N(g_k\mathbf{m}) \succeq 0, k = 1, \ldots, d
\end{array} \tag{3}$$

C. Exploiting the Sparse Structure

The problems considered in this paper exhibit a special sparse structure that can be exploited to reduce the computational complexity entailed in solving (P1).

Definition 1. *Consider problem (P1) and let $I_k \subset \{1, \ldots, n\}$ be the set of indices of variables such that each $g_k(\mathbf{x})$ contains variables only from some I_k. Assume that the objective function $p(\mathbf{x})$ can be partitioned as $p(\mathbf{x}) = p_1(\mathbf{x}) + \ldots + p_l(\mathbf{x})$ where each p_k contains only variables from I_k. Problem (P1) is said to satisfy the running intersection property if there exists a reordering $I_{k'}$ of I_k such that for every $k' = 1, \ldots, l-1$:*

$$I_{k'+1} \cap \bigcup_{j=1}^{k'} I_j \subseteq I_s \quad \text{for some} \quad s \le k' \tag{4}$$

It can be shown [15] that when this property holds, it is possible to construct a convergent hierarchy of semidefinite programs of smaller size:

$$\begin{array}{rll}
p_N^* = & \min_{\mathbf{m}} & \sum_{j=1}^l \sum_{\alpha(j)} p_{j,\boldsymbol{\alpha}(j)} m_{\boldsymbol{\alpha}(j)} \\
& \text{s.t.} & \mathbf{M}_N(\mathbf{m}_{I_k}) \succeq 0, k = 1, \ldots, d, \\
& & \mathbf{L}_N(g_k\mathbf{m}_{I_k}) \succeq 0, k = 1, \ldots, d,
\end{array} \tag{5}$$

where $p_{j,\boldsymbol{\alpha}(j)}$ is the coefficient of the $\boldsymbol{\alpha}(j)^{th}$ monomial in the polynomial p_j, $\mathbf{M}_N(\mathbf{m}_{I_k})$ denotes the moment matrix and $\mathbf{L}_N(g_k\mathbf{m}_{I_k})$ is the localizing matrix for the subset of variables in I_k. Thus, for a given N, this approach requires considering moments and localization matrices containing $O(\kappa^{2N})$ variables, where κ is the maximum cardinality of I_k, rather than $O(n^{2N})$. Since in the problems considered in this paper $\kappa \ll n$ this leads to substantial computational complexity reduction.

D. Rank Minimization and Relaxations

Many of the problems discussed in this paper can be reduced to a constrained rank minimization of the form:

$$\min_{\mathbf{x}} \{\text{rank}[\mathbf{V}(\mathbf{x})]\} \text{ subject to } \mathbf{L}(\mathbf{x}) \succeq 0$$

where the matrices $\mathbf{V}$ and $\mathbf{L}$ depend affinely on $\mathbf{x}$. Although this problem is generically NP–hard, it can be relaxed to a convex optimization by using the fact that $\|.\|_*$ is the convex envelope (e.g the tightest convex relaxation) of rank [8], leading to a problem of the form:

$$\min_{\mathbf{x}} \|\mathbf{V}(\mathbf{x})\|_* \text{ subject to } \mathbf{L}(\mathbf{x}) \succeq 0 \tag{6}$$

It has been shown [28] that under certain conditions on the constraint set, the problem above indeed recovers the minimum rank solution. Unfortunately, in most of the problems of interest in this paper, these conditions do not hold, due to structural constraints. Nevertheless, good solutions can be obtained by using the following iterative re-weighted heuristic [19]:

Algorithm 1 Reweighted $\|.\|_*$ based rank minimization

```
Initialize: $k = 0, \mathbf{W}_y(0) = \mathbf{I}, \mathbf{W}_z(0) = \mathbf{I}$, $\delta_o$ small
repeat
    Solve
        $\min_{\mathbf{X}^{(k)},\mathbf{Y}^{(k)},\mathbf{Z}^{(k)}} \text{Trace} \begin{bmatrix} \mathbf{W}_y^{(k)}\mathbf{Y}^{(k)} & 0 \\ 0 & \mathbf{W}_z^{(k)}\mathbf{Z}^{(k)} \end{bmatrix}$
        subject to: $\begin{bmatrix} \mathbf{Y}^{(k)} & \mathbf{X}^{(k)} \\ \mathbf{X}^{(k)} & \mathbf{Z}^{(k)} \end{bmatrix} \succeq 0$
        $\mathbf{X}^{(k)} \in \mathcal{S}$
    where $\mathcal{S}$ is the feasible set in (6).
    Decompose $\mathbf{X}^{(k)} = \mathbf{UDV}^T$.
    Set $\delta \leftarrow \min[\text{diag}(\mathbf{D})] + \delta_0$.
    Set $\mathbf{W}_y^{(k+1)} \leftarrow \left(\mathbf{Y}^{(k)} + \delta\mathbf{I}\right)^{-1}$
    Set $\mathbf{W}_z^{(k+1)} \leftarrow \left(\mathbf{Z}^{(k)} + \delta\mathbf{I}\right)^{-1}$
    Set $k \leftarrow k + 1$.
until a convergence criterion is reached.
return $\mathbf{X}^{(k)}$
```

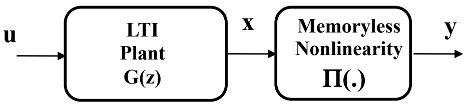


Fig. 2. Wiener System Model

III. Manifold Embedding of Dynamic Data as a Wiener Identification Problem

Conventional dimensionality reduction methods exploit spatial correlations in the data to substantially reduce its dimensionality by embedding it in low dimensional manifold, via non-linear projections. However, while these techniques preserve local spatial neighborhoods, they typically fail to exploit temporal information and thus are not well suited for dealing with dynamic data. The starting point to obtain embeddings that respect dynamical constraints is the realization that, since the projections to/from the embedding manifolds can be modeled as memoryless (possibly, time varying) nonlinearities, the observed (high-dimensional) data can be considered to be trajectories of a Wiener system of the form illustrated in Fig. 2, where the signals $\mathbf{y} \in R^{n_y}$ and $\mathbf{x} \in R^{n_x}$, with $n_y \gg n_x$, represent the raw data and its projection. In this context, the embedding manifold is implicitly described by the output space of G and the data is associated to the pair $\{G, \Pi\}$. Note that while robust identification of Wiener systems is known to be generically NP–hard [33], computationally tractable relaxations can be obtained by recasting the problem into a rank minimization form, proceeding as in [39].

The approach above assumes that n_x, the dimension of the embedding manifold, is known, as is the case in several popular machine learning methods such as Locally Linear Embeddings [31]. However, in many cases of practical interest, this information is not a-priori available. Examples of this situation are computer vision applications such as target tracking or activity recognition [17], [36], where the output $\mathbf{y}$ consists of the vectorized frames of a video sequence, and $\mathbf{x}$ is a small set of independent parameters that encapsulate the correlations between the different pixels. In these cases the dimension of $\mathbf{x}$ must also be identified from the experimental data, a situation that cannot be handled by conventional Wiener systems identification techniques. Motivated by this difficulty, [37] recently introduced a new approach, briefly outlined below, that allows for both, finding an embedding manifold such that the data can indeed be explained as a trajectory of a Wiener system, and identifying its linear and non-linear portions. A salient feature of this approach (common in machine learning, but until recently not used in the identification community), is the ability to use both positive and negative samples, that is, experimental data generated both by the system to be identified and by other systems. This is a situation commonly encountered in applications such as activity classification, where sample clips of different activities are available, or in tracking, where often a segmentation separating the target of interest from other targets and the background is known.

Briefly (see [37] for details), the goal is to use a nonlinear projection $\mathbf{x}_t = \Pi(\mathbf{y}_t)$ to embed a given ordered temporal sequence $\{\mathbf{y}_t\}$, in a manifold where its evolution $\{\mathbf{x}_t\}$ can be (locally) explained by a linear model of the form:

$$\mathbf{x}_t = \sum_{i=1}^{n_a} a_i \mathbf{x}_{t-i} + \sum_{i=1}^{n_b} b_i \mathbf{u}_{t-i} + \eta_t, \quad |\eta_t| \le \epsilon_t$$

where η_t accounts for approximation error[1]. In the sequel, for simplicity we will assume that $\mathbf{u}_t = 0$, that is, the data has been generated by an ARX model driven by noise and initial condition. This assumption captures the case, common in many applications such as computer vision, that lack controlled exogenous inputs. Further, the framework can be easily extended to encompass control inputs by considering the associated matrix $\mathbf{H}_u$ (see [37] for details). In this context, to each embedded time series $\mathbf{x}_t$, we can associate its Hankel matrix $\mathbf{H}_\mathbf{x}$. Since the vector $\mathbf{w} \doteq \{a_1, \ldots, a_{n_a}, -1\}$ satisfies $\mathbf{H}_\mathbf{x}\mathbf{w} = 0$, it follows that the dynamic data is completely characterized by the null space of $\mathbf{H}_x$. Keeping in mind that the goal is to find representations that are optimally suited for classifying time series and detecting anomalies, [37] proposed to use the extra degrees of freedom available to optimize the margin between classes[2]. Specifically, given two sets of training sequences, $\{\mathbf{y}_t^+\}$ and $\{\mathbf{y}_t^-\}$ corresponding to nominal and anomalous scenarios, this approach jointly seeks for embeddings $\mathbf{x}_t^+, \mathbf{x}_t^-$ and a vector $\mathbf{w}$ such that minimizes γ subject to $\|\mathbf{H}_{\mathbf{x}^+}\mathbf{w}\|_2^2 \le \gamma$ and $\|\mathbf{H}_{\mathbf{x}^-}\mathbf{w}\|_2^2 > 1 + \gamma$. Intuitively, it seeks a vector $\mathbf{w}$ such that (i) it approximately lies in the null space of the Hankel matrices of all the positive examples dynamic sequences, and (ii) it maximizes the margin between the residue $\|\mathbf{H}_{\mathbf{x}^+}\mathbf{w}\|_2^2$

[1] Such a representation always exists (locally) since hinging hyperplanes are universal approximators [32], [5].

[2] Recall that due to the fact that the experimental data record is finite and corrupted by noise, there exist multiple models that interpolate the data within the noise margin. The set of all such models constitutes the consistency set (see for instance [7] or Chapter 10 in [30]). The algorithm described here selects a model from this set and thus is interpolatory.

for the nominal and anomalous sequences. The problem outlined above can be formalized as:

$$\begin{array}{l} \min_{\mathbf{G_i}\succeq 0,\mathbf{w},\gamma\geq 0} \frac{1}{2}||\mathbf{w}||_2^2 + C\gamma \\ \text{subject to: } \mathbf{w}^T\mathbf{G}_i\mathbf{w} \leq \gamma \quad ,\forall \mathbf{G}_i \in \mathbf{G}_+ \\ \mathbf{w}^T\mathbf{G}_i\mathbf{w} + \gamma \geq 1 \quad ,\forall \mathbf{G}_i \in \mathbf{G}_- \end{array} \tag{7}$$

where $\mathbf{G}_i \doteq \mathbf{H}_i^T\mathbf{H}_i$, $\mathbf{G}_+$ and $\mathbf{G}_-$ denote the positive (or in-class) and negative (out of class) sequences, respectively, and where the regularization term $\frac{1}{2}||\mathbf{w}||_2^2$ is added to the objective to prevent trivial solutions with unbounded $\mathbf{w}$.

Note that the formulation above tries to maximize the separation between classes, without taking into consideration the dimension of the resulting embedding manifold. Indeed, proceeding as in [35], it can be shown that this dimension is given by rank($\mathbf{K}$), the Kernel (or Gram) matrix defined by its submatrices

$$\mathbf{K}_{i,j} = \begin{bmatrix} \mathbf{x}_j\mathbf{x}_j & \mathbf{x}_j\mathbf{x}_{j+1} & \cdots & \mathbf{x}_j\mathbf{x}_{j+c} \\ \mathbf{x}_{j+1}\mathbf{x}_j & \mathbf{x}_{j+1}\mathbf{x}_{j+1} & \cdots & \mathbf{x}_{j+1}\mathbf{x}_{j+c} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_{j+c}\mathbf{x}_j & \mathbf{x}_{j+c}\mathbf{x}_{j+1} & \cdots & \mathbf{x}_{j+c}\mathbf{x}_{j+c} \end{bmatrix}$$

It follows that the dimension of the embedding manifold can be minimized by minimizing the rank of $\mathbf{K}$. Further, as argued in [35], this can be accomplished by maximizing its trace. Finally, noting that $\mathbf{G}_i = \mathbf{H}_i^T\mathbf{H}_i = \sum_{j=1}^{T-c+1}\mathbf{K}_{i,j}$, leads to the following formulation that balances separation margin against the dimension of the embedding manifold through the tuning parameter λ:

$$\begin{array}{l} \min_{\mathbf{K},\mathbf{w},\gamma} \frac{1}{2}||\mathbf{w}||_2^2 + C\gamma - \lambda\text{Trace}(\mathbf{K}) \\ \text{subject to: } \mathbf{w}^T\mathbf{G}_i\mathbf{w} \leq \gamma \quad ,\forall \mathbf{G}_i \in \mathbf{G}_+ \\ \mathbf{w}^T\mathbf{G}_i\mathbf{w} + \gamma \geq 1 \quad ,\forall \mathbf{G}_i \in \mathbf{G}_- \\ \mathbf{G}_i = \sum_{j=1}^{T-c+1}\mathbf{K}_{i,j} \\ \mathbf{K} \succeq 0,\ \gamma \geq 0 \\ (1-\epsilon)||\mathbf{y}_i - \mathbf{y}_j||^2 \leq k_{ii} + k_{jj} - 2k_{ij} \\ \leq (1+\epsilon)||\mathbf{y}_i - \mathbf{y}_j||^2 \end{array} \tag{8}$$

where the last constraint approximately enforces preservation of the local spatial geometry. Since the problem above is semi-algebraic, it can be solved proceeding as in [15]. Finally, once the manifold projections $\mathbf{x}$ have been found, if needed, the non-linearity can be found for instance by approximating it by a semi-algebraic function and solving an interpolation problem. It is worth emphasizing that the algorithm outlined above, rather than working with the potentially high dimensional data $\mathbf{y}$, uses the inner products $\langle \mathbf{y}_i, \mathbf{y}_j \rangle$, resulting in substantial dimensionality reduction in the matrices involved. Thus, it can comfortably handle large-sized problems. The effectiveness of this approach is illustrated in Fig. 3 and Table I showing the results of an experiment where it was used to classify human activities using video data from the Weizmann dataset [12]. The data consisted of 13 frames with 2400 pixels/frame from 18 videos with an equal number of "running" and "skipping" activities, performed by different actors. 10 of these sequences were used for training and the remaining 8 sequences for testing. As shown in Table I, the algorithm achieved perfect classification, even though there are substantial differences between the appearance of the subjects. For comparison, a similar approach but using only positive (e.g. in-class) training data had a substantial misclassification rate, due to the similarity between the two activities considered.

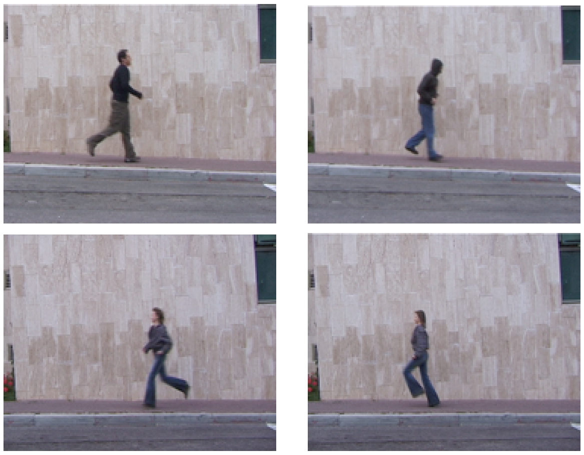

Running Skipping

Fig. 3. Sample frames from Weizmann data set. Top: training data. Bottom: testing data.

TABLE I

CLASSIFICATION RESULTS FOR THE WEIZMANN DATASET, WITH $\lambda = 0.1$ AND $C = 1 \times 10^2$: 100% ACCURACY.

Train $\|\mathbf{G}_i\mathbf{w}\|_2^2$		Test $\|\mathbf{G}_i\mathbf{w}\|_2^2$	
+	−	+	−
0.0116	1.9365	0.1912	2.6805
0.0104	17.9540	0.1025	0.6422
0.0169	1.0007	0.0521	0.5266
0.0063	1.8692	0.0985	0.7232
0.0239	1.0279		
$\mu = 0.0138$	$\mu = 4.7577$	$\mu = 0.1111$	$\mu = 1.1431$
$\sigma = 0.0061$	$\sigma = 6.6102$	$\sigma = 0.0503$	$\sigma = 0.8903$

IV. INFORMATION EXTRACTION AS A HYBRID SYSTEMS IDENTIFICATION/(IN)VALIDATION PROBLEM

As outlined in the introduction, the main idea driving this paper is to treat the observed data as the output of an underlying switched dynamical system, with events

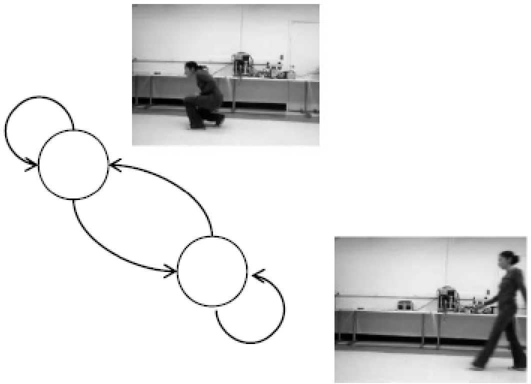


Fig. 4. Activity segmentation as a piecewise affine identification problem. Here each node in the graph corresponds to a single activity. The goal is to identify both the switching instants and the model corresponding to each activity, as a first step towards detecting contextually abnormal sequence of activities.

indicated by changes in invariants associated with each subsystem. In the sequel, we will assume, by using the embedding procedure outlined in the previous section, if necessary, that the data record has been generated by a piecewise affine model of the form[3]

$$f\left(\mathbf{p}_{\sigma(t)}, \left\{\mathbf{x}(k), \boldsymbol{\eta}_f(k)\right\}_{k=t-i}^{t+j}\right) = 0 \tag{9}$$

where f is an affine function, the parameter vector $\mathbf{p}(t)$ takes values from a finite set indexed by piecewise constant function $\sigma(t)$ and where $\boldsymbol{\eta}_f(t)$ represents bounded noise. In this context, the information is encapsulated in the parameter vector $\mathbf{p}$. For instance, events are indicated by changes in $\mathbf{p}(t)$ (an identification problem). Similarly, two time series can be considered to have been generated by the same process if they can be explained by the same $\mathbf{p}$ (e.g. a model (in)validation problem). While both, identification and model (in)validation of switched affine systems are known to be NP-hard problems, in the past few years several computational tractable relaxations have been developed (see for instance [10], [25], [3], [22], [23], [24], [9], [16], [2] and references therein). In particular, in the sequel we will cover a class of algorithms based on recasting the problem into a sparsification form. These approaches are attractive since they allow for exploiting the inherently sparse structure of the problem and provide a sequence of increasingly tighter relaxations, along with means of certifying both when an optimal solution has been obtained or the underlying optimization problem is infeasible.

Before proceeding, note that the problem of identifying a model of the form (9) that explains a given data record is ill-posed, in the sense that it admits infinitely many solutions. For instance, it is always possible to satisfy (9) by fitting an hyperplane to each data point, or, alternatively, a single model, of sufficiently high order that perfectly interpolates the given data. In order to avoid this ambiguity, additional criteria should be imposed. In particular, in the sequel we consider two different scenarios (a) identification with minimum number of switches, and (b) identification with a given number of subsystems. The first scenario arises for instance in fault detection, where the goal is to minimize the number of false alarms, and in segmentation problems in image processing and computer vision, where it is often desirable to maximize the size of regions (roughly equivalent to minimizing the number of boundaries). The second situation arises for instance in cases where it is known a-priori that the system switches a finite number of states and the goal is to characterize and recognize these. Examples of this scenario include recognizing human activities in video sequences (Fig. 4) or metabolic stages from gfp activated genomic data.

A. *Identification with minimum number of switches*

Formally, this problem can be stated as: Given input/output data $\{u_t, y_t\}_{t_0}^{T}$ 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 switched affine model of the form:

$$y_t = \sum_{i=1}^{n_a} a_i(\sigma_t) y_{(t-i)} + \sum_{i=1}^{n_c} c_i(\sigma_t) u_{(t-i)} + \eta_t \tag{10}$$

where u, y and η denote the input, output and noise, respectively, that explains the experimental data with the minimum number of switches.

In order to solve this problem, start by defining the sequence of first order differences $\boldsymbol{\delta}_t \doteq \mathbf{p}_t - \mathbf{p}_{(t+1)}$. Since $\boldsymbol{\delta}_t = 0 \iff \mathbf{p}_t = \mathbf{p}_{(t+1)}$, it follows that the number of switches can be minimized by maximizing the sparsity of the sequence $\{\boldsymbol{\delta}_t\}$, or, equivalently, by solving the following sparsification problem:

$$\begin{array}{c} \min_{\mathbf{p}_t} \|\mathbf{p}_t - \mathbf{p}_{(t+1)}\|_0 \\ \text{subject to } y_t - \mathbf{r}_t^T \mathbf{p}_t \in \mathcal{N}\ \forall t \end{array} \tag{11}$$

While the problem above is non-convex, convex relaxations can be obtained by relaxing $\|.\|_o$ to $\|.\|_{\infty,1}$ [6]. Notably, the main result in [24] shows that when the

[3] Note that this can be assumed without loss of generality, since piece wise affine models are universal approximators [32], [5]. Nevertheless, modeling the data as the output of a switched Wiener system allows leveraging the nonlinearity to obtain more compact descriptions.

noise is characterized in terms of its ℓ_∞ norm, that is $\mathcal{N} = \{\eta \colon \|\eta\|_\infty \leq \epsilon\}$, then an exact solution can be found by solving a sequence of Linear Programming problems, proceeding as outlined in Algorithm 2.

```
Algortihm 2: Identification with
 minimum number of switches
$k = 0$
$t_0 = \max(n_y, n_u)$
$\tau_k = t_0$
FOR $i = t_0 : T$
    Solve the following feasibility problem in $\mathbf{p}$:
        $\mathcal{F} : \{ \; |y_t - \mathbf{r}_t^T \mathbf{p}| \leq \epsilon \quad \forall t \in [\tau_k, i] \; \}$
    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 $k$
```

Application: Segmentation of Video Sequences. In this example we illustrate the application of Algorithm 2 to a non-trivial computer vision problem: segmentation of video-sequences. The data consisted of 250 frames of the sequence *family.avi*, available from http://www.open-video.org. A low order representation $\mathbf{y}_t \in \mathbb{R}^{35}$ of this sequence was obtained by first converting each frame to gray scale, vectorizing, subtracting the sample mean and projecting onto a $d = 35$ manifold. Fig. 5 compares the performance of Algorithm 2, using 3rd order models, to that of three standard methods: GPCA [34], a histogram based algorithm (bin to bin difference (B2B) with 256 bin histograms and window average thresholding [11], and MPEG [38]. As shown there and in Table II, recasting video segmentation as a switched systems identification yields the best performance. It is also worth emphasizing that the second best performer (MPEG) requires manual tuning, by trial and error, of up to seven parameters and its performance is highly sensitive to these values. Additional comparisons are given in [24].

TABLE II

RAND INDICES [27] FOR THE FAMILY SEQUENCE SEGMENTATION EXAMPLE. A RAND INDEX OF 1 INDICATES PERFECT SEGMENTATION.

Sequence	Sparsification	MPEG	GPCA	B2B
family	0.9946	0.9480	0.8220	0.9078

B. Identification with bounded number of subsystems.

In this case, the problem can be formally stated as: Given input/output data over the interval $[t_0; T]$, a bound on the ℓ_∞ norm of the noise (i.e. $\|\eta\|_\infty \leq \epsilon$) find a switched ARX model of the form (10), with no more than s subsystems, that interpolates the experimental data. Although in principle this problem is NP-hard, in the noise free case (i.e. $\eta_t = 0 \; \forall t$), it can be reduced to finding the null space of a suitable constructed matrix, followed by polynomial differentiation [34]. The starting point to accomplish this is to rewrite (10) as

$$\mathbf{b}(\sigma_t)^T \mathbf{r}_t = 0 \tag{12}$$

where $\mathbf{r}_t = \left[-y_t, y_{t-1}, \ldots, y_{t-n_a}, u_{t-1}, \ldots, u_{t-n_c}\right]^T$ and $\mathbf{b}(\sigma_t) = \left[1, a_1(\sigma_t), \ldots, a_{n_a}(\sigma_t), c_1(\sigma_t), \ldots\right]^T$, denote the regressor and (unknown) coefficients vectors at time t, respectively. The idea behind the Generalized Principal Component Analysis (GPCA) method [34] is to decouple the identification of model parameters from the identification of the switching sequence by noting that (12) holds for some σ_t if and only if

$$p_s(\mathbf{r}) = \Pi_{i=1}^{s}(\mathbf{b}_i^T \mathbf{r}_t) = \mathbf{c}_s^T \nu_s(\mathbf{r}_t) = 0 \tag{13}$$

holds for all t independent of which of the s submodels is active at time t, where $\mathbf{b}_i \in R^{n_a+n_c+1}$ and $\nu_s(.)$, denote the parameter vector corresponding to the i^{th} submodel and the Veronese map of degree s, respectively. Collecting all data into a matrix form leads to:

$$\mathbf{V}_s \mathbf{c}_s \doteq \begin{bmatrix} \nu_s(\mathbf{r}_{t_o})^T \\ \vdots \\ \nu_s(\mathbf{r}_T)^T \end{bmatrix} \tag{14}$$

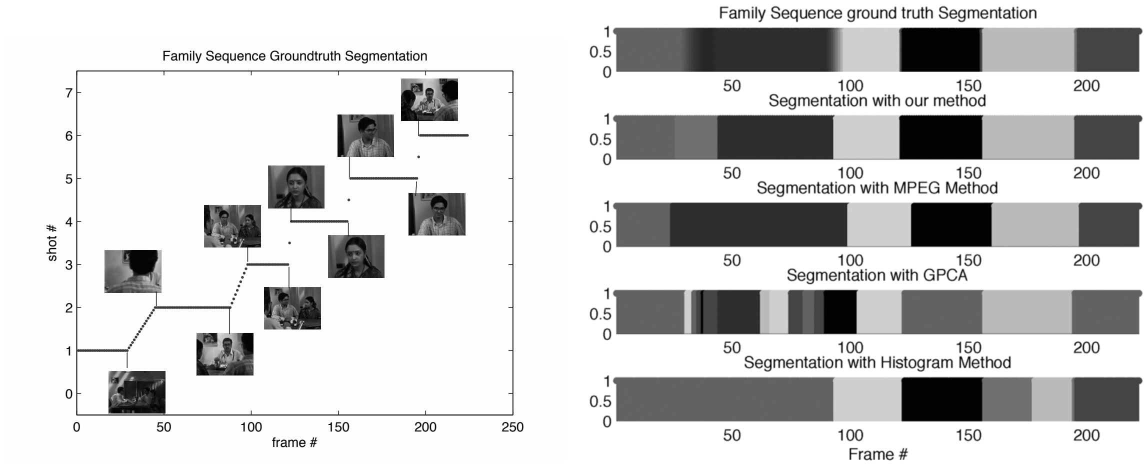
Hence, one can solve for a vector $\mathbf{c}_s$ in the null space of $\mathbf{V}_s$ to find the coefficients of the multivariate polynomial $p_s(\mathbf{r})$. Unfortunately, this approach breaks down in the presence of noise, since (13) no longer holds. Rather, we have the following (noisy) equivalent

$$p_s(\tilde{\mathbf{r}}_t) \doteq \prod_{i=1}^{s}(\mathbf{b}_i^T \tilde{\mathbf{r}}_t) = \mathbf{c}_s^T \nu_s(\tilde{\mathbf{r}}_t) = 0 \tag{15}$$

where $\tilde{\mathbf{r}}_t = [-y_t + \eta_t, y_{t-1}, \ldots, u_{t-1}, \ldots, u_{t-n_c}]^T$, and its associated "noisy" data matrix $\mathbf{V}_s(\mathbf{r}, \boldsymbol{\eta}) \doteq \mathbf{V}_s(\tilde{\mathbf{r}})$. The main difficulty here is that finding the coefficients of the polynomial $p_s(\tilde{\mathbf{r}}_t)$ requires now finding both an admissible noise sequence $\boldsymbol{\eta}^o$ and a vector $\mathbf{c}^o$ such that

$$\mathbf{V}_s(\tilde{\mathbf{r}}^o)\mathbf{c}^o = 0 \tag{16}$$

Since $\mathbf{V}_s(\tilde{\mathbf{r}})$ is now a matrix polynomial function of the unknown noise sequence η_t, this is a computationally



(a)

(b)

Fig. 5. 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.

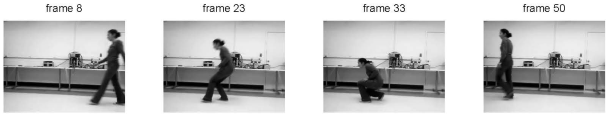


Fig. 6. Sample frames of a video sequence with a human performing two activities: walking and bending.

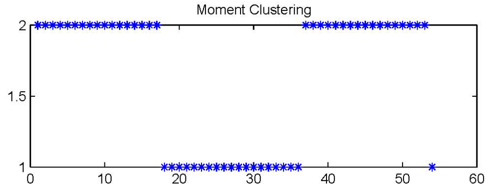


Fig. 7. Activity segmentation via a moments based method.

very challenging problem. Nevertheless, as briefly outlined below, the use of the polynomial optimization tools described in section II-B, allows for transforming (16) to the minimization of the rank of a matrix that is affine in the optimization variables. This can be accomplished by using the key fact, established in Ozay et al. [23], that there exists an admissible noise sequence η^o such that (16) is satisfied for some vector $\mathbf{c}^o$ if and only if there exists an admissible moments sequence $\mathbf{m}$ such that

$$\begin{aligned}\tilde{\mathbf{V}}_s(\mathbf{r}_t,\mathbf{m}^{(t)})\mathbf{c}^o = 0\\ \text{subject to } \mathbf{M}(\mathbf{m})\succeq 0 \text{ and } \mathbf{L}(\mathbf{m})\succeq 0\end{aligned} \tag{17}$$

where $\tilde{\mathbf{V}}_s(\mathbf{r}_t,\mathbf{m}^{(t)})$ is the matrix obtained by replacing each k^{th} degree monomial η_t^k in $\mathbf{V}_s(\mathbf{r}_t,\eta_t)$ with the corresponding k^{th} order moment $m_k^{(t)}$, and where $\mathbf{M}(\mathbf{m})$ and $\mathbf{L}(\mathbf{m})$ denote the (truncated) moment and localization matrices corresponding to the constraint $\|\eta_t\|_\infty \leq \epsilon$. Since $\tilde{\mathbf{V}}_s(\mathbf{r}_t,\mathbf{m}^{(t)})$ is affine in $\mathbf{m}$, it follows that a suitable pair $\{\mathbf{m},\mathbf{c}^o\}$ can be found by (approximately) minimizing rank $\tilde{\mathbf{V}}_s(\mathbf{r}_t,\mathbf{m}^{(t)})$, using for instance the algorithm outlined in section II-D. Once a rank deficient $\tilde{\mathbf{V}}_s(\mathbf{r}_t,\mathbf{m}^{(t)})$ is found, the parameters of each subsystem can be obtained by simply finding a vector $\mathbf{c}^o$ in its null space and then proceeding as in the noise free case.

Example: Human activity segmentation. Next, we illustrate an application of the ideas outlined above to the problem of human activity analysis from video data. In this case, the goal is to segment the video sequence shown in Fig. 6 into its constituent activities: walking and bending. In this case, the data was pre-processed to estimate the location of the center of mass of the person in each frame. The horizontal[4] position of the center of mass was then modeled as the output of a first order switched affine system:

$$x_t = a(\sigma_t)x_{t-1} + b(\sigma_t) + \eta_t$$

where $a(\sigma_t)$ and $b(\sigma_t)$ are unknown. Finally, the noise level was estimated to be $\|\eta_t\|_\infty = 3$, (e.g 3 pixels). Running the identification algorithm outlined above with these priors lead to the segmentation shown in Figure

[4] It may seem more natural to use the vertical position. However, this led to 3 segments, corresponding to no vertical motion, downward and upward motion, while there are only two different activities involved.

7, illustrating the ability of this approach to correctly classify the frames according to the underlying activity[5]. Finally, it is worth emphasizing that this technique can be easily extended to handle outliers and corrupted data, proceeding as shown in [20].

C. Model (In)Validation

Classically, model (in)validation has been used as an intermediate step following identification and prior to use the identified models for control synthesis. Interestingly, as illustrated in the sequel, the same ideas can be used in the context of information extraction to identify contextually abnormal sequences. Formally, the problem of interest can be stated as establishing whether a noisy input/ouput sequence could have been generated by a given model of the form[6]:

$$\begin{array}{rcl} y_t & = & \sum_{i=1}^{n_a} a_i(\sigma_t) y_{t-i} + \sum_{i=1}^{n_c} c_i(\sigma_t) u_{t-i} \\ \tilde{y}_t & = & y_t + \eta_t,\ \sigma_t \in \{1,\ldots,s\},\ \|\eta_t\|_\infty \leq \epsilon \end{array} \tag{18}$$

where $\tilde{y}_t$ denotes the measured output corrupted by the noise η_t. As in the identification case, this problem is known to be generically NP-hard, due to the presence of noise and the fact that the mode variable σ_t is not directly measurable. Cases where σ_t takes only a small number of discrete values (for instance a system switching between two known modes), can be handled by simply considering all possible switching sequences. Clearly, due to its combinatorial nature, this approach becomes infeasible for cases involving relatively small number of subsystems. On the other hand, this combinatorial complexity can be avoided by appealing to semi-algebraic geometry tools. To this effect, begin by noting that, as in section IV-B, (18) holds if and only if:

$$p_r(\eta_{t:t-n_c}) \doteq \prod_{i=1}^{s} \left[g_{t,i}(\eta_{t:t-n_c})\right]^2 = 0 \tag{19}$$

where:

$$\begin{array}{l} g_{t,i}(\eta_{t:t-n_c}) \doteq a_1(i)(\tilde{y}_{t-1} - \eta_{t-1}) + \ldots + \\ a_{n_a}(i)(\tilde{y}_{t-n_a} - \eta_{t-n_a}) - (\tilde{y}_t - \eta_t) + c_1(i)u_{t-1} \\ + \ldots + c_{n_c}(i)u_{t-n_c} \end{array} \tag{20}$$

Similarly, the norm constraint on the noise sequence η_t is equivalent to the polynomial constraint $h_t(\eta_t) \doteq \epsilon^2 - \eta_t^2 \geq 0$. Hence, there exists noise and switching sequences such that (19) holds if and only if the semi-algebraic set

$$\begin{array}{rcl} \mathcal{T}(\eta) & \doteq & \{\eta \mid f_t(\eta_t) \geq 0\ \forall t \in [t_o, T] \text{and} \\ & & p_t(\eta_{t:t-n_a}) = 0\ \forall t \in [n_a, T]\}\,. \end{array} \tag{21}$$

is not empty. Thus, an (in)validation certificate can be obtained by using semi-algebraic geometry techniques to establish whether $\mathcal{T} = \emptyset$. In particular, the use of the *Positivstellensatz* [26] allows to obtain a hierarchy of convex relaxations. Alternatively, proceeding as in [21], leads to the following optimization problem:

$$\begin{array}{l} o^* = \min_\eta \sum_{t=n_a}^{T} p_t(\eta_{t:t-n_a}) \\ \text{s.t. } f_t(\eta_t) \geq 0\ \ \forall t \in [0,T] \end{array} \tag{22}$$

Note that $o^* > 0 \iff \mathcal{T}'(\boldsymbol{\eta}) = \emptyset$. While computing o^* requires solving a computationally challenging polynomial optimization problem, a convergent sequence of lower bounds can be obtained using the tools described in Section II-B as follows: Let d_N^* denote the solution to the following convex optimization:

$$\begin{array}{rl} d_N^* = & \min_{\mathbf{m}} \sum_{t=n_a}^{T} l_t(\mathbf{m}_{t-n_a:t}) \\ & \text{s.t.} \\ & \mathbf{M}_N(\mathbf{m}_{t-n_a:t}) \succeq 0\ \forall t \in [n_a, T] \\ & \mathbf{L}_N(f_t \mathbf{m}_{t-n_a:t}) \succeq 0\ \ \forall t \in [n_a+1, T] \end{array} \tag{23}$$

where each l_t is the linear functional of moments defined as $l_t(\mathbf{m}_{t-n_a:t}) \doteq \mathbf{E}\left\{p_t(\boldsymbol{\eta}_{t:t-n_a})\right\}$, and where $\mathbf{M}_N$ and $\mathbf{L}_N$ are the corresponding (truncated) moments and localization matrices. Then, from the results in Section II-B, it follows that $d^* \uparrow o^*$. Moreover, if for some N_o, $d_{N_o}^* > 0$, then $\mathcal{T} = \emptyset$. It is worth emphasizing that this reformulation allows for exploiting the inherently sparse structure of the problem. It can be shown (see [21] for details) that (22) satisfies the running intersection property, and hence it can be solved considering smaller subproblems as outlined in Section II-C.

Application: detecting contextually abnormal activities: The goal here is to detect activities that do not belong to a database of known, safe activities. Examples of application include monitoring assisted living facilities and public spaces. A difficulty here is that typically a video clip contains several activities, e.g. walking for two minutes, standing for one, and then resuming walking, and thus, in principle, parsing of the sequence is required, a hard task if each segment is only a few frames long. On the other hand, the need for explicit parsing can be eliminated by recasting the problem into a model (in)validation form. In this context, a model is associated to each of the activities in the database, and abnormal sequences are those that cannot be explained as the trajectory of a system that switches amongst these, precisely the situation addressed by the

[5] The single misclassification in the last frame of the sequence is due to an inaccurate estimate of the centroid of the person as she starts to leave the field of view of the camera.

[6] For simplicity, we consider here SISO models. A treatment of the MIMO case can be found in [21].

switched (in)validation framework described above. An example of application of these ideas is shown in Fig. 8.

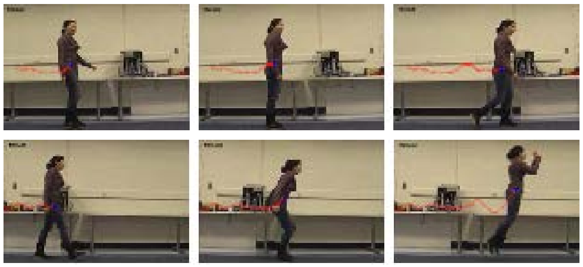


Fig. 8. Anomalous behavior detection as a switched (in)validation problem. The activity database consists of models of two activities, walk and wait. The top sequence (walkwaitwalk) is not (in)validated since both activities are in the database. The bottom sequence (walk-jump) is flagged as abnormal since it cannot be generated by switching amongst models in the database.

V. Computational Complexity Considerations

In the previous sections we have shown that recasting information extraction problems into an identification/model (in)validation form leads to semi-definite optimization problems. Since these problems are convex, they can be solved in polynomial time, using for instance interior point methods implemented in the freely available software package CVX [13]. Nevertheless, while these methods work well for moderately sized problems, they have poor scaling properties (computational complexity scales at least as $\mathcal{O}(n^3)$ and memory requirements as $\mathcal{O}(n^4)$, where n is the number of decision variables). Thus, even when using the methods described in Section III to reduce the dimensionality of the data, using interior point methods is feasible for relatively short sequences (typically a few hundred data points). Motivated by this difficulty, during the past few years considerable interest has been devoted to Augmented Lagrangian Type first order methods [4]. These methods are both computationally (since they can exploit closed form solutions to partial problems) and memory efficient (since they do not require computing Hessians). Note that the type of problems of interest in this paper are special cases of a more general class, structurally constrained nuclear norm minimization, of the form:

$$\begin{aligned}&\min_{\mathbf{m}} \|\mathbf{WA}(\mathbf{m})\|_* + \lambda_1 \|\mathbf{E_1}(\mathbf{m})\|_1 + \lambda_2 \|\mathbf{E_2}(\mathbf{m})\|_2 \\ &\text{subject to semidefinite constraints}\end{aligned} \tag{24}$$

where $\mathbf{W}, \lambda_1$ and λ_2 are fixed weights, and $\mathbf{A}(.), \mathbf{E}_1(.)$ and $\mathbf{E}_2(.)$. are structured matrices that depend affinely on the optimization variable $\mathbf{m}$. The main result in [1], shows that these problems can be solved using an Alternating Directions Methods of Multipliers (ADMM) type algorithm that requires performing only thresholding and eigenvalue decomposition steps. Notably, this algorithm typically requires computing only a few singular values, hence avoiding the $\mathcal{O}(n^3)$ scaling of full blown SVDs, resulting in up to two orders of magnitude improvement in computational time, vis-a-vis conventional SDP solvers, with far lower memory requirements [1].

VI. Conclusions

The development of low cost, ultra low power sensors and the parallel development in actuation capabilities have rendered feasible a spectrum of new control applications, ranging from zero emission buildings to reconfigurable, self aware environments. that can profoundly impact society. However, realizing this potential, requires endowing controllers with the ability to timely extract actionable information from the very large data streams generated by the sensors, a goal that challenges the capabilities of existing techniques. As shown in this paper, embedding the problem of actionable information extraction into a dynamical systems identification form opens up a large knowledge base developed in the systems and control community, leading to computationally attractive algorithms. In this context, data is associated with a few dynamic invariants that describe the underlying dynamical model, allowing for leveraging the inherent sparsity of these representations to robustly solve challenging problems such as data segmentation and interpretation, even in the presence of corrupted data. An issue only partially addressed in this paper is the computational complexity of the resulting methods. While, as briefly indicated in Section V, the use of ADMM based methods can partially alleviate this issue, these methods are still limited to medium sized sequences (few thousands of elements), due to the need to perform SVDs, and, in some cases, relatively slow convergence. A promising approach for identifying systems subject to sparsity constraints, based on the concept of atomic norm minimization using a modified Frank-Wolfe type algorithm has been recently proposed in [40], where it was shown to outperform ADMM based methods. Further, this approach only entails computing inner products and thus its complexity grows linearly with the data. However, at this point it is not clear how to extend this approach to deal with switching systems, with the main difficulty here being the development of Frank-Wolfe type algorithms for systems that can switch arbitrarily fast.

REFERENCES

[1] M. Ayazoglu and M. Sznaier. An algorithm for fast constrained nuclear norm minimization and applications to systems identification. In *2012 IEEE CDC*, pages 3469–3475, December 2012.

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

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

[4] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.*, 3(1):1–122, 2011.

[5] L. Breiman. Hinging hyperplanes for regression, classification and function approximation. *IEEE Trans. Inf. Theory*, pages 999–1013, 1993.

[6] E. J. Candes, M. Wakin, and S. Boyd. Enhancing sparsity by reweighted l1 minimization. *Journal of Fourier Analysis and Applications*, 14(5):877–905, December 2008.

[7] J. Chen and G. Gu. *Control Oriented System Identification, An $\mathcal{H}_\infty$ Approach*. John Wiley, New York, 2000.

[8] M. Fazel, H. Hindi, and S. P. Boyd. Log-det heuristic for matrix rank minimization with applications to hankel and euclidean distance matrices. In *Proceedings of American Control Conf. 2003*, volume 3, pages 2156–2162. AACC, 2003.

[9] C. Feng, C. Lagoa, N. Ozay, and M. Sznaier. Hybrid system identification: An SDP approach. In *Proc. 2010 IEEE Conf. on Dec. and Control (CDC)*, pages 1546–1552, Dec. 2010.

[10] G. Ferrari-Trecate, M. Muselli, D. Liberati, and M. Morari. A clustering technique for the identification of piecewise affine and hybrid systems. *Automatica*, 39:205–207, 2003.

[11] 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.

[12] Lena Gorelick, Moshe Blank, Eli Shechtman, Michal Irani, and Ronen Basri. Actions as space-time shapes. *Transactions on Pattern Analysis and Machine Intelligence*, 29(12):2247–2253, December 2007.

[13] M. Grant and S. Boyd. CVX: Matlab software for disciplined convex programming, 2009. `http://stanford.edu/~boyd/cvx`.

[14] Jean B Lasserre. Global optimization with polynomials and the problem of moments. *SIAM Journal on Optimization*, 11(3):796–817, 2001.

[15] Jean B Lasserre. Convergent sdp-relaxations in polynomial optimization with sparsity. *SIAM Journal on Optimization*, 17(3):822–843, 2006.

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

[17] H. Lim, O. I. Camps, M. Sznaier, and V. Morariu. Dynamic appearance modelling for human tracking. In *IEEE Conf. Comp. Vision and Pattern Recognition (CVPR)*, pages 751–757, 2006.

[18] R.S. Lin, C.B. Liu, M.H. Yang, N. Ahuja, and S. Levinson. Learning nonlinear manifolds from time series. In *ECCV*, volume LNCS 3952, pages 245–256. Springer-Verlag, 2006.

[19] Karthik Mohan and Maryam Fazel. Iterative reweighted algorithms for matrix rank minimization. *J. of Machine Learning Research*, 13:3441–3473, 2012.

[20] N. Ozay and M. Sznaier. Hybrid system identification with faulty measurements and its application to activity analysis. In *Proc. 2011 IEEE CDC*, pages 5011–5016, December 2011.

[21] N. Ozay, M. Sznaier, and C. Lagoa. Model (in)validation of switched ARX systems with unknown switches and its application to activity monitoring. In *Proc. 2010 IEEE Conf. on Dec. and Control (CDC)*, pages 7624–7630, Dec. 2010.

[22] N. Ozay, M. Sznaier, C. Lagoa, and O. Camps. A sparsification approach to set membership identification of a class of affine hybrid system. In *Proc. 47^{th} IEEE Conf. Dec. Control (CDC)*, pages 123–130, Dec 2008.

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

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

[25] 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.

[26] P. A. Parrilo. Semidefinite programming relaxations for semialgebraic problems. *Mathematical Programming Ser. B*, 96(2):293–320, 2003.

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

[28] B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Rev.*, 52:471–501, August 2010.

[29] S. Roweis, L. Saul, and G.E. Hinton. Global coordination of local linear models. *Advances in Neural Information Processing Systems*, 14:889–896, 2001.

[30] R. Sánchez Peña and M. Sznaier. *Robust Systems Theory and Applications*. Wiley & Sons, Inc., 1998.

[31] Lawrence K. Saul and Sam T. Roweis. Think globally, fit locally: unsupervised learning of low dimensional manifolds. *Journal on Machine Learning Research*, 4:119–155, 2003.

[32] E. Sontag. Nonlinear regulation: The piecewise linear approach. *IEEE Trans. Aut. Contr.*, pages 346–358, 1981.

[33] M. Sznaier. Computational complexity of set membership Hammerstein and Wiener systems identification. *Automatica*, 45(3):701–705, 2009.

[34] R. Vidal, Y. Ma, and S. Sastry. Generalized principal component analysis (GPCA). *PAMI*, 27(12):1945–1959, December 2005.

[35] K.Q. Weinberger and L.K. Saul. Unsupervised learning of image manifolds by semidefinite programming. In *CVPR*, volume 2, pages 988 – 995, June 2004.

[36] F. Xiong, O. Camps, and M. Sznaier. Low order dynamics embedding for high dimensional time series,. In *2012 IEEE ICCV*, pages 2368–2374, 2012.

[37] F. Xiong, Y. Cheng, O. Camps, M. Sznaier, and C. Lagoa. Hankel based maximum margin classifiers: A connection between machine learning and wiener systems identification. In *Proc. 2013 IEEE CDC*, December 2013.

[38] 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.

[39] B. Yilmaz, M. Ayazoglu, M. Sznaier, and C. Lagoa. Convex relaxations for robust identification of wiener systems and applications. In *Proc. 2011 IEEE Conf. Dec. Control*, pages 2812–2818, 2011.

[40] B. Yilmaz, C. Lagoa, and M. Sznaier. An efficient atomic norm minimization approach to identification of low order models. In *2013 IEEE CDC*, pages 5834 – 5839, December 2013.