Dynamics-based Information Extraction: A Hybrid Systems Approach

Necmiye Ozay
EECS, University of Michigan

Joint work with Mario Sznaier, Octavia Camps (Northeastern), Constantino Lagoa (Penn State), Farshad Harirchi (Michigan)

IMA Annual Program Year Workshop Optimization and Parsimonious Modeling

Research partly funded by DARPA
Outline
Dynamics-based Information Extraction: A Hybrid Systems Approach

• Motivation
  – Control + Dynamics + Data + Information + Parsimony??

• Hybrid dynamical models (hybrid systems)
  – and their use in information extraction

• Tools from optimization

• Three concrete problems
  – Identification of hybrid models
  – Model (in)validation for hybrid models
  – Fault detection for hybrid models
Motivation

Control + Dynamics + Data + Information
+ Parsimony??
Motivation

Control + Dynamics + Data + Information + Parsimony??

Observation 1: Dynamics enable parsimonious modeling
Dynamics enable parsimony

A sparse set of features suffices for identifying and understanding dynamic events!
Dynamics enable parsimony
Dynamics enable parsimony

A sparse set of features suffices for identifying and understanding dynamic events!
Motivation

Control + Dynamics + Data + Information + Parsimony??

Observation 1: Dynamics enable parsimonious modeling

Observation 2: Control requires parsimonious modeling
Control requires parsimony

Complex models

Useful/actionable models for (i) control design, (ii) fast simulations (iii) system monitoring, (iv) anomaly detection, etc.

Big data
Overview of mixture models

Common in many fields:
• Gaussian mixtures
• Subspace arrangements
• Hybrid systems

Collection of simple models that can explain complex objects!

• Two fold difficulty in learning such models:
  • Data association
  • Parameter estimation
Hybrid systems

• “mixture models” for dynamical systems
• Switched systems
\[ y(t + 1) = G_{\sigma_t}(y(t : t - n_a), u(t : t - n_c)) \]
where mode signal \( \sigma_t \in \{1, \ldots, s\} \)
• For this talk, \( G_i \)'s are polynomial (or affine)

• Global approximators even when G is affine!
• Identification from data is not easy!
• Two fold difficulty:
  • Estimation of the mode signal (data association)
  • Estimation of the parameters (identification)
Information extraction as an Identification Problem

- Hybrid Dynamical Models
  - Simple models for complex phenomena
- Two-fold difficulty
  - Estimation of mode signal (data association)
  - Estimation of parameters (identification)

- Model data streams as outputs of switched linear systems
- “Interesting” events $\leftrightarrow$ Changes in model invariants
- “Homogenous” segments $\leftrightarrow$ Output of a single submodel
A Simple Problem: Event Detection

- Key observation: as new modes get excited, complexity (order) of the system increases

Look for changes in model complexity.
A Simple Problem: Event Detection

- Key observation: as new modes get excited, complexity (order) of the system increases
- Order of the system is given by the rank of the Hankel matrix

\[ H_y = \begin{bmatrix} y_1 & y_2 & \cdots & y_n \\ y_2 & y_3 & \cdots & y_{n+1} \\ \vdots & \vdots & \ddots & \vdots \\ y_m & y_{m+1} & \cdots & y_{m+n-1} \end{bmatrix} \]

Look for changes in the rank of the Hankel matrix. (no need to explicitly find the model!)
Fast Event Detection

Use SVD to estimate the rank of the Hankel Matrix. (five lines of Matlab code, runs on a laptop)
A Simple Problem: Event Detection

- A few issues: delays, fast switching, noise and outliers

\[ H_y = \begin{bmatrix} y_1 & y_2 & \cdots & y_n \\ y_2 & y_3 & \cdots & y_{n+1} \\ \vdots & \vdots & \ddots & \vdots \\ y_m & y_{m+1} & \cdots & y_{m+n-1} \end{bmatrix} \]

- How to more rigorously reason about noisy data?
- What if we want to learn individual dynamics?
The tool (my big hammer)

Moments-based convex relaxations to polynomial optimization (Lasserre’s hierarchy).
Moments-based convex relaxations to polynomial optimization (Lasserre’s hierarchy).

Multivariate polynomial optimization problem on a compact basic semi algebraic set $K$

$$p^*_K := \min_{x \in K} p(x) \quad (P1)$$

A functional optimization problem over the set of probability measures $\mu$ with support $K$

$$\tilde{p}^*_K := \min_{\mu \in \mathcal{P}(K)} \mathbb{E}_\mu[p(x)] \quad (P2)$$

**Theorem** (Lasserre 01): P1 is equivalent to P2

$$\tilde{p}^*_K = p^*_K$$
The tool (my big hammer)

Moments-based convex relaxations to polynomial optimization (Lasserre’s hierarchy).

Multivariate polynomial optimization problem on a compact basic semi algebraic set \( K \)

\[
p^*_K := \min_{x \in K} p(x) \tag{P1}
\]

A functional optimization problem over the set of probability measures \( \mu \) with support \( K \)

\[
\tilde{p}^*_K := \min_{\mu \in \mathcal{P}(K)} \mathbb{E}_\mu [p(x)] \tag{P2}
\]

**Theorem** (Lasserre 01): P1 is equivalent to P2

\[
\tilde{p}^*_K = p^*_K
\]
The tool (my big hammer 🛠️)

Moments-based convex relaxations to polynomial optimization (Lasserre’s hierarchy).

Multivariate polynomial optimization problem on a compact basic semi algebraic set $K$

$$p^*_K := \min_{x \in K} p(x) \quad (P1)$$

A functional optimization problem over the set of probability measures $\mu$ with support $K$

$$\tilde{p}^*_K := \min_{\mu \in \mathcal{P}(K)} \mathbb{E}_\mu [p(x)] \quad (P2)$$

**Theorem** (Lasserre 01): P1 is equivalent to P2

$$p^*_K = \tilde{p}^*_K$$
The tool (my big hammer)

Moments-based convex relaxations to polynomial optimization (Lasserre’s hierarchy).

Multivariate polynomial optimization problem on a compact basic semi algebraic set $K$

$$p^*_K := \min_{x \in K} p(x) \quad (P1)$$

A functional optimization problem over the set of probability measures $\mu$ with support $\tilde{K}$

$$\tilde{p}^*_K := \min_{\mu \in \mathcal{P}(K)} \mathbb{E}_{\mu}[p(x)] \quad (P2)$$

Theorem (Lasserre 01): P1 is equivalent to P2

$$\tilde{p}^*_K = p^*_K$$
Moment-based relaxations for polynomial optimization

A functional optimization problem over the set of probability measures $\mu$ with support $K$

$$\tilde{p}_K^* := \min_{\mu \in \mathcal{P}(K)} \mathbb{E}_\mu[p(x)] \quad \text{(P2)}$$

Equivalent to an SDP with countably infinite variables, where variables are moments of the distribution $\mu$. 
A functional optimization problem over the set of probability measures $\mu$ with support $K$

$$\tilde{p}_K^* := \min_{\mu \in \mathcal{P}(K)} \mathbb{E}_\mu[p(x)] \quad (P2)$$

Equivalent to an SDP with countably infinite variables, where variables are moments of the distribution $\mu$.

If $K = [a, b]$ univariate polynomial on an interval, there is a finite exact SDP (Hausdorff moment problem).
Moment-based relaxations for polynomial optimization

A functional optimization problem over the set of probability measures $\mu$ with support $K$

$$\tilde{p}_K^* := \min_{\mu \in \mathcal{P}(K)} \mathbb{E}_\mu[p(x)]$$  \hspace{1cm} (P2)

Equivalent to an SDP with countably infinite variables, where variables are moments of the distribution $\mu$.

If $K = [a, b]$ univariate polynomial on an interval, there is a finite exact SDP (Hausdorff moment problem).

If there is a sparse structure in the polynomial and constraints defining $K$, possible to get a hierarchy where we have smaller LMIs at each relaxation order (Lasserre, Nie, Waki, Kojima).
Three problems

• Identification of switched affine systems (SARX Id)
• Model (in)validation for switched affine systems (SARX invalidation)
• Fault/anomaly detection for systems with polynomial state-space models

Given that we will be using polynomial optimization affine or polynomial or switched or non-switched makes a little difference.
Switched System Identification

• Particular interest to switched linear models in control and system identification communities.

• Problem Formulation:
  • Given experimental input/output data, and bounds on noise and submodel orders \((n_a, n_c)\)
  • Find a switched linear autoregressive model with exogenous inputs (ARX) of the form:

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

\[
y(t) = p(\sigma_t)^T r(t) + \eta(t) \quad \|\eta(t)\| \leq \epsilon
\]

• Ill-posed, always have a trivial solution!
Hybrid System Identification

• Particular interest to switched linear models in control and system identification communities.

• Problem Formulation:
  • Given experimental input/output data, and bounds on noise and submodel orders \((n_a, n_c)\)
  • Find a switched linear autoregressive model with exogenous inputs (ARX) of the form:

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

\[
y(t) = p(\sigma_t)^T r(t) + \eta(t) \quad \|\eta(t)\| \leq \epsilon
\]

Possible objectives:
• Minimum # of switches
• Minimum # of submodels
• Fixed # of submodels
Hybrid System Identification

- Particular interest to switched linear models in control and system identification communities.

**Problem Formulation:**
- Given experimental input/output data, and bounds on noise and submodel orders \((n_a, n_c)\)
- Find a switched linear autoregressive model with exogenous inputs (ARX) of the form:

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

\[
y(t) = p(\sigma_t)^T r(t) + \eta(t) \quad ||\eta(t)|| \leq \epsilon
\]

Possible objectives:
- Minimum # of switches (non-convex – polytime exact algorithms exist)
- Minimum # of submodels
- Fixed # of submodels
SARX Id in noise-free case

- GPCA: an algebraic geometric method due to Vidal et al.
- Main Idea:

\[
b(\sigma_t)^T r_t = 0, \quad \sigma_t \in \{1, \ldots, s\}
\]

Neither the mode signal nor the parameters, b, are known!

Independent of mode signal, linear in parameters, c!
SARX Id in noise-free case

- GPCA: an algebraic geometric method due to Vidal et al.

- Main Idea:
  - Embed the data in a higher dim. space via Veronese map
    \[ \nu_s([x_1, \ldots, x_n]^T) = [\ldots, \xi^s, \ldots]^T \]
    where
    \[ \xi^s = x_1^{s_1} x_2^{s_2} \ldots x_n^{s_n}, \quad \sum s_i = s \]

- hybrid decoupling constraint

\[
p_s(r) = \prod_{i=1}^s (b_i^T r_t) = c_s^T \nu_s(r_t) = 0
\]
What happens when data is noisy?

- **Veronese map** $\nu_s(r_t)$:
  - Polynomial mapping
  - Lifts the data to higher dimensional space where parameter vector is in the nullspace of embedded data matrix

- **Noisy case** $\nu_s(r_t, \eta_t)$:
  - Lifted data depends on noise polynomially!
  - Need to find an admissible noise sequence to estimate the nullspace

\[
V_s c_s = \begin{bmatrix}
\nu_s(r_{t_0})^T \\
\vdots \\
\nu_s(r_T)^T \\
\end{bmatrix} c_s = 0
\]
Noisy embedded data matrix $V_s$

- 1$^{st}$ order system: $n_a = n_c = 1$, with 2 modes: $s=2$

$$r_t = [-y_t, y_{t-1}, u_{t-1}]^T$$

$$\nu_2(r_t, \eta_t)^T = \begin{bmatrix} y_t^2 - 2y_t \eta_t + \eta_t^2 \\ -y_t y_{t-1} + y_{t-1} \eta_t \\ -y_t u_{t-1} + u_{t-1} \eta_t \\ y_{t-1}^2 \\ y_{t-1} u_{t-1} \\ u_{t-1}^2 \end{bmatrix}^T$$

$V_s$ is polynomial in noise

Need to find a rank deficient $V_s$

Optimization Problem 1:

 minimize $\eta_t$ subject to $\ker V_s(r_t, \eta_t)$

$$\min_{\eta_t} \text{rank} V_s(r_t, \eta_t)$$

subject to $\|\eta_t\|_\infty \leq \epsilon$
Optimization Problem

• Rank is not a polynomial function. Can we use ideas from polynomial optimization?
  – YES.
• Can we utilize the problem structure to find an efficient formulation?
  – YES. Main Idea: Noise is independent. Define one dimensional distributions for each noise term.

\[
\text{Optimization Problem 1:} \\
\begin{align*}
\text{minimize}_{\eta_t} & \quad \text{rank} \mathbf{V}_s(\mathbf{r}_t, \eta_t) \\
\text{subject to} & \quad \|\eta_t\|_\infty \leq \epsilon
\end{align*}
\]
Noisy embedded data matrix $V_s$

$$\nu_2 (r_t, \eta_t)^T = \begin{bmatrix} y_t^2 - 2y_t \eta_t + \eta_t^2 \\ -y_t y_{t-1} + y_{t-1} \eta_t \\ -y_t u_{t-1} + u_{t-1} \eta_t \\ y_{t-1}^2 \\ y_{t-1} u_{t-1} \\ u_{t-1}^2 \end{bmatrix}^T$$

$$E_\mu [\nu_2 (r_t, \eta_t)^T] =$$

$$m^{(t)} = [m_1^{(t)}, \ldots, m_s^{(t)}]$$

$$m_i^{(t)} = E_{\mu^i} (\eta_t^i)$$

$$\begin{bmatrix} y_t^2 - 2y_t m_{1}^{(t)} + m_2^{(t)} \\ -y_t y_{t-1} + y_{t-1} m_{1}^{(t)} \\ -y_t u_{t-1} + u_{t-1} m_{1}^{(t)} \\ y_{t-1}^2 \\ y_{t-1} u_{t-1} \\ u_{t-1}^2 \end{bmatrix}^T$$
Optimization Problem

• **Theorem** (O., Lagoa, Sznaier Automatica 15):

  – There exists a rank deficient solution for Problem 2 if and only if there exists a rank deficient solution for Problem 1.

  – If \( c \) belongs to the nullspace of the solution of Problem 2, there exists a noise value \( \eta^* \) with \( \| \eta^* \|_{\infty} \leq \epsilon \) such that \( c \) belongs to the nullspace of \( V_s(r, \eta^*) \)

Optimization Problem 1:

minimize_{\eta_t} \quad \text{rank} V_s(r_t, \eta_t)

subject to \quad \| \eta_t \|_{\infty} \leq \epsilon

Optimization Problem 2:

minimize_{m^{(t)}} \quad \text{rank} \tilde{V}_s(r_t, m^{(t)})

subject to \quad \text{each } m^{(t)} \text{ is a moment sequence}

Convex constraint set:

- **Finite** Hankel matrix of moments should satisfy to LMIs

no relaxation!!
Optimization Problem

Problem 2
- Matrix rank minimization
- Subject to LMI constraints

- Use a convex relaxation (e.g. log-det heuristic of Fazel et al.) to solve Problem 2
- Find a vector \( c \) in the nullspace
- Estimate noise by root finding (\( V_s c = 0 \) polynomials of one variable)
- Proceed as in noise-free case

Optimization Problem 1:

\[
\begin{align*}
\text{minimize}_{\eta_t} & \quad \text{rank} V_s(r_t, \eta_t) \\
\text{subject to} & \quad \|\eta_t\|_\infty \leq \epsilon
\end{align*}
\]

Optimization Problem 2:

\[
\begin{align*}
\text{minimize}_{m^{(t)}} & \quad \text{rank} \tilde{V}_s(r_t, m^{(t)}) \\
\text{subject to} & \quad \text{each } m^{(t)} \text{ is a moment sequence}
\end{align*}
\]

Can also “handle” missing data, outliers
SARX (In)validation Problem

• Given:
  – A nominal hybrid model of the form:
    \[ y_t = \sum_{k=1}^{n_a} A_k(\sigma_t)y_{t-k} + \sum_{k=1}^{n_c} C_k(\sigma_t)u_{t-k} + f(\sigma_t) \]
    \[ \tilde{y}_t = y_t + \eta_t \]
  – A bound on the noise \( ||\eta||_{\infty} \leq \epsilon \)
  – Experimental input/output data \( \{u_t, \tilde{y}_t\}_{t=t_0}^T \)

• Determine:
  – whether there exist noise and switching sequences consistent with a priori information and experimental data

Unknown switches: Consistency set is non-convex!
Semialgebraic Consistency Set

• If $i^{th}$ submodel is active at time $t$

$$A_1(i)(\tilde{y}_{t-1} - \eta_{t-1}) + \ldots + A_{n_a}(i)(\tilde{y}_{t-n_a} - \eta_{t-n_a}) - (\tilde{y}_t - \eta_t) + C_1(i)u_{t-1} + \ldots + C_{n_c}(i)u_{t-n_c} + f(i) = 0$$

– all components of the output evolve with $i^{th}$ submodel (logical AND)

$$[h_{t,i}^{(1)}(\eta_{t:t-n_a}) = 0] \land \ldots \land [h_{t,i}^{(n_y)}(\eta_{t:t-n_a}) = 0] \iff g_{t,i}(\eta_{t:t-n_a}) \doteq \sum_{j=1}^{n_y} [h_{t,i}^{(j)}(\eta_{t:t-n_a})]^2 = 0$$

• One of the submodels is active at time $t$ (logical OR)

$$[g_{t,1}(\eta_{t:t-n_a}) = 0] \lor \ldots \lor [g_{t,s}(\eta_{t:t-n_a}) = 0] \iff p_t(\eta_{t:t-n_a}) \doteq \prod_{i=1}^{s} g_{t,i}(\eta_{t:t-n_a}) = 0$$
Semialgebraic Consistency Set

• The model is invalid if and only if

\[ T'(\eta) = \{ \eta \mid \epsilon^2 - [\eta_t^{(j)}]^2 \geq 0 \ \forall t \in [0, T], j \in \mathbb{N}_{ny} \text{ and} \]
\[ p_t(\eta_{t:t-n_a}) = 0 \ \forall t \in [n_a, T] \}\]

is empty.

• Structured polynomial optimization problem:

\[ o^* = \min_{\eta} \sum_{t=n_a}^{T} p_t(\eta_{t:t-n_a}) \]

\[ \text{subject to} \]
\[ f_{t,j}(\eta_t^{(j)}) \geq 0 \ \forall t \in [0, T], j \in \mathbb{N}_{ny}. \]
Polynomial Optimization

- Problem has a sparse structure (*running intersection property* holds)

\[
o^* = \min_{\eta} \sum_{t=n_a}^{T} p_t(\eta_{t:t-n_a})
\]

\[\text{s.t.} \quad f_{t,j}(\eta_t^{(j)}) \geq 0 \quad \forall t \in [0, T], j \in N_n_y.\]

- We can create a convergent SDP hierarchy with \(O((n_a n_y)^{2N})\) variables using structure (instead of \(O((T n_y)^{2N})\) variables), where \(N\) is the relaxation order.

- **Theorem (O., Sznaier, Lagoa, TAC 14):** The hierarchy converges latest at \(N = s^{T-n_a+1}+1.\)

  where \(s:\) # of submodels, \(n_a:\) regressor order, \(T:\) time horizon
A fun example in information extraction

Normal behaviors: walking and waiting
Walking dynamics are learnt from training data using sys id, waiting dynamics are trivial

Example: Activity monitoring via model invalidation
A priori hybrid model: walking (learned from data) and waiting, 4% noise

WALK, WAIT
RUN
WALK, JUMP

Not Invalidated
Invalidated
Invalidated
Model Invalidation – fault detection?

- Can be easily extended to uncertain models:

\[ \sum : x(t + 1) = f_{\sigma(t)}(x(t), u(t), \delta(t), \Delta) \]
\[ y(t) = g_{\sigma(t)}(x(t), u(t), \delta(t), \Delta) \]

There is a basic semialgebraic consistency set.

- Can be used to:
  - **Run-time:** do anomaly detection (abnormal with respect to model and spec)
  - **Design-time:** find tight *provable error bounds* on uncertain parameters

No need to have explicit fault models (complex systems can fail infinitely many different ways!)
Can handle missing data!
Fault detection

• Model invalidation directly applies but the problem size increases with time...

• What if we have fault models?

\[
\begin{align*}
\sum : & \quad x(t + 1) = f_{\sigma(t)}(x(t), u(t), \delta(t), \Delta) \\
& \quad y(t) = g_{\sigma(t)}(x(t), u(t), \delta(t), \Delta) \\
\sum_F : & \quad x(t + 1) = f^F_{\sigma(t)}(x(t), u(t), \delta(t), \Delta) \\
& \quad y(t) = g^F_{\sigma(t)}(x(t), u(t), \delta(t), \Delta)
\end{align*}
\]

• Can we use the models to bound the amount of data needed to do fault detection?
T-detectability

- Given a system model and fault model with associated state, input and noise bounds, if there exists a T such that for any initial condition and any input/noise realization the “T-length behaviors” deviate, the fault is said to be T-detectable for the system.

- Intersection of the (semi-algebraic) behavior sets for the system and fault models for horizon T should be empty!

- For fixed T, polynomial optimization problem (need to iterate on T)
T-detectability

- intersection of the consistency sets for the system and fault models for horizon $T$ should be empty!
- For fixed $T$, polynomial optimization problem (need to iterate on $T$) – sufficient conditions for T-detectability

"Theorem": If T-detectability certificate is obtained with a relaxation order $N$, then using the same relaxation order for model invalidation problem gives a N&S condition for online fault detection.

$$
\epsilon^* := \min_{\{u(k), y(k)\}_{k=t}^{t+T}} \epsilon_o(\{u(k), y(k)\}_{k=t}^{t+T})
$$

s.t. \quad \{u(k), y(k)\}_{k=t}^{t+T} \in B_{poly}^T(G^f) \ .
Anomaly detection in building control

Switched affine model:
-- switching due to control actions
-- six states (room temperatures, pipe temperatures)
-- only a sensor measuring pipe temperature
-- noisy sensor measurements

Boiler fails at time 8:00 (supply temp drops)

Invalidation algorithm detects the failure in 2 steps:

\[
C_r \dot{T}_c = \sum_{i=1}^{2} K_{r,i}(T_i - T_c) + K_w(T_w - T_c),
\]

\[
C_i \dot{T}_i = K_{r,i}(T_c - T_i) + \sum_{j \neq i} K_{i,j}(T_j - T_i)
\]
Summary

Goal: go from data to information to control in a rigorous way with correctness guarantees.

- **Dynamics based information extraction:**
  - Hybrid dynamical models as compact representation for complex data streams
  - Lots of structure in problems involving dynamics
  - Optimization is a good lens to look at these problems
  - ★ connections between system identification/invalidation and information extraction/machine learning

Computational efficiency through
- Convex Relaxations
- Structural decompositions