A constrained clustering approach to bounded-error identification of switched and piecewise affine systems

Federico Bianchi[a,*], Alessandro Falsone[a], Luigi Piroddi[a], Maria Prandini[a]

[a] *Politecnico di Milano, 20133 Milano (Italy) – (e-mail: name.surname@polimi.it)*

Abstract

This paper proposes a novel clustering-based approach to the bounded-error identification of switched and piecewise affine autoregressive exogenous systems. We address the problem of determining a minimal collection of linear-in-the-parameters models (called modes) fitting with a given accuracy ε a set of input-output data while complying with the switched or piecewise affine nature of the system. The problem is tackled by suitably clustering the data according to their preferences with respect to a pool of candidate models identified on subsets of the available data. The preference of a data point for a model is assessed based on the extent to which that model fits that data point and is set to zero if the fit is worse than ε. A two-level clustering with outliers isolation is employed, first grouping data based on their preferences subject to suitable time/space adjacency conditions depending on the nature of the switching mechanism, and then collecting together non-adjacent clusters that can be described by the same mode. The performance of the proposed method is demonstrated via comparative numerical examples and on experimental data from an electronic component placement process in a pick-and-place machine.

Keywords: Bounded-error identification; Clustering; Switched affine models; Piecewise affine models.

1. Introduction

This paper addresses the identification of switched systems expressed in input-output form and characterized by a set of affine AutoRegressive with eXogenous inputs (ARX) systems, called modes. Specifically, we consider the classes of Switched (SARX) and PieceWise affine ARX (PWARX) systems, depending on the nature of the switching mechanism. An exogenous switching signal, possibly subject to some dwell-time condition, is assumed for the first class, whereas in PWARX systems switching is triggered by transitions of the regressor vector from one region to another of a polyhedral partition in the regressor space. The switching signal is not known in advance so that one needs to infer from the available input/output data not only the model parameters of each mode, but also the number of modes and the sample-mode assignment.

The identification problem can be formulated as a mixed integer optimization program with continuous (the model parameters) and discrete (the number of modes and the sample-mode assignment) variables, which is NP-hard [1], and hence challenging to solve in practice. Various approaches have been proposed in the literature to address the identification of switched systems, see *e.g.,* the surveys [12, 15] and the papers referenced in [6, 7] for a comprehensive review. We next briefly review the methods to which the present approach is closer to. The k-RANSAC algorithm [13] is a greedy approach based on a repeated application of the scheme in [11]: at each iteration a new model is identified that fits most of the remaining data, until all data are covered. This method copes well with outliers, but is negatively affected by the incremental nature of the procedure. In [10, 9] data are clustered based on the assumption that samples well fitted by models with similar parameterizations are likely associated to the same mode. A similar assumption is exploited in [8], where suitable probability distributions are introduced to model the influence of each datum on the cluster membership of its neighbors, and in [19, 17] to construct the weight matrix appearing in the regularization term. In [3] a bounded-error condition is used to define a set of linear inequalities derived from the data, which are then partitioned in subsets, each one leading to a different subsystem. A greedy partitioning scheme is adopted, whereby at each iteration a model is obtained that satisfies the maximum number of remaining inequalities. Differently from the k-RANSAC method, a refinement step is applied to *a posteriori* correct the sample-mode assignment. Even so, the presence of noise in the data and outliers may cause misclassifications. Notably, this method does not require prior information on the number of modes, as this follows indirectly from imposing a threshold on the error. However, various parameters need to be appropriately tuned in order to correctly estimate the number of modes. Exploiting prior knowledge on the switching mechanism can greatly enhance the ability of the identification method to handle

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

noise and outliers, thereby improving the sample classification accuracy. For example, [4] adds a term in the cost function of the optimization problem, that suitably penalizes mode switchings. The method of [5], instead, adopts an heuristic post-processing of the sample assignment that exploits information such as the minimum dwell time.

In this work, we present a novel clustering-based method for the identification of SARX and PWARX systems that builds upon the bounded-error approach of [3] and extends to the class of switched dynamical systems the clustering method described in [21, 18], originally introduced for fitting multiple instances of a static model to noisy data in computer vision applications. Differently from [10, 9], the clustering is not carried out in the regressor domain, which relies on the similarity of the parameters of the local models and is consequently badly affected by noise in the data. Data are instead clustered in groups sharing similar "preferences" with respect to a set of models, implying that data sharing similar preferences are likely to be explained by the same mode. Furthermore, along with [4] and [5], we exploit prior knowledge on the switching mechanism to improve the mode assignment, integrating spatial (for PWARX) and temporal (for SARX) adjacency relations between data points in the data clustering mechanism. Key to the effectiveness of the approach is the introduction of a graph encoding these adjacency relations.

The proposed method operates two clustering steps, the first one grouping adjacent samples that are compatible with the same model, and the second one further collecting together non-adjacent data segments that can be assigned to the same model. These two steps are iterated until convergence, which is guaranteed to occur in a finite number of iterations. When no bound on the additive noise is known, different ε are explored for a fine tuning.

2. Problem formulation

We consider a single-input single-output switched system with input u_t and output y_t, characterized by s° affine dynamics (called modes) and a signal $\sigma_t \in \mathcal{M} = \{1, \dots, s^\circ\}$ governing the switching among them. More precisely, the system is described by

$$y_t = \boldsymbol{\varphi}_t^\top \boldsymbol{\vartheta}_{\sigma_t}^\circ + e_t, \tag{1}$$

where $\boldsymbol{\varphi}_t = [1 \ \boldsymbol{x}_t]^\top \in \mathbb{R}^n$, with $\boldsymbol{x}_t = [y_{t-1} \ \cdots \ y_{t-n_y} \ u_{t-1} \ \cdots \ u_{t-n_u}]$ and $n = n_u + n_y + 1$, is the (extended) regression vector, $\boldsymbol{\vartheta}_{\sigma_t}^\circ \in \mathbb{R}^n$ is the parameter vector associated with the mode σ_t that is active at time t, and e_t is a disturbance input. In this work, we assume that

Assumption 1. *The orders n_y and n_u of the switched system* (1) *are known.* □

If σ_t is an exogenous signal, then (1) is referred to as a SARX system. If, instead, σ_t is an endogenous signal whose value depends on which element of a polyhedral partition of $\mathbb{R}^{n_u+n_y}$ $\boldsymbol{x}_t$ belongs to, then (1) is a PWARX system. For PWARX systems, samples that are close in the regressor space are likely to belong to the same mode. For SARX systems, this is the case for samples that are consecutive in time. This information can be exploited to constrain the identification procedure.

Given a set $\mathcal{D} = \{(y_t, \boldsymbol{x}_t)\}_{t=1}^N$ of time-ordered consecutive data collected from (1), we consider the optimization problem

$$\begin{aligned} \min_{s\in\mathbb{N}, \{\boldsymbol{\vartheta}_i \in \mathbb{R}^n\}_{i=1}^s} \quad & s \\ \text{subject to:} \quad & \max_{t=1,\dots,N} \min_{i=1,\dots,s} |y_t - \boldsymbol{\varphi}_t^\top \boldsymbol{\vartheta}_i| \leq \varepsilon, \end{aligned} \tag{2}$$

where $\varepsilon > 0$ is a bound on the disturbance amplitude. Problem (2) entails finding a minimal collection of modes

$$\hat{y}_t = \boldsymbol{\varphi}_t^\top \boldsymbol{\vartheta}_i, \ i = 1, \dots, s, \tag{3}$$

fitting $\mathcal{D}$ with a Maximum Absolute Error (MAE) that is below the threshold ε. The optimal solution is then characterized by a minimal number $s^\star$ of modes and the corresponding parameterizations $\{\boldsymbol{\vartheta}_i^\star\}_{i=1}^{s^\star}$, identifying the system that generated the data in $\mathcal{D}$. The associated switching signal can be reconstructed by selecting

$$\sigma_t^\star \in \{i : \ |y_t - \boldsymbol{\vartheta}_i^{\star\top} \boldsymbol{\varphi}_t| \leq \varepsilon\} \subseteq \{1, \dots, s^\star\}, \ t = 1, \dots, N. \tag{4}$$

Such a signal has to comply with the underlying SARX/PWARX nature of the system. In particular, in the PWARX case, the partition of $\mathbb{R}^{n_u+n_y}$ originating the switching signal can be further identified by solving a supervised classification problem with $s^\star$ classes on the data set $\{(\boldsymbol{x}_t, \sigma_t^\star)\}_{t=1}^N$.

Problem (2) is hard to solve in practice since it is equivalent to the NP-hard minimum-size partition of feasible subsystems problem, [1]. Here, we propose a novel constrained clustering method to approximately solve it.

3. The proposed constrained clustering method

The proposed approach is inspired by [18] and consists in performing data clustering based on the closeness of data in a suitably defined *preference space*, built using the preferences expressed by each datum for all models in a pool constructed from small sets of data. Intuitively, data that share similar preferences are likely to be well explained by the same model and can then be grouped together.

Prior knowledge on the SARX or PWARX model class to which the underlying system belongs is integrated within the construction of the initial pool of local models and the clustering procedure. This is achieved by describing the temporal (SARX) or spatial (PWARX) relation between data through an undirected graph, on which clusters of coherent data can be easily identified.

The clustering process is separated into two stages, the first one aiming to find the largest temporally or spatially

coherent subsets using the introduced graph, and the second one to find the minimum number of modes, by aggregating non-adjacent clusters with the same dynamics. This two-stage procedure is also beneficial in increasing the accuracy of the model identification, as it enlarges the set of data on which each individual mode is estimated.

By iterating this clustering process using the identified modes as model pool for the next iteration, we allow the algorithm to further reduce the number of modes and refine the associated models. The properties of the resulting iterative procedure are discussed in Section 3.6.

3.1. Graph encoding prior knowledge on the model class

Depending on the nature of the switched system, samples that are temporally (SARX) or spatially (PWARX) close are likely to belong to the same mode. In order to make our approach exploit this a-priori information, we associate an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ to the data set $\mathcal{D} = \{(y_t, \boldsymbol{x}_t)\}_{t=1}^N$ so that two data that are close are mapped into two adjacent vertices of the graph, *i.e.*, two vertices that belong to an edge in $\mathcal{E}$. More precisely, each datum is coded through its time index t and is represented by a vertex in the set $\mathcal{V} = \{1, \dots, N\}$. As for the edge set, if $\mathcal{D}$ is generated by a SARX system, then the unordered pair $\{t_1, t_2\}$ is an edge in $\mathcal{E}$ if and only if $t_2 = t_1 + 1$. Instead, if it is generated by a PWARX system, then $\{t_1, t_2\} \in \mathcal{E}$ if and only if $\boldsymbol{x}_{t_1}$ and $\boldsymbol{x}_{t_2}$ are connected by an edge in the Delaunay triangulation [16] of $\{\boldsymbol{x}_t\}_{t=1}^N$ within $\mathbb{R}^{n_u+n_y}$. Indeed, since in the Delaunay triangulation each $\boldsymbol{x}_t$ is joined to neighboring points in the regression space, we can use it to model spatial relations among $\{\boldsymbol{x}_t\}_{t=1}^N$. Alternatively, one can employ the K-nearest neighbor graph, which is simpler to build, but requires to specify the number of neighbors K of a datum as a design parameter.

Note that, since data in $\mathcal{D}$ are assumed to be consecutive in time, then, in both the SARX and PWARX cases the resulting graph $\mathcal{G}$ is connected by construction, *i.e.,* for any two vertices v_0 and v_n, there exists a sequence of vertices $v_1, \dots, v_{n-1}$, such that $\{v_{k-1}, v_k\} \in \mathcal{E}$, $k = 1, \dots, n$. Given a subset of vertices $\mathcal{V}' \subset \mathcal{V}$, the subgraph $\mathcal{G}_{\mathcal{V}'} = (\mathcal{V}', \mathcal{E}')$ is the graph containing the vertices in $\mathcal{V}'$ and the edges connecting only the vertices in $\mathcal{V}'$ (*i.e.*, $\mathcal{E}' = \mathcal{E} \cap \{\{v_1, v_2\}, v_1, v_2 \in \mathcal{V}'\}$). Let $\mathcal{G}_1 = (\mathcal{V}_1, \mathcal{E}_1)$ and $\mathcal{G}_2 = (\mathcal{V}_2, \mathcal{E}_2)$ be two connected subgraphs of $\mathcal{G}$. Then, the subgraph of $\mathcal{G}$ with $\mathcal{V}_1 \cup \mathcal{V}_2$ as set of vertices is connected only if $\mathcal{G}_1$ and $\mathcal{G}_2$ are adjacent, *i.e.*, there exist $v_1 \in \mathcal{V}_1$ and $v_2 \in \mathcal{V}_2$ such that v_1 and v_2 are adjacent.

3.2. Initial model pool and preference space generation

The method starts with the selection of sets containing a given (small) number α of data points that are close to each other in time (for SARX) or in space (for PWARX), which are then used to estimate the local models. These sets are built using the introduced graph representation $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ of the dataset $\mathcal{D}$ and constructing $N_p \leq N$ connected subgraphs $\tilde{\mathcal{G}}_j = (\tilde{\mathcal{V}}_j, \tilde{\mathcal{E}}_j)$, $j = 1, \dots, N_p$, each one with α vertices.

More precisely, the N_p subgraphs are built by choosing N_p vertices $\{v_1, v_2, \dots, v_{N_p}\} \subseteq \mathcal{V}$ and performing for each $j = 1, 2, \dots, N_p$ a breadth-first search on $\mathcal{G}$ starting from v_j until $\alpha - 1$ vertices are discovered. $\tilde{\mathcal{V}}_j$ is then given by the set of these $\alpha - 1$ vertices and v_j, whereas $\tilde{\mathcal{E}}_j = \{\{v_1, v_2\} \in \mathcal{E} : v_1, v_2 \in \tilde{\mathcal{V}}_j\}$.

In order to guarantee a coverage of the available data, one can set $N = N_p$, so that a subgraph for each vertex in $\mathcal{V}$ (and hence for each datum in $\mathcal{D}$) is obtained. If N is too large, one can still obtain a wide data coverage by choosing the vertices in $\mathcal{V}$ corresponding to $N_p < N$ data that are equally spaced in time (for SARX models) or obtained by uniformly gridding the convex hull containing all available regressors $\boldsymbol{x}_t$'s and then picking the data associated with N_p regressors that are closest to the grid points (for PWARX models).

For each subgraph $\tilde{\mathcal{G}}_j = (\tilde{\mathcal{V}}_j, \tilde{\mathcal{E}}_j)$, $j = 1, \dots, N_p$, a model $\tilde{\boldsymbol{\vartheta}}_j \in \mathbb{R}^n$ is identified by minimizing the MAE, *i.e.*,

$$\tilde{\boldsymbol{\vartheta}}_j \in \arg \min_{\boldsymbol{\vartheta}_j \in \mathbb{R}^n} \max_{t \in \tilde{\mathcal{V}}_j} |y_t - \boldsymbol{\varphi}_t^\top \boldsymbol{\vartheta}_j|, \tag{5}$$

which amounts to solving the linear programming problem

$$\begin{aligned} \min_{\boldsymbol{\vartheta}, h} \quad & h \\ \text{subject to:} \quad & y_t - \boldsymbol{\varphi}_t^\top \boldsymbol{\vartheta} \leq h \quad \forall t \in \tilde{\mathcal{V}}_j \\ & \boldsymbol{\varphi}_t^\top \boldsymbol{\vartheta} - y_t \leq h \quad \forall t \in \tilde{\mathcal{V}}_j. \end{aligned} \tag{6}$$

Clearly, a necessary condition for the solution to (6) to be uniquely defined, is $\alpha \geq n$, where n is the size of $\boldsymbol{\vartheta}$.

Once the model pool has been generated, following [18], one can assess the preference of the t-th datum for the j-th model as

$$p_{tj} = \begin{cases} 0 & r_{tj} > \varepsilon \\ e^{-5\frac{r_{tj}}{\varepsilon}} & r_{tj} \leq \varepsilon \end{cases} \tag{7}$$

where $r_{tj} = |y_t - \boldsymbol{\varphi}_t^\top \tilde{\boldsymbol{\vartheta}}_j|$ is the absolute value of the residual and ε the user-given threshold in (2). Note that $(y_t, \boldsymbol{x}_t)$ has a strictly positive preference for model j if and only if model j fits $(y_t, \boldsymbol{x}_t)$ with an absolute error within ε. Expression (7) provides a normalized preference value between 0 and 1 and the chosen coefficient 5 allows to fully exploit the variability range of the exponential for values of r_{tj} within $[0, \varepsilon]$. Notice that this exponential preference function is only instrumental to the clustering procedure, and it is not meant to provide any ranking between local models. Indeed, all the local models with $r_{tj} \leq \varepsilon$ are equally good according to the problem formulation (2).

A data point $(y_t, \boldsymbol{x}_t)$ can then be represented by its *preference vector* $\boldsymbol{p}_t = [p_{t1} \cdots p_{tN_p}]^\top$ in the preference space $[0, 1]^{N_p}$. Intuitively, if the model pool is rich enough, data points generated by the same mode will have a similar preference vector and one can leverage this information to cluster these points together.

If the model pool is sufficiently rich and the disturbance absolute value is upper bounded by ε, then $\boldsymbol{p}_t \neq \mathbf{0}$, $\forall t$. If this were not the case, then all data such that $\boldsymbol{p}_t = \mathbf{0}$ (*i.e.*, no model fits them within an ε error) are considered as outliers and left aside in a set O. One then needs to define the reduced dataset $\mathcal{D}_\varepsilon = \mathcal{D} \setminus O$ and construct its (not necessarily connected) graph representation $\mathcal{G}_\varepsilon$ to feed into the two-stage constrained clustering.

3.3. Two-stage constrained clustering

To measure the likelihood that two data points have been generated by the same mode, following [18] we consider their preference vectors and evaluate the complement of the Tanimoto similarity coefficient, which is defined for every $a, b \in [0,1]^{N_p}$ that are not both equal to zero as

$$S_T^\circ(a,b) = 1 - \frac{a^\top b}{\|a\|^2 + \|b\|^2 - a^\top b}, \tag{8}$$

where $\|\cdot\|$ denotes the standard Euclidean norm, [22]. $S_T^\circ(a,b)$ ranges from 0 ($a = b$) to 1 ($a^\top b = 0$). According to (7) and (8), $S_T^\circ(\boldsymbol{p}_t, \boldsymbol{p}_{t'}) < 1$ implies that there exists j such that p_{tj} and $p_{t'j}$ are both positive. Note that removing the data with $\boldsymbol{p}_t = \mathbf{0}$ prevents the case $a = b = \mathbf{0}$, for which (8) is not defined.

Resting on (8) one could perform data clustering to partition the dataset into groups of data with similar preferences, along the lines of [18]. However, the resulting classification performance is not entirely satisfactory, especially regarding the "undecidable" points, *i.e.* those samples that (*e.g.* due to noise) exhibit a small residual (and thus a positive preference) for more modes than just the correct one. As a result, the identified switched model switches too frequently and displays many isolated samples (see also the illustrative example in Section 3.4).

Intuitively, as also acknowledged in [3], the ambiguity of these points may be solved by exploiting temporal/spatial localization information of these points. However, differently from [3], where this information is used only *a posteriori* with the aim of reducing misclassifications in the already computed data partition, we here propose to use it to explicitly assist the clustering procedure. More precisely, we first apply a "low-level clustering" phase, that aims to group together samples that are both temporally or spatially coherent, and share common preference vectors (*i.e.*, are well described by the same model). Then, after a pruning phase, a second "high-level clustering" is operated, to aggregate non-adjacent data clusters obtained in the first stage that can be described by the same model.

The low-level clustering phase (Algorithm 1) is initialized assigning each data point index in the set $\mathcal{I} \subseteq \{1, \dots, N\}$ associated with $\mathcal{D}_\varepsilon$ to a different cluster $\mathcal{V}_i$, which inherits the preference vector of the corresponding data point. Then, the procedure looks for the closest pair of *adjacent* clusters according to (8), adjacency being assessed with reference to graph $\mathcal{G}_\varepsilon$, and if they share a positive preference it merges them into a new cluster with

Algorithm 1 Low-level clustering

Require: Data indices $\mathcal{I}_\varepsilon = \{t : (y_t, \boldsymbol{x}_t) \in \mathcal{D}_\varepsilon\}$, preference vectors $\{\boldsymbol{p}_t\}_{t \in \mathcal{I}_\varepsilon}$, graph $\mathcal{G}_\varepsilon$

Ensure: Data partition $\{\mathcal{V}_i\}$, preference vectors $\{\boldsymbol{\rho}_i\}$

$\mathcal{V}_t = \{t\}$, $t \in \mathcal{I}_\varepsilon$ ▷ Initialize clusters
$\boldsymbol{\rho}_t = \boldsymbol{p}_t$, $t \in \mathcal{I}_\varepsilon$ ▷ Initialize cluster preference
repeat
  $(t^\star, \tau^\star) \leftarrow \underset{t,\tau \in \mathcal{I}_\varepsilon, t \neq \tau}{\arg\min} \; S_T^\circ(\boldsymbol{\rho}_t, \boldsymbol{\rho}_\tau)$ ▷ Find closest pair
    subject to: $\mathcal{V}_t$ and $\mathcal{V}_\tau$ adjacent in $\mathcal{G}_\varepsilon$
  if $S_T^\circ(\boldsymbol{\rho}_{t^\star}, \boldsymbol{\rho}_{\tau^\star}) < 1$ **then**
    $\mathcal{V}_{t^\star} \leftarrow \mathcal{V}_{t^\star} \cup \mathcal{V}_{\tau^\star}$ ▷ Merge cluster pair
    $\boldsymbol{\rho}_{t^\star} \leftarrow \min\{\boldsymbol{\rho}_{t^\star}, \boldsymbol{\rho}_{\tau^\star}\}$ ▷ Update preference set
    $\mathcal{I}_\varepsilon \leftarrow \mathcal{I}_\varepsilon \setminus \{\tau^\star\}$
  end if
until $S_T^\circ(\boldsymbol{\rho}_{t^\star}, \boldsymbol{\rho}_{\tau^\star}) = 1$

preference vector equal to the component-wise minimum among the preference vectors of the merged clusters. The procedure is iterated while there are at least two adjacent clusters sharing a positive preference, otherwise it stops. Intuitively, if an undecidable point has a positive preference for both modes j and i, but it is adjacent to data points preferring i, then it is more likely to be clustered with them despite its (possibly higher) preference for j.

Following the first clustering stage, we perform a pruning phase, in which we start from the partition $\{\mathcal{V}'_i, \boldsymbol{\rho}'_i\}_{i=1}^{s'}$ resulting from the low-level clustering and look for all the clusters $\mathcal{V}'_i$ with cardinality less than a user-defined threshold $\beta > 0$, discarding their data as outliers and transferring them to the set O. The rationale underlying this pruning step is that a few isolated points calling for an additional cluster are likely to be outliers. The value of β can be inferred from prior knowledge on the underlying system, if available. In the SARX case, for example, β can be set equal to the minimum dwell time, if this is known. In both the SARX and PWARX cases, $\beta \geq n$, since problem (6) has multiple solutions and, hence, model parameters are not uniquely defined for a cluster with less than n data. After pruning, up to a re-indexing of the clusters, one obtains $\{\mathcal{V}''_i, \boldsymbol{\rho}''_i\}_{i=1}^{s''}$, with $s'' \leq s'$.

Finally, in the second high-level clustering stage we run Algorithm 2, using as input the indices $\mathcal{I} = \{1, \dots, s''\}$ of the clusters returned by the pruning phase and $\boldsymbol{p}_t = \boldsymbol{\rho}''_t$, $t \in \mathcal{I}$, for their preference vectors. The clustering criterion is the same as before, except that the adjacency constraint has been removed. Algorithm 2 aggregates those data portions $\{\mathcal{V}''_i\}_{i=1}^{s''}$ that are not adjacent. This allows to associate to the same mode data that are apart (either in time or in space), allowing at the same time a more accurate estimation of the model parameters (since more data can be used for identification purposes). The outcome $\{\mathcal{V}_i, \boldsymbol{\rho}_i\}_{i=1}^{\hat{s}}$ of the high-level clustering phase provides both the estimated data partition and the estimated number of modes $\hat{s}$. Notice that, for PWARX systems, one can detect if a mode has been associated to different non-adjacent re-

Algorithm 2 High-level clustering

Require: Indices set $\mathcal{I}$, preference vectors $\{\boldsymbol{p}_t\}_{t\in\mathcal{I}}$
Ensure: Data partition $\{\mathcal{V}_i\}$, preference vectors $\{\boldsymbol{\rho}_i\}$

$\mathcal{V}_t = \{t\}$, $t \in \mathcal{I}$ ▷ Initialize clusters
$\boldsymbol{\rho}_t = \boldsymbol{p}_t$, $t \in \mathcal{I}$ ▷ Initialize cluster preference
repeat
  $(t^\star, \tau^\star) \leftarrow \arg\min_{t,\tau\in\mathcal{I}, t\neq\tau} S_T^\circ(\boldsymbol{\rho}_t, \boldsymbol{\rho}_\tau)$ ▷ Find closest pair
  if $S_T^\circ(\boldsymbol{\rho}_{t^\star}, \boldsymbol{\rho}_{\tau^\star}) < 1$ **then**
    $\mathcal{V}_{t^\star} \leftarrow \mathcal{V}_{t^\star} \cup \mathcal{V}_{\tau^\star}$ ▷ Merge cluster pair
    $\boldsymbol{\rho}_{t^\star} \leftarrow \min\{\boldsymbol{\rho}_{t^\star}, \boldsymbol{\rho}_{\tau^\star}\}$ ▷ Update preference set
    $\mathcal{I} \leftarrow \mathcal{I} \setminus \{\tau^\star\}$
  end if
until $S_T^\circ(\boldsymbol{\rho}_{t^\star}, \boldsymbol{\rho}_{\tau^\star}) = 1$

gions of $\mathbb{R}^{n_u+n_y}$ by inspecting the graph associated with that mode and checking whether or not it is connected.

Given the estimated data partition $\{\mathcal{V}_i\}_{i=1}^{\hat{s}}$, we can trivially reconstruct the switching signal as

$$\hat{\sigma}_t = i, \ t \in \mathcal{V}_i, t \in \{1, \dots, N\}, t \notin \mathcal{I}_O, \tag{9}$$

where $\mathcal{I}_O$ is the set of time indices associated with the set O of outliers.

The partition returned by the explained two-stage clustering scheme satisfies the following properties [18]:

P1: all data belonging to a cluster share at least a positive preference for some model, *i.e.*, for each cluster $\mathcal{V}_i$ there exists some model j such that $0 < \rho_{ij} = \min_{t\in\mathcal{V}_i} p_{tj}$;

P2: two distinct clusters do not share a positive preference for any model, *i.e.*, their preference vectors are orthogonal: $\boldsymbol{\rho}_i^\top \boldsymbol{\rho}_{i'} = 0$ for all $i \neq i'$.

Remark 1. *Notice that the low-level clustering phase only satisfies property P1, as clusters sharing a similar preference vector are merged together only if they are adjacent.*

Remark 2. *Alternative approaches to data-segmentation could be adopted in place of Algorithm 1, e.g., [20]. Note, however, that in our approach data-segmentation is functional to the next clustering stage where data corresponding to different segments that are generated by the same mode are aggregated in the same cluster.*

3.4. Illustrative example

The next example illustrates the importance of introducing adjacency relations in the clustering process. Consider the following SARX system [2]

$$\begin{aligned}
\text{mode 1: } y_t &= -0.0322y_{t-1} + 0.8017y_{t-2} \\
&\quad - 1.2878u_{t-1} - 1.1252u_{t-2} + e_t \\
\text{mode 2: } y_t &= -0.1921y_{t-1} + 0.5917y_{t-2} \\
&\quad + 1.1050u_{t-1} + 0.0316u_{t-2} + e_t \\
\text{mode 3: } y_t &= +1.4746y_{t-1} - 0.5286y_{t-2} \\
&\quad - 0.4055u_{t-1} + 0.2547u_{t-2} + e_t
\end{aligned} \tag{10}$$

where e_t and u_t are zero mean white Gaussian processes with variance $\zeta^2 = 0.01$ and 1, respectively. The system switches randomly 20 times in the interval $(1, 2000)$, with a minimum dwell time τ_D of 10 time instants.

Suppose that the true model parameters $\boldsymbol{\vartheta}_i^\circ$, $i = 1, 2, 3$, are accessible. Then, if one tries to associate samples to modes using the intuitive heuristic

$$\hat{\sigma}_t = \arg\min_k |y_t - \boldsymbol{\varphi}_t^\top \boldsymbol{\vartheta}_k^\circ|, \tag{11}$$

6% of the samples (a subset of the undecidable ones) are misclassified. This result is only marginally improved if one applies the preference-based clustering *without* the adjacency constraint (*i.e.*, using [18]), even if the preferences are computed using the true models. Indeed, with $\varepsilon = 4\zeta = 0.4$, one obtains a misclassification rate of 5.6%. Figure 1 provides a pictorial representation of the unsatisfactory mode classification in both cases. Note that the misclassified points are not always located at the boundaries of data segments associated with different modes.

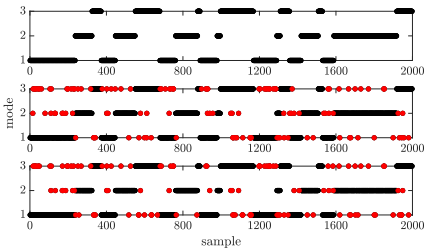

Figure 1: True switching signal (top), switching signal estimated according to rule (11) (middle), and switching signal estimated with the preference-based clustering approach without the adjacency constraint (bottom). Red dots represent misclassified data points.

Using Algorithm 1 instead, one obtains the clustering reported in Figure 2 (middle plot). As can be seen from the picture, consecutive data points generated by the same mode are correctly identified, except for two (close) misclassified points. Notice that the algorithm has correctly estimated the number of switchings and their locations as well, except for one case. This example provides also evidence for the observation in Remark 1. Indeed, inspecting Figure 3, where each column represents the preference of a cluster for the model indexed by the row number (dark color reads as high preference), it is easy to see that there are non-adjacent clusters (columns) with similar preferences. For comparison purposes, we also applied the data-segmentation scheme in [20] and obtained a comparable data segmentation result but with a much (almost 1000 times) higher computational effort.

Performing the pruning procedure with $\beta = 10$ (equal to the minimum dwell time) on the outcome of Algorithm 1 results in no outliers being detected. If one then applies Algorithm 2, the switching signal $\hat{\sigma}_t$ shown in Figure 2 (bottom plot) is obtained. Apparently, the estimated switching signal almost perfectly matches the true one (0.1% misclassification rate).

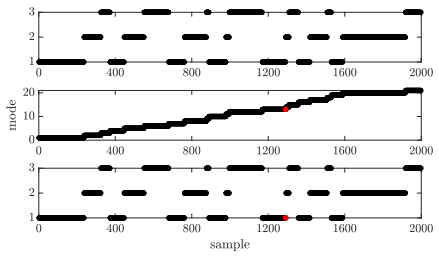

Figure 2: True switching signal (top), output of the low-level clustering (middle), and switching signal estimated after the high-level clustering (bottom). Red dots represent misclassified data points.

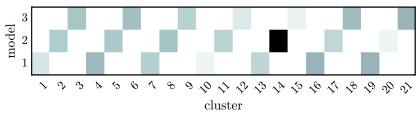

Figure 3: Preference matrix. Each column is a cluster and each row is a model. The preference of a cluster for each model is represented as shades of color (dark color reads as high preference).

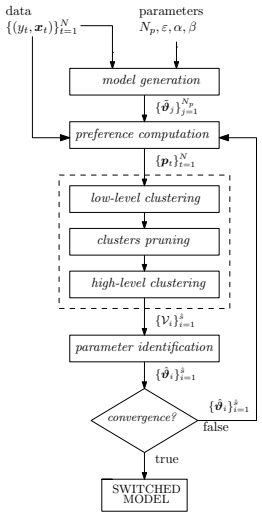

Figure 4: Flow chart of the proposed algorithm. Notice that while iterating, the identified modes $\{\hat{\boldsymbol{\vartheta}}_i\}_{i=1}^{\hat{s}}$ are used as model pool.

3.5. Procedure iteration and stopping criteria

The estimated data partition $\{\mathcal{V}_i\}_{i=1}^{\hat{s}}$ may be improved, possibly re-including also those data points marked as outliers, by iterating the presented clustering procedure on the whole data set $\mathcal{D}$ according to the flow chart in Figure 4.

In particular, based on the clustered data $\{\mathcal{V}_i\}_{i=1}^{\hat{s}}$, the parameters of each mode i are identified as follows

$$\hat{\boldsymbol{\vartheta}}_i \in \arg \min_{\boldsymbol{\vartheta}_i \in \mathbb{R}^n} \max_{t \in \mathcal{V}_i} |y_t - \boldsymbol{\varphi}_t^\top \boldsymbol{\vartheta}_i|, \tag{12}$$

and the procedure to generate the preference vectors described in Section 3.2 is repeated using the identified modes as model pool (*i.e.*, setting $N_p = \hat{s}$ and $\tilde{\boldsymbol{\vartheta}}_j = \hat{\boldsymbol{\vartheta}}_j$, $j = 1, \dots, \hat{s}$), followed by the two-stage clustering scheme described in Section 3.3. The procedure is iterated until the estimated switching signal $\hat{\sigma}_t$ computed according to (9) is equal to a previous estimate. Note that, in the PWARX case, there is no guarantee that the returned clusters are linearly separable.

3.6. Algorithm properties

We can now state the properties of the proposed constrained clustering iterative algorithm in terms of convergence and performance.

Theorem 3.1. *The proposed iterative algorithm is guaranteed to terminate, and, upon termination, each data $(y_t, \boldsymbol{x}_t)$, $t = 1, \dots, N$, is either classified as an outlier or it satisfies*

$$|y_t - \boldsymbol{\varphi}_t^\top \hat{\boldsymbol{\vartheta}}_{\hat{\sigma}_t}| \leq \varepsilon, \tag{13}$$

where $\hat{\sigma}_t$ and $\{\hat{\boldsymbol{\vartheta}}_i\}_{i=1}^{\hat{s}}$ are the switching signal and mode parameters in (9) *and* (12) *obtained at convergence.*

Proof. Termination is guaranteed since a) the reconstructed switching signal can assume only a finite number of values, given that the number of possible data clusters is finite (there can be at most N clusters, one for each sample), and b) the stopping criterion prevents from cycling back to previously found solutions. Since both the low-level and the high-level clustering operations satisfy property P1 and the pruning procedure only removes clusters, one has that for each of the resulting clusters $\mathcal{V}_i$, $i = 1, \dots, \hat{s}$, there exists at least one model $\bar{\jmath} \in \{1, \dots, N_p\}$ such that $0 < \rho_{i\bar{\jmath}} = \min_{t \in \mathcal{V}_i} p_{t\bar{\jmath}}$. This implies

$$r_{t\bar{\jmath}} = |y_t - \boldsymbol{\varphi}_t^\top \tilde{\boldsymbol{\vartheta}}_{\bar{\jmath}}| \leq \varepsilon,\ t \in \mathcal{V}_i. \tag{14}$$

Now, if $t \in \mathcal{I}_O$, then $(y_t, \boldsymbol{x}_t)$ is considered as an outlier, otherwise there exists a cluster i such that $t \in \mathcal{V}_i$ and the following chain of inequalities holds

$$\begin{aligned} |y_t - \boldsymbol{\varphi}_t^\top \hat{\boldsymbol{\vartheta}}_i| &\leq \max_{t \in \mathcal{V}_i} |y_t - \boldsymbol{\varphi}_t^\top \hat{\boldsymbol{\vartheta}}_i| = \min_{\boldsymbol{\vartheta}_i \in \mathbb{R}^n} \max_{t \in \mathcal{V}_i} |y_t - \boldsymbol{\varphi}_t^\top \boldsymbol{\vartheta}_i| \\ &\leq \max_{t \in \mathcal{V}_i} |y_t - \boldsymbol{\varphi}_t^\top \tilde{\boldsymbol{\vartheta}}_{\bar{\jmath}}| \leq \varepsilon, \end{aligned}$$

where the first equality and the second inequality are due to $\hat{\boldsymbol{\vartheta}}_i$ being a minimizer according to (12), and the last inequality is due to (14). If $t \in \mathcal{V}_i$, by (9), $\hat{\sigma}_t = i$, thus

$$|y_t - \boldsymbol{\varphi}_t^\top \hat{\boldsymbol{\vartheta}}_{\hat{\sigma}_t}| = |y_t - \boldsymbol{\varphi}_t^\top \hat{\boldsymbol{\vartheta}}_i| \leq \varepsilon.$$

These results are valid at each iteration and, specifically, at the last one, which concludes the proof. □

3.7. Selection of parameter ε

The output of the proposed procedure obviously depends on the value of the threshold ε. Large values of ε cause many data points to express positive preferences for a large number models in the pool, as large absolute errors are tolerated, and this results in a data partition with few clusters (under-estimation of the number of modes).

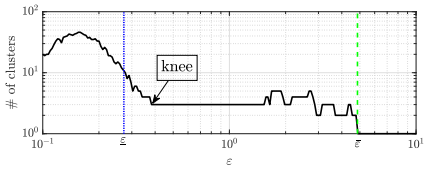

Figure 5: Estimated number of clusters as a function of ε.

In particular, if the value of ε is set too large, *i.e.*,

$$\varepsilon \geq \overline{\varepsilon} = \min_{j=1,\dots,N_p} \max_{t=1,\dots,N} r_{tj}, \tag{15}$$

then a single mode is identified, since there exists one model in the pool that fits all data points within an absolute error ε. Conversely, small values of ε cause many data points to express positive preferences for only a few models in the pool (only small errors are tolerated), thus leading to an over-estimation of the number of modes s and overfitting issues. Note, however, that the number of clusters may decrease as ε gets smaller (see, *e.g.*, Figure 5), since the chance of getting outliers grows. Indeed, if

$$\varepsilon < \underline{\varepsilon} = \max_{t=1,\dots,N} \min_{j=1,\dots,N_p} r_{tj}, \tag{16}$$

then there exists at least one data point $\bar{t}$ for which $r_{\bar{t}j} > \varepsilon$ for all models $j = 1, \dots, N_p$, thus implying $\boldsymbol{p}_{\bar{t}} = \mathbf{0}$ and $(y_{\bar{t}}, \boldsymbol{x}_{\bar{t}}) \in O$. In case $\mathcal{D}$ is known not to contain outliers, then $\underline{\varepsilon}$ and $\overline{\varepsilon}$ appear to be reasonably good bounds for ε.

The *a priori* information on the noise characteristics can be used to set a sensible value for ε. For example, if the noise is Gaussian with zero mean and standard deviation ζ, then one can set $\varepsilon \geq 3\zeta$. If no information is available, one can execute the proposed algorithm for increasing values of ε. Then, a sensible value of ε can be taken at the knee of the curve that plots the identified number of modes $\hat{s}$ as a function of ε, as discussed also in [3].

Applying the overall procedure to Example 3.4 for $\varepsilon \in [0.1, 10]$ one obtains the curve reported in Figure 5, where it is apparent that any value of ε between $3.3\zeta = 0.33$ and $11\zeta = 1.1$ leads to the correct identification of the true number of modes s°.

4. Numerical and experimental examples

In this section, we perform a comparative analysis with various competitor algorithms. The methods are evaluated in terms of fitting and classification accuracy, based on the indices defined next. The fitting accuracy is measured according to

$$FIT = 100 \left(1 - \frac{\sum_{t=1}^{N} \|y_t - \hat{y}_t\|_2^2}{\sum_{t=1}^{N} \|y_t - \bar{y}\|_2^2}\right), \tag{17}$$

where $\hat{y}$ denotes the estimated model output sequence, and $\bar{y}$ is the average of the true output sequence y, so as to normalize with respect to the size and variability of the output data. Unless specified differently, $\hat{y}$ denotes the one-step ahead prediction. Detected outliers are ruled out from the computation of (17). This consideration applies only to our method and the one in [3], since they alone include an outlier detection procedure.

The classification accuracy is the percentage of correctly classified data points and is given by

$$\mathcal{C}_N^{true} = \frac{100}{N} \sum_{t=1}^{N} \mathbb{1}_{[\hat{\sigma}_t = \sigma_t]}, \tag{18}$$

where $\mathbb{1}_{[\hat{\sigma}_t = \sigma_t]} = 1$, if $\hat{\sigma}_t = \sigma_t$, and 0, otherwise, with $\hat{\sigma}_t$ set to 0 if the datum at t is identified as an outlier. Since the value of index (18) depends on the labeling of the modes of the identified switched model, we consider the labeling that maximizes it in all methods.

The comparative analysis is performed on data realizations of 2000 samples each associated with 100 independent extractions of the exogenous inputs, which include the switching signal in the SARX case. Indices (17) and (18) are computed on training data and averaged over the 100 runs. Table 1 lists the parameters adopted in the different methods, which are finely tuned on one additional data realization of length 4000 so as to get the largest FIT. Recall that in the algorithm proposed in this paper and in the bounded-error method [3] the number of modes s° of the system (1) generating data is not assumed to be known and the goal is to identify a switched model that guarantees an estimation error smaller than or equal to ε (δ in [3]). In order to carry out a fair comparison, for each Monte Carlo run, ε is chosen so that the correct number s° of modes is identified. Finally, we assume that the model orders n_y and n_u are known for all methods.

Table 1: Parameter settings for each method (please, refer to the cited papers for the parameter definitions and notations).

Method	**SARX**	**PWARX**
Proposed method	$N_p = 100$, $\alpha = 20$, $\beta = \tau_D$	$N_p = 100$, $\alpha = 20$, $\beta = 20$
Bounded-error [3]	$C = 10$, $\rho = 0.7$, $T_0 = 100$, $\gamma = 0.001$, $c = 5$, $\beta = \frac{\tau_D}{N}$, α not used	$C = 10$, $\rho = 0.7$, $T_0 = 100$, $\gamma = 0.001$, $c = 90$, $\beta = \frac{70}{N}$, α not used
k-RANSAC [13]	$\varepsilon = 0.1$, $\lvert\mathcal{S}_k\rvert = 10$	$\varepsilon = 0.3$, $\lvert\mathcal{S}_k\rvert = 10$
Clustering [10]	– –	$c = 35$ K-means runs: 15
Framework [4]	$\ell = \|y_t - \boldsymbol{x}_t^T \boldsymbol{\vartheta}_{s_t}\|^2$ $r(\boldsymbol{\vartheta}_k) = 0.00001$ $\mathcal{L}(S)$ see Section 5.1 in [4], with $\tau = 0.02$, $\pi = 0.01$	$\ell = \|y_t - \boldsymbol{x}_t^T \boldsymbol{\vartheta}_{y,s_t}\|^2 + 0.01\|\boldsymbol{x}_t - \boldsymbol{\vartheta}_{x,s_t}\|^2$ $r(\boldsymbol{\vartheta}_k) = 0$ $\mathcal{L}(S) = 0$
Method [5]	$w = 19$	$p = 9$

4.1. SARX identification

With reference again to the example discussed in Section 3.4, Table 2 reports the aggregated results of the comparative analysis for the methods in Table 1. The methods

that account explicitly for information about the switching mechanism, *i.e.*, [4] and [5], as well as the proposed method, outperform the others, especially in terms of classification accuracy. Among this restricted group of methods, the proposed one has the highest classification and fitting accuracy. Also, the low standard deviation of the evaluation indices indicates that the proposed approach is robust with respect to the exogenous inputs realizations. A maximum of 5 algorithm iterations is required.

Table 2: SARX comparative analysis: mean ± standard deviation.

Method	$\mathcal{C}_N^{true}$[%]	***FIT*** [%]	**Comp. time** [s]
k-RANSAC [13]	85.42 ± 6.59	81.76 ± 8.44	0.48 ± 0.05
Framework [4]	99.85 ± 3.99	92.88 ± 3.30	1.10 ± 0.59
Method [5]	99.41 ± 0.19	91.80 ± 1.59	0.50 ± 0.19
Bounded-error [3]	85.85 ± 1.72	91.85 ± 0.86	6.63 ± 1.82
Bounded-error [3]†	91.85 ± 0.86	99.66 ± 0.15	2.78 ± 0.12
Proposed method	93.74 ± 0.67	99.86 ± 0.15	2.74 ± 0.07

† We exploit temporal adjacency instead of spatial adjacency when refining the assignment of undecidable points.

4.2. PWARX identification

Consider the following PWARX system [3]:

$$y_t = \begin{cases} -0.4y_{t-1} + u_{t-1} + 1.5 + e_t \\ \quad \text{if } 4y_{t-1} - u_{t-1} + 10 < 0 \\ 0.5y_{t-1} - u_{t-1} - 0.5 + e_t \\ \quad \text{if } 4y_{t-1} - u_{t-1} + 10 \geq 0 \\ \quad \text{and } 5y_{t-1} + u_{t-1} - 6 \leq 0 \\ -0.3y_{t-1} + 0.5u_{t-1} - 1.7 + e_t \\ \quad \text{if } 5y_{t-1} + u_{t-1} - 6 > 0 \end{cases} \tag{19}$$

where u_t is a sequence of independent random variables uniformly distributed in the range $[-4, 4]$, while e_t is a zero mean white Gaussian noise with variance $\zeta^2 = \left(\frac{1}{15}\right)^2$. A data realization of length 2000 was used for estimation. The data clustering obtained by applying the proposed approach with $\varepsilon = 0.3$ (see Figure 6) is consistent with the true region boundaries except for a few data points (red dots). This is reflected by the final $\mathcal{C}_N^{true}$ and FIT values which amount to 99.85% and 97.58%, respectively. A deeper analysis points out that the few misclassified points lie in regions where two modes exhibit comparable output values y_t. Consequently, since the regressor vectors $\boldsymbol{x}_t$ are also close to each other, in these cases the spatial adjacency information does not help in solving the ambiguity between the two modes.

Figure 7 shows the curves for the MAE, the number of outliers, the number of clusters, and the classification error as a function of the error bound ε. The shaded areas represent the envelopes of the curves relative to 100 data realizations. Observe that there is wide range of ε values for which the correct number of modes is exactly retrieved irrespective of the realization, thus confirming that the actual number of modes can be identified by appropriately tuning ε. Also, as discussed in Section 3.7 (see (15)), an increased variance in the retrieved number of clusters is experienced beyond $\varepsilon \simeq 2.3$, until only one cluster results

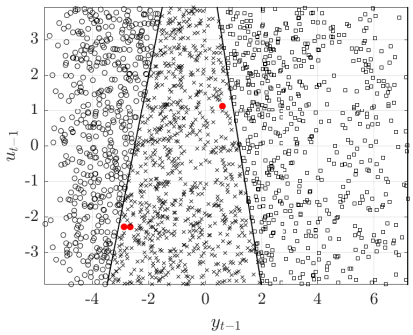

Figure 6: PWARX Identification: Obtained data classification. Different markers denote different modes. Solid lines denote the true region boundaries. Misclassified data are plotted in red.

from the identification procedure. Finally, observe that the MAE is always below the error bound (dashed red line) and that this holds for all the data, provided that ε is chosen so as no outliers arise (see (16)).

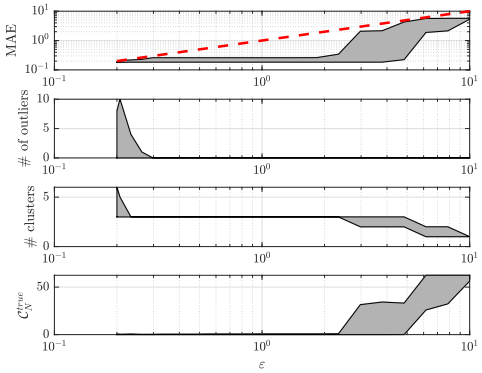

Figure 7: PWARX Identification: sensitivity analysis to ε. The shaded areas represent the envelopes of the curves relative to 100 different data realizations.

A comparative analysis (see Table 3) shows the supremacy of the methods that consider time/spatial localization information, as in the SARX case. Conversely, the clustering approach of [10] is performing poorly in this example. This is due to the fact that in [10] data clustering is performed in the model parameter space and, hence, is badly affected by noise and requires a careful tuning of the size of the local datasets. Instead, performing data clustering in the preference space robustifies the proposed method against noise and makes the choice of the size of the local data sets less critical as clearly shown in Figure 8. Regarding the proposed method, the larger computation time needed with respect to the SARX example is mainly due to the computation of the Delaunay triangulation required to derive the graph $\mathcal{G}_\varepsilon$ at each iteration (at most 4 in this example, 5 in the SARX one).

4.3. Robustness to outliers

In this subsection we first analyze the robustness of the proposed method and the bounded-error approach with respect to the presence of outliers. To this purpose, we

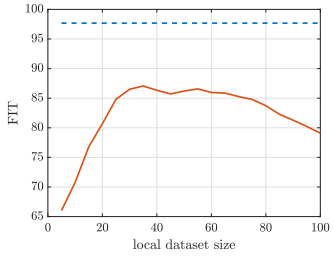

Figure 8: Proposed method (dashed blue line), clustering method in [10] (solid red line).

Table 3: PWARX comparative analysis: mean ± standard deviation.

Method	$\mathcal{C}_N^{true}$[%]	FIT [%]	Comp. time [s]
k-RANSAC [13]	97.47 ± 0.40	97.44 ± 0.82	0.43 ± 0.03
Clustering [10]	95.39 ± 1.07	86.65 ± 3.27	5.87 ± 0.22
Framework [4]	99.91 ± 0.07	97.84 ± 0.06	1.73 ± 0.38
Method [5]	99.90 ± 0.09	97.83 ± 0.06	0.71 ± 0.37
Bounded-error [3]	97.43 ± 0.76	96.98 ± 0.29	3.52 ± 1.12
Proposed method	99.91 ± 0.09	97.68 ± 0.11	11.02 ± 1.70

consider the previous PWARX example, but this time we corrupt a fraction r of the data of each realization by adding a zero mean white Gaussian noise δ_t with variance 25. Note that not all the corrupted data can be considered as outliers since for some index t the resulting noise term $e_t + \delta_t$ might be compatible with the assumed range of variability for e_t. Accordingly, we consider as outliers only the corrupted data for which the resulting noise term $|e_t + \delta_t| > 0.2 = 3\zeta$ and we assign $\sigma_t = 0$ to them. In this way, when we compute index (18) for the proposed method and the bounded-error approach in [3], correctly detected outliers count as correctly classified data points. Figure 9 shows the $\mathcal{C}_N^{true}$, FIT, and percentage of the detected outliers, for $r = 0.05, 0.10, 0.15, 0.20, 0.25$. Apparently, the proposed method outperforms the bounded-error approach in all considered performance indexes. This confirms its effectiveness in detecting and isolating the outliers.

For the sake of completeness, Table 4 reports the aggregated results of the comparative analysis for all the methods in Table 1, in the case when $r = 0.10$. Since only the proposed method and the bounded-error approach are able to detect and isolate the outliers, it is not surprising that the other methods provide relatively poor results both in terms of classification accuracy and fit.

Table 4: PWARX comparative analysis in the presence of outliers: mean ± standard deviation. Case 1: all data are used to compute $\mathcal{C}_N^{true}$ and FIT. Case 2: data outliers are ruled out.

	Case 1		*Case 2*	
Method	$\mathcal{C}_N^{true}$ [%]	FIT [%]	$\mathcal{C}_N^{true}$ [%]	FIT [%]
k-RANSAC [13]	86.43 ± 1.15	48.43 ± 2.15	95.54 ± 1.27	88.24 ± 2.09
Clustering [10]	85.79 ± 1.47	54.29 ± 1.94	94.83 ± 1.63	84.73 ± 3.81
Framework [4]	81.16 ± 11.01	66.85 ± 1.70	89.71 ± 12.19	86.90 ± 4.30
Method [5]	84.28 ± 6.91	65.18 ± 1.89	93.16 ± 7.62	87.78 ± 3.82
Bounded-error [3]	95.51 ± 0.89	95.68 ± 0.41	96.51 ± 0.96	95.78 ± 0.43
Proposed method	99.83 ± 0.13	97.59 ± 0.27	99.87 ± 0.12	97.59 ± 0.26

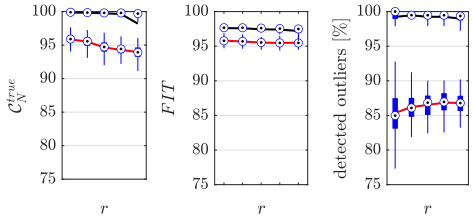

Figure 9: PWARX Identification in the presence of r% outliers, $r = 0.05, 0.10, 0.15, 0.20, 0.2$. Comparative analysis: proposed method (black line) and bounded-error method [3] (red line). Solid lines denote mean values.

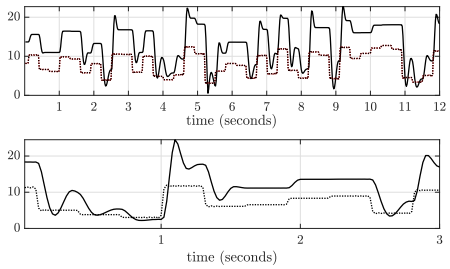

Figure 10: Experimental example: data set used for identification (top) and validation (bottom). The solid and dashed lines represent the system output (position of the mounting head) and the scaled input (voltage applied to the motor), respectively.

4.4. Experimental example: pick-and-place machine

The proposed method is applied to a benchmark example presented in [14] and discussed in [3, 8, 19, 4]. The example consists of a component placement process operated by a pick-and-place machine. Specifically, this process is characterized by switchings between two main operational modes, the *free* and the *impact mode* (the mounting head in contact with the board). Available data represent measurements of the voltage applied to the motor driving the mounting head, u_t, and the vertical position of the mounting head, y_t, (see Figure 10).

The objective is to identify a PWA model in the form of (1), with regression vector $\boldsymbol{\varphi}_t = [1\ y_{t-1}\ y_{t-2}\ u_{t-1}]^\top$. The model performance is evaluated by the FIT criterion (17) considering the model simulation output as $\hat{y}_t$, *i.e.,* the model output computed based on previous model estimations and the measured system input only. The Multicategory Robust Linear Programming (MRLP) linear separation technique is employed to reconstruct the polyhedral partition of the regressor space which is needed to map each $x_t = [\hat{y}_{t-1}\ \hat{y}_{t-2}\ u_{t-1}]$ to one of the identified modes, for use in the validation.

By choosing $\varepsilon = 0.75$, 0.65, and 0.3, models with $s = 2$, 3, and 4 discrete modes, respectively, are identified from the training data, the other parameter settings being $N_p = N$ (one local model per datum), $\alpha = 10$, and $\beta = 30$. Also, a single ARX model with the same model orders is identified for comparison purposes. Table 5 reports the FIT values for the four identified models. These values clearly show that a single ARX model is not adequate, and

that better models can be obtained with a switched model with multiple modes. Further analysis, not reported for brevity, proved that no significant improvements in terms of fitting accuracy can be obtained by further reducing ε.

Table 5: FIT values for the identified models.

s	1	2	3	4
FIT	67%	78%	81%	84%

The simulated responses are graphically compared to the measured one in Figure 11. Note that all the models failed in tracking the system output in the interval between 1.5s and 2.5s, as experienced also in [14]. In this period, the system response is not affected by some input variations, possibly due to the presence of friction, whereas the simulated responses are. As for the reconstructed switching signal, the model with $s = 2$ apparently captures well the physical operational modes. Indeed, mode 1 and mode 2 can be respectively mapped on the impact and free modes. In the model with $s = 3$, the extra mode seems to account for transitions between the two physical modes, which can be still mapped on mode 1 and 2, respectively. Further reducing the value of ε increases the number of modes, hampering the physical interpretation of the reconstructed switching signal. The presented results have been obtained within a maximum of 14 algorithm iterations and 13.21 seconds on average.

5. Conclusions

This paper introduces a unified framework to address bounded-error identification of SARX and PWARX systems by an iterative two-level clustering with outliers isolation integrating the a-priori information on the model class through an adjacency graph and adopting as clustering criterion the preferences expressed by data for some candidate models rather than the similarity of the candidate model parameters, which is extremely sensitive to noise and subject to overparameterization issues.

The resulting algorithm proved to be superior to alternative state-of-the-art approaches in solving some identification problem benchmarks, being also robust to outliers.

References

[1] Edoardo Amaldi and Marco Mattavelli. The MIN PFS problem and piecewise linear model estimation. *Discrete Applied Mathematics*, 118(1-2):115–143, 2002.

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

[3] Alberto Bemporad, Andrea Garulli, Simone Paoletti, and Antonio Vicino. A bounded-error approach to piecewise affine system identification. *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] Federico Bianchi, Alessandro Falsone, Luigi Piroddi, and Maria Prandini. An alternating optimization method for switched linear systems identification. *IFAC-PapersOnLine*, 53(2):1071–1076, 2020.

[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, February 2020.

[7] Federico Bianchi, Valentina Breschi, Dario Piga, and Luigi Piroddi. Model structure selection for switched NARX system identification: A randomized approach. *Automatica*, 125: 109415, 2021.

[8] Khaled Boukharouba, Laurent Bako, and S Lecoeuche. Identification of piecewise affine systems based on dempster-shafer theory. *IFAC Proceedings Volumes*, 42(10):1662–1667, 2009.

[9] Giancarlo Ferrari-Trecate and Marco Muselli. Single-linkage clustering for optimal classification in piecewise affine regression. *IFAC Proceedings Volumes*, 36(6):33–38, 2003.

[10] 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.

[11] Martin A. Fischler and Robert C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[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] A. L. Juloski, W. P. M. H Heemels, and G. Ferrari-Trecate. Data-based hybrid modelling of the component placement process in pick-and-place machines. *Control Engineering Practice*, 12(10):1241–1252, 2004.

[15] Fabien Lauer and Gérard Bloch. *Hybrid System Identification*. Springer International Publishing, 2019.

[16] Der-Tsai Lee and Bruce J. Schachter. Two algorithms for constructing a Delaunay triangulation. *International Journal of Computer & Information Sciences*, 9(3):219–242, 1980.

[17] Fredrik Lindsten, Henrik Ohlsson, and Lennart Ljung. *Just relax and come clustering!: A convexification of k-means clustering*. Linköping University Electronic Press, 2011.

[18] Luca Magri and Andrea Fusiello. T-linkage: A continuous relaxation of j-linkage for multi-model fitting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3954–3961, 2014.

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

[20] 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.

[21] Roberto Toldo and Andrea Fusiello. Robust multiple structures estimation with j-linkage. In *European conference on computer vision*, pages 537–547, 2008.

[22] Peter Willett, John M Barnard, and Geoffrey M Downs. Chemical similarity searching. *Journal of chemical information and computer sciences*, 38(6):983–996, 1998.

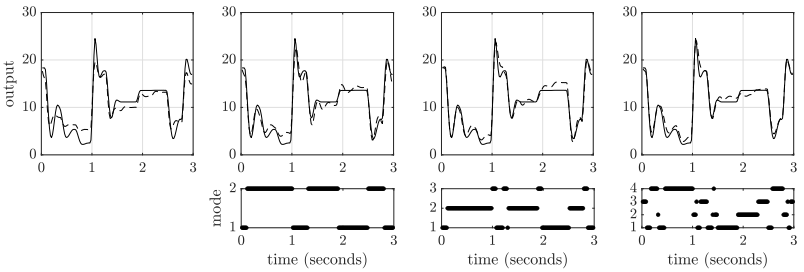

Figure 11: Experimental example: simulation results for the models with $s = 1$, 2, 3, and 4 modes, from left to right. Top: simulated output (dashed line) and system output (solid line). Bottom: switching signal.