Identification of Switched Wiener Systems Based on Local Embedding

X. Zhang Y. Cheng Y. Wang M. Sznaier O. Camps

Abstract— **This paper considers the problem of switched Wiener system identification from a Kernel based manifold embedding perspective. Our goal is to identify both the Kernel mapping and the dynamics governing the evolution of the data on the manifold from noisy output measurements and with minimal assumptions about the nonlinearity and the affine portion of the systems. While in principle this is a very challenging problem, the main result of the paper shows that a computationally efficient solution can be obtained using a polynomial optimization approach that allows for exploiting the underlying sparse structure of the problem and provides optimality certificates. As an alternative, we provide a low complexity algorithm for the case where the affine part of the system switches only between 2 sub models.**

I. INTRODUCTION

During the past few years a large research effort has been devoted to the problem of identifying switched affine systems from experimental data (see for instance [18], [8], [1], [2], [6], [7], [13], [15], [16], [17], [21] and references therein for a summary of the difficulties involved in this problem and different approaches to overcome them).

Switched affine systems arise in a wide spectrum of applications, ranging from manufacturing processes, biology and communication systems to computer vision. In addition, since piece-wise affine models are known to be universal approximators [3], they provide a tractable "poor man's" non-linear identification framework. Note however, in many cases obtaining low error piece-wise affine approximations of non-linear systems requires a large number of sub-models. Thus, while conceptually appealing, since the computational complexity of most algorithms scales exponentially with the number of models, from a practical stand-point, this idea is limited to relatively small-sized problems. On the other hand, identification of full blown, switched non-linear systems is an intractable problem. As a compromise between these two extremes one can consider certain classes of non-linear systems, with the expectation that they will lead to tractable problems while still retaining features not easily captublack by piece-wise affine models. In this paper we will consider a specific class of non-linear switched systems, Wiener systems, composed of the cascade of a piece-wise affine switched system and a memoryless non-linearity. These systems are interesting in their own, not only in control theory, but also in related fields such as machine-learning. For example, in activity recognition applications, the information about the activity being performed is usually encapsulated in the piece-wise affine dynamics (each sub-model corresponding to a sub-activity), while the nonlinearity accounts for nuisance factors, such as view-points. A similar situation arises in computer vision when tracking targets across multiple cameras with different viewpoints. In addition, when approximating non-linear processes, using switched Wiener models (rather than linear ones), allows for leveraging the properties of the underlying nonlinear systems while leading to problems that still retain some of the computationally attractive properties of the piecewise affine case.

Identification of Wiener systems has been a very active research topic, but very few of these results apply to switched systems. Indeed, to the best of our knowledge, this case has been explicitly considered in [12], [23], [28], with [23] and [28] considering only switching non-linearities, while [12] proposed a general kernel based method for identifying switched non-linear systems. While the latter has been shown to be efficient in many scenarios, it requires knowledge of a set of suitable kernel functions and does not exploit the specific Wiener structure. As an alternative, in this paper we propose a polynomial optimization based method for identifying switched Wiener systems from experimental data and some very general information about the structure of the non-linearity. Motivated by the work in [26], [27] for identification of time-invariant Wiener systems, we will search for a manifold embedding of the observed data that preserves the local geometry and such that the embedded data evolves according to piece-wise affine dynamics. Note that allowing for switching (rather than time invariant) dynamics on the manifold leads to considerably more complex, generically NP hard problems. However, as shown in the paper, the problem can be recast as a (generically non-convex) polynomial optimization problem that can be relaxed to a sequence of computationally tractable semi-definite programs. Finally, when compared against [12], while both methods are kernel based, the main differences are the facts that, rather than postulating a specific kernel, our method identifies it from the experimental data, and that it explicitly exploits the piece-wise linear nature of the manifold dynamics.

II. PRELIMINARIES

A. Notation

$\mathbb{R}$	set of real numbers
$\mathbb{N}$	set of non-negative integer numbers
$\mathbf{I}$	identity matrix

This work was supported in part by NSF grant ECCS-1404163 and AFOSR grant FA9550-15-1-0392. The authors are with the ECE Department, Northeastern University, Boston, MA 02115.

$\mathbf{x}_{t:t+q}$	short notation for $[\mathbf{x}_t, \cdots, \mathbf{x}_{t+q}]$, where $\mathbf{x}_t$ is the sequence $\mathbf{x}$ at time t
$\tilde{\mathbf{r}}^{(t)}$	estimated model parameter at time t
$\hat{\mathbf{r}}_i$	estimated model parameter of system i
$\mathbf{K}$	kernel matrix of $\{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$, $\mathbf{K} \in \mathbb{R}^{n \times n}$, $\mathbf{K}_{ij} = \mathbf{x}_i^\top \mathbf{x}_j$
$\text{vec}(\text{tril}(\mathbf{K}))$	vectorized lower triangular portion of $\mathbf{K}$
$v_n(\mathbf{x})$	Veronese map of $\mathbf{x}$ at degree-n, a vector containing all degree-n monomials of $\mathbf{x}$.

B. Background on polynomial optimization

Next, we briefly summarize some key results in polynomial optimization that will be used to obtain tractable relaxations of the switched Wiener identification problem. The interested reader is refered to [11] for details. Consider a polynomial optimization of the form:

$$p_C^* := \underset{\mathbf{v} \in C}{\text{minimize}} p(\mathbf{v}),\ p(\mathbf{v}) \doteq \sum_\alpha p_\alpha \mathbf{v}^\alpha \tag{1}$$

where $\mathbf{v}^\alpha = v_1^{\alpha_1} v_2^{\alpha_2} \cdots v_n^{\alpha_n}$ and the semi-algebraic set $C \subset \mathbb{R}^n$ is defined by the polynomial inequalities $g_k(\mathbf{v}) \doteq \sum_\beta g_{k,\beta} \mathbf{v}^\beta \geq 0,\ k = 1, \cdots, d$. While this problem is nonconvex, [10] showed that is equivalent to the following (infinite-dimensional) convex one:

$$\tilde{p}_C^* = \underset{\mu \in \mathscr{P}(C)}{\text{minimize}} \int p(\mathbf{v}) \mu(dv) = \underset{\mu}{\text{minimize}} \sum_\alpha p_\alpha m_\alpha \tag{2}$$

where $\mathscr{P}(C)$ is the space of probability measures on C and

$$m_\alpha \doteq \int_C \mathbf{v}^\alpha \mu(dv) \tag{3}$$

is the moment of $\mathbf{v}^\alpha$ with respect to the distribution μ. Note that the objective function in (2) is affine in $\mathbf{m} \doteq \{m_\alpha, \forall \alpha\}$ and, as shown in [10], the condition for the existence of a measure μ supported in C such that (3) holds can be written as (infinite dimensional) semi-definite constraints $\mathbf{M}(\mathbf{m}) \succeq \mathbf{0}$ and $\mathbf{L}(g_k \mathbf{m}) \succeq \mathbf{0}$, $k = 1, \ldots, d$, where $\mathbf{M}$ and $\mathbf{L}$ (the moment and localizing matrices, respectively) are affine in $\mathbf{m}$. A finite dimensional sequence of approximations to problem (2) can be obtained by considering truncated versions of these matrices are given by [10]:

$$\begin{aligned} \mathbf{M}_N(\mathbf{m})(i,j) =& m_{\alpha^{(i)} + \alpha^{(j)}}, \forall i, j = 1, \cdots, S_N \\ \mathbf{L}_{N - \lceil \frac{\text{degree}(g_k)}{2} \rceil}(g_k \mathbf{m})(i,j) =& \sum_\beta g_{k,\beta} m_{\beta + \alpha^{(i)} + \alpha^{(j)}}, \\ & \forall i, j = 1, \cdots, S_{N - \lceil \frac{\text{degree}(g_k)}{2} \rceil} \end{aligned} \tag{4}$$

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 $\mathbf{0} = \alpha^{(1)} < \ldots < \alpha^{(S_N)}$.

It follows that problem (1) can be reduced to a sequence of Linear Matrix Inequalities optimization problems of the form

$$\begin{array}{rll} p_N^* = & \underset{\mathbf{m}}{\text{minimize}} & \sum_\alpha p_\alpha m_\alpha \\ & \text{s.t.} & \mathbf{M}_N(\mathbf{m}) \succeq \mathbf{0}, \\ & & \mathbf{L}_{N - \lceil \frac{\text{degree}(g_k)}{2} \rceil}(g_k \mathbf{m}) \succeq \mathbf{0}, \forall_{k=1}^d \end{array} \tag{5}$$

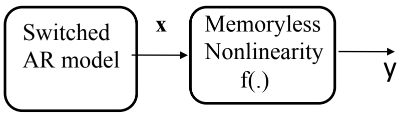

Fig. 1. Switched Wiener system model

Further, as N increases, $p_N^* \uparrow p_C^*$ from below monotonically. Finally, if for some finite N, $\mathbf{M}_N \succeq \mathbf{0}$, $\mathbf{M}_{N+1} \succeq \mathbf{0}$, and $\text{rank}(\mathbf{M}_N) = \text{rank}(\mathbf{M}_{N+1})$, then $p_M^* = p_C^*$.

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 (1).

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

$$I_{k'+1} \cap \bigcup_{j=1}^{k'} I_j \subseteq I_s \text{ for some } s \leq k' \tag{6}$$

As shown in [9], when this property holds, it is possible to construct a convergent hierarchy of semidefinite programs of smaller size:

$$\begin{array}{rll} p_N^* = & \underset{\mathbf{m}}{\text{minimize}} & \sum_{j=1}^l \sum_{\alpha(j)} p_{j,\alpha(j)} m_{\alpha(j)} \\ & \text{s.t.} & \mathbf{M}_N(\mathbf{m}^{(k)}) \succeq \mathbf{0}, \forall_{k=1}^l, \\ & & \mathbf{L}_{N - \lceil \frac{\text{degree}(g_k)}{2} \rceil}(g_k \mathbf{m}^{(k)}) \succeq \mathbf{0}, \forall_{k=1}^d, \end{array} \tag{7}$$

where $p_{j,\alpha(j)}$ is the coefficient of the $\alpha(j)^{th}$ monomial in the polynomial p_j, $\mathbf{M}_N(\mathbf{m}^{(k)})$ denotes the moment matrix and $\mathbf{L}_{N - \lceil \frac{\text{degree}(g_k)}{2} \rceil}(g_k \mathbf{m}^{(k)})$ is the localizing matrix associated with the constraint $g_k(\mathbf{v}) \geq 0$. Thus, for a given N, this approach requires considering moments and localizing matrices containing $O(\kappa^{2N})$ variables, where κ is the maximum cardinality of $\mathbf{v}^{(k)}$, rather than $O(n^{2N})$. Since in the problems considered in this paper $\kappa \ll n$ this leads to substantial complexity reduction.

III. PROBLEM FORMULATION

Consider the model shown in Figure 1 consisting of a system described by a switched AutoRegressive (AR) model followed by an unknown memoryless nonlinearity. We will only assume that this nonlinearity is locally (but not globally) invertible and that bounds on its local gain and that of its inverse are available. In this context, our goal is to identify the parameters of each AR submodel from the observed data $\mathbf{y}$. Specifically:

Problem 1: Given:

1.- A sequence of measurements $\{\mathbf{y}_t\}_{i=1}^n$, possibly corrupted by additive noise with known bound ε.

2.- A neighborhood parameter k, which defines a neighborhood matrix $\mathbf{F} \in \mathbb{R}^{n\times n}$ with entries $\mathbf{F}_{ij} \in \{0,1\}$, where $\mathbf{F}_{ij} = 1$ when $\mathbf{y}_j$ is within the k nearest neighbors of $\mathbf{y}_i$.

3.- Upper and lower bounds on the local gain of the nonlinearity, that is constants c_1 and c_2 such that

$$\frac{1}{c_1}\|\mathbf{x}_i - \mathbf{x}_j\|_2 \geq \ \|\mathbf{y}_i - \mathbf{y}_j\|_2 \geq \frac{1}{c_2}\|\mathbf{x}_i - \mathbf{x}_j\|_2 \quad \text{for all } (i,j) \text{ such that } \mathbf{F}_{ij} = 1 \text{ or } [\mathbf{F}^\top\mathbf{F}]_{ij} = 1. \tag{8}$$

4.- Bounds p on the number of affine subsystems and q on their order.

Find the internal signal sequence $\{\mathbf{x}_t\}$, and coefficient vector $\hat{\mathbf{r}}_i \in \mathbb{R}^{q+1}, i \in \{1,\cdots,p\}$ such that for each data segment $\mathbf{x}_{(t:t+q)}$ there exists at least one vector $\tilde{\mathbf{r}}^{(t)} \in \{\hat{\mathbf{r}}_1,\cdots,\hat{\mathbf{r}}_p\}$ such that $\mathbf{x}_{(t:t+q)}\tilde{\mathbf{r}}^{(t)} = 0$ holds.

IV. Main Results

In this section we present the main results of the paper: a computationally tractable algorithm for solving Problem 1, with optimality certificates. This algorithm will be obtained by firstly recasting Problem 1 into a polynomial optimization form and then using the results from section II-B to obtain a sequence of convex relaxations. Finally, the desired computationally efficient algorithm will be obtained by exploiting the underlying sparse structure of the problem and noting that the resulting relaxations are exact provided that certain submatrix, involving only a relatively small subset of the total number of variables, is rank 1. For simplicity, we will consider first the noiseless case and defer the treatment of noisy measurements until Section IV-D.

A. A polynomial optimization reformulation

Note that, by introducing binary indicator variables $s_i \in \{0,1\}$, the condition that there exists at least one $\tilde{\mathbf{r}}_t \in \{\hat{\mathbf{r}}_1,\cdots,\hat{\mathbf{r}}_p\}$ such that $\mathbf{x}_{t:t+q}\tilde{\mathbf{r}}_t = 0$ holds can be expressed as feasibility of the constraint set

$$\begin{array}{l} \mathbf{x}_{(t:t+q)}\tilde{\mathbf{r}}^{(t)} = 0,\ \tilde{\mathbf{r}}_t = \sum_{i=1}^p s_i^{(t)}\hat{\mathbf{r}}_i,\ \sum_{i=1}^p s_i^{(t)} = 1 \\ s_{i,t}^2 = s_{i,t}, \forall_{i=1}^p \end{array} \tag{9}$$

Clearly, with this observation Problem 1 is equivalent to establishing non-emptiness of the semi-algebraic set $C(\{\mathbf{x}_t\}_{t=1}^n, \{\hat{\mathbf{r}}_i\}_{i=1}^p, \{\tilde{\mathbf{r}}_t\}_{t=1}^{n-q}, \{s_{i,t}\}_{i=1,t=1}^{p,n-q})$ defined by:

$$\begin{cases} \hat{\mathbf{r}}_i^\top\hat{\mathbf{r}}_i = 1, \forall_{i=1}^p & \text{(10a)} \\ \hat{\mathbf{r}}_1(1) \geq \hat{\mathbf{r}}_2(1) \geq \cdots \geq \hat{\mathbf{r}}_p(1) \geq 0 & \text{(10b)} \\ c_2^2\|\mathbf{y}_i - \mathbf{y}_j\|_2^2 \geq \|\mathbf{x}_i - \mathbf{x}_j\|_2^2 \geq c_1^2\|\mathbf{y}_i - \mathbf{y}_j\|_2^2 \text{ for all} \\ \qquad (i,j) \text{ such that } \mathbf{F}_{ij} = 1 \text{ or } [\mathbf{F}^\top\mathbf{F}]_{ij} = 1 & \text{(10c)} \\ \mathbf{x}_{t:t+q}\tilde{\mathbf{r}}_t = 0,\ \tilde{\mathbf{r}}_t = \Sigma_{i=1}^p s_{i,t}\hat{\mathbf{r}}_i,\ \Sigma_{i=1}^p s_{i,t} = 1,\ \forall_{t=1}^{n-q} & \text{(10d)} \\ s_{i,t} = s_{i,t}^2, \forall_{i=1}^p\forall_{t=1}^{n-q} & \text{(10e)} \end{cases}$$

where (10a) and (10b) were added to avoid the trivial solution and to eliminate the ambiguity caused by the symmetry of the solution ($\hat{\mathbf{r}}_i(j)$ denotes the j-th entry of $\hat{\mathbf{r}}_i$).

A potential difficulty with the formulation above is that it explicitly uses the (unknown) manifold data $\mathbf{x}$. While this approach will work well for relatively small data sets, it quickly becomes intractable, due to the potentially large number of variables involved. To avoid this difficulty, rather than using (10) we will use the kernel based reformulation presented below. Since this formulation uses the inner products $\mathbf{x}_i^T\mathbf{x}_j$, it can comfortably handle situations where the dimension of the embedded data is not small. To this effect, consider a matrix $\mathbf{K}$ with entries $\mathbf{K}_{ij} = \mathbf{x}_i^T\mathbf{x}_j$ and note that:

$$\|\mathbf{x}_i - \mathbf{x}_j\|_2^2 = \mathbf{x}_i^\top\mathbf{x}_i + \mathbf{x}_j^\top\mathbf{x}_j - 2\mathbf{x}_i^\top\mathbf{x}_j = \mathbf{K}_{ii} + \mathbf{K}_{jj} - 2\mathbf{K}_{ij} \tag{11}$$

and that

$$\begin{array}{ll} \mathbf{x}_{t:t+q}\tilde{\mathbf{r}}_t = 0 & \Longrightarrow \mathbf{x}_l^\top\mathbf{x}_{t:t+q}\tilde{\mathbf{r}}_t = 0, \forall_{t'=1}^n \\ & \Longrightarrow \mathbf{K}_{l,t:t+q}\tilde{\mathbf{r}}_t = 0, \forall_{t'=1}^n \end{array} \tag{12}$$

for $t = 1,\cdots,n-q$. These observations lead to the following problem in the elements of $\mathbf{K}$ rather than $\mathbf{x}$:

$$\begin{cases} \hat{\mathbf{r}}_i^\top\hat{\mathbf{r}}_i = 1, \forall_{i=1}^p & \text{(13a)} \\ \hat{\mathbf{r}}_1(1) \geq \hat{\mathbf{r}}_2(1) \geq \cdots \geq \hat{\mathbf{r}}_p(1) \geq 0 & \text{(13b)} \\ \quad \mathbf{K}_{ii} + \mathbf{K}_{jj} - 2\mathbf{K}_{ij} \geq c_1^2\|\mathbf{y}_i - \mathbf{y}_j\|_2^2, \text{ and} \\ \mathbf{K}_{ii} + \mathbf{K}_{jj} - 2\mathbf{K}_{ij} \leq c_2^2\|\mathbf{y}_i - \mathbf{y}_j\|_2^2 \text{ for all} & \text{(13c)} \\ \quad (i,j) \text{ such that } \mathbf{F}_{ij} = 1 \text{ or } [\mathbf{F}^\top\mathbf{F}]_{ij} = 1 \\ \mathbf{K} \succeq \mathbf{0},\ \mathbf{K}_{ij} = \mathbf{K}_{ji}, \forall_{i=1}^n\forall_{j=1}^n & \text{(13d)} \\ \mathbf{K}_{t',t:t+q}\tilde{\mathbf{r}}_t = 0, \forall_{t'=1}^n\forall_{t=1}^{n-q} & \text{(13e)} \\ \tilde{\mathbf{r}}_t = \Sigma_{i=1}^p s_{i,t}\hat{\mathbf{r}}_i,\ \Sigma_{i=1}^p s_{i,t} = 1,\ \forall_{t=1}^{n-q} & \text{(13f)} \\ s_{i,t} = s_{i,t}^2, \forall_{i=1}^p, \forall_{t=1}^{n-q} & \text{(13g)} \end{cases}$$

where the constraint (13d) has been added to guarantee that $\mathbf{K}$ is indeed a kernel matrix.

In principle, the problem above could be solved using the moments-based techniques mentioned in Section II-B. However, this approach quickly gets intractable due to the complexity entailed in enforcing the positive semidefiniteness of $\mathbf{K}$, which requires the determinants of all its leading principal minors to be nonnegative, leading to polynomials constraints of high degree, which in turn require considering high order relaxations, since the order of the relaxation must be at least as large as 0.5×(the highest degree of monomials in the problem). To circumvent this difficulty, next we develop computationally tractable algorithms by exploiting the sparse structure of the problem.

B. A sparse reformulation

In this section we show that problem (13) above exhibits the running intersection property and thus, as indicated in Section II-B can be solved by considering a reduced set of constraints. To establish this fact, denote all the variables in (10) by $\mathbf{v} \doteq \{\text{vec}(\text{tril}(\mathbf{K})), \mathbf{R}, \tilde{\mathbf{R}}, \mathbf{S}\}$, with $\mathbf{R} = \{\hat{\mathbf{r}}_i, i = 1,\ldots,p\}$, $\tilde{\mathbf{R}} = \{\tilde{\mathbf{r}}_t,\ t = 1,\ldots,n-q\}$, and $\mathbf{S} = \{\mathbf{s}_t, t = 1,\ldots,n-q\}$, $\mathbf{s}_t = \{s_{i,t}, i = 1,\ldots,p\}$. Next, partition $\mathbf{v}$ into the $n-q+1$ sets as follows

$$\begin{array}{l} \mathbf{v}^{(0)} = \{\text{vec}(\text{tril}(\mathbf{K})), \mathbf{R}\}, \\ \mathbf{v}^{(t)} = \{\text{vec}(\text{tril}(\mathbf{K})), \mathbf{R}, \tilde{\mathbf{r}}_t, \mathbf{s}_t\}, \forall_{t=1}^{n-q}, \end{array} \tag{14}$$

and partition the set C_K in (13) into

$$C^{(0)} : \{(13\text{a}) - (13\text{d})\}$$

$$\forall_{t=1}^{n-q} : C^{(t)} : \begin{cases} \mathbf{K}_{l,t:t+q}\tilde{\mathbf{r}}_t = 0, \forall_{l=1}^{n} \\ \tilde{\mathbf{r}}_t = \sum_{i=1}^{p} s_{i,t}\hat{\mathbf{r}}_i \\ s_{i,t} = s_{i,t}^2, \forall_{i=1}^{p} \\ \sum_{i=1}^{p} s_{i,t}^2 = 1 \end{cases} \tag{15}$$

It is easy to check that for each $t = 0, 1, \ldots, n-q$, the constraints in set $C^{(t)}$ contain variables only in $\mathbf{v}^{(t)}$ and that

$$\mathbf{v}^{(j)} \cap \left(\cup_{k=1}^{j-1} \mathbf{v}^{(k)} \right) = \{\text{vec}(\text{tril}(\mathbf{K})), \mathbf{R}\} = \mathbf{v}^{(0)}, \tag{16}$$

holds for each $j = 1, \ldots, n-q$. Thus, the *running intersection property* holds. From the results in [9] and Section II-B, it follows that problem (10) can be solved via finding a feasible solution $\mathbf{m}^{(t)}$ to the following reduced-sized relaxation:

$$\begin{cases} \mathbf{M}_N(\mathbf{m}^{(t)}) \succeq \mathbf{0}, \forall_{t=0}^{n-q}, \\ \mathbf{L}_{N-1}(g_{k,t}\mathbf{m}^{(t)}) \succeq \mathbf{0}, \forall_{t=0}^{n-q}, \end{cases} \tag{17}$$

where $g_{k,t}$ denotes the constraints in the set $C^{(t)}$ (15) and $\mathbf{m}^{(t)}$ represents the moments sequence associated with variables in $\mathbf{v}^{(t)}$ up to order $2N$. Comparing to using the moments matrix associated with all the variable $\mathbf{v}$, (17) reduces the size of the positive semidefinite matrices dramatically, specially for problems involving long data records (large n).

C. A Computationally Tractable Relaxation

An additional reduction in computational complexity can be achieved by exploiting the fact that, for a rank-1 moment matrix, there always exists an associated probability measure composing of a single atom, precisely at the location given by the first order moments of the distribution. This fact allows for considering only the first order relaxation, subject to an additional rank constraint on the moment matrix.

Theorem 1: The nonconvex problem (13) is equivalent to finding a feasible solution to the following set

$$\begin{cases} \mathbf{M}_1(\mathbf{m}^{(t)}) \succeq \mathbf{0}, \forall_{t=0}^{n-q}, \\ \mathbf{L}_0(g_{k,t}\mathbf{m}^{(t)}) \succeq \mathbf{0}, \forall_{t=0}^{n-q}, \\ \text{Rank}\{\mathbf{M}_1(\mathbf{m}^{(t)})\} = 1, \forall_{t=0}^{n-q}. \end{cases} \tag{18}$$

Proof: Suppose that $\{\mathbf{m}^{(t)*}\}_{t=0}^{n-q}$ is a feasible solution to (18). (18)⇒(13) follows from the fact that the elements corresponding to the first order moments of the variable of $\mathbf{v}^{(t)}$ in $\mathbf{m}^{(t)*}$ are a feasible solution to (13). Suppose now that $\mathbf{v}^*$ is a feasible solution to (13). Partitioning $\mathbf{v}^*$ into $\{\mathbf{v}^{(t)*}\}_{t=0}^{n-q}$ as in (14), the sequence $\mathbf{m}^{(t)*}$ consisting of all the monomials of $\mathbf{v}^{(t)*}$ up to order 2 is a feasible solution to (18), therefore, (13)⇒ (18) holds. ■

The result above, while leading to substantial computational complexity reduction, still requires enforcing a rank-1 constraint on $n-q+1$ relatively large matrices. As we show next, surprisingly, rather than enforcing this condition, it suffices to enforce sparsity of the solution to the binary variables $s_{i,t}$, together with a rank-1 constraint on a single matrix involving a reduced set of variables.

Lemma 1: For a 3×3 symmetric matrix $\mathbf{M}$ denoted by

$$\mathbf{M} = \left[\begin{array}{c|c} \mathbf{M}_{11} & \mathbf{M}_{12} \\ \hline \mathbf{M}_{21} & \mathbf{M}_{22} \end{array} \right] = \left[\begin{array}{cc|c} 1 & m_1 & m_2 \\ m_1 & m_{11} & m_{12} \\ \hline m_2 & m_{12} & m_{22} \end{array} \right], \tag{19}$$

if $\mathbf{M} \succeq 0$ and $\text{rank}\{\mathbf{M}_{11}\} = 1$, then $m_{12} = m_1 m_2$ holds.

Proof: Since $\text{rank}\{\mathbf{M}_{11}\} = 1$, then $m_{11} = m_1^2$. Since $\mathbf{M} \succeq \mathbf{0}$, $\det(\mathbf{M}) \geq 0$. On the other hand, $\det(\mathbf{M}) = -(m_{12} - m_1 m_2)^2 \leq 0$. Thus, $\det(\mathbf{M}) = 0$, and $m_{12} = m_1 m_2$. ■

Theorem 2: Problem (18) is equivalent to finding a feasible solution to

$$\begin{cases} \mathbf{M}_1(\mathbf{m}^{(t)}) \succeq \mathbf{0}, \forall_{t=0}^{n-q}, \\ \mathbf{L}_0(g_{k,t}\mathbf{m}^{(t)}) \succeq \mathbf{0}, \forall_{t=0}^{n-q}, \\ \text{Rank}\{\mathbf{M}_1(\mathbf{m}^{(0)})\} = 1, \forall_{t=0}^{n-q}, \\ m(s_i^{(t)}) \in \{0,1\}, \forall_{i=1}^{p} \forall_{t=1}^{n-q}. \end{cases} \tag{20}$$

Proof: Omitted due to space constraints ■

The nonconvex sparsity and rank-1 constraint in (20) can be handled by resorting to an iterative algorithm [4], [20], leading to Algorithm 1.

Algorithm 1 Moments-Based Convex Certificates for (10)

```
Initialize: $j = 0$, $0 < \delta \ll 1, \mathbf{W}^{(0)} = \mathbf{I}$, $w_{i,t}^{(0)} = 1, \forall_{i=1}^{p} \forall_{t=1}^{n-q}$;
repeat
    Solve
        $\underset{\mathbf{m}^{(t)}}{\text{minimize}} \quad \text{trace}\{\mathbf{W}^{(j)}\mathbf{M}_1(\mathbf{m}^{(0)})\} + \lambda \sum_{t=1}^{n-q} \sum_{i=1}^{p} w_{i,t}^{(j)} m(s_{i,t})$
        $\text{s.t.} \quad \mathbf{M}_1(\mathbf{m}^{(t)}) \succeq \mathbf{0}, \forall_{t=0}^{n-q},$
        $\quad \mathbf{L}_0(g_{k,t}\mathbf{m}^{(t)}) \succeq \mathbf{0}, \forall_{t=0}^{n-q}.$    (21)
    Update
        $\mathbf{W}^{(j+1)} = [\mathbf{M}_1(\mathbf{m}^{(0)})^{(j)} + \sigma_2(\mathbf{M}_1(\mathbf{m}^{(0)})^{(j)})\mathbf{I}]^{-1}$
        $w_{i,t}^{(j+1)} = [m(s_{i,t})^{(j)} + \delta]^{-1}$
        $j = j + 1$
until $\sigma_2(\mathbf{M}_1(\mathbf{m}^{(0)})^{(j)}) \ll 1$ and $s_{i,t}^{(j)} \approx 1$ or 0.
```

In Algorithm 1, $\sigma_2(\bullet)$ denotes the second largest singular value of the matrix $\bullet$. If the optimization above converges to a rank-1 matrix $\mathbf{M}(\mathbf{m}^{(0)})$ and variables $s_{i,t} \in \{0,1\}$, we automatically get $\mathbf{K}$, $\hat{\mathbf{r}}_{1:p}$ and $\mathbf{s}_{1:n-q}$. The indicator variables $\mathbf{s}_{1:n-q}$ give the discrete labels of each data point, that is, which subsystem generated it. $\hat{\mathbf{r}}_i$, gives the model parameter of the ith subsystem, where $i = 1, \cdots, p$. Finally, if needed, the embedded data $\mathbf{x}$ can be recovered either by performing a spectral factorization on $\mathbf{K}$ as in SDE [25] or a Cholesky decomposition. In the former case, let $[\mathbf{U}, \mathbf{S}, \mathbf{V}] = svd(\mathbf{K})$ and let $\mathbf{R}$ be the thresholded $\mathbf{S}$ which keeps only the large singular values. Then, we get the embedded data: $\mathbf{x} = \mathbf{R}^{\frac{1}{2}}\mathbf{V}$. In the latter case, let $\mathbf{L}^T\mathbf{L} = \mathbf{K}$, then we get $\mathbf{x}$ by eliminating the all-zero rows of $\mathbf{L}$.

D. The Noisy Case

Next, we briefly indicate how to modify the proposed algorithm to handle noisy measurements. Assume that the measured data $\mathbf{y}$ is corrupted by additive ℓ_∞ bounded noise $\mathbf{e}$, with $\|\mathbf{e}\|_\infty \leq \varepsilon$; In this case we have:

$$\begin{aligned}
\|\mathbf{y}_i+\mathbf{e}_i-\mathbf{y}_j-\mathbf{e}_j\|_2^2 &= \|\mathbf{y}_i-\mathbf{y}_j\|_2^2+\|\mathbf{e}_i-\mathbf{e}_j\|_2^2 \\
&\quad +2(\mathbf{y}_i-\mathbf{y}_j)^T(\mathbf{e}_i-\mathbf{e}_j) \\
&\leq \|\mathbf{y}_i-\mathbf{y}_j\|_2^2+4d\varepsilon^2+4\varepsilon\|\mathbf{y}_i-\mathbf{y}_j\|_1 \\
\|\mathbf{y}_i+\mathbf{e}_i-\mathbf{y}_j-\mathbf{e}_j\|_2^2 &\geq \|\mathbf{y}_i-\mathbf{y}_j\|_2^2-4\varepsilon\|\mathbf{y}_i-\mathbf{y}_j\|_1
\end{aligned} \tag{22}$$

where d is the dimension of the output data. Proceeding as in Section IV-A leads to solving the feasibility problem similar to (13) except for replacing the constraint (13c) by

$$\begin{aligned}
&\mathbf{K}_{ii}+\mathbf{K}_{jj}-2\mathbf{K}_{ij} \geq c_1^2\|(\|\mathbf{y}_i-\mathbf{y}_j\|_2^2-4\varepsilon\|\mathbf{y}_i-\mathbf{y}_j\|_1)) \text{ and} \\
&\mathbf{K}_{ii}+\mathbf{K}_{jj}-2\mathbf{K}_{ij} \leq c_2^2(\|\mathbf{y}_i-\mathbf{y}_j\|_2^2+4d\varepsilon^2+4\varepsilon\|\mathbf{y}_i-\mathbf{y}_j\|_1) \\
&\text{for all } (i,j) \text{ such that } \mathbf{F}_{ij}=1 \text{ or } [\mathbf{F}^\top\mathbf{F}]_{ij}=1
\end{aligned} \tag{23}$$

Clearly, the problem above can be solved using the same algorithm outlined in Section IV-C.

V. A Low Complexity Method for the 2-Subsystems Case

In this section we derive a low complexity algorithm for the case where the model only switches between two candidate systems.

A. The Noiseless Case

In this section we derive a low complexity algorithm for the case where the model only switches between two candidate systems. Note that when $p=2$, searching for a feasible solution to (13) is equivalent to looking for a feasible solution to the constraint set $C_{K,2}(\mathbf{K},\hat{\mathbf{r}}_1,\hat{\mathbf{r}}_2,\mathbf{x})$ given by

$$\begin{cases}
\hat{\mathbf{r}}_1 \neq \mathbf{0}, \hat{\mathbf{r}}_2 \neq \mathbf{0} & \text{(24a)} \\
\text{Constraint (13c)} & \text{(24b)} \\
\text{Constraint (13d)} & \text{(24c)} \\
\mathbf{K}_{ij} = \mathbf{x}_i^T\mathbf{x}_j, \forall_{i=1}^n \forall_{j=1}^n & \text{(24d)} \\
(\mathbf{x}_{t:t+q}\hat{\mathbf{r}}_1)^T(\mathbf{x}_{t:t+q}\hat{\mathbf{r}}_2) = 0, \forall_{t=1}^{n-q} & \text{(24e)}
\end{cases}$$

Rewriting the last condition in terms of the Veronese map yields:

$$(\mathbf{x}_{t:t+q}\hat{\mathbf{r}}_1)^T(\mathbf{x}_{t:t+q}\hat{\mathbf{r}}_2) = v_2(\mathbf{x}_{t:t+q})\hat{\mathbf{r}} = [\cdots,\mathbf{K}_{ij},\cdots]\hat{\mathbf{r}} = 0 \tag{25}$$

where $i,j=t,\cdots,t+q$ and $\hat{\mathbf{r}} \in \mathbb{R}^{\frac{(q+2)(q+1)}{2}}$ is the vector consisting of the outer product of $\hat{\mathbf{r}}_1$ and $\hat{\mathbf{r}}_2$.

Define the matrix

$$\mathbf{V}_2 = \begin{bmatrix} v_2(\mathbf{x}_{1:1+q})^T & v_2(\mathbf{x}_{2:2+q})^T & \cdots & v_2(\mathbf{x}_{n-q:n})^T \end{bmatrix}^T \tag{26}$$

From (25) we note that the entries of $\mathbf{V}_2$ are a subset of the elements of $\mathbf{K}$ and that $\mathbf{V}_2\hat{\mathbf{r}}=0$, which implies that $\mathbf{V}_2$ is rank deficient. Thus, in principle, in this case Problem 1 can be solved by seeking rank deficient $\mathbf{V}_2$ to (24). However, this approach can lead to high rank kernel matrices, which in turn implies high dimensionality of the embedded data $\mathbf{x}$. To avoid this situation, we will add a regularization term that penalizes the rank of $\mathbf{K}$, leading to the following optimization problem:

$$\begin{aligned}
\underset{\mathbf{K}}{\text{minimize}} \quad & \text{rank}(\mathbf{V}_2)+\lambda\text{rank}(\mathbf{K}) \\
\text{s.t.} \quad & \text{Constraints (13c) and (13d).}
\end{aligned} \tag{27}$$

B. The Noisy Case

As before, the noisy case can be handled by simply modifying the constraint (13c) to include the noise effects, leading to the optimization problem:

$$\begin{aligned}
\underset{\mathbf{K}}{\text{minimize}} \quad & \text{rank}(\mathbf{V}_2)+\lambda\text{rank}(\mathbf{K}) \\
\text{s.t.} \quad & \text{Constraints (23) and (13d).}
\end{aligned} \tag{28}$$

We can use reweighted rank minimization algorithm in [5] and [20] to solve the above problems (27) and (28) iteratively. As mentioned in Section IV-C, we can either perform spectral factorization on $\mathbf{K}$ or Cholesky decomposition, to get the embedded data $\mathbf{x}$. Since we do not have indicator variables in the model, we have to get the subsystem labels from either $\mathbf{K}$ or $\mathbf{x}$. This can be accomplished by applying the same post-processing method proposed in [24] in the context of GPCA: Form the embedded data matrix $\mathbf{V}_2$ from $\mathbf{K}$, take derivatives with respect to $\mathbf{x}$ and perform a normalized cuts on the derivative vectors.

Remark 1: The approach outlined above can be applied to the case of more than two candidate systems by considering a sliding window. If in each sliding window there are no more than two systems, then the problem can be solved locally for the two subsystems active in each window.

VI. Illustrative Example

In this example, we generate 20 data points from a switched Wiener system consisting of two order-2 subsystems cascaded with a sigmoid function. The regressors of the two systems are $\mathbf{r}_1=[-1,-1.1329,-1]$ and $\mathbf{r}_2=[-1,-0.1250,-1]$. The first two data points are initialized randomly. Then we use $\mathbf{r}_1$ to generate data points 3 to 4, then $\mathbf{r}_2$ to generate 5 to 11, then $\mathbf{r}_1$, 12 to 17, then $\mathbf{r}_2$, 18 to 20. The data are normalized to range between -1 and 1. Then, each element of the data points is passed through a sigmoid function $f(t)=\frac{1}{1+e^{-t}}$. If we define the sigmoid function only in the domain of $[-1,1]$, it is Lipschitz continuous with Lipschitz constant 0.25. Similarly, the inverse sigmoid function is also Lipschitz continuous with Lipschitz constant 5.09. Therefore, $c_1=4$ and $c_2=5.09$. Then, we add uniform noise with bound 0.01 to the output signal. This data serve as the observations given to identify the switched system.

The identification result using the moment method is shown in Figure 2. The subsystem identity is obtained from the indicator variables. The recovered models are $\hat{\mathbf{r}}_1=[1,0.1168,1]$ and $\hat{\mathbf{r}}_2=[1,-1.1316,1]$. If we flip the sign, they are very close to the ground truth regressors. The error of the parameters are $\|(-\hat{\mathbf{r}}_1)-\mathbf{r}_1\|_2=0.0012$ and $\|(-\hat{\mathbf{r}}_2)-\mathbf{r}_2\|_2=0.0081$. The identification result using the low complexity method described in Section V is shown in Figure 3. The recovered models are $\hat{\mathbf{r}}_1=[-1,-0.0564,-0.9922]$ and $\hat{\mathbf{r}}_2=[-1,1.0997,-1.0098]$. The error of the parameters are $\|\hat{\mathbf{r}}_1-\mathbf{r}_1\|_2=0.0345$ and $\|\hat{\mathbf{r}}_2-\mathbf{r}_2\|_2=0.0690$.

VII. CONCLUSIONS

This paper considers the problem of switched Wiener system identification from a Kernel based manifold embedding perspective. The goal here is to jointly identify the

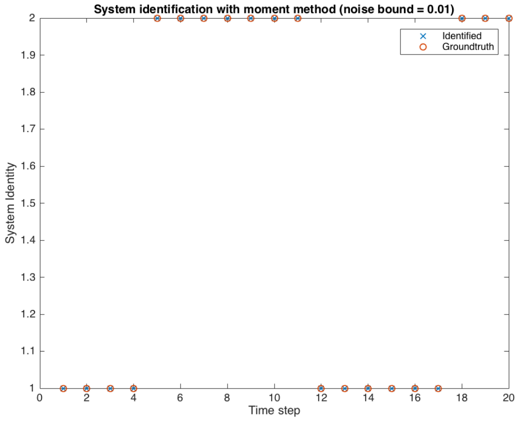

Fig. 2. Identification using the moment based method with noisy data

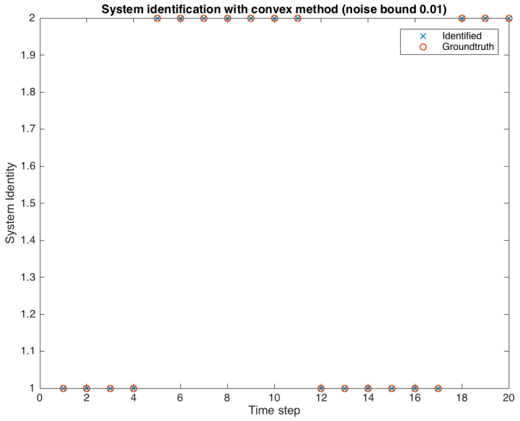

Fig. 3. Identification using the method from section V with noisy data

Kernel mapping and the dynamics governing the evolution of the data on the manifold. While in principle this is a very challenging non-convex optimization problem, the main result of this paper shows that a computationally efficient solution can be obtained by recasting the problem into a polynomial optimization form that can be solved using moment-based techniques. Notably, this approach allows for exploiting the underlying sparse structure of the problem and guarantees that an optimal solution has been found if a matrix involving only a subset of the variables has rank one. As an alternative, we provide a low complexity algorithm for the case where the affine part of the system switches only between 2 sub models.

References

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

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

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

[4] 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.

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

[6] C. Feng, C. M. Lagoa, N. Ozay, and M. Sznaier. Hybrid system identification: An sdp approach. In *Decision and Control (CDC), 2010 49th IEEE Conference on*, pages 1546–1552. IEEE, 2010.

[7] G. Ferrari-Trecate, M. Muselli, D. Liberati, and M. Morari. A clustering technique for the identification of piecewise affine systems. *Automatica*, 39(2):205–217, 2003.

[8] A. Garulli, S. Paoletti, and A. Vicino. A survey on switched and piecewise affine system identification. In *System Identification*, volume 16, pages 344–355, 2012.

[9] J. Lasserre. Convergent SDP-relaxations in polynomial optimization with sparsity. *SIAM J. on Optimization*, 17(3):822–843, Oct. 2006.

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

[11] J. B. Lasserre. *Moments, positive polynomials and their applications*, volume 1. Imperial College Press, 2010.

[12] F. Lauer and G. Bloch. Switched and piecewise nonlinear hybrid system identification. In *Hybrid Systems: Computation and Control*, pages 330–343. Springer, 2008.

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

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

[15] N. Ozay, C. Lagoa, and M. 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.

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

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

[18] B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

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

[20] M. A. M. Sznaier, C. Lagoa, and O. Camps. A moments-based approach to estimation and data interpolation for a class of Wiener systems. In *Proc. 2010 IEEE Conf. on Dec. and Control (CDC)*, Dec. 2010.

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

[22] K. Q. Weinberger and L. K. Saul. Unsupervised learning of image manifolds by semidefinite programming. *International Journal of Computer Vision*, 70(1):77–90, 2006.

[23] F. Xiong, O. I. Camps, and M. Sznaier. Low order dynamics embedding for high dimensional time series. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2368–2374. IEEE, 2011.

[24] 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.

[25] 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.