Model structure selection for switched NARX system identification: a randomized approach

Federico Bianchi[a,*], Valentina Breschi[a], Dario Piga[b], Luigi Piroddi[a]

[a]*Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, 20133 Milano (Italy)*
(e-mail: {federico.bianchi, valentina.breschi, luigi.piroddi}@polimi.it).
[b]*IDSIA - Dalle Molle Institute for Artificial Intelligence Research - USI/SUPSI, CH-6928 Manno (Switzerland)*
(e-mail: dario.piga@supsi.ch).

Abstract

The identification of switched systems is a challenging problem, which entails both combinatorial (sample-mode assignment) and continuous (parameter estimation) features. A general framework for this problem has been recently developed, which alternates between parameter estimation and sample-mode assignment, solving both tasks to global optimality under mild conditions. This article extends this framework to the nonlinear case, which further aggravates the combinatorial complexity of the identification problem, since a model structure selection task has to be addressed for each mode of the system. To solve this issue, we reformulate the learning problem in terms of the optimization of a probability distribution over the space of all possible model structures. Then, a randomized approach is employed to tune this distribution. The performance of the proposed approach on some benchmark examples is analyzed in detail.

Keywords: NARX systems; Switched models; Structure selection; Randomized algorithms.

1. Introduction

In many modeling problems the system under study is characterized by the presence of some heterogeneity arising from changes in the operational conditions, so that both continuous (physical) and discrete (logical) dynamics are observed. Examples range from stock market analysis [27, 11], to human motion [28, 10, 31] and speech recognition [30, 36], just to cite a few. This heterogeneity is hardly captured with a single model, and typically requires to switch among multiple models (modes), each associated with a different system condition. The resulting learning problem is particularly complex, in that, besides having to fit multiple models, no prior information is usually available on the switching mechanism, which must also be inferred from the data.

In this work, we address the identification of a general class of switched systems, where the continuous dynamics is described by a set of *linear-in-the-parameters* regression models defining the relationship between the regression vector $\boldsymbol{\varphi} \in \mathbb{R}^n$ and the output $y \in \mathbb{R}$. To fully estimate a model of this class from data, one needs to jointly perform data clustering (*i.e.* assign each sample to a mode) and multi-model identification (*i.e.* estimate the parameters of the model associated to each mode). The induced optimization problem is therefore of the mixed-integer type, since it involves the identification of discrete variables representing the mapping of the samples to the modes, as well as continuous ones describing the model parameters.

Many approaches have been proposed to solve this problem over the last two decades (see, *e.g.*, [33], [12], and [17], for a comprehensive review). These methods can be roughly classified into two categories, depending on how the optimization problem is tackled. Some methods adopt a solution strategy which addresses the problem in one shot, optimizing simultaneously over both the continuous and discrete variables, [3], [37], [24], [26], [1], [32], [29], [25], [19], [34], while others deal separately with the sample-mode assignment and the parameter estimation tasks, [4], [9], [14], [35], [13], [8].

Some of these works have also been extended to the case of nonlinear modes, by resorting to the Nonlinear AutoRegressive with eXogenous input (NARX) modeling framework [22], [23]. For example, a framework based on kernel functional expansions to represent the nonlinear functions and on the minimization of a cost function involving only the continuous parameters of the model as variables is introduced in [19] (and later extended in [18], [20]). In [16], the authors propose an extension of the sum-of-norms approach described in [29] to piecewise systems with nonlinear dynamics, based again on kernel functional expansions. In [2], [21] the identification problem is first formulated as a sparse optimization problem and then relaxed in a convex form by approximating the ℓ_0 norm with the ℓ_1 norm. Although these approaches do allow to learn switching models with nonlinear sub-models, they might result in a model that is rather difficult to interpret due to the use of nonparametric techniques.

Alternatively, one can pursue a parametric approach, *e.g.,* by approximating the nonlinear functions through finite-dimensional parametrized polynomial expansions. This is in-

[*] Corresponding author at: Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, 20133 Milano (Italy)

deed a popular approach in black-box nonlinear model identification [7], provided the identification procedure includes a *model structure selection* (MSS) process to tackle the curse of dimensionality that is inherent to polynomial expansions. This consists in selecting the smallest subset of model terms that yields a prescribed level of accuracy. MSS requires the solution of a combinatorial problem of exponential complexity in the number of candidate model terms, and is often tackled with heuristic techniques, such as greedy incremental model building methods.

If, as in the present work, the objective is to identify a switched NARX (SNARX) model, the combinatorial complexity of the problem is further aggravated, since a MSS task must be addressed for each of the NARX sub-models, not to mention the sample-mode assignment problem. For these reasons the identification of SNARX models is a rather challenging problem. A first attempt to address this problem is presented in [6], where an iterative two-stage randomized approach for the identification of SNARX models with time-ordered data is proposed. This method relies on the definition of a probability distribution over the space of possible SNARX model structures, including in this concept not only the structure of the NARX sub-models but also the switching signal. This distribution represents the likelihood of each structure being the actual one and is progressively refined through a sample-and-evaluate strategy. More in detail, several structure samples are extracted from the distribution, and for each of them the NARX sub-models are identified (exploiting the segmentation of the dataset induced by the extracted switching signal). Then, the probability distribution is updated based on the aggregate evaluation of the obtained SNARX models, favoring the structure choices resulting in better models. To manage the combinatorial complexity of the sample-mode assignment problem, switching is only allowed in a small number of pre-assigned time instants, which motivates the introduction of a second stage devoted to the refinement of the number and location of the switching times, based on the evidence gathered in the first step. While relatively effective, this method heavily depends on the initial choice of the switching time instants and on the heuristic nature of the refinement stage.

This article presents a new approach for the identification of parametric SNARX models, which incorporates some features of the described method of [6], and specifically the randomized approach to MSS, but addresses the sample-mode assignment in a completely different way, that does not require to limit *a priori* the number of switching time instants, and consequently avoids the necessity to resort to a refinement stage. More precisely, the proposed method builds on the general framework of [4], that alternates between parameter estimation and sample-mode assignment, using a cost function that accounts for both tasks, and incorporates constraints on the switching mechanism directly within the objective. Under certain conditions, both tasks can be solved to global optimality using convex optimization and dynamic programming, respectively.

To deal with the MSS task in the framework of [4], which is not naturally equipped with this ability[1], we reformulate the learning problem in terms of the optimization of a probability distribution over the space of all possible model structures, along the lines of [6]. This distribution is progressively tuned via a sample-and-evaluate strategy, where each extracted structure is used to run an instance of the method of [4]. More precisely, the extracted mode sequence is used as initialization for the sample-mode assignment optimization, whereas the parameter estimation phase assumes the extracted NARX sub-model structures. In summary, the outer algorithm addresses the MSS problem, using the approach of [4] to estimate the parameters of the given model structures and to optimize the sample-mode assignment. The method is iterated until convergence to a limit distribution concentrated on the best switched model of the system generating the observed data. Besides addressing the MSS task in a structured way, this approach provides an efficient *warm-startup* for the sample-mode assignment task, the initialization of which is a crucial point in the method of [4]. Indeed, multiple guesses for the initial mode sequence have to be considered there, to alleviate the dependence of the resulting model on the initialization.

The rest of the paper is organized as follows. The SNARX identification problem is formalized in Section 2. In Section 3 the model structure selection problem is reformulated in a probabilistic setting. The identification algorithm is detailed in Section 4. Finally, some examples are presented in Section 5, followed by concluding remarks.

1.1. Notation

The following notation will be used throughout the paper. The set of integers is denoted by $\mathbb{N}$ and the set of real numbers by $\mathbb{R}$. Given $r \in \mathbb{R}$, let $\lfloor r \rfloor$ denote the largest value in $\mathbb{N}$ that is not greater than r.

Let $S \subset \{1, 2, \ldots\}$ be a finite set of integers and denote by $\# S$ the cardinality of S. Then, $\mathbb{1}_{[s=i]}$ represents an indicator function defined on the set S which has value 1 when $s \in S$ is equal to i and 0 for all the remaining values in S.

Given a vector $\boldsymbol{a} \in \mathbb{R}^n$, a^i denotes the i-th entry of $\boldsymbol{a}$ and $\|\boldsymbol{a}\|_2$ is the Euclidean norm of $\boldsymbol{a}$. Given a matrix $\boldsymbol{A} \in \mathbb{R}^{n \times m}$, $\boldsymbol{A}^T$ denotes the transpose of $\boldsymbol{A}$, and A^{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, m$, the element of $\boldsymbol{A}$ at row i and column j. The identity matrix of size n is denoted as $\boldsymbol{I}_n$. Given an ordered collection $\boldsymbol{C}$ of N elements, let $\boldsymbol{c}_i$, $i = 1, \ldots, N$, denote the i-th element of $\boldsymbol{C}$.

Given a random variable x with probability distribution $\mathcal{P}_x$, $\mathbb{E}_{\mathcal{P}_x}[x]$ denotes the expected value of x w.r.t. $\mathcal{P}_x$. Let x be a discrete random variable defined on domain $\mathcal{X}$. Then, with some abuse of notation, we will refer to $\mathcal{P}_x(\bar{x})$ as the value of the probability mass function of the distribution $\mathcal{P}_x$ evaluated at $\bar{x} \in \mathcal{X}$. Let x be a random binary variable which takes the value 1 with probability μ and the value 0 with probability $(1 - \mu)$. We say that x is distributed according to a Bernoulli distribution with parameter μ, *i.e.*, $x \sim \mathtt{Be}(\mu)$. Let x be a random variable that can take one of K possible values, the probability of each value being separately specified by $\boldsymbol{\eta} = [\eta_1, \ldots, \eta_K]$, with $\sum_{i=1}^{K} \eta_i = 1$.

[1] The approach of [4] encompasses the use of regularization techniques, which can only partially address the MSS task, and may require careful tuning.

Then, we say that x is distributed according to a Categorical distribution with parameter $\boldsymbol{\eta}$, *i.e.*, $x \sim \texttt{Cat}(\boldsymbol{\eta})$.

2. Problem setting

2.1. System description

Consider a *single-input single-output* (SISO) nonlinear switched system with $K \in \mathbb{N}$ modes in the form

$$y_t = g_{\sigma_t}(\boldsymbol{x}_t) + e_t, \tag{1}$$

where $\boldsymbol{x}_t \in \mathcal{X} \subseteq \mathbb{R}^{n_y+n_u}$ is the vector $[y_{t-1}, \cdots, y_{t-n_y}, u_{t-1}, \cdots, u_{t-n_u}]$; $u_t \in \mathbb{R}$ and $y_t \in \mathbb{R}$ are the input and output signals at time $t \in \mathbb{N}$, respectively; the model orders n_y and n_u describe the dynamical order of the system; and $g_k : \mathcal{X} \to \mathbb{R}$, with $k = 1, \ldots, K$, is a nonlinear function of x_t. The term e_t is a white noise independent of u_t, assumed to be Gaussian distributed with zero-mean and variance ζ^2. The latent variable $\sigma_t \in \mathcal{K} = \{1, \ldots, K\}$ indicates the active mode at time t.

2.2. SNARX identification

Given a dataset $\mathcal{D} = \{(u_t, y_t)\}_{t=1}^N$ of time-ordered samples generated by system (1), we aim at fitting a model to the data $\mathcal{D}$ by approximating the *unknown* nonlinear maps $\{g_k\}_{k=1}^K$ with polynomial expansions up to a given order n_d. This choice has the advantage of making the model *linear-in-the-parameters*, and, as argued in [7], more amenable to interpretation and analysis.

Accordingly, the identification problem requires both to reconstruct the mode sequence $\boldsymbol{\sigma} = \{\sigma_t\}_{t=1}^N \in \mathcal{K}^N$ and to estimate the parameters of the following model:

$$\hat{y}_t(\boldsymbol{\vartheta}_{\sigma_t}) = \boldsymbol{\varphi}_t^T \boldsymbol{\vartheta}_{\sigma_t} = \sum_{j=1}^{n} \varphi_t^j \vartheta_{\sigma_t}^j, \tag{2}$$

where $\boldsymbol{\varphi}_t = \left[\varphi^1(\boldsymbol{x}_t), \ldots, \varphi^n(\boldsymbol{x}_t)\right] \in \mathcal{F} \subseteq \mathbb{R}^n$ is the regressor vector, the *regressor* $\varphi^i(\boldsymbol{x}_t)$, $i = 1, \ldots, n$, being a mapping that projects $\boldsymbol{x}_t$ onto a finite-dimensional space, and $\boldsymbol{\vartheta}_k \in \mathbb{R}^n$ is the parameter vector defining the submodel associated to the k-th mode, with $k = 1, \ldots, K$.

Under the assumption that the number of modes K and the model orders n_y and n_u are fixed *a priori*, the estimation of the parameters $\boldsymbol{\Theta} = [\boldsymbol{\vartheta}_1 \ldots \boldsymbol{\vartheta}_K]$ and of the sequence $\boldsymbol{\sigma}$ is addressed by minimizing the following cost function [4]:

$$\tilde{\mathcal{J}}(\boldsymbol{\Theta}, \boldsymbol{\sigma}) = \sum_{t=1}^{N} (y_t - \hat{y}_t(\boldsymbol{\vartheta}_{\sigma_t}))^2 + \beta \sum_{k=1}^{K} \|\boldsymbol{\vartheta}_k\|_2^2 + \mathcal{L}(\boldsymbol{\sigma}), \tag{3a}$$

where $\beta > 0$ is a tunable regularization parameter. The first two terms of the cost function account for the model precision and size. The term $\mathcal{L} : \mathcal{K}^N \to \mathbb{R}$ is introduced to explicitly account for the switching nature of the underlying system and is defined as

$$\mathcal{L}(\boldsymbol{\sigma}) = \sum_{t=2}^{N} \mathcal{L}^{trans}(\sigma_t, \sigma_{t-1}), \tag{3b}$$

where the *mode transition cost* $\mathcal{L}^{trans} : \mathcal{K}^2 \to \mathbb{R}$ accounts for changes in the operating condition.

2.3. Model structure selection (MSS)

The setting described in the previous subsection allows one to completely identify the switching model. However, using the full polynomial expansion of the map $\boldsymbol{\varphi}_t$ in model (2) is typically a recipe for overparametrization. On the other hand, an incorrect or incomplete map $\boldsymbol{\varphi}_t$ might result in a poorly performing and structurally biased approximation of the underlying system. Thus, a MSS procedure must be put in place, as discussed in the following.

Let $S = [\boldsymbol{s}_1, \ldots, \boldsymbol{s}_K] \in \mathcal{S} = \{0, 1\}^{n \times K}$ be an $n \times K$ matrix coding the K NARX structures associated to the local linear-in-the-parameter models, such that $s_k^j = 1$ if the j-th regressor φ^j belongs to the k-th sub-model structure and $s_j^k = 0$ otherwise. As the main information on the switching mechanism is retained in the active mode sequence $\boldsymbol{\sigma}$, the overall structure of the model in (2) is fully embedded into the pair $\lambda = (\boldsymbol{\sigma}, S)$, taking values in $\Lambda = \mathcal{K}^N \times \mathcal{S}$. The performance of a given model structure λ can thus be measured as the optimal value of the cost $\tilde{\mathcal{J}}(\boldsymbol{\Theta}, \boldsymbol{\sigma})$, *i.e.*,

$$J(\lambda) = \min_{\boldsymbol{\Theta}} \; \tilde{\mathcal{J}}(\boldsymbol{\Theta}, \boldsymbol{\sigma}), \tag{4a}$$

$$\text{s.t.} \quad \vartheta_k^j = 0 \text{ if } s_k^j = 0, j = 1, \ldots, n, \; k = 1, \ldots, K, \tag{4b}$$

where the constraints in (4b) take into account the model structure. Accordingly, the MSS problem is discussed below.

First, the following useful definition is introduced.

Definition (Z-score)**.** *The Z-score z_k^j associated to the estimate of the parameter ϑ_k^j is given by the ratio:*

$$z_k^j = \frac{\vartheta_k^j}{\hat{\zeta}_k \sqrt{V^{jj}}}, \tag{5}$$

where

$$\hat{\zeta}_k = \sqrt{\frac{1}{N_k - n} \sum_{\sigma_t = k} (y_t - \hat{y}_t(\boldsymbol{\vartheta}_{\sigma_t}))^2},$$

is the sampled estimate of the noise standard deviation ζ, $N_k = \#\{\sigma_t = k\}$ is the number of samples assigned to mode k, and V^{jj} denotes the j-th diagonal element of the matrix

$$V = \left(\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \beta \boldsymbol{I}_n\right)^{-1} \boldsymbol{\Phi}^T\boldsymbol{\Phi} \left[\left(\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \beta \boldsymbol{I}_n\right)^{-1}\right]^T, \tag{6}$$

with $\boldsymbol{\Phi} \in \mathbb{R}^{N_i \times n}$ stacking on its rows the regression vectors $\boldsymbol{\varphi}_t$, for $t \in \{t : \sigma_t = k\}$. □

In this work, regressor redundancy is tackled by applying an *a posteriori* t-test on the estimated parameter vectors to detect terms that are statistically indistinguishable from 0, which are then pruned from the corresponding local model structure. The t-test relies on the computation of the local Z-scores z_k^j, for $j = 1, \ldots, n$ and $k = 1, \ldots, K$.

Based on the considerations above, the MSS problem is formalized as follows.

Problem 1 (Model Structure Selection). *Under the assumption that the number of modes K and the model orders n_y and n_u are fixed, the MSS problem consists in finding $\lambda^\star = (\boldsymbol{\sigma}^\star, S^\star)$ that solves to optimality the fitting problem in* (4)*, without introducing redundant terms. Formally:*

$$
\begin{aligned}
\lambda^\star = & \arg\min_{\lambda\in\Lambda} J(\lambda) \\
& \text{s.t. } s_k^j = 0 \text{ if } \left|z_k^j\right| < t_{\alpha/2, N_k - n}, \\
& \qquad k = 1, \dots, K, j = 1, \dots, n,
\end{aligned} \tag{7}
$$

where z_k^j is the Z-score associated to the estimate of the parameter ϑ_k^j and $t_{\alpha/2,N_k-n}$ is the critical value of a t-Student distribution with $N_k - n$ degrees of freedom and confidence level α. ■

Assumption 1 (Uniqueness of the solution). *Let $\Lambda^\star \subset \Lambda$ be the set of the $K!$ optimizers for* (7) *that are equivalent under a permutation of the mode labels $\boldsymbol{\sigma}^\star$. We assume that a lexicographic rule is specified to chose $\lambda^\star \in \Lambda^\star$ such that the solution in* (7) *is unique.* □

3. Continuous reformulation of the MSS problem

Problem (7) is a mixed-integer programming problem involving N categorical variables $\boldsymbol{\sigma}$ taking values in $\mathcal{K}$ and $n \times K$ binary variables S. Thus, the optimization problem (7) can quickly become computationally intractable for mixed-integer numerical solvers using *e.g.*, *branch-and-bound* methods. To overcome this limitation, we look at problem (7) from a probabilistic perspective which allows us to reformulate the MSS problem using only continuous optimization variables.

Let $\boldsymbol{\gamma}$ be a discrete random variable which takes values in the set of model structures Λ, according to some probability distribution $\mathcal{P}_{\boldsymbol{\gamma}}$. The expected performance of $\boldsymbol{\gamma}$ can be measured as:

$$
\mathbb{E}_{\mathcal{P}_{\gamma}}[J(\boldsymbol{\gamma})] = \sum_{\lambda\in\Lambda} J(\lambda)\mathcal{P}_{\boldsymbol{\gamma}}(\lambda), \tag{8}
$$

and if we let $\mathcal{P}_{\boldsymbol{\gamma}}$ span all possible distributions over the set Λ, the minimum value of (8) with respect to $\mathcal{P}_{\boldsymbol{\gamma}}$ is obtained by making all the probability mass concentrate on the optimizer $\lambda^\star$ of the original MSS problem (7). Formally, by introducing

$$
\mathcal{P}_{\boldsymbol{\gamma}}^\star = \arg\min_{\mathcal{P}_{\gamma}} \mathbb{E}_{\mathcal{P}_{\gamma}}[J(\boldsymbol{\gamma})], \tag{9}
$$

it thus holds that under Assumption 1 $\mathcal{P}_{\boldsymbol{\gamma}}^\star(\lambda^\star) = 1$ and $\mathcal{P}_{\boldsymbol{\gamma}}^\star(\lambda) = 0$ for all $\lambda \in \Lambda \setminus \{\lambda^\star\}$.

In order to tackle the optimization of (8) with respect to $\mathcal{P}_{\boldsymbol{\gamma}}$, it is necessary to adopt a suitable parametrization of $\mathcal{P}_{\boldsymbol{\gamma}}$. To this end, we introduce the following assumption.

Assumption 2 (Independence assumptions). *We work under the following probabilistic assumptions:*

A2.1 *The mode sequence $\boldsymbol{\sigma}$ and the K NARX structure S are independent random variables.*

A2.2 *The local model structures $\mathbf{s}_k$, $k = 1, \dots, K$, are mutually independent and the elements s_k^j, $j = 1, \dots, n$, of the k-th sub-model are also mutually independent, $k = 1, \dots, K$.*

A2.3 *The mode activation elements σ_t, $t = 1, \dots, N$ are mutually independent.* □

Remark 1. *The independence assumptions 2 are not meant to provide any statistical insight on the data-generating system, but they are only functional to the sampling of the model structures λ when applying the randomized method discussed in this paper. Stated otherwise, no information regarding the interdependence among elements of λ is taken into account when sampling from $\mathcal{P}_{\boldsymbol{\gamma}}$.* ■

Under Assumption A2.1 , the probability density function of $\mathcal{P}_{\boldsymbol{\gamma}}$ can be expressed as

$$
\mathcal{P}_{\boldsymbol{\gamma}}(\lambda) = \mathcal{P}_{\xi}(\boldsymbol{\sigma}) \cdot \mathcal{P}_{\rho}(S), \tag{10a}
$$

where $\mathcal{P}_{\xi}(\boldsymbol{\sigma})$ accounts for the mode sequence and $\mathcal{P}_{\rho}(S)$ for the K local model structures.

Let us associate a Bernoulli random variable $\rho_k^j \sim \mathtt{Be}(\mu_k^j)$ to each element s_k^j, for $k = 1, \dots, K$ and $j = 1, \dots, n$. The success probability μ_k^j thus represents the belief that the regressor φ^j belongs to the k-th local model, and thus μ_k^j is referred to as *Regressor Inclusion Probability* (RIP). Based on the independence assumption A2.2 , $\mathcal{P}_{\rho}$ can be factorized as

$$
\mathcal{P}_{\rho}(S) = \prod_{k\in\mathcal{K}} \prod_{j=1}^{n} \left(\mu_k^j\right)^{\mathbb{1}_{[s_k^j=1]}} \left(1 - \mu_k^j\right)^{\mathbb{1}_{[s_k^j=0]}}. \tag{10b}
$$

Similarly, $\mathcal{P}_{\xi}(\boldsymbol{\sigma})$ is defined by associating to each σ_t a Categorical random variable $\xi_t \sim \mathtt{Cat}\,(\boldsymbol{\eta}_t)$, where $\boldsymbol{\eta}_t = \left[\eta_t^1, \dots, \eta_t^K\right]$ and η_t^k denotes the probability of σ_t taking value k. In the following, we refer to η_t^k as the *Mode Extraction Probability* (MEP), for which it holds that

$$
\sum_{k=1}^{K} \eta_t^k = 1.
$$

Under assumption A2.3 $\mathcal{P}_{\xi}(\boldsymbol{\sigma})$ is factorized as

$$
\mathcal{P}_{\xi}(\boldsymbol{\sigma}) = \prod_{t=1}^{N} \prod_{k\in\mathcal{K}} \left(\eta_t^k\right)^{\mathbb{1}_{[\sigma_t=k]}}. \tag{10c}
$$

Using parametrization (10) of $\mathcal{P}_{\boldsymbol{\gamma}}$, the MSS problem is addressed as discussed in the next section.

4. The proposed algorithm

To solve the MSS problem through the continuous formulation presented in Section 3, we employ a randomized strategy, that operates by sampling and evaluating model structures from the distribution $\mathcal{P}_{\boldsymbol{\gamma}}$. This procedure employs the gathered information to iteratively update both the RIPs $\{\mu_k^j\}_{j=1}^n$ in (10b), for all sub-models $k = 1, \dots, K$, and the MEPs $\{\eta_t^k\}_{k=1}^K$ in (10c), for $t = 1, \dots, N$, thus modifying the sampling distribution to increase the probability of selecting good model structures.

The algorithm iteratively repeats the following four steps:

1. *Structure extraction* – A population of N_p model structures λ_p, $p = 1, \dots, N_p$, is extracted according to the distribution $\mathcal{P}_{\boldsymbol{\gamma}}$. Each extracted $\lambda_p = (\boldsymbol{\sigma}_p, S_p)$ represents a fixed model structure.
2. *Model fitting* – For each model structure λ_p, the fitting cost $\tilde{\mathcal{J}}(\boldsymbol{\Theta}, \boldsymbol{\sigma})$ in (3) is minimized using the coordinate descent approach [4], alternating optimization w.r.t. $\boldsymbol{\Theta}$ and $\boldsymbol{\sigma}$. The extracted mode sequence $\boldsymbol{\sigma}_p$ is used as initial guess in the coordinate descent algorithm.
3. *Redundancy check* – A redundancy check is performed on the parameters $\boldsymbol{\Theta}$ estimated at stage 2. Redundant parameters (if any) are pruned from the model and the parameters $\boldsymbol{\Theta}$ of the pruned model are re-estimated by solving problem (4) with the reduced model structure. The optimal fitting loss $J(\lambda_p)$ (with $p = 1, \dots, N_p$) is also computed.
4. *Distribution update* – The RIPs and MEPs are updated based on the optimal fitting losses computed at stage 3.

The overall identification procedure is sketched in Figure 1. Each stage is discussed in detail in the next sections. Note that the algorithm requires the initialization of the MEPs η_t^k, $t = 1, \dots, N$, $k = 1, \dots, K$, and the RIPs μ_k^j, $k = 1, \dots, K$, $j = 1, \dots, n$. To encourage the extraction of sparse models at the early stages, a convenient choice is to set $\mu_k^j = \epsilon$, for small values of $\epsilon \in (0, 1)$. In the absence of any *a priori* assumption on the switching signal, one can attribute equal probabilities $\eta_t^k = 1/K$, $\forall t$ to all modes.

The algorithm ends when a stopping criterion is met. This can either be associated with a maximum number of iterations, or a practical convergence of the MEPs and RIPs parameters, which is achieved when the relative difference between the η_t^k and μ_k^j calculated at subsequent iterations is lower than a given threshold.

4.1. Structure extraction

According to the continuous formulation presented in Section 3, the definition of a structure λ amounts to assigning each input-output pair $(\boldsymbol{\varphi}_t, y_t)$ to a mode $\sigma_t \in \mathcal{K}$ and establishing if the regressor φ^j belongs to the k-th local model structure, for $k = 1, \dots, K$ and $j = 1, \dots, n$. The former task is carried out by extracting a random variable $\boldsymbol{\sigma}_p$ from the Categorical distribution $\mathcal{P}_\xi(\boldsymbol{\sigma})$ in (10c). The latter involves sampling a random variable S_p from the Bernoulli distribution $\mathcal{P}_\rho(S)$ in (10b). This procedure is performed N_p times until a population of N_p candidate structures $\{\lambda_p\}_{p=1}^{N_p}$ is extracted, with $\lambda_p = (\boldsymbol{\sigma}_p, S_p)$.

4.2. Model fitting

For each extracted structure λ_p, the corresponding SNARX model is identified via an instance of the procedure proposed in [4], which alternates between the minimization of the cost $\tilde{\mathcal{J}}(\boldsymbol{\Theta}, \boldsymbol{\sigma})$ in (3) with respect to $\{\boldsymbol{\vartheta}_k\}_{k=1}^K$, for a fixed mode sequence $\boldsymbol{\sigma}$, and the optimization of $\tilde{\mathcal{J}}(\boldsymbol{\Theta}, \boldsymbol{\sigma})$ with respect to $\boldsymbol{\sigma}$, this time keeping the parameters of the local models fixed. The procedure ends when either the identified optimal mode sequence computed at two consecutive iterations does not change or a stopping criterion on the performance of the fitted model is met.

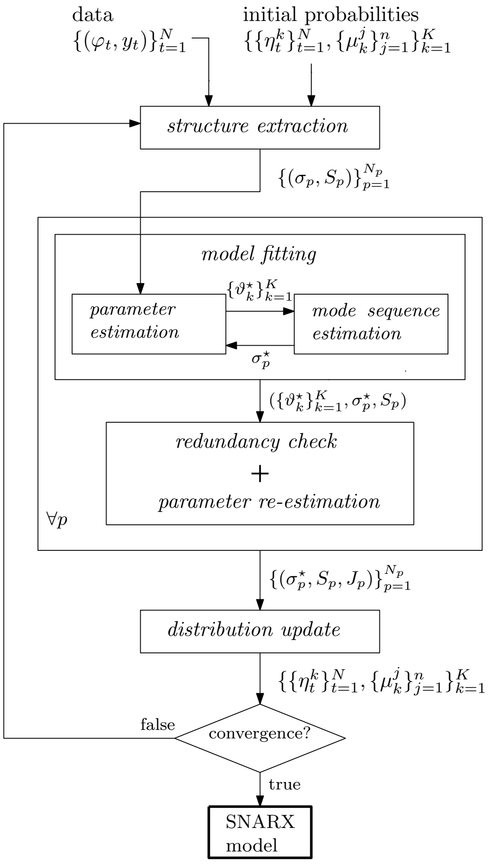

data
$\{(\varphi_t, y_t)\}_{t=1}^N$
initial probabilities
$\{\{\eta_t^k\}_{t=1}^N, \{\mu_k^j\}_{j=1}^n\}_{k=1}^K$
structure extraction
$\{(\sigma_p, S_p)\}_{p=1}^{N_p}$
model fitting
$\{\vartheta_k^\star\}_{k=1}^K$
parameter estimation
mode sequence estimation
$\sigma_p^\star$
$(\{\vartheta_k^\star\}_{k=1}^K, \sigma_p^\star, S_p)$
redundancy check
+
parameter re-estimation
$\forall p$
$\{(\sigma_p^\star, S_p, J_p)\}_{p=1}^{N_p}$
distribution update
$\{\{\eta_t^k\}_{t=1}^N, \{\mu_k^j\}_{j=1}^n\}_{k=1}^K$
false
convergence?
true
SNARX model

Figure 1: Flow chart of the proposed SNARX identification algorithm.

As the outcome of this iterative procedure depends on the chosen initial conditions, we propose to exploit the extracted mode sequence $\boldsymbol{\sigma}_p$ to initialize $\boldsymbol{\sigma}$. This choice is shown to speed up the convergence of the coordinate descent approach in practice, since the optimization starts from an educated initial guess of the mode sequence as the SNARX identification algorithm proceeds.

Thanks to the separability of the cost in (3), the estimation of the local model parameters when $\sigma_t = \sigma_t^\star$ amounts to solving K separate constrained Least Squares (LS) problems:

$$\boldsymbol{\vartheta}_k^\star = \arg\min_{\boldsymbol{\vartheta}_k} \sum_{k=1}^K \sum_{t:\sigma_t^\star = k} (y_t - \hat{y}_t(\boldsymbol{\vartheta}_k))^2 + \beta \|\boldsymbol{\vartheta}_k\|_2^2, \tag{11}$$
$$\text{subject to } \vartheta_k^j = 0 \text{ if } s_k^j = 0, j = 1, \dots, n,$$

for $k = 1, \dots, K$, with the constraint accounting for the extracted model structure.

Once the parameters $\{\boldsymbol{\vartheta}_k^\star\}_{k=1}^K$ are computed, the mode sequence is updated by solving the following optimization problem:

$$\boldsymbol{\sigma}^{\star} = \arg\min_{\sigma\in\Sigma} \sum_{t=1}^{N} \left(y_t - \hat{y}_t(\boldsymbol{\vartheta}^{\star}_{\sigma_t})\right)^2 + \mathcal{L}(\boldsymbol{\sigma}), \tag{12}$$

that can be solved via standard discrete *dynamic programming* (DP) [5], as briefly summarized in the following.

Let $Q \in \mathbb{R}^{K\times N}$ be a matrix that stores the cost of assigning each data pair $(\boldsymbol{\varphi}_t, y_t)$, $t = 1,\ldots,N$ to each mode $k \in \mathcal{K}$. By setting $Q_{k,N}$ as the cost of assigning the last data pair to the k-th mode, *i.e.,*

$$Q_{k,N} = (y_N - \hat{y}_N(\boldsymbol{\vartheta}^{\star}_k))^2, \tag{13a}$$

for $k = 1,\ldots,K$, the elements of Q can be computed backwards for $t = N-1,\ldots,1$ as follows:

$$Q_{k,t} = (y_t - \hat{y}_t(\boldsymbol{\vartheta}^{\star}_k))^2 + Q_{\mathcal{V}_{k,t},t+1} + \mathcal{L}^{trans}(k, \mathcal{V}_{k,t}), \tag{13b}$$

with the index $\mathcal{V}_{k,t}$ retaining information on the optimal backward path followed to reach the k-th mode, namely

$$\mathcal{V}_{k,t} = \arg\min_{i\in\mathcal{K}} \left(Q_{i,t+1} + \mathcal{L}^{trans}(k,i)\right), \quad k = 1,\ldots,K. \tag{13c}$$

Once the initial costs

$$Q_{k,1} = \min_{i\in\mathcal{K}} \left(Q_{i,2} + \mathcal{L}^{trans}(k,i)\right) \tag{13d}$$

are computed for $k = 1,\ldots,K$, the optimal mode sequence can then be retrieved forwards, from 1 to N, by setting

$$\sigma^{\star}_1 = \arg\min_{k} Q_{k,1}, \tag{13e}$$

$$\sigma^{\star}_t = \mathcal{V}_{\sigma^{\star}_t,t}, \quad t = 1,\ldots,N. \tag{13f}$$

It is worth remarking that selecting the transition loss $\mathcal{L}^{trans}$ usually requires several attempts that involve fitting and cross-validation. A simple approach proposed in [4] consists in updating the mode transition loss $\mathcal{L}^{trans}$ after the fitting phase based on the computed best sequence $\boldsymbol{\sigma}^{\star}$, and then run the fitting algorithm again.

More specifically, given a set of relative weights $\tau_1,\ldots,\tau_K$, $\mathcal{L}^{trans}$ is updated by computing the empirical switching frequencies (with Laplace smoothing) from mode j to mode i:

$$\pi_{ij} = \frac{1 + \#\{t \in \{2,\ldots,N\} :\ \sigma^{\star}_t = i,\ \sigma^{\star}_{t-1} = j\}}{N + K^2} \tag{14a}$$

and the empirical frequency of being in mode j:

$$\pi_j = \frac{1 + \#\{t \in \{2,\ldots,N\} :\ \sigma^{\star}_{t-1} = j\}}{N + K}. \tag{14b}$$

and then setting

$$\mathcal{L}^{trans}(i,j) = \quad -\tau_i \frac{\log\left(\frac{\pi_{ij}}{\pi_j}\right)}{\sum_{j=1}^{K} \log\left(\frac{\pi_{ij}}{\pi_j}\right)}, \quad i,j = 1,\ldots,K, \tag{14c}$$

$$\tau_i = \quad \sum_{j=1}^{K} \mathcal{L}^{trans}(i,j), \quad i = 1,\ldots,K. \tag{14d}$$

The proposed update of τ_i (with $i = 1,\ldots,K$) in (14d), preserves the initial relative weight between the components of the fitting cost in (3).

4.3. Redundancy check

The model fitting step described in Section 4.2 does not take into account possible redundancies in the model parametrization. To this end, in the third step, the estimated parameters $\{\boldsymbol{\vartheta}^{\star}_k\}_{k=1}^{K}$ undergo a statistical t-test based on the Z-score defined in (5), computed for fixed $\sigma_t = \sigma^{\star}_t$ and for $N_k - \sum_{j=1}^{n} s^j_k$ degrees of freedom.

Let $\boldsymbol{\vartheta}^{\circ}_k$ be the true parameter vector describing the k-th local model of the data-generating system in (1), with $k = 1,\ldots,K$. For each $j \in \{j : s^j_k = 1\}$, we test the null hypothesis

$$\mathcal{H}_0 : \vartheta^{\circ\, j}_k = 0, \tag{15}$$

by analyzing the Z-score z^j_k associated with $\vartheta^{\star\, j}_k$. A large absolute value of z^j_k leads to reject the null hypothesis, while small values of the Z-score indicate that there is no sufficient evidence in the data to reject the null hypothesis.

The regressors for which the t-test fails to reject the null hypothesis $\mathcal{H}_0$ are removed from the NARX model structure $\boldsymbol{s}_k$. Then, a new least-squares parameter estimation is carried out as in (11), assuming the reduced structure S and the previous identified optimal mode sequence $\boldsymbol{\sigma}^{\star}$.

4.4. Distribution update

Finally, the *mode extraction* and *regression inclusion* probabilities are updated, with the purpose of encouraging the extraction of “good” samples from the distribution $\mathcal{P}_{\gamma}$, *i.e.* samples corresponding to SNARX model structures yielding good performance. The update rules are based on the aggregate comparison of the sub-populations of extracted structures λ obtained for the different values of each element σ^k_t and s^j_k. To facilitate such comparisons, we adopt the following exponential performance index (to be maximized) to characterize the structure λ:

$$\mathcal{J}(\lambda) = e^{-K_{\lambda} J(\lambda)}, \tag{16}$$

where $K_{\lambda} > 0$ is a tunable scaling parameter. This choice can help in discriminating between models with similar performance by amplifying their differences [38]. The parameter K_{λ} can be tuned in the first iteration of the algorithm as suggested in [6]:

$$K_{\lambda} = 10^{-(\min(OM(\mathcal{J}(\lambda)))+1)}, \tag{17}$$

where $OM(x) = \lfloor \log_{10}(x) \rfloor$ denotes the order of magnitude of a non-negative number x.

According to the approach proposed in [6], MEPs η_t^k and RIPs μ_k^j are updated as discussed next. Given $\xi_t \sim \mathtt{Cat}(\boldsymbol{\eta}_t)$ (with $t \in \mathbb{N}$), under the independence assumptions 2, for the *total expectation theorem* it holds that:

$$\mathbb{E}_{\mathcal{P}_\gamma}[\mathcal{J}(\boldsymbol{\gamma})] = \eta_t^k \mathbb{E}_{\mathcal{P}_\gamma}[\mathcal{J}(\boldsymbol{\gamma}) \mid \xi_t = k] + \left(1 - \eta_t^k\right) \mathbb{E}_{\mathcal{P}_\gamma}[\mathcal{J}(\boldsymbol{\gamma}) \mid \xi_t \neq k], \tag{18}$$

for all $k \in \mathcal{K}$. Taking the derivative of (18) w.r.t. η_t^k, one obtains:

$$\frac{\partial E_{\mathcal{P}_\gamma}[\mathcal{J}(\boldsymbol{\gamma})]}{\partial \eta_t^k} = \mathbb{E}_{\mathcal{P}_\gamma}\left[\mathcal{J}(\boldsymbol{\gamma}) \mid \xi_t = k\right] - \mathbb{E}_{\mathcal{P}_\gamma}\left[\mathcal{J}(\boldsymbol{\gamma}) \mid \xi_t \neq k\right]. \tag{19a}$$

Similarly, for $\rho_k^j \sim \mathtt{Be}(\mu_k^j)$ (for all $j = 1, \ldots, n$), the following expression is obtained for the gradient:

$$\frac{\partial E_{\mathcal{P}_\gamma}[\mathcal{J}(\boldsymbol{\gamma})]}{\partial \mu_k^j} = \mathbb{E}_{\mathcal{P}_\gamma}\left[\mathcal{J}(\boldsymbol{\gamma}) \mid \rho_k^j = 1\right] - \mathbb{E}_{\mathcal{P}_\gamma}\left[\mathcal{J}(\boldsymbol{\gamma}) \mid \rho_k^j = 0\right]. \tag{19b}$$

The information provided by (19) is used to update the mode extraction and regression inclusion probabilities as follows:

$$\eta_t^k \leftarrow \eta_t^k + \chi \frac{\partial E_{\mathcal{P}_\gamma}[\mathcal{J}(\boldsymbol{\gamma})]}{\partial \eta_t^k} \tag{20a}$$

$$\mu_j^k \leftarrow \mu_j^k + \chi \frac{\partial E_{\mathcal{P}_\gamma}[\mathcal{J}(\boldsymbol{\gamma})]}{\partial \mu_k^j}, \tag{20b}$$

where $\chi > 0$ is the learning rate. As suggested in [6], $\chi > 0$ is chosen according to the following adaptive rule:

$$\chi = \frac{1}{10\left(\mathcal{J}_{\text{best}} - \overline{\mathcal{J}}\right) + 0.1}, \tag{21}$$

where $\mathcal{J}_{\text{best}}$ and $\overline{\mathcal{J}}$ are the best and the mean value for $\mathcal{J}$ computed on the extracted samples for $\boldsymbol{\gamma}$, respectively.

A saturation is applied *a posteriori* to ensure that the MEP and RIP values resulting from (20) still represent valid probabilities *i.e.,* $\eta_t^k \in [0, 1]$ and $\mu_k^j \in [0, 1]$, $k = 1, \ldots, K$ and $j = 1, \ldots, n$. In addition, the MEPs are normalized to impose $\sum_{k=1}^K \eta_t^k = 1$.

Remark 2. *The MEPs updated in* (20a) *are used at the next iteration of Algorithm 1 to extract a set of* N_p *mode sequences* $\boldsymbol{\sigma}_p$. *Although this population of sequences is solely used to initialize the coordinate descent approach described in Section 4.2, the iterative update of the MEPs causes a progressive restriction and improvement of the set of initial conditions tested when learning the model. This, in turn, is likely to reduce the learning time and improve the accuracy of the resulting SNARX model.* ■

Remark 3. *In practice, only an approximate sampled version of the gradients in* (19) *can be computed, since the exact computation of the conditional expectations would require to exhaustively explore the entire solution space* Σ. *Therefore, at each iteration, the expected values are approximated by their corresponding sample estimates, computed based on the extracted candidate solutions* λ_p, $p = 1, \ldots, N_p$. ■

4.5. Algorithm convergence

The coordinate approach described in Section 4.2 always terminates in a finite number of steps, since at each stage the cost $\tilde{\mathcal{J}}(\boldsymbol{\Theta}, \boldsymbol{\sigma})$ in (3) is non-increasing and the number of possible mode sequences $\boldsymbol{\sigma}$ is finite for a fixed number K of system modes.

The convergence of the randomized *sample-and-evaluate* procedure proposed to solve the MSS problem 7 within the continuous framework described in Section 3 has been shown in [6]. In particular, when $\mathcal{P}_\gamma$ is sufficiently close to $\mathcal{P}_\gamma^\star$ in (9), then the sign of the gradients in (19) provide a reliable information for tuning the mode extraction and regression inclusion probabilities towards $\mathcal{P}_\gamma^\star$ by the iterative application of the update rules in (20). Practical experiences show that the algorithm converges even when it is iterated with $\mathcal{P}_\gamma$ randomly initialized.

5. Examples

We tested the proposed algorithm on both simulation examples using synthetic data and on an experimental case study addressing the unsupervised segmentation of honeybee dances.

5.1. Algorithm settings and performance indices

We recall that Algorithm 1 works under the assumption that number of modes K and the model orders n_y and n_u are fixed. In all the analysis which follows, the RIPs are initially set equal to $\mu_k^j = 0.1$ (for $k = 1, \ldots, K$, $j = 1, \ldots, n$), and equal MEPs $\eta_t^k = 1/K$ (for $t = 1, \ldots, N$) are assumed for all modes. The number of candidate model structures λ_p extracted at each iteration is set to $N_p = 100$. During the model redundancy step, each estimated model undergoes a t-test with *confidence level* $\alpha = 10^{-3}$. The coordinate descent approach tackling the model fitting step 4.2 terminates when either the identified optimal mode sequence or the performance of the fitted model does not change over two consecutive iterations (we employ a threshold of 10^{-8} for the performance). The regularization parameter β in (3) is set to 10^{-5}, while the following mode transition cost $\mathcal{L}^{trans}(\sigma_t^\star, \sigma_{t-1}^\star)$ in (3b) is considered:

$$\mathcal{L}^{trans}(\sigma_t^\star, \sigma_{t-1}^\star) = \begin{cases} -\tau \log\left(1 - (K-1)\pi\right) & \text{if } \sigma_t^\star = \sigma_{t-1}^\star \\ -\tau \log \pi & \text{if } \sigma_t^\star \neq \sigma_{t-1}^\star \end{cases} \tag{22}$$

where $\pi \in [0, 1]$ denotes the transition probability and $\tau > 0$ is a weight.

In the following, whenever the true mode sequence $\boldsymbol{\sigma}$ is available, it is used to evaluate the accuracy of the reconstructed mode sequence $\boldsymbol{\sigma}^\star$, which is measured through the following *clustering accuracy index*:

$$C_N^{true} = \frac{100}{N} \sum_{t=1}^{N} \mathbb{1}_{[\sigma_t^\star = \sigma_t]}. \tag{23}$$

Likewise, when available, the true K NARX structure S is used to asses the correctness of the selected structure $S^\star$.

The quality of the predicted output is measured in terms of the *fit rate index* defined as:

$$FIT = 100\left(1 - \frac{\sum_{t=1}^{N} \|y_t - \hat{y}_t\|_2^2}{\sum_{t=1}^{N} \|y_t - \bar{y}\|_2^2}\right), \tag{24}$$

where $\bar{y}$ denotes the average of the output $y_1, \ldots, y_N$. For Algorithm 1, the predicted output values $\hat{y}_t$, $t = 1, \ldots, \tilde{N}$ are computed via the *recursive inference algorithm* presented in Section 4.2.2 [4] for the case of *one-step ahead prediction*, with $\sigma_t^\star$ predicted by exploiting $\mathcal{L}^{trans}$ in (3b) and the fitting losses computed up to time $t-1$ with the observed past output values $\tilde{y}_1, \ldots, \tilde{y}_{t-1}$. We stress that validation is not performed through open-loop simulations, since it would require to test all possible switching paths.

All tests have been performed in a MATLAB 2019b environment, on an HP ProBook 650 G1 CORE i7-4702MQ CPU @2.20 GHz with 8GB of RAM.

5.2. Simulation example 1: SNARX system

We first apply the proposed MSS procedure to the example in [15], which switches with probability $\pi = 2.5\%$ between a linear mode 1:

$$y_t = -0.905 y_{t-1} + 0.9 u_{t-1} + e_t,$$

and a nonlinear mode 2:

$$y_t = -0.4 y_{t-1}^2 + 0.5 u_{t-1} + e_t,$$

where e_t is a zero mean Gaussian noise of variance ζ^2 and u_t is uniformly distributed in the interval $[0, 1]$. Regarding the NARX model structure selection, the candidate regressor set is defined by $n_d = 2$, $n_y = n_u = 3$, amounting to $n = 28$ regressors.

The algorithm was applied for several levels of the output noise ζ and weights τ in (22). For each value of ζ, a Monte Carlo (MC) analysis was carried out for each possible τ over the same data realization, consisting of 2000 samples for identification purposes and 2000 samples for validation. The aggregated results are summarized in Figure 2. The best values for τ obtained by cross-validation, corresponding to the maxima of the displayed curves, confirm the optimal theoretical values $\tau^\star = 2\zeta^2$ estimated according to the statistical interpretation of the mode transition cost $\mathcal{L}^{trans}$ (see [4, Section 3]).

From now on, we consider $\zeta^2 = 0.012$ and $\tau = 0.0240$. An MC analysis was carried out over 100 different data realizations of 4000 samples each, 2000 for the training and 2000 for validation. Table 1 reports the aggregated results, showing that both local structures have been always estimated successfully, and that the algorithm proved to be very accurate in reconstructing the mode sequence. It is all the more remarkable that the algorithm achieved such results by exploring a fairly small fraction of the overall solution space.

Additionally, we ran a comparative analysis with the non-parametric approach of [21], extending the one presented in [15] by fixing the submodel size and limiting the number of optimization variables. As in [15], the two modes are modeled via a linear kernel and a RBF kernel[2], respectively. Note that the method requires that the true model orders are known, *i.e.*, $n_y = n_u = 1$. Among the four methods proposed in [21] to fix the submodel size, we chose the Feature Vector Selection (FVS) method. The method of [21] selects the active mode based on the following criterion:

$$\sigma_t^\star = \arg\min_{k=1,\ldots,K} \left(y_t - \hat{y}_{t,k}\right)^2, \tag{25}$$

where $\hat{y}_{t,k}$ is the output predicted by the k-th mode. Notice that this clustering rule requires the measurement of the output at time t to select the active mode, whereas the proposed approach relies only on past data and the transition cost in (3b).

Figure 3 reports the results of this comparative study, showing that we manage to outperform [21] in terms of clustering accuracy, but apparently obtain a lower fit rate index. As discussed earlier, the proposed algorithm is capable of reconstructing quite accurately both the models and the switching signal (see also Figure 4), and this provides nearly equivalent one-step-ahead prediction performance compared to the true system. In view of this, at first glance, the better fitting performance of [21] is somewhat surprising. Observe, however, that

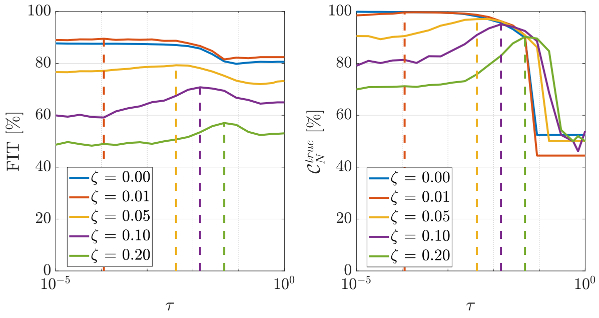


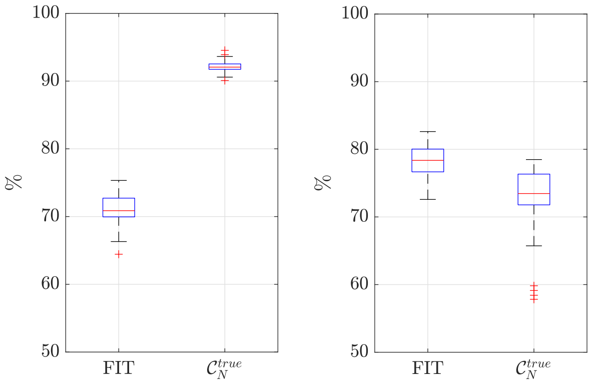
Figure 2: Simulation example 1: fit rate (left) and clustering accuracy (right) indices as a function of τ for different values of ζ. Optimal theoretical values $\tau^\star = 2\zeta^2$ are emphasized by vertical dashed lines. Results obtained on validation data using recursive one-step ahead prediction.

Table 1: Simulation example 1: Aggregated results from the MC analysis (computed on the training data).

Average elapsed time [s]	29.13
Average # of iterations	50.00
Average clustering accuracy [%]	97.09
Average # of explored sequences	$1.96{\cdot}10^3$
Percentage of correct selection of s_1 [%]	100
Average # of explored models for mode 1	673.63
Total # of possible model structures for mode 1	$2.68{\cdot}10^8$
Percentage of correct selection of s_2 [%]	100
Average # of explored model structures for mode 2	756.23
Total # of possible model structures for mode 2	$2.68{\cdot}10^8$

[2] By cross-validation, we set the width of the RBF kernel to 0.3 and $C = 10^3$, with C/N governing the trade-off between model complexity and model accuracy.

the method of [21] assigns the data to the modes based only on a fitting error criterion. This ultimately leads to a clustering which is quite loosely related to the true sample-mode subdivision, but allows to achieve a better overall fitting performance. Indeed, a closer inspection of the sample-mode assignment of the method of [21] (see Figure 4) reveals a high level of fragmentation, which implies a much higher number of switchings compared to the true system. Stated otherwise, the discrete dynamics of the underlying system is completely lost. Notice that this implies some level of distortion on the part of the submodel identification as well.



(a) Algorithm 1

(b) Algorithm [21]

Figure 3: Simulation example 1: Comparative analysis with the method of [21]. Results obtained on the validation data using one-step ahead prediction.

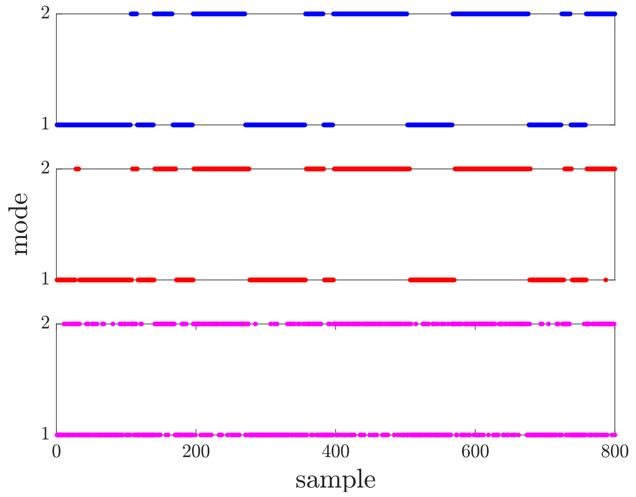


Figure 4: Simulation example 1: true mode sequence (top), reconstructed sequence by Algorithm 1 (middle), reconstructed sequence by Algorithm [21] (bottom).

We further analyzed the impact of an increasing number of mode commutations on the algorithm performance, by varying the transition probability π while keeping the data set size fixed. Figure 5 displays the trend of the clustering accuracy index for increasing π values, computed on training data sets of size $N = 2000$. The corresponding number of switching time instants ranges from 2 to 1400. The NARX structure selection sub-task seems to be unaffected by the increase in π, with the two structures correctly selected 96% (mode 1) and 93% (mode 2) of the times, respectively. On the other hand, for π values greater than 5%, a loss on the clustering accuracy is apparent due to the increasing complexity of the discrete dynamics. This reflects also on the fit rate index, which drops with a similar trend from 80.5% to 66%, but in our tests it seemed not to impact on the elapsed time required to reach algorithm convergence.

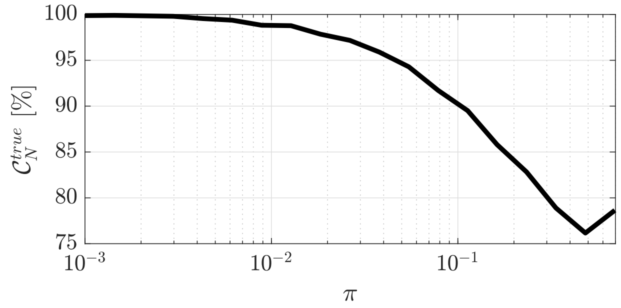


Figure 5: Simulation example 1: clustering accuracy index for varying transition probability π. Results obtained on the training data using one-step ahead prediction.

Finally, the computational time required to run Algorithm 1 was evaluated for training sets of increasing length N (for each N, 100 MC runs were carried out). As expected, the elapsed time (ET) increases with N as shown in Figure 6, proving that the computational burden for a fixed number of modes is mainly linked to the dimension of the data set. We stress that the clustering accuracy is almost 97% in all the performed experiments and the linear model structure is always correctly identified, while the structure selected for the nonlinear mode is correct with a rate always greater than 96%.

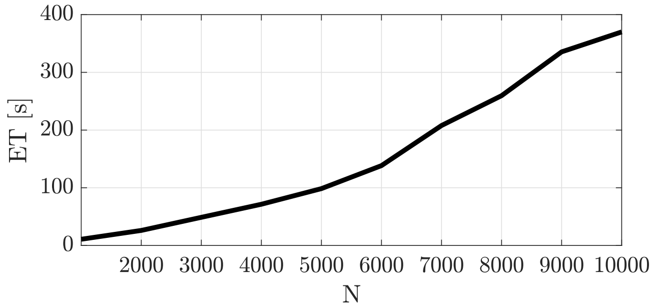


Figure 6: Simulation example 1: computational time for increasing training set size.

5.3. Simulation example 2: SARX system

The aim of this example is to assess how the proposed algorithm copes with local models that share the same structure. To this end, consider the system presented in [13]:

$$y_t = \vartheta_k^1 y_{t-1} - 0.7y_{t-2} + u_{t-1} - 0.5u_{t-2} + e_t, \tag{26}$$

that switches every 100 samples between $K = 4$ local models described by $\vartheta_1^1 = 1.5$, $\vartheta_2^1 = 1$, $\vartheta_3^1 = 0.5$, and $\vartheta_4^1 = -0.5$, respectively. The input signal $u(t)$ is a ±1 Pseudo-Random Binary Sequence (PRBS), while the noise is an i.i.d. Gaussian process, $e(t) \sim \mathcal{N}(0, 0.25)$. The regressor set is defined by $n_d = 1$, $n_y = n_u = 10$, amounting to $n = 21$ regressors. We set τ and π in (22) equal to 0.5 and 0.01, respectively.

Table2 reports the aggregated results of an MC study which has been carried out by executing the algorithm 100 times on different data realizations of size $N = 2000$. The obtained results show that the overall accuracy is satisfactory, in that all the NARX structures have been correctly selected with a rate greater than 80%. Nonetheless, they clearly show that the algorithm has struggled in distinguishing the first three modes, due to the common structure and equal parametrization of all local models except for one term. Indeed, the inspection of Table 3, which displays the parameter estimates associated to the true regressors, indicates a non-negligible dispersion of the results, although the mean values are quite accurate.

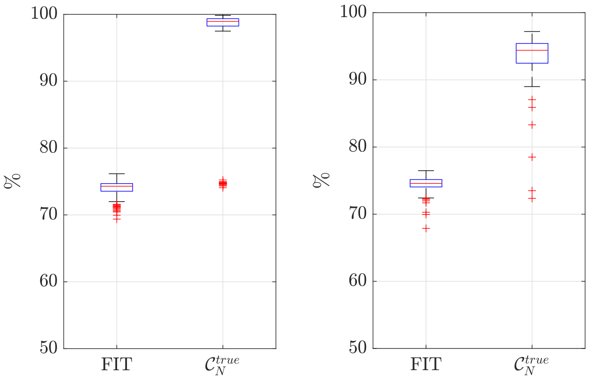
Table 2: Simulation example 2: Aggregated results from the Monte Carlo analysis (computed on the training data).

Average elapsed time [s]	54.36
Average # of iterations	114.93
Average clustering accuracy [%]	94.36
Median clustering accuracy [%]	98.94
Percentage of correct selection of s_1 [%]	88
Percentage of correct selection of s_2 [%]	80
Percentage of correct selection of s_3 [%]	87
Percentage of correct selection of s_4 [%]	98

We further used the same example to compare our algorithm with the SON-EM method described in [13]. The results of this comparison are reported in Figure 7 in terms of fit rate index and clustering accuracy, showing that the overall performance of the estimated models is comparable. In this respect, observe that the SON-EM method does not perform the MSS task (the NARX structures are fixed to the correct form), as opposed to the proposed algorithm. It is worth remarking that the optimization problem solved by the SON-EM accounts for mode transitions through the regularization term[3] $\sum_{t=2}^{N} \|\hat{\boldsymbol{\vartheta}}(t) - \hat{\boldsymbol{\vartheta}}(t-1)\|$, where $\hat{\boldsymbol{\vartheta}}(t)$ is the estimate of the parameter vector of the mode active at time $t \in \mathbb{N}$. This term acts similarly to the mode transition cost $\mathcal{L}^{trans}$ in the proposed approach, thus leading to a fairly good clustering accuracy when exploiting SON-EM. This confirms the importance of accounting explicitly for the discrete dynamics when training the model, as we do through $\mathcal{L}^{trans}$.

Table 3: Simulation example 2: Parameter estimates: mean value and standard deviation. Aggregated results from the Monte Carlo analysis (computed on the training data).

Parameter	Mode			
	1	2	3	4
ϑ^1	1.4994 (0.0270)	-0.7024 (0.0507)	1.0028 (0.0225)	-0.5010 (0.0510)
ϑ^2	0.9907 (0.1045)	-0.6926 (0.0818)	0.9846 (0.1171)	-0.4954 (0.0667)
ϑ^3	0.4914 (0.0599)	-0.6971 (0.0317)	1.0015 (0.0234)	-0.4943 (0.0650)
ϑ^4	-0.5009 (0.0143)	-0.7012 (0.0131)	1.0009 (0.0244)	-0.5008 (0.0249)



(a) Algorithm 1

(b) Algorithm [13]

Figure 7: Simulation example 2: Comparative analysis with the SON-EM method [13]. Results obtained on the training data using one-step ahead prediction.

5.4. Experimental case study: segmenting the honeybee dance

The effectiveness of the proposed approach is further assessed on the same example considered in [10]. Given 6 honeybee dance sequences, each comprising the $2D$ coordinates of the bee's body (x_t, y_t) and its head angle θ_t, our goal is to segment them into the $K = 3$ modes characterizing the bee dance, namely "turn left", "turn right" and "waggle". Manually assigned labels are available for all six sets, that are only used as ground truth to assess the performance of the approach.

In this example, the MSS task is unnecessary, as the model structure is fixed and equal for each mode, in the form:

$$\hat{cos}(\theta_t) = \left[1 \;\; \cos(\theta_{t-1}) \;\; \sin(\theta_{t-1}) \;\; x_{t-1} \;\; y_{t-1}\right] \boldsymbol{\vartheta}_{\sigma_t} + e_t. \qquad (27)$$

However, the application of the proposed algorithm is still valuable for estimating the segmentation of the data in the three modes. In particular, we are here interested in the potential benefits introduced by the initialization strategy resulting from the application of the proposed SNARX identification approach, w.r.t. the plain application of the method of [4].

When running Algorithm 1, for each honeybee dance sequence the transition probability π in (22) is obtained from the true mode sequence $\boldsymbol{\sigma}$ as:

$$\pi = \frac{1}{N} \sum_{t=2}^{N} \mathbb{1}_{[\sigma_t \neq \sigma_{t-1}]}, \qquad (28)$$

while the weight τ in (22) is chosen by testing different values and selecting the maximizing C_N^{true} in (23), see Table 4.

[3] The regularization parameter is set to $\lambda = 1$ by cross-validation.

Table 4: Honeybee dance: optimal τ values.

Sequence	$\tau^\star$
1	0.110
2	0.057
3	0.075
4	0.080
5	0.027
6	0.020

The actual and estimated motion patterns for the six sequences are compared in Figure 8, where the segments associated with different modes are depicted with different colors. As confirmed by the results reported in Figure 9 and Table 5, better segmentation performance is obtained for the 4-th, 5-th and 6-th sequences, with the least accuracy achieved for the 2-nd subset. This result is due to the consistent variations of the head angle during waggle dances, that makes it more challenging to distinguish between different modes. Comparing the results with the approaches of [4]and [10], we observe a significant improvement in the segmentation performance (see Table 5). Indeed, the proposed approach achieves an average *clustering accuracy index* of 81.0%, against the 76.7% and 66.9% obtained with the other methods. This improvement is remarkable, in consideration of the fact that in this particular experiment the main difference between Algorithm 1 and Algorithm [4] lies in the choice of the initial mode sequences.

Table 5: Honeybee dance: data clustering accuracy C_N^{true} [%] .

	Sequence					
	1	2	3	4	5	6
Algorithm 1	79.6	69.2	75.3	88.4	87.5	86.0
Algorithm [4]	74.7	65.0	64.1	88.0	85.6	82.6
Algorithm [10]	46.5	44.1	45.6	83.2	93.2	88.7

6. Conclusions

A novel algorithm for the identification of general switched nonlinear models in a parametric setting has been discussed. The proposed approach blends the features of two different methods. In particular, it exploits the randomized scheme for the estimation of the discrete part of the switched model (sample-mode assignment and model structure selection) of [6] and the optimization approach of [4] that alternates between parameter estimation and sample-mode assignment. In doing so, the limitations of [6] regarding the solution of the sample-mode assignment task are removed, while the approach of [4] is extended with the capability of performing MSS. Furthermore, the issue of the sensitivity of the method of [4] to the initial conditions is greatly alleviated. The resulting algorithm is capable of solving challenging nonlinear switched model identification problems, as illustrated in the experimental section. In particular, it displays a remarkable accuracy both in associating the data to the modes and in the identification of the nonlinear models representing the modes. Different comparisons with state-of-the-art methods are also discussed, which emphasize the potential of the presented approach.

References

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

[2] Laurent Bako, Khaled Boukharouba, and Stéphane Lecoeuche. An l_0–l_1 norm based optimization procedure for the identification of switched nonlinear systems. In 49^{th} *IEEE Conference on Decision and Control*, pages 4467–4472, 2010.

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

[4] Alberto Bemporad, Valentina Breschi, Dario Piga, and Stephen P Boyd. Fitting jump models. *Automatica*, 96:11–21, 2018.

[5] D. P. Bertsekas. *Nonlinear programming*. Athena scientific Belmont, 1999.

[6] Federico Bianchi, Maria Prandini, and Luigi Piroddi. A randomized two-stage iterative method for switched nonlinear systems identification. *Nonlinear Analysis: Hybrid Systems*, 35:1–23, 2020.

[7] S. A. Billings. *Nonlinear System Identification: NARMAX Methods in the Time, Frequency, and Spatio-Temporal Domains*. Wiley, 2013.

[8] Valentina Breschi, Dario Piga, and Alberto Bemporad. Piecewise affine regression via recursive multiple least squares and multicategory discrimination. *Automatica*, 73:155–162, 2016.

[9] Giancarlo Ferrari-Trecate, Marco Muselli, Diego Liberati, and Manfred Morari. A clustering technique for the identification of piecewise affine systems. *Automatica*, 39(2):205–217, 2003.

[10] E. B. Fox, M. C. Hughes, E. B. Sudderth, and M. I. Jordan. Joint modeling of multiple time series via the beta process with application to motion capture segmentation. *The Annals of Applied Statistics*, 8(3):1281–1313, 2014.

[11] T. Fu, F. Chung, V. Ng, and R. Luk. Evolutionary segmentation of financial time series into subsequences. In *Proceedings of the 2001 Congress on Evolutionary Computation*, volume 1, pages 426–430, Seoul, South Korea, 2001.

[12] Andrea Garulli, Simone Paoletti, and Antonio Vicino. A survey on switched and piecewise affine system identification. In 16^{th} *IFAC Symposium on System Identification*, pages 344–355, Brussels, Belgium, July 11-13 2012.

[13] András Hartmann, João M Lemos, Rafael S Costa, João Xavier, and Susana Vinga. Identification of switched ARX models via convex optimization and expectation maximization. *Journal of Process Control*, 28:9–16, 2015.

[14] Aleksandar Lj Juloski, Siep Weiland, and WPMH Heemels. A Bayesian approach to identification of hybrid systems. *IEEE Transactions on Automatic Control*, 50(10):1520–1533, 2005.

[15] Fabien Lauer and Gérard Bloch. Switched and piecewise nonlinear hybrid system identification. In *International Workshop on Hybrid Systems: Computation and Control*, pages 330–343, 2008.

[16] Fabien Lauer and Gérard Bloch. Piecewise smooth system identification in reproducing kernel hilbert space. In *53rd IEEE Conference on Decision and Control*, pages 6498–6503, 2014.

[17] Fabien Lauer and Gérard Bloch. *Hybrid System Identification*. 2019.

[18] Fabien Lauer, Gérard Bloch, and René Vidal. Nonlinear hybrid system identification with kernel models. In 49^{th} *IEEE Conference on Decision and Control*, pages 696–701, 2010.

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

[20] Van Luong Le, Gérard Bloch, and Fabien Lauer. Reduced-size kernel models for nonlinear hybrid system identification. *IEEE Transactions on Neural Networks*, 22(12):2398–2405, 2011.

[21] Van Luong Le, Fabien Lauer, Laurent Bako, and Gérard Bloch. Learning nonlinear hybrid systems: from sparse optimization to support vector regression. In *Proceedings of the* 16^{th} *International Conference on Hybrid systems: computation and control*, pages 33–42, 2013.

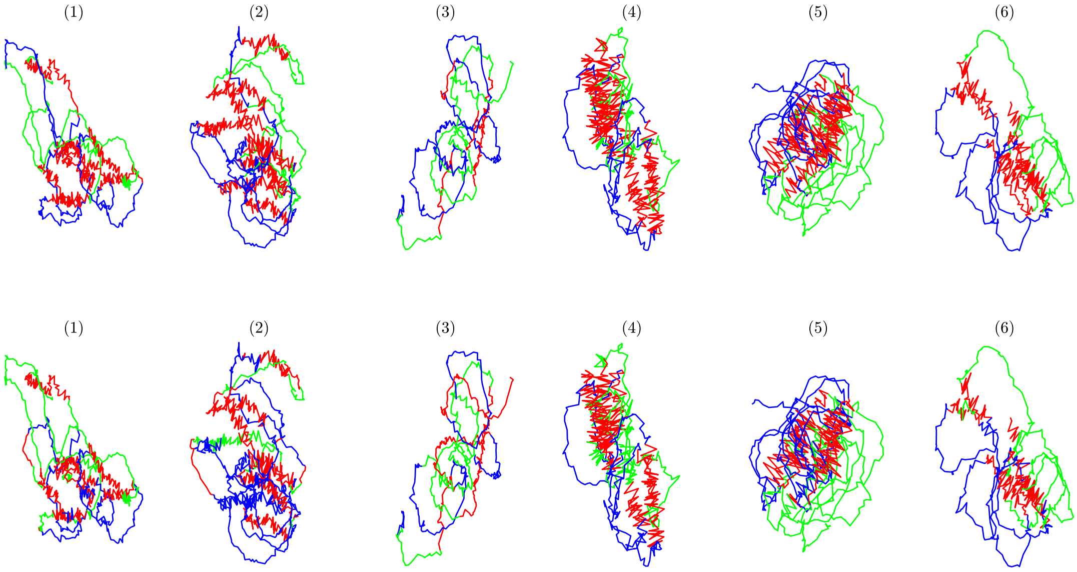


Figure 8: Honeybee dance: true (top) vs estimated (bottom) trajectories segmentation for sequences 1 to 6, with “turn right” motion (blue), “waggle” dance (red) and “turn left” motion (green).

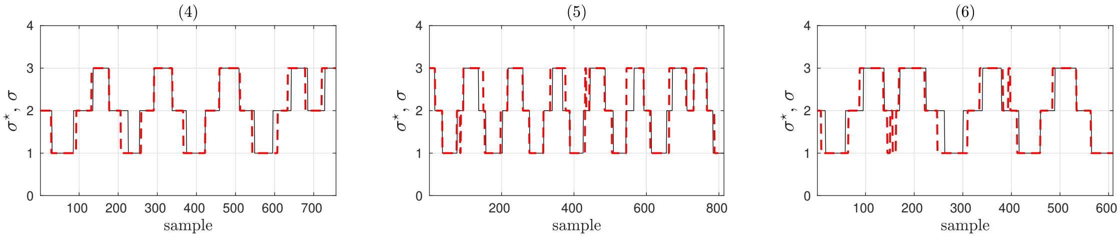


Figure 9: Honeybee dance: true (black solid) vs estimated (red dashed) *mode sequence* for sequences 4 to 6.

[22] I.J. Leontaritis and S.A. Billings. Input-output parametric models for non-linear systems part I: deterministic non-linear systems. *International Journal of Control*, 41(2):303–328, 1985.

[23] I.J. Leontaritis and S.A. Billings. Input-output parametric models for non-linear systems part II: stochastic non-linear systems. *International Journal of Control*, 41(2):329–344, 1985.

[24] Yi Ma and René Vidal. Identification of deterministic switched ARX systems via identification of algebraic varieties. In *International Workshop on Hybrid Systems: Computation and Control*, pages 449–465, 2005.

[25] Ichiro Maruta, Toshiharu Sugie, and Tae-Hyoung Kim. Identification of multiple mode models via distributed particle swarm optimization. In *Proceedings of the* 18th *IFAC World Congress*, pages 7743–7748, Milano, Italy, Aug. 28 – Sept. 2 2011.

[26] Sohail Nazari, Qing Zhao, and Biao Huang. An improved algebraic geometric solution to the identification of switched ARX models with noise. In *Proceedings of the American Control Conference*, pages 1230–1235, 2011.

[27] N. Nguyen. Hidden Markov Model for Stock Trading. *International Journal of Financial Studies*, 6(2):36, 2018.

[28] Sang Min Oh, James M. Rehg, Tucker Balch, and Frank Dellaert. Learning and inferring motion patterns using parametric segmental switching linear dynamic systems. *International Journal of Computer Vision*, 77 (1–3):103–124, 2008.

[29] Henrik Ohlsson and Lennart Ljung. Identification of switched linear regression models using sum-of-norms regularization. *Automatica*, 49(4): 1045–1050, 2013.

[30] M. Ostendorf, V. V. Digalakis, and O. A. Kimball. From HMM's to segment models: A unified view of stochastic modeling for speech recognition. *IEEE Transactions on speech and audio processing*, 4(5):360–378, 1996.

[31] N. Ozay, C. Lagoa, and M. Sznaier. Set membership identification of switched linear systems with known number of subsystems. *Automatica*, 51:180–191, 2015.

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

[33] Simone Paoletti, Aleksandar Lj Juloski, Giancarlo Ferrari-Trecate, and René Vidal. Identification of hybrid systems a tutorial. *European Journal of Control*, 13(2–3):242–260, 2007.

[34] Dario Piga, Alberto Bemporad, and Alessio Benavoli. Rao-Blackwellized sampling for batch and recursive Bayesian inference of Piecewise Affine models. *Automatica*, 117, 2020.

[35] Gianluigi Pillonetto. A new kernel-based approach to hybrid system identification. *Automatica*, 70:21–31, 2016.

[36] L. R. Rabiner. Readings in speech recognition. chapter A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, pages 267–296. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1990.

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

[38] J. Speyer, J. Deyst, and D. Jacobson. Optimization of stochastic linear systems with additive measurement and process noise using exponential performance criteria. *IEEE Transactions on Automatic Control*, 19(4): 358–366, 1974.