A Bounded-Error Approach to Piecewise Affine System Identification

Alberto Bemporad, Andrea Garulli, Simone Paoletti, and Antonio Vicino

***Abstract*— This paper proposes a three-stage procedure for parametric identification of PieceWise affine AutoRegressive eXogenous (PWARX) models. The first stage simultaneously classifies the data points and estimates the number of submodels and the corresponding parameters by solving the MIN PFS problem (Partition into a MINimum Number of Feasible Subsystems) for a suitable set of linear complementary inequalities derived from data. Second, a refinement procedure reduces misclassifications and improves parameter estimates. The third stage determines a polyhedral partition of the regressor set via two-class or multi-class linear separation techniques. As a main feature, the algorithm imposes that the identification error is bounded by a quantity δ. Such a bound is a useful tuning parameter to trade off between quality of fit and model complexity. The performance of the proposed PWA system identification procedure is demonstrated via numerical examples and on experimental data from an electronic component placement process in a pick-and-place machine.**

***Index Terms*— Nonlinear identification, piecewise affine autoregressive exogenous models, bounded error, MIN PFS problem.**

I. Introduction

When linear models are not appropriate for describing accurately the dynamics of a system, nonlinear identification must be employed. Several nonlinear model structures have been considered and their properties investigated in the literature, see, *e.g.*, the survey papers [1], [2], and references therein. This paper focuses on the problem of identifying PieceWise Affine (PWA) models of discrete-time nonlinear and hybrid systems from input-output data. PWA systems are obtained by partitioning the state and input set into a finite number of polyhedral regions, and by considering linear/affine subsystems sharing the same continuous state in each region [3]. In other words, the state and output maps of a PWA system are both piecewise affine. PWA models represent an attractive model structure for system identification. Thanks to the universal approximation properties of PWA maps [4], [5], PWA models form a nonlinear black-box structure, *i.e.* a model structure that is prepared to describe virtually any nonlinear dynamics [1]. In addition, given the equivalence between PWA systems and several classes of hybrid systems [6], [7], PWA system identification is useful for estimating hybrid models from data.

The identification of a PWA model involves the estimation of the parameters of the affine submodels and the hyperplanes defining the partition of the state and input set (or the regressor set, for models in regression form). This issue clearly underlies a classification problem, namely each data point must be associated to the most suitable submodel. As long as partitioning is concerned, two alternative approaches can be distinguished: 1) the partition is fixed a priori; 2) the partition is estimated together with the submodels. In the first case, data classification is very simple, and estimation of the submodels can be carried out by resorting to standard linear identification techniques. In the second case, the regions are shaped to the clusters of data, and the strict relation among data classification, parameter estimation and region estimation makes the identification problem very hard to cope with. The problem is even more complicated when also the number of submodels must be estimated. A number of approaches dealing with the estimation of PWA models of nonlinear dynamical systems can be found in different fields, such as neural networks, electrical networks, time-series analysis, function approximation. See [8] for a nice overview and classification. Recently, novel contributions to this topic have been proposed in both the hybrid systems and the nonlinear identification communities. In [9] PieceWise affine ARX (PWARX) models are considered and the combined use of clustering, linear identification, and pattern recognition techniques is exploited in order to identify both the affine submodels and the polyhedral partition of the regressor set. In [10] the authors propose an algebraic geometric solution to the identification of PieceWise Linear (PWL) models which establishes a connection between PWL system identification, polynomial factorization, and hyperplane clustering. [11] describes an iterative algorithm that sequentially estimates the parameters of the model and classifies the data through the use of adapted weights. In [12] the identification problem is formulated for two subclasses of PWA models, namely Hinging Hyperplane ARX (HHARX) and Wiener PWARX (W-PWARX) models, and solved via mixed-integer linear or quadratic programs.

In this paper, a different approach inspired by ideas from set-membership identification (see [13], [14] and references therein) is proposed. The main feature is to impose that the identification error is bounded by a given quantity δ for all the samples in the estimation data set. In order to meet this condition, the estimation of the number of submodels, data classification and parameter estimation are performed simultaneously by partitioning a suitable set of linear complemen-

The authors are with Dipartimento di Ingegneria dell'Informazione, Università di Siena, Via Roma 56, 53100 Siena, Italy, and with Centro per lo Studio dei Sistemi Complessi, Università di Siena, Via T. Pendola 37, 53100 Siena, Italy. E-mail: {bemporad,garulli,paoletti,vicino}@dii.unisi.it.

This work was partially supported by Ministero dell'Istruzione, dell'Università e della Ricerca through the research grants MIUR-PRIN 2004, and by the European Community through the HYCON Network of Excellence, contract number FP6-IST-511368, and project "Computation and Control", contract number FP5-IST-2001-33520.

tary inequalities derived from data into a minimum number of feasible subsystems (MIN PFS problem). A suboptimal solution to the MIN PFS problem (which is an NP-hard problem) is obtained by applying a modified version of the greedy algorithm proposed in [15]. A refinement procedure is also employed in order to reduce misclassifications and to improve parameter estimates. Region estimation is lastly performed via two-class [16], [17], or multi-class [18], [19], linear separation techniques. The bound δ is used as a tuning knob to trade off between quality of fit and model complexity: The larger δ, the smaller the required number of submodels at the price of a worse fit of the data. Another interesting feature of the approach is that a set of feasible parameters can be associated to each submodel according to the bounded-error condition, thus allowing the evaluation of the related parametric uncertainty [13].

Preliminary versions of the proposed identification technique appeared in [20], [21]. The present version contains further new material, including improvements of the greedy algorithm used to initialize the identification procedure, and a way to associate undecidable data points to submodels by a suitable reassignment in the classification process. In addition, a case-study is presented, where the identification technique is tested on real data from the electronic component placement process in a pick-and-place machine [22].

II. Problem Statement

Given a discrete-time nonlinear dynamical system with input $\mathbf{u}_k \in \mathbb{R}^p$, output $y_k \in \mathbb{R}$, and possibly discontinuous dynamics, let $\mathbf{u}^{k-1}$ and $\mathbf{y}^{k-1}$ be, respectively, past inputs and outputs generated by the system up to time $k-1$. A PWARX model establishes a relationship between past observations $(\mathbf{u}^{k-1}, \mathbf{y}^{k-1})$ and future outputs y_k in the form

$$y_k = f(\mathbf{x}_k) + \varepsilon_k \,, \tag{1}$$

where $\varepsilon_k \in \mathbb{R}$ is the error term, $\mathbf{x}_k \in \mathbb{R}^n$ is the regression vector with fixed structure depending only on past n_a outputs and n_b inputs:

$$\mathbf{x}_k = [\; y_{k-1} \;\; \dots \;\; y_{k-n_a} \;\; \mathbf{u}'_{k-1} \;\; \dots \;\; \mathbf{u}'_{k-n_b} \;]' \tag{2}$$

(hence, $n = n_a + p \cdot n_b$), and $f : \mathcal{X} \to \mathbb{R}$ is the PWA map:

$$f(\mathbf{x}) = \begin{cases} \varphi'\theta_1 & \text{if} \quad \mathbf{x} \in \mathcal{X}_1 \\ \vdots & \vdots \\ \varphi'\theta_s & \text{if} \quad \mathbf{x} \in \mathcal{X}_s \,, \end{cases} \tag{3}$$

which is defined over the regressor set $\mathcal{X} \subseteq \mathbb{R}^n$ where the PWARX model is valid. In (3), s is the number of submodels (or discrete modes), φ is the extended vector $\varphi = [\; \mathbf{x}' \;\; 1]'$, and $\theta_i \in \mathbb{R}^{n+1}$, $i = 1, \dots, s$, are the parameter vectors of each affine ARX submodel. The regions $\mathcal{X}_i$ form a *complete partition* of $\mathcal{X}$ (*i.e.* $\bigcup_{i=1}^{s} \mathcal{X}_i = \mathcal{X}$ and $\mathcal{X}_i^\circ \bigcap \mathcal{X}_j^\circ = \emptyset$, $\forall\, i \neq j$, where $\mathcal{X}_i^\circ$ denotes the interior of $\mathcal{X}_i$), and are assumed to be convex polyhedra, described by

$$\mathcal{X}_i = \left\{ \mathbf{x} \in \mathbb{R}^n \,:\, H_i \, \varphi \preceq \mathbf{0} \right\}, \tag{4}$$

where $H_i \in \mathbb{R}^{q_i \times (n+1)}$, $i = 1, \dots, s$, and "$\preceq$" denotes componentwise inequality. Since the PWA map (3) is not assumed to be continuous, with definition (4) f could be multi-valued over common boundaries of the regions $\mathcal{X}_i$. This issue can be easily overcome by making some of the inequalities strict in the definitions of the polyhedra $\mathcal{X}_i$.

Remark 1: In (4), q_i is the number of linear inequalities defining the i-th polyhedral region. As will be clarified in Section V, $q_i \leq s - 1$ in the identified model. □

The identification of a PWARX model (1)-(4) from a finite data set $(y_k, \mathbf{x}_k)$, $k = 1, \dots, N$, is a very complex problem involving data classification and the estimation of s, $\{\theta_i\}_{i=1}^s$ and $\{\mathcal{X}_i\}_{i=1}^s$. When the number of discrete modes s is fixed, the problem amounts to the reconstruction of the PWA map f, and identification can be in principle carried out by minimizing with respect to θ_i and H_i, $i = 1, \dots, s$, the cost function

$$V_N(\theta_i, H_i) = \frac{1}{N} \sum_{k=1}^{N} \ell\big(y_k - f(\mathbf{x}_k)\big) \,, \tag{5}$$

where ℓ is a given error penalty function, such as $\ell(\varepsilon) = \varepsilon^2$, or $\ell(\varepsilon) = |\varepsilon|$. Note that, if the regions $\mathcal{X}_i$, $i = 1, \dots, s$, are fixed a priori and $\ell(\varepsilon) = \varepsilon^2$, the minimization of (5) is carried out only with respect to θ_i, $i = 1, \dots, s$, and reduces to ordinary least-squares. When both θ_i and H_i, $i = 1, \dots, s$, must be estimated, the problem is in general nonconvex, and hence much harder to solve. If the number s of submodels must be also estimated, the optimization problem should include additional terms in the objective and/or additional constraints (*e.g.*, bounds on s), in order to limit the number of submodels and avoid overfit. Several heuristic and suboptimal approaches that are applicable, or at least related, to the identification of PWARX models, have been proposed in the literature (see [8] for an overview). Most of them look for good suboptimal solutions of the minimization of (5), except the one in [12], where the global optimum can be attained for two subclasses of PWA models by reformulating the problem as a mixed integer linear or quadratic program. As regards the number of submodels, most approaches either assume a fixed s, or adjust s iteratively (*e.g.*, by adding one submodel at a time) until the quality of fit is acceptable.

Inspired by ideas from set-membership identification (see [13], [14] and references therein), the approach presented in this paper is based on imposing a bound $\delta > 0$ on the error term ε_k in (1) for all the samples in the estimation data set. Feasible solutions of the identification problem are thus all PWARX models (1)-(4) satisfying

$$|y_k - f(\mathbf{x}_k)| \leq \delta \,, \quad \forall\, k = 1, \dots, N \,, \tag{6}$$

for the given $\delta > 0$. The focus here is on providing a particular feasible solution. Since the number s of submodels is neither assumed to be known, nor fixed a priori, in order to obtain a model which is as simple as possible (where "simplicity" is measured in terms of the number of submodels) the minimum s allowing to satisfy (6) is sought. Hence, the considered identification problem is as follows:

Problem 1: Given N data points $(y_k, \mathbf{x}_k)$, $k = 1, \dots, N$, and $\delta > 0$, estimate a minimum positive integer s, a set of parameter vectors $\{\theta_i\}_{i=1}^s$, and a polyhedral partition $\{\mathcal{X}_i\}_{i=1}^s$

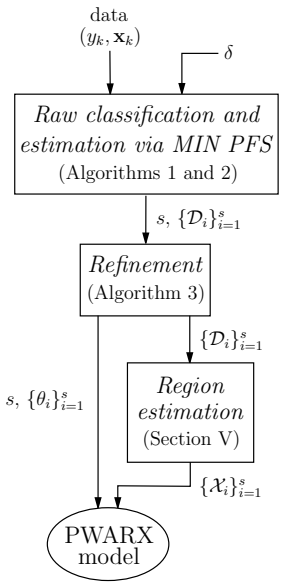

Fig. 1. Flow diagram of the proposed PWA identification procedure. The boxes correspond to the procedure steps, and the labels on the edges describe the inputs and the outputs of each step.

of the regressor set $\mathcal{X}$, such that the corresponding PWARX model (1)-(4) satisfies condition (6).

Note that solving Problem 1 involves to classify the available data points into clusters $\{\mathcal{D}_i\}_{i=1}^s$ such that $(y_k, \mathbf{x}_k) \in \mathcal{D}_i$ if and only if $(y_k, \mathbf{x}_k)$ is attributed to the i-th mode.

The procedure proposed in this paper to solve Problem 1 consists of three steps:

1) *Raw classification and estimation via MIN PFS*. Data classification and parameter estimation are carried out simultaneously, together with the estimation of the number of submodels, by partitioning a suitable set of linear inequalities derived from data into a minimum number of feasible subsystems (MIN PFS problem).
2) *Refinement*. Misclassifications are reduced and parameter estimates are improved through an iterative procedure alternating between data reassignment and parameter update.
3) *Region estimation*. The clusters of regression vectors are linearly separated via two-class or multi-class linear separation techniques.

The first two steps will be described in Sections III and IV. Region estimation will be addressed in Section V. A flow diagram clarifying the links between the three steps is shown in Fig. 1.

It is worthwhile to point out that the bound δ is not necessarily given a priori, rather it is used as a tuning knob of the identification procedure. As discussed in Section III-C, δ can be adjusted in order to find the desired trade off between model complexity and quality of fit, because the smaller δ, the larger is typically the number of submodels needed to fit the data to a PWA map (3), while on the other hand, the larger δ, the worse is the quality of fit, since larger errors are allowed. The case of different bounds for each data point can be always cast into (6) by suitably scaling the data.

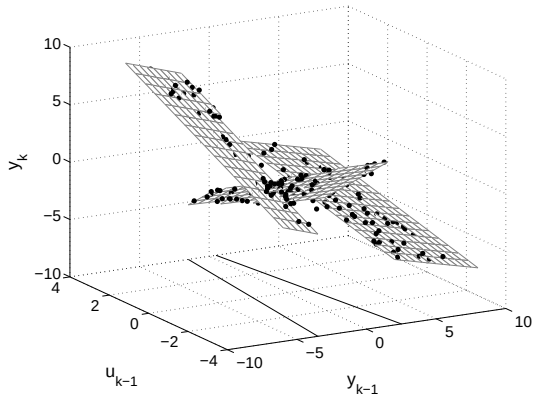

Fig. 2. Affine subsystems, partition of the regressor set, and available data points in Example 1.

In the following, pointwise parameter estimates will be computed through the ℓ_∞ projection estimator [13]. Given a set $\mathcal{D}$ of data points $(y_k, \mathbf{x}_k)$, the ℓ_∞ projection estimate is defined as:

$$\phi_p(\mathcal{D}) = \arg\min_{\theta} \max_{(y_k,\mathbf{x}_k)\in\mathcal{D}} |y_k - \varphi_k'\theta| \,, \tag{7}$$

where $\varphi_k = [\ \mathbf{x}_k'\ 1\ ]'$. Problem (7) can be solved via linear programming. The ℓ_∞ projection estimate is preferred because it has favorable properties in the refinement procedure, as shown in Section IV-A. However, it can be replaced by any other projection estimate, such as least squares.

Remark 2: Problem 1 can be easily extended to multi-output models (or models in state-space form for which the whole state is measurable). In this case, the output of the system is $\mathbf{y}_k \in \mathbb{R}^q$, the PWA map f is a q-valued function, and (6) is replaced by:

$$\|\mathbf{y}_k - f(\mathbf{x}_k)\|_\infty \le \delta \,, \quad \forall\, k = 1, \ldots, N \,. \tag{8}$$

The approach to the solution of Problem 1 presented in this paper is also applicable to the case $q > 1$, provided that small amendments to the procedures described in Sections III and IV are introduced. The reader is referred to [23] for a detailed description. □

The following example will be used throughout the paper to clarify the different steps of the identification procedure.

Example 1: Let $N = 200$ data points $(y_k, \mathbf{x}_k)$ be generated by the following PWARX system [20]:

$$y_k = \begin{cases} -0.4y_{k-1} + u_{k-1} + 1.5 + e_k & \text{if } 4y_{k-1} - u_{k-1} + 10 < 0 \\ 0.5y_{k-1} - u_{k-1} - 0.5 + e_k & \text{if } 4y_{k-1} - u_{k-1} + 10 \ge 0 \\ & \text{and } 5y_{k-1} + u_{k-1} - 6 \le 0 \\ -0.3y_{k-1} + 0.5u_{k-1} - 1.7 + e_k & \text{if } 5y_{k-1} + u_{k-1} - 6 > 0 \,. \end{cases}$$

where $\mathbf{x}_k = [\ y_{k-1}\ u_{k-1}\ ]'$. The number of modes is $\tilde{s} = 3$. The input signal u_k and the noise signal e_k are uniformly distributed on $[-4, 4]$ and on $[-0.2, 0.2]$, respectively. The data points available for estimation are shown in Fig. 2. From left to right, 54, 83 and 63 data points were respectively generated by the three affine subsystem.

III. Raw classification and estimation via MIN PFS

For the moment, let us not address the estimation of the hyperplanes defining the polyhedral partition of the regressor

set, and focus our attention on determining a suitable number of submodels, classifying the data points, and estimating the affine submodels. In view of condition (6), this is accomplished by solving the following problem.

Problem 2: Given $\delta > 0$, find the smallest number s of vectors θ_i, $i = 1, \ldots, s$, and a mapping $k \rightarrowtail i(k)$ such that $|y_k - \varphi_k'\theta_{i(k)}| \leq \delta$ for all $k = 1, \ldots, N$.

Problem 2 consists in finding a *Partition* of the system of linear complementary inequalities:

$$|y_k - \varphi_k'\theta| \leq \delta\,, \quad k = 1, \ldots, N\,, \tag{9}$$

into a *Minimum* number of *Feasible Subsystems* (MIN PFS problem). Given any solution of Problem 2, the partition of the linear complementary inequalities (9), *i.e.* the mapping $k \rightarrowtail i(k)$, provides the classification of the data points, whereas according to the bounded error condition each feasible subsystem defines the set of feasible parameter vectors for the corresponding affine submodel [13]. Note that each inequality in (9) is termed a *linear complementary* inequality because it corresponds to the pair of linear inequalities:

$$\begin{cases} \varphi_k'\theta \leq y_k + \delta \\ \varphi_k'\theta \geq y_k - \delta\,. \end{cases} \tag{10}$$

The MIN PFS problem is NP-hard. Hence, in [20] Problem 2 was tackled by resorting to the greedy randomized algorithm proposed in [15]. The basic idea of the algorithm is to find a vector θ that makes the inequalities in (9) true for as many k as possible (MAX FS problem), then remove those satisfied inequalities and repeat over the remaining ones, until all inequalities have been accounted for. In the following, a modified version of the algorithm [15] is proposed in order to obtain a number s of feasible subsystems which is typically closer to be minimal.

A. A greedy algorithm for the MIN PFS problem

The greedy approach [15] to the MIN PFS problem divides the overall partition problem into a sequence of MAX FS subproblems, each one consisting in finding a vector θ that satisfies the maximum number of linear complementary inequalities of the system at hand. Starting from (9), feasible subsystems of maximum cardinality are iteratively extracted (and the corresponding inequalities removed), until the remaining subsystem is feasible.

Finding a *Feasible Subsystem* of *MAXimum* cardinality of a system of linear complementary inequalities (MAX FS problem) is still an NP-hard problem [24], [25], [26]. Thus, a randomized and thermal relaxation method providing (suboptimal) solutions with a limited computational burden is also proposed in [15].

However, due to both the suboptimality of the greedy approach to the MIN PFS problem and the randomness of the algorithm for the MAX FS problem, the greedy randomized algorithm [15] is not guaranteed to find the minimum number of feasible subsystems. In particular, it has been observed in extensive trials that both the variance of the results may be quite large (*i.e.* the number of extracted subsystems may differ considerably from trial to trial), and the average number of extracted subsystems may be rather far from the minimum (see Example 2).

TABLE I

MODIFIED GREEDY ALGORITHM FOR THE MIN PFS PROBLEM WITH COMPLEMENTARY INEQUALITIES

Algorithm 1

Let $I_1 = \{1, \ldots, N\}$ and $\ell = 0$
REPEAT
  Let $\ell = \ell + 1$ and $\Sigma_\ell = \{|y_k - \varphi_k'\theta| \leq \delta \,:\, k \in I_\ell\}$
  Find a solution θ_ℓ to the MAX FS problem for Σ_ℓ (*Algorithm 2*)
  Let $i = 1$
  WHILE $i < \ell$
    Let $K_{i\ell} = \{k \in I_i \,:\, |y_k - \varphi_k'\theta_\ell| \leq \delta\}$
    IF $\#K_{i\ell} > \#K_i$ THEN let $\theta_i = \theta_\ell$ and $\ell = i$
    Let $i = i + 1$
  END WHILE
  Let $K_\ell = \{k \in I_\ell \,:\, |y_k - \varphi_k'\theta_\ell| \leq \delta\}$ and $I_{\ell+1} = I_\ell \backslash K_\ell$
UNTIL $I_{\ell+1} = \emptyset$
RETURN $s = \ell$ and K_i, $i = 1, \ldots, s$

Based on the above discussion, the algorithm in [15] has been modified as shown in Table I, where $\#A$ denotes the cardinality of a finite set A, and $A \backslash B$ denotes the difference of two sets A and B. The enhanced algorithm differs from the original version for the addition of the WHILE loop. Let Σ_ℓ be the system consisting of the remaining inequalities after having extracted $\ell - 1$ feasible subsystems from (9), and let θ_ℓ be a (suboptimal) solution of the MAX FS problem for system Σ_ℓ (see Algorithm 2 in Table II). The solution θ_ℓ is applied to the systems Σ_i with $i < \ell$ (WHILE loop). Note that Σ_ℓ is a subsystem of Σ_i for all $i < \ell$, so that θ_ℓ satisfies at least as many complementary inequalities in Σ_i as in Σ_ℓ. Let i^* be the smallest index i, if any, such that θ_ℓ satisfies a larger number of complementary inequalities in Σ_i than those satisfied by θ_i. Then, the best solution θ_{i^*} for system Σ_{i^*} is set equal to θ_ℓ, and ℓ is reset to i^*. Since the number of data points is finite, the algorithm terminates in a finite number of steps.

Improvements obtained by the proposed modification to the original algorithm are twofold. First, the cardinalities of successively extracted subsystems are not increasing, as one would expect if all MAX FS problems were solved exactly. Second, it favors the construction of subsystems with larger cardinality (*e.g.*, by making it possible to merge subsystems of complementary inequalities that might be satisfied by the same parameter vector, but were extracted at different MAX FS iterations due to the suboptimality of Algorithm 2). The second improvement is also pursued by suitably modifying the algorithm for the MAX FS problem in [15], as will be described in the next subsection.

B. A relaxation algorithm for the MAX FS problem

Given a system of complementary inequalities like (9), the problem of finding a vector θ that makes the inequalities true for as many k as possible, is an extension of the combinatorial problem of finding a feasible subsystem of maximum cardinality of an infeasible system of linear inequalities, which is known as MAX FS problem. Since the MAX FS problem is NP-hard, its extension with complementary inequalities is

TABLE II

MODIFIED RANDOMIZED AND THERMAL RELAXATION ALGORITHM FOR THE MAX FS PROBLEM WITH COMPLEMENTARY INEQUALITIES

Algorithm 2

GIVEN: $C > 0$, $T_0 > 0$, $\theta^{(0)} \in \mathbb{R}^{n+1}$, $\rho \in (0,1)$
Let $j = 0$, $\theta_{best} = \theta^{(0)}$ and $I_{best} = \{k \in I_\ell : |y_k - \varphi_k' \theta_{best}| \leq \delta\}$
FOR $c = 0$ TO $C-1$ DO
  Let $I = I_\ell$ and $T = (1 - c/C)T_0$
  REPEAT
    Let $j = j + 1$
    Pick an index k from I according to the prescribed rule
    Compute v_j^k and λ_j
    Update $\theta^{(j)}$ and let $I^{(j)} = \{k \in I_\ell : |y_k - \varphi_k' \theta^{(j)}| \leq \delta\}$
    IF $\#I^{(j)} > \#I_{best}$ THEN let $\theta_{best} = \theta^{(j)}$ and $I_{best} = I^{(j)}$
    Let $I = I \backslash \{k\}$
  UNTIL $I = \emptyset$
  IF $c > \rho C$ THEN
    Let $\mathcal{D} = \{(y_k, \mathbf{x}_k) : k \in I_{best}\}$
    Let $\theta_{best} = \phi_p(\mathcal{D})$ and $\theta^{(j)} = \theta_{best}$
    Let $I_{best} = \{k \in I_\ell : |y_k - \varphi_k' \theta_{best}| \leq \delta\}$
  END IF
END FOR
RETURN θ_{best}

tackled in [15] by resorting to a randomized and thermal variant of the classical Agmon-Motzkin-Schoenberg relaxation method for solving systems of linear inequalities. In this section, some modifications to the original algorithm [15] are proposed in order to get a feasible subsystem whose cardinality is typically closer to be maximal.

The modified algorithm for the MAX FS problem with complementary inequalities is shown in Table II. It differs from the original version for the addition of the final IF statement. The algorithm requires to define a maximum number of cycles $C > 0$, an initial temperature parameter $T_0 > 0$, an initial estimate $\theta^{(0)} \in \mathbb{R}^{n+1}$ (*e.g.*, randomly generated or computed through least squares), and a coefficient $\rho \in (0,1)$. It consists in a simple iterative procedure generating a sequence $\theta^{(j)}$ of estimates, where $j = 1, \ldots, CN_\ell$ is the iteration counter, and N_ℓ is the number of complementary inequalities of the current subsystem Σ_ℓ of (9) (see Algorithm 1 in Table I). During each of the C outer cycles, all the N_ℓ complementary inequalities of Σ_ℓ are selected in the order defined by a prescribed rule (*e.g.*, cyclicly or uniformly at random without replacement). If k is the index of the complementary inequality considered at iteration j, the current estimate is updated as follows:

$$\theta^{(j)} = \theta^{(j-1)} - \text{sign}(v_j^k)\lambda_j \, \varphi_k, \tag{11}$$

where v_j^k is the violation of the k-th complementary inequality:

$$v_j^k = \begin{cases} \varphi_k'\theta^{(j-1)} - y_k - \delta & \text{if } \varphi_k'\theta^{(j-1)} > y_k + \delta \\ \varphi_k'\theta^{(j-1)} - y_k + \delta & \text{if } \varphi_k'\theta^{(j-1)} < y_k - \delta \\ 0 & \text{otherwise}, \end{cases} \tag{12}$$

and the step size λ_j decreases exponentially with $\left|v_j^k\right|$:

$$\lambda_j = \frac{T}{T_0} \exp^{-\frac{\left|v_j^k\right|}{T}}. \tag{13}$$

Geometrically, the inequality $|y_k - \varphi_k'\theta| \leq \delta$ defines a hyperstrip in the parameter space (see Fig. 3). If the current estimate $\theta^{(j-1)}$ belongs to the hyperstrip (*i.e.* $\theta^{(j-1)}$ satisfies the k-th complementary inequality), then $\theta^{(j)}$ is set equal to $\theta^{(j-1)}$. Otherwise, $\theta^{(j)}$ is obtained by making a step toward the hyperstrip along the line orthogonal to the hyperstrip and passing through $\theta^{(j-1)}$. The basic idea of the algorithm is to favor updates of the current estimate which aim at correcting unsatisfied inequalities with a relatively small violation. Decreasing attention to unsatisfied inequalities with large violations (whose correction is likely to corrupt other inequalities that the current estimate satisfies) is obtained by introducing the decreasing temperature parameter T to which the violations are compared.

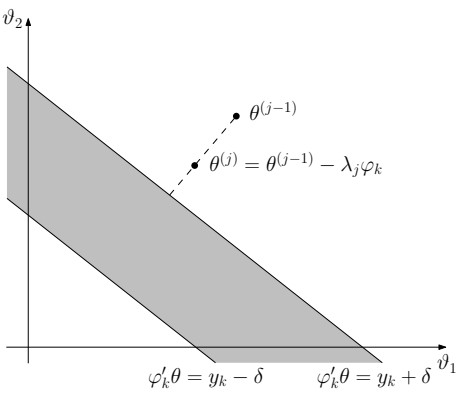

Fig. 3. Geometric interpretation in the parameter space of a single iteration of the relaxation algorithm for the MAX FS problem with complementary inequalities ($\theta \in \mathbb{R}^2$).

If the cycle counter c is greater than ρC (last IF statement), the current best solution θ_{best} (*i.e.* the one that has satisfied the largest number of complementary inequalities so far), is replaced by the ℓ_∞ projection estimate (7). More precisely, denoting by $\mathcal{D}$ the set of data points $(y_k, \mathbf{x}_k)$ such that the corresponding inequality $|y_k - \varphi_k'\theta| \leq \delta$ is in Σ_ℓ and is satisfied by the current θ_{best} (θ_{old} in the following), θ_{best} is updated as follows:

$$\theta_{best} = \arg\min_\theta \max_{(y_k, \mathbf{x}_k) \in \mathcal{D}} |y_k - \varphi_k'\theta|. \tag{14}$$

The new θ_{best} satisfies at least as many complementary inequalities in Σ_ℓ as θ_{old}, since:

$$\max_{(y_k, \mathbf{x}_k) \in \mathcal{D}} |y_k - \varphi_k'\theta_{best}| \leq \max_{(y_k, \mathbf{x}_k) \in \mathcal{D}} |y_k - \varphi_k'\theta_{old}| \leq \delta, \tag{15}$$

and could possibly satisfy more complementary inequalities than θ_{old}, thus providing a better solution of the MAX FS problem for system Σ_ℓ. It was found experimentally that suitable values for ρ lie between 0.7 and 0.8. Indeed, the current solution $\theta^{(j)}$ (and hence the number of satisfied complementary inequalities of Σ_ℓ) would not change significantly as c approaches C, because the temperature parameter T to which the violations are compared becomes smaller and smaller. By resetting $\theta^{(j)}$ to the current best solution (14) at the exit of a cycle when c approaches C, one focuses the future search in a neighborhood of θ_{best}, where it is more likely to satisfy a larger number of complementary inequalities.

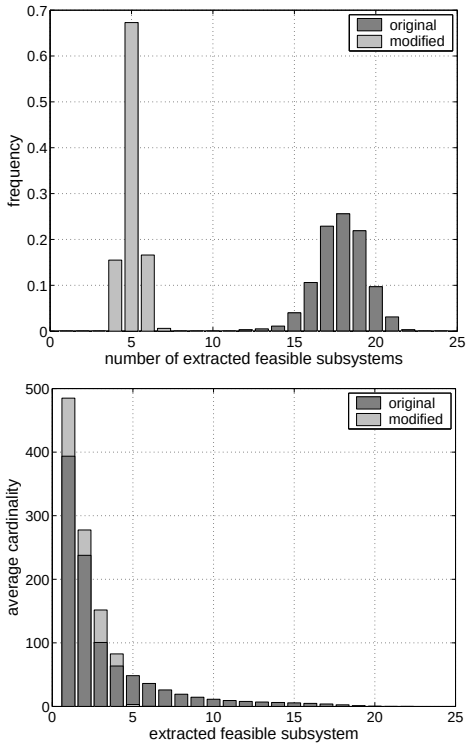

Fig. 4. Results of the application of the original and the modified version of the randomized greedy algorithm for the MIN PFS problem in Example 2. Top: Frequency of the number of extracted feasible subsystems. Bottom: Average cardinality of the extracted feasible subsystems.

The solution θ_{best} returned by the algorithm is the one that, during the overall process, has satisfied the largest number of complementary inequalities. It is however not guaranteed to be optimal, due to the randomness of the search.

For the choice of T_0, the reader is referred to [15]. In general, the larger the value of C, the better the solution, at the price of a longer computation time. Typical good choices for C are $C = 10 \div 20$ [23].

Example 2: In order to show the improvements of the modified randomized greedy algorithm for the MIN PFS problem (Table I and Table II) with respect to the original version [15], the two algorithms are applied to a system of complementary inequalities derived from $N = 1000$ data points generated by a PWARX system with orders $n_a = 2$ and $n_b = 2$, $\tilde{s} = 4$ submodels, and zero-mean gaussian noise with variance $\sigma^2 = 0.2$. $M = 1000$ independent runs of the two algorithms are carried out with the same choice of the parameters $C = 10$, $T_0 = 100$ and $\delta = 1.4 = 3.13\sigma$. In addition, $\rho = 0.7$ is used in the modified algorithm. The frequency of the number of extracted feasible subsystems and their average cardinality over the M trials are computed for both algorithms. The results are shown in Fig. 4. The original algorithm extracts a number of feasible subsystems varying between 12 and 22 over the M trials. The average is 18 subsystems, which is very far from the minimum, namely $\tilde{s} = 4$. Moreover, the variance of the results is quite large. On the other hand, the modified algorithm extracts an average number of 5 feasible subsystems, with the number of feasible subsystems varying between 4 and 7 over the M trials. In addition, the average cardinality of the subsystems extracted after the fourth one is less than the 0.3% of the total number of complementary inequalities. These subsystems, that account for very few data points, can be easily discarded (this issue is addressed in Section IV-B). It is clear from Fig. 4(bottom) that the better performance of the modified algorithm is due to the fact that it is able to extract feasible subsystems of larger cardinality in the first iterations. On the contrary, after extracting some large subsystems, the original algorithm starts to extract many small subsystems. It is worthwhile to note that the average computation time was 7.53 sec for the original algorithm and 6.35 sec for the modified algorithm by running Matlab 6.5 on a 1GHz Intel Pentium III.

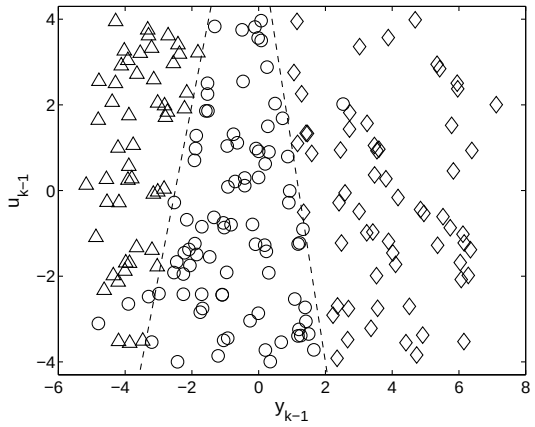

Fig. 5. Initial classification of the regression vectors in Example 1. Each mark corresponds to a different cluster, for a total of three clusters. From left to right, the three clusters consist of 51, 87 and 62 points, respectively. The dashed lines represent the true partition of the regressor set, which is unknown during the identification process.

*Example 1 (*cont'd*):* The initialization of the identification procedure provides $s = 3$ submodels for the given data set with the choice $\delta = 0.2$ (equal to the true bound on the noise). The other parameters of the greedy algorithm are $C = 10$, $T_0 = 100$, $\rho = 0.7$, and cyclic selection of the complementary inequalities is used. Note that the estimated number of submodels equals the true one. The corresponding three clusters of regression vectors are shown in Fig. 5, where some data points marked with circles clearly look as misclassified. They are *undecidable* data points (*i.e.* consistent with more than one submodel), that have been associated by the greedy strategy to the compatible submodel corresponding to the largest feasible subsystem extracted from (9).

C. On the choice of δ

For too large values of δ, very large subsystems of (9) are feasible, and beyond a certain value the whole system (9) becomes feasible, which corresponds to fitting a linear model to the data set. Hence, for large δ the identified PWARX model is simple because it contains very few affine submodels, but the submodels do not fit well the corresponding data points, as large errors are tolerated. Conversely, small values of δ may lead to a very large number of subsystems. In this case overfit

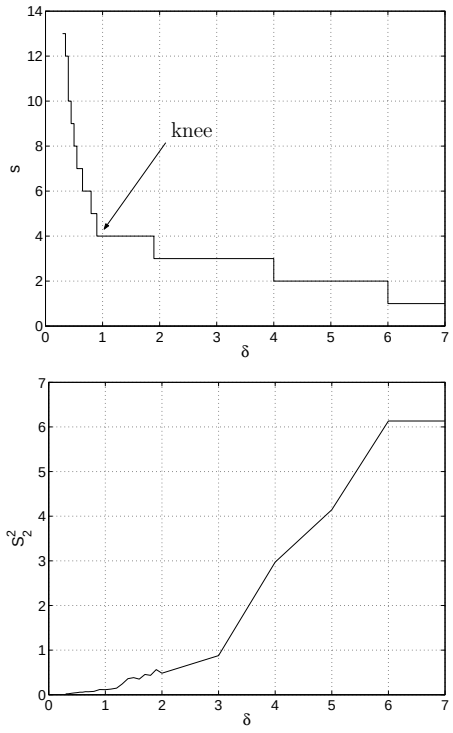

Fig. 6. Number of submodels (top) and average quadratic error (bottom) versus the error bound δ in Example 3.

may occur, *i.e.* the model starts to adjust to the particular noise realization.

When a priori information on the system structure and the noise characteristics is not available, an appropriate value of δ can be selected by solving Problem 2 for different values of δ. Given the limited computational burden of the randomized greedy algorithm for the MIN PFS problem, the curves expressing the number of feasible subsystems of (9) and the average quadratic error:

$$S_2^2 = \frac{1}{N} \sum_{i=1}^{s} \sum_{k \in K_i} |y_k - \varphi_k' \theta_i|^2 \tag{16}$$

as a function of δ, can be easily plotted. Typically, when δ increases starting from a very small value, the number of feasible subsystems decreases first sharply, and then more smoothly after a certain value of δ. Conversely, the average quadratic error increases with δ. An appropriate value of δ should be chosen close to the knee of the first curve, trying to keep the average quadratic error small, as shown in the following example.

Example 3: $N = 500$ data points generated by a PWARX system composed by $\tilde{s} = 4$ subsystems with orders $n_a = 1$ and $n_b = 1$, are considered. The additive noise is normally distributed with zero mean and variance $\sigma^2 = 0.1$, and the signal-to-noise ratio is about 10. The number of feasible subsystems of (9), and the corresponding average quadratic error are plotted as a function of δ in Fig. 6. For values of δ below 0.9 ($\simeq 2.85\sigma$), the average quadratic error is small, but the large number of submodels clearly indicates overfit of the data. For values of δ between 0.9 and 1.9 ($\simeq 6\sigma$), the number of submodels remains constant and equal to the true number $\tilde{s} = 4$, whereas the average quadratic error grows moderately with δ. For values beyond $\delta = 6$, system (9) becomes feasible, and only one submodel is sufficient. It is clear in Fig. 6 that the best trade-off between model accuracy and model complexity is achieved in this example for δ ranging from 0.9 to 1.1.

IV. Refinement

The raw classification and estimation step described in Section III returns an estimate s of the number of submodels, and the sets of indices K_i, $i = 1, \ldots, s$, characterizing the s feasible subsystems extracted from (9). These provide the initial classification of the N data points $(y_k, \mathbf{x}_k)$ into the s clusters $\mathcal{D}_i^{(0)} = \left\{(y_k, \mathbf{x}_k) : k \in K_i\right\}$, $i = 1, \ldots, s$.

Such estimate of the number of affine submodels and classification of the data points may suffer two drawbacks. The major one is that it is not guaranteed to yield minimum partitions, *i.e.* due to both the suboptimality of the greedy approach and the randomness of the algorithm used to tackle each MAX FS problem, the number s of feasible subsystems extracted from (9) might be not minimal. The second drawback is related to a kind of ambiguity that is inherent with the data. Some data points may be consistent with more than one affine submodel, *i.e.* they may satisfy $|y_k - \varphi_k' \theta_i| \leq \delta$ for more than one $i = 1, \ldots, s$. These data points are termed *undecidable*. Due to the undecidable data points, the cardinality and the composition of the feasible subsystems could depend on the order in which they are extracted from (9), as shown in Example 1.

In order to cope with the above drawbacks, a procedure for the refinement of the estimates is presented in Table III. It consists of a basic procedure (steps 2, 4, 5 and 6) whose aim is to improve iteratively both data classification and quality of fit by properly reassigning the data points and updating the parameter estimates. The basic procedure is illustrated in Section IV-A. The additional steps 1 and 3 allow one to reduce the number of submodels by exploiting parameter similarities and cluster cardinalities. These are described in Section IV-B.

A. Dealing with undecidable data

As discussed above, there may exist undecidable data points $(y_k, \mathbf{x}_k)$ that satisfy $|y_k - \varphi_k' \theta_i| \leq \delta$ for more than one index i. Undecidable data points could be classified correctly only by exploiting the partition of the regressor set, which is however not available at this stage of the identification process. When solving Problem 2 via the greedy strategy described in Section III-A, undecidable data points are classified depending on the order in which the feasible subsystems are extracted from (9), as it was also clear in Example 1. As an alternative, each undecidable data point $(y_k, \mathbf{x}_k)$ could be associated a posteriori to the submodel i^* such that the identification error is minimized, namely:

$$i^* = \arg \min_{i=1,\ldots,s} |y_k - \varphi_k' \theta_i| \,. \tag{17}$$

Both criteria may lead to misclassifications when the partition of the regressor set is estimated (see Fig. 7). Thus, in [20],

TABLE III

ALGORITHM FOR THE REFINEMENT OF THE ESTIMATES

Algorithm 3

GIVEN: $\alpha \geq 0$, $\beta \geq 0$, $\gamma > 0$, $c > 0$

Let $t = 1$ and $\theta_i^{(1)} = \phi_p(\mathcal{D}_i^{(0)})$, $i = 1, \ldots, s$

1) *Merge submodels*
   Compute $(i^*, j^*) = \arg \min_{1 \leq i < j \leq s} \alpha_{i,j}^{(t)}$, where $\alpha_{i,j}^{(t)} = \mu(\theta_i^{(t)}, \theta_j^{(t)})$
   IF $\alpha_{i^*,j^*}^{(t)} \leq \alpha$
   THEN merge submodels i^* and j^*, and let $s = s - 1$
2) *Data point reassignment*
   For each data point $(y_k, \mathbf{x}_k)$, $k = 1, \ldots, N$:
   - IF $\left|y_k - \varphi_k' \theta_i^{(t)}\right| \leq \delta$ for only one $i = 1, \ldots, s$ THEN assign $(y_k, \mathbf{x}_k)$ to $\mathcal{D}_i^{(t)}$ and mark it as *feasible*
   - IF $\left|y_k - \varphi_k' \theta_i^{(t)}\right| \leq \delta$ for more than one $i = 1, \ldots, s$ THEN mark $(y_k, \mathbf{x}_k)$ as *undecidable*
   - OTHERWISE mark $(y_k, \mathbf{x}_k)$ as *infeasible*
3) *Discard submodels*
   Compute $i^* = \arg \min_{i=1,\ldots,s} \beta_i^{(t)}$, where $\beta_i^{(t)} = \#\mathcal{D}_i^{(t)}/N$
   IF $\beta_{i^*}^{(t)} \leq \beta$
   THEN discard submodel i^*, let $s = s - 1$ and go to step 2
4) *Assignment of undecidable data points*
   For each undecidable data point $(y_k, \mathbf{x}_k)$:
   Compute $\mathcal{C}_i(\mathbf{x}_k)$, $i = 1, \ldots, s$, and $i^* = \arg \max_{i=1,\ldots,s} \#\mathcal{C}_i(\mathbf{x}_k)$
   IF $\left|y_k - \varphi_k' \theta_{i^*}^{(t)}\right| \leq \delta$
   THEN assign $(y_k, \mathbf{x}_k)$ to $\mathcal{D}_{i^*}^{(t)}$ and mark it as *feasible*
5) *Parameter update*
   Compute $\theta_i^{(t+1)} = \phi_p(\mathcal{D}_i^{(t)})$, $i = 1, \ldots, s$
6) *Termination*
   IF $\left\|\theta_i^{(t+1)} - \theta_i^{(t)}\right\| \leq \gamma \left\|\theta_i^{(t)}\right\|$ for all $i = 1, \ldots, s$
   THEN RETURN s, $\theta_i = \theta_i^{(t+1)}$ and $\mathcal{D}_i = \mathcal{D}_i^{(t)}$, $i = 1, \ldots, s$
   ELSE let $t = t + 1$ and go to step 1

undecidable data points were discarded during the classification process. Although this approach works well in many cases, a non-negligible amount of information is lost when a large number of undecidable data points shows up. Hence, a modification to the classification procedure is proposed here in order to attribute undecidable data points to submodels by exploiting spatial localization. This improves both the data classification (in view of the estimation of the regions) and the parameter estimates.

Initial parameter estimates for each submodel are computed through the ℓ_∞ projection estimator (7). Then, at each iteration indexed by $t = 1, 2, \ldots$, all data points are processed in step 2 of the refinement procedure, and classified as *feasible*, *infeasible* or *undecidable* according to the current estimated parameter vectors $\theta_i^{(t)}$, $i = 1, \ldots, s$. A feasible data point $(y_k, \mathbf{x}_k)$ satisfies the complementary inequality:

$$\left|y_k - \varphi_k' \theta_i^{(t)}\right| \leq \delta \tag{18}$$

for only one $i = 1, \ldots, s$, say i^*. Hence, it can be uniquely associated to the i^*-th submodel, and assigned to the corresponding cluster $\mathcal{D}_{i^*}^{(t)}$. Note that the classification of the feasible data points induces also a classification of the (feasible) regression vectors $\mathbf{x}_k$ into the clusters

$$\mathcal{R}_i^{(t)} = \left\{\mathbf{x}_k \,:\, (y_k, \mathbf{x}_k) \in \mathcal{D}_i^{(t)}\right\}, \quad i = 1, \ldots, s\,. \tag{19}$$

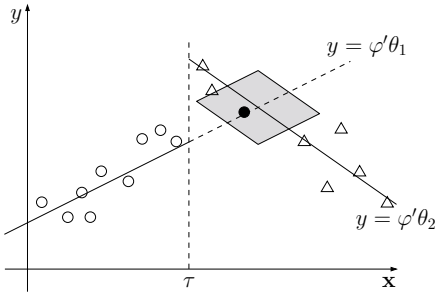

Fig. 7. PWA model with two discrete modes, $\mathbf{x} \in \mathbb{R}$. The gray set represents the region of all possible undecidable data points for a fixed δ. By applying both the greedy strategy for the MIN PFS problem and the criterion (17), the only undecidable data point in the data set (the black circle) is attributed to the first submodel. This yields two non-linearly separable clusters of points.

Infeasible data points do not satisfy (18) for any $i = 1, \ldots, s$. If the corresponding violations are large, they are most likely outliers and are therefore neglected. Undecidable data points satisfy (18) for more than one $i = 1, \ldots, s$, *i.e.* they are consistent with more than one submodel.

Step 4 tries to solve the ambiguity concerned with undecidable data points by exploiting spatial localization in the regressor set. The feasible points around $\mathbf{x}_k$ are indeed expected to provide useful information for correctly classifying the undecidable data point $(y_k, \mathbf{x}_k)$. To this aim, let $\mathcal{C}(\mathbf{x}_k)$ be the set of the c feasible regression vectors nearest to $\mathbf{x}_k$, where c is a fixed positive integer and the Euclidean distance is used. If all points in $\mathcal{C}(\mathbf{x}_k)$ belong to the same cluster $\mathcal{R}_{i^*}^{(t)}$, then $(y_k, \mathbf{x}_k)$ can be most likely associated to submodel i^*, provided that also

$$\left|y_k - \varphi_k' \theta_{i^*}^{(t)}\right| \leq \delta \tag{20}$$

is satisfied. However, $\mathcal{C}(\mathbf{x}_k)$ may in general contain regression vectors from different sets $\mathcal{R}_i^{(t)}$. A candidate submodel is then selected by computing the sets $\mathcal{C}_i(\mathbf{x}_k) = \mathcal{C}(\mathbf{x}_k) \bigcap \mathcal{R}_i^{(t)}$ for all $i = 1, \ldots, s$, and the index i^* such that the cardinality of $\mathcal{C}_i(\mathbf{x}_k)$ is maximized, *i.e.*

$$i^* = \arg \max_{i=1,\ldots,s} \#\mathcal{C}_i(\mathbf{x}_k)\,. \tag{21}$$

If $(y_k, \mathbf{x}_k)$ satisfies (20), then it is associated to the i^*-th submodel and assigned to $\mathcal{D}_{i^*}^{(t)}$, otherwise it is left as undecidable. Extensive tests have shown that this heuristic criterion is very effective in reducing the number of undecidable data points, thus improving data classification. However, it is still conservative, because $(y_k, \mathbf{x}_k)$ is left as undecidable if (20) is not satisfied.

New parameter estimates for each submodel are computed in step 5 through the ℓ_∞ projection estimator (7). The use of the ℓ_∞ projection estimate is favorable here because it guarantees that no feasible data point at refinement t becomes infeasible at refinement $t + 1$, since for all $i = 1, \ldots, s$:

$$\max_{(y_k, \mathbf{x}_k) \in \mathcal{D}_i^{(t)}} \left|y_k - \varphi_k' \theta_i^{(t+1)}\right| \leq \max_{(y_k, \mathbf{x}_k) \in \mathcal{D}_i^{(t)}} \left|y_k - \varphi_k' \theta_i^{(t)}\right| \leq \delta. \tag{22}$$

Good choices for the parameter c in step 4 depend on the density of the data set. A small c may originate sets $\mathcal{C}(\mathbf{x}_k)$

which do not contain enough points for correct classification. On the other hand, for large values of c, a set $\mathcal{C}(\mathbf{x}_k)$ might contain points distant from $\mathbf{x}_k$. In this case, the data point $(y_k, \mathbf{x}_k)$ could be badly assigned to a "far" cluster, or left undecidable. Indeed, if many data points are still classified as undecidable at the exit of the refinement procedure, one can reduce c, and repeat Algorithm 3. The parameter $\gamma > 0$ is the tolerance used to check the termination condition in step 6. A default value could be $\gamma = 0.001$.

B. Reducing the number of submodels

If the initialization procedure provides an overestimation of the number of submodels needed to fit the data, this number can be reduced by exploiting parameter similarities and cluster cardinalities. Two submodels i^* and j^* characterized by similar parameter vectors can be merged in step 1, where

$$\mu(\theta_1, \theta_2) = \frac{\|\theta_1 - \theta_2\|}{\min\{\|\theta_1\|, \|\theta_2\|\}} \tag{23}$$

is used as a measure of the similarity of vectors θ_1 and θ_2. The joined parameter vector at iteration t is computed as $\phi_p(\mathcal{D}_{i^*}^{(t-1)} \bigcup \mathcal{D}_{j^*}^{(t-1)})$. Note that a large number of undecidable data points is likely to show up in step 2 when two parameter vectors are very close. If the cardinality of a cluster of feasible data points is too small, the corresponding submodel (which accounts only for few data) can be discarded in step 3.

The nonnegative thresholds α and β in steps 1 and 3 should be suitably chosen in order to reduce the number of submodels still preserving a good fit of the data. Tentative values for α and β may be chosen after computing $\alpha_{i^*,j^*}^{(1)}$ and $\beta_{i^*}^{(1)}$ in the first iteration of Algorithm 3. If $\alpha_{i^*,j^*}^{(1)} < 0.2$, a rule of thumb is to take $\alpha \simeq 0.8\,\alpha_{i^*,j^*}^{(1)}$. Similarly, if $\beta_{i^*,j^*}^{(1)} < 0.1$, one may take $\beta \simeq 0.8\,\beta_{i^*}^{(1)}$. The user may also choose larger values of α and β so as to impose the reduction of the number of submodels. However, for too large values of α and β, a large number of infeasible data points will typically show up as the number of submodels decreases and some significant submodel is neglected. One can use this information in order to adjust α and β, and then repeat Algorithm 3.

*Example 1 (*cont'd*):* Fig. 8 shows the classification of the regression vectors provided by the refinement procedure. The parameters α and β are not used (a reduction of the number of submodels would result deleterious for the fit), and the other parameters of the procedure are $\gamma = 0.001$ and $c = 5$. The termination condition is reached after three refinements. All data points are correctly classified, and no data point is left undecidable or infeasible. In particular, all undecidable data points are correctly associated to submodels by exploiting spatial localization in the regressor set (compare Fig. 5 and Fig. 8). The parameter vectors estimated for the three submodels are shown in Table IV.

TABLE IV
TRUE ($\tilde{\theta}_i$) AND ESTIMATED (θ_i) PARAMETER VECTORS IN EXAMPLE 1

$\tilde{\theta}_1$	θ_1	$\tilde{\theta}_2$	θ_2	$\tilde{\theta}_3$	θ_3
-0.4	-0.3961	0.5	0.5018	-0.3	-0.2989
1	0.9903	-1	-0.9980	0.5	0.5045
1.5	1.5472	-0.5	-0.4994	-1.7	-1.7072

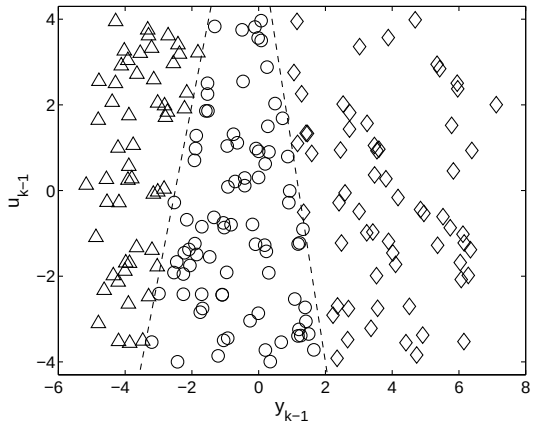

Fig. 8. Classification of the regression vectors (*triangles*, *circles*, *diamonds*) in Example 1 after the refinement. The dashed lines represent the true partition of the regressor set. All data points are correctly classified.

V. REGION ESTIMATION

Given the clusters $\mathcal{D}_i$, $i = 1, \ldots, s$, of feasible data points returned by the refinement procedure, consider the corresponding sets of regression vectors:

$$\mathcal{R}_i = \left\{\mathbf{x}_k \,:\, (y_k, \mathbf{x}_k) \in \mathcal{D}_i\right\}, \quad i = 1, \ldots, s\,. \tag{24}$$

The region estimation problem consists in finding a complete polyhedral partition $\left\{\mathcal{X}_i\right\}_{i=1}^{s}$ of the regressor set $\mathcal{X}$ such that $\mathcal{R}_i \subseteq \mathcal{X}_i$ for all $i = 1, \ldots, s$. The polyhedral regions (4) are defined by hyperplanes. Hence, the problem of region estimation is equivalent to that of separating s sets of points by means of linear classifiers (hyperplanes). Note that a hyperplane separating without errors the points in $\mathcal{R}_i$ from those in $\mathcal{R}_j$, $i \neq j$, might not exist because the sets $\mathcal{R}_i$ and $\mathcal{R}_j$ have intersecting convex hulls. In this case, one will look for a separating hyperplane that minimizes some misclassification index.

Linear separation of the s sets $\mathcal{R}_1, \ldots, \mathcal{R}_s$ can be tackled in two different ways:

$a)$ Construct a linear classifier for each pair $(\mathcal{R}_i, \mathcal{R}_j)$, with $i \neq j$.
$b)$ Construct a piecewise linear classifier which is able to discriminate among s classes.

In the first approach a separating hyperplane is constructed for each pair $(\mathcal{R}_i, \mathcal{R}_j)$, with $i \neq j$. This amounts to solve $s(s-1)/2$ *two-class* linear separation problems. Redundant hyperplanes (*i.e.*, not contributing to the boundary of the corresponding region) can be eliminated a posteriori through standard linear programming techniques, so that the number of linear inequalities defining the i-th polyhedral region is $q_i \leq s-1$. Linear separation of two sets can be accomplished by resorting to, *e.g.*, Robust Linear Programming (RLP) [16] or Support Vector Machines (SVM) [17] methods. Both RLP and SVM look for a separating hyperplane of two sets that

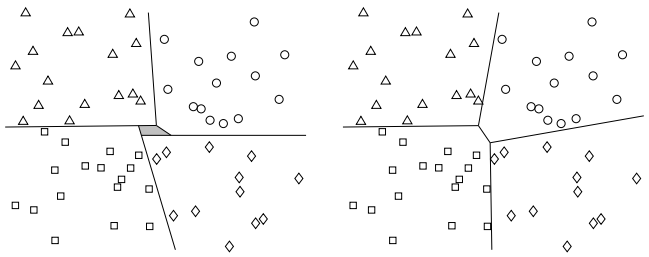

Fig. 9. Linear separation of four sets of points: pairwise linear separation (left) and piecewise linear separation (right). The partition on the left is not complete (the gray area is not covered).

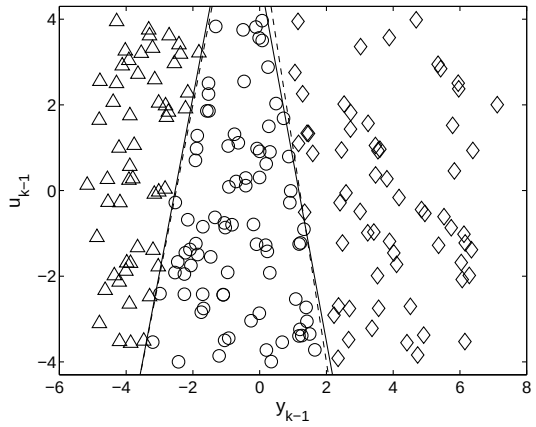

Fig. 10. Final classification of the regression vectors (*triangles*, *circles*, *diamonds*), and true (dashed lines) and estimated (solid lines) partition of the regressor set in Example 1.

additionally minimizes a weighted sum of the misclassification errors. An alternative method is to look for a separating hyperplane that minimizes the number of misclassified points. As detailed in [20], this is equivalent to solving a MAX FS problem. Approach $a)$ is computationally appealing, since it does not involve all the data simultaneously. A major drawback is that the estimated regions are not guaranteed to form a complete partition of the regressor set when $n > 1$, as shown in Fig. 9(left). This drawback is quite important, since it causes the model to be not completely defined over the whole regressor set.

If the presence of "holes" in the partition is not acceptable, approach $b)$ can be employed to solve a *multi-class* linear separation problem, where a piecewise-linear classifier is constructed as the maximum of s linear classification functions. A first way [19], [27], to tackle the multi-class problem is to compute the s linear classifiers by separating each set $\mathcal{R}_i$ from the union of all the others. This requires the solution of s two-class linear separation problems. Unless each set $\mathcal{R}_i$ is linearly separable from the union of the remaining sets, this approach has the drawback that multiply classified points or unclassified points may occur, when all s classifiers are applied to the original data set. This ambiguity is avoided by assigning a point to the class corresponding to the classification function that is maximal at that point. A second way to tackle the multi-class problem is to directly construct s classification functions such that, at each data point, the corresponding class function is maximal. Classical two-class separation methods such as SVM and RLP have been extended to this multi-class case [18], [19]. The resulting methods are called *Multicategory* SVM (M-SVM) or *Multicategory* RLP (M-RLP), to stress their ability of dealing with problems involving more than two classes. Multi-class linear separation problems involve all the available data, and therefore approach $b)$ is computationally more demanding than approach $a)$. For a more detailed overview of several linear separation techniques, see [23].

If a large number of misclassified points shows up when linearly separating two sets $\mathcal{R}_i$ and $\mathcal{R}_j$, it probably means that at least one of the two clusters corresponds to either a nonconvex region (which then needs to be split into convex polyhedra), or nonconnected regions where the submodel is the same. Recall that the classification procedure groups together all the data points that are fitted by the same affine submodel. Efficient techniques for detecting and splitting the clusters corresponding to such situations, are currently under investigation.

Once the regions $\mathcal{X}_i$ have been estimated, all the data points can be finally classified by exploiting both the partition and the bounded-error condition (6). For $k = 1, \ldots, N$, if $\mathbf{x}_k \in \mathcal{X}_i$ for some $i = 1, \ldots, s$, and $|y_k - \varphi_k' \theta_i| \leq \delta$, then $(y_k, \mathbf{x}_k)$ is assigned to the cluster $\mathcal{D}_i$, otherwise it is marked as infeasible. A feasible parameter set:

$$FPS_i = \left\{ \theta \in \mathbb{R}^{n+1} \, : \, |y_k - \varphi_k' \theta| \leq \delta \, , \forall (y_k, \mathbf{x}_k) \in \mathcal{D}_i \right\} \quad (25)$$

can be also associated to the i-th submodel, thus allowing the evaluation of the related parametric uncertainty [13].

*Example 1 (*cont'd*):* The final classification of the regression vectors, and the estimated partition of the regressor set are shown in Fig. 10. The partition is estimated through SVM. The line separating triangles and diamonds is not drawn, since it is redundant, while the two solid lines are defined by the coefficients vectors:

$$\begin{aligned} h_1 &= (3.9591, \, -0.9665, \, 10.0196) \\ h_2 &= (5.0513, \, 1.1876, \, -5.9223) \, , \end{aligned}$$

that are very close to the true ones.

VI. Examples and applications

In this section, the performance of the proposed PWA identification procedure is demonstrated on a numerical example, and on experimental data from an electronic component placement process in a pick-and-place machine. The effective use of the parameter δ of the procedure as a tuning knob to trade off between model complexity and quality of fit is shown.

A. A numerical example

The proposed identification procedure is applied to fit the data generated by a discontinuous PWARX system with orders $n_a = 2$ and $n_b = 2$, and $\tilde{s} = 4$ regions. The input signal is generated according to a uniform distribution on $[-5, 5]$, and the noise signal from a normal distribution with zero mean and variance $\sigma^2 = 0.2$. The estimation data set contains

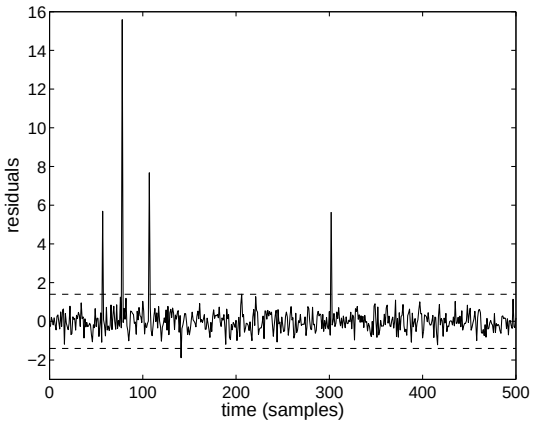

Fig. 11. Plot of the residuals on validation data in the numerical example. The dashed lines limit the interval $[-\delta, \delta]$. Spikes are due to regression vectors incorrectly classified, and to discontinuity of the PWA map.

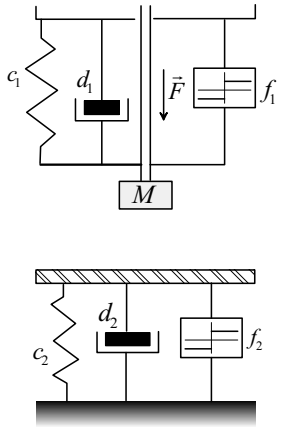

Fig. 12. Identification of the mounting head: Physical model of the experimental setup.

$N = 1000$ data points, of which 292, 234, 361 and 113 are generated by each of the four subsystems, respectively. The SNR is about 17. The bound δ is chosen equal to 1.4, approximately 3.13σ. Since the noise is normally distributed, pointwise estimates of the parameters are computed by least squares. The initialization with $C = 10$, $T_0 = 100$ and $\rho = 0.7$ provides the correct number $s = 4$ of submodels, and clusters containing 363, 287, 229 and 121 data points, respectively. The refinement procedure is run with $c = 10$, and terminates after 5 iterations. The estimated parameter vectors after the refinement are shown in Table V. At this stage, the classification of the data points consists of clusters with 293, 207, 365 and 101 data points, respectively. One data point is infeasible, and only 33 data points out of ≈ 350 are left undecidable. Then, the regions are estimated by M-RLP. The final reassignment of the data points provides clusters with 291, 235, 360 and 113 data points. Only one data point is left infeasible. The 99.7% of the data points are correctly classified.

The model is validated by computing the residuals on $N_{\mathrm{V}} = 500$ validation data. The plot of the residuals is shown in Fig. 11. They are mostly contained in the interval $[-\delta, \delta]$. Recall that the noise follows a normal distribution, and that δ is taken $\approx 3\sigma$. Spikes are due to discontinuity of the PWA map and to regression vectors incorrectly classified because of errors in estimating the regions. For them, the wrong parameter vector is used to compute the prediction. Errors in the estimation of the switching surfaces from a finite data set are in general inevitable. This example shows that such errors can be detected and corrected a posteriori during the validation of the model. When distinct spikes show up in the plot of the residuals, the corresponding data points can be re-attributed, *e.g.*, to the nearest region with a compatible submodel, and the augmented data set used to re-estimate the regions.

Note that this example is quite challenging due to the high number of parameters to be estimated with respect to the available data, and the high number ($\approx 35\%$) of undecidable data. The solution is determined in about 15 s by running Matlab 6.5 on a 1GHz Intel Pentium III.

TABLE V

NUMERICAL EXAMPLE: TRUE ($\tilde{\theta}_i$) AND ESTIMATED (θ_i) PARAMETER VECTORS

$\tilde{\theta}_1$	θ_1	$\tilde{\theta}_2$	θ_2	$\tilde{\theta}_3$	θ_3	$\tilde{\theta}_4$	θ_4
-0.05	-0.0593	1.21	1.2208	1.49	1.4939	-1.20	-1.1838
0.76	0.7818	-0.49	-0.4957	-0.50	-0.4995	-0.72	-0.7275
1.00	1.0081	-0.30	-0.3007	0.20	0.2115	0.60	0.5716
0.50	0.5054	0.90	0.9035	-0.45	-0.4481	-0.70	-0.7013
-0.50	-0.4782	0	0.0242	1.70	1.7451	2.00	1.8076

B. A case study

We apply the proposed identification procedure to model the electronic component placement process described in [22]. The process consists of a mounting head carrying the electronic component. The component is pushed down until it comes in contact with the circuit board, and then is released. A real experimental setup consisting of a mounting head and an impacting surface simulating the printed circuit board, is used to gather the input-output data used for identification. A physical model of the experimental setup is shown in Fig. 12. The mounting head is represented by the mass M, whose movement is only enabled along the vertical axis. The springs c_1 and c_2 simulate elasticity. The dampers d_1 and d_2 provide linear friction, while the blocks f_1 and f_2 provide dry friction. The input to the system is the voltage applied to the motor driving the mounting head, represented by the force F in Fig. 12. The output of the system is the position of the mounting head. The reader is referred to [22] for a more detailed description of the experimental setup.

A data record over an interval of 15 s is available. The considered data set is sampled at 150 Hz. Two modes of operation are excited. In the *free* mode, the mounting head

TABLE VI

IDENTIFICATION OF THE MOUNTING HEAD: FIT BETWEEN THE MEASURED AND THE SIMULATED RESPONSES WITH THE IDENTIFIED PWARX MODELS

	$s = 1$	$s = 2$	$s = 3$	$s = 4$
FIT	78.22%	81.33%	90.18%	93.48%

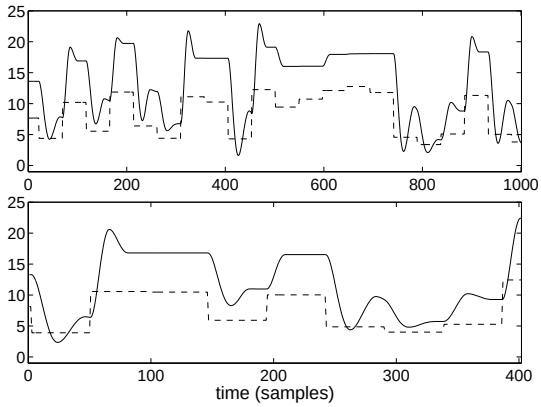

Fig. 13. Data sets used for estimation (top) and validation (bottom) in the identification of the mounting head. The solid and the dashed lines represent the system output and the scaled input, respectively.

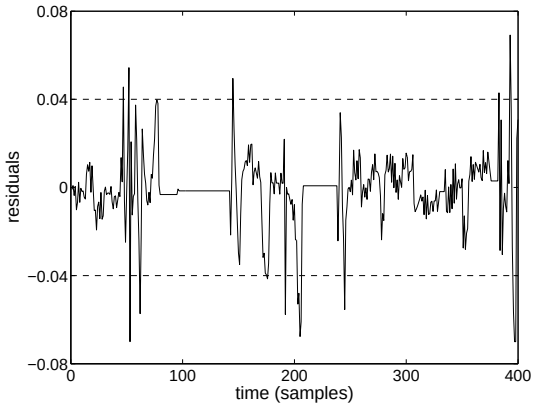

Fig. 14. Identification of the mounting head: Plot of the residuals on validation data using the identified PWARX model with $s = 4$ discrete modes. The dashed lines limit the interval $[-\delta, \delta]$.

moves unconstrained, *i.e.* without being in contact with the impacting surface. In the *impact* mode, the mounting head moves in contact with the impacting surface. Input-output data used for identification and validation are plotted in Fig. 13. Nonlinear phenomena due to dry friction damping are evident in both data sets, *e.g.*, in the upper plot of Fig. 13 on the interval $(500, 750)$. A PWARX model structure with orders $n_a = 2$ and $n_b = 2$ is considered. By choosing $\delta = 0.06$, $\delta = 0.05$, and $\delta = 0.04$, models with $s = 2$, $s = 3$, and $s = 4$ discrete modes, respectively, are identified from $N = 1000$ estimation data. For completeness, also a single ARX model with the same model orders is identified. For $s = 3$ and $s = 4$, M-RLP linear separation techniques are applied in the region estimation step in order to avoid "holes" in the partition. Validation is then carried out by evaluating the fit between the measured and the simulated responses using $N_v = 400$ validation data. By letting $\mathbf{y} = (y_1, \ldots, y_{N_v})$ be the vector of system outputs, $\bar{y}$ the mean value of $\mathbf{y}$, and $\hat{\mathbf{y}} = (\hat{y}_1, \ldots, \hat{y}_{N_v})$ the vector of simulated outputs, the values of the following measure of fit [28]:

$$FIT = 100 \cdot \left(1 - \frac{\|\hat{\mathbf{y}} - \mathbf{y}\|}{\|\mathbf{y} - \bar{y}\|}\right) \% \tag{26}$$

are shown in Table VI for the four identified models. These values demonstrate that the fit improves as the number of submodels increases, *i.e.* as smaller and smaller values of δ are chosen in the identification procedure. We stress that ARX models are also identified for all combinations of the model orders $n_a = 1, \ldots, 20$ and $n_b = 1, \ldots, 20$. The best value of the fit obtained on validation data is $\simeq 80.00\%$ for $n_a = 2$ and $n_b = 18$.

In Fig. 15, the plots of the simulated responses are graphically compared to the measured response. Fig. 15(left) clearly shows that two affine submodels are not sufficient for accurately reproducing the system dynamics. Very good accordance between the measured and the simulated responses is instead obtained with $s = 3$ and $s = 4$ submodels. Difficulties of the identified models in reproducing the nonlinear phenomena on the interval $(210, 240)$ are likely to be due to incomplete information provided by the estimation data. Indeed, in the estimation data set (upper plot of Fig. 13), all significant transitions of the output from low to high values show an overshoot. Consequently, an overshoot shows up in the simulated responses on the intervals $(60, 140)$ and $(210, 240)$, that are both generated by the same sequence of affine submodels, and are caused by large variations of the input signal. It is interesting to note that the identified model with $s = 4$ discrete modes is able to reproduce very faithfully the peak in the interval $(60, 140)$. The discrete mode evolution in Fig. 15(left) clearly shows that one of the two submodels is active in situations of high incoming velocity of the mounting head (*i.e.* rapid transitions from low to high values of the mounting head position). One submodel modelling the same situation is also present in the identified models with $s = 3$ and $s = 4$ discrete modes.

For completeness, the plot of the residuals on validation data using the identified PWARX model with $s = 4$ discrete modes is shown in Fig. 14. Note that the residuals are mostly contained in the interval $[-\delta, \delta]$, although the bound δ on the error cannot be guaranteed on data not used for estimation.

VII. Conclusions

In this paper, a novel procedure for the identification of PWARX models from input-output data has been presented and discussed. The key approach is the selection of a bound δ on the identification error, that enables one to address simultaneously the three issues of data classification, parameter estimation and estimation of the number of submodels via the solution of the MIN PFS problem for a suitable set of linear inequalities derived from data. A refinement procedure improves both data classification and parameter estimation by alternating between data point reassignment and parameter update. In this phase, outliers may be detected and discarded, as well as the ambiguity concerned with undecidable data points may be solved. The final step is the estimation of the partition of the regressor set, that is carried out by resorting to either two-class or multi-class linear separation techniques. The performance of the proposed identification procedure with respect to noise, overestimated model orders and classification

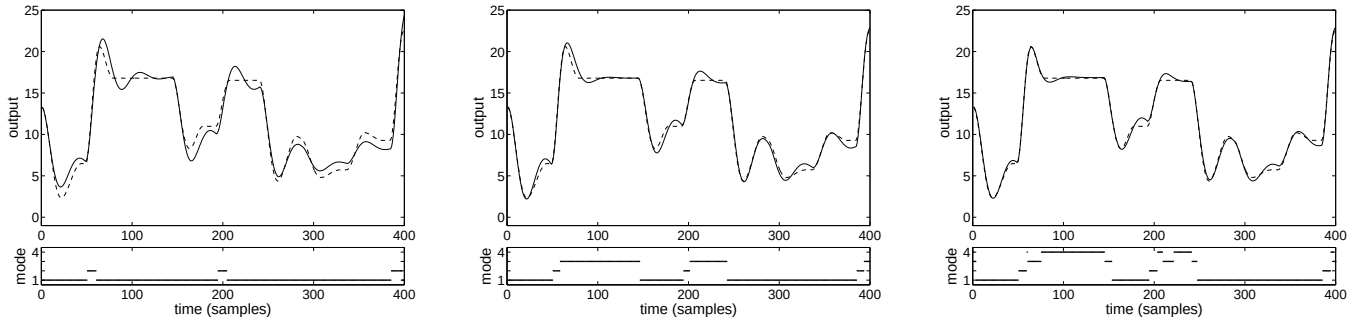

Fig. 15. Simulation results for the mounting head using the identified PWARX models with $s = 2$ (left), $s = 3$ (center), and $s = 4$ (right) discrete modes. Top: Simulated output (solid line) and system output (dashed line). Bottom: Evolution of the discrete mode.

accuracy is analyzed in the recent comparison papers [29], [30] through extensive testing.

Current research concerns the possibility to include in the identification procedure the a priori knowledge on the system to be identified (*e.g.*, saturations, thresholds, deadzones, Wiener or Hammerstein structures), as well as to identify submodels of different orders for each discrete mode. Techniques for efficiently detecting and handling non-convex regions, or non-connected regions where the parameter vector is the same, are also currently investigated.

An interesting issue would be to define suitable criteria for validating the identified PWA models. Classical criteria like residual analysis and whiteness tests could be not satisfactory for this class of models. Since the switching surfaces cannot be determined exactly from a given finite estimation data set, even small errors in estimating the boundaries of the regions might determine large residuals, if the system dynamics is discontinuous. In this respect, it would be also useful to provide bounds on the errors when reconstructing the regions. Experiment design and order selection are other issues of interest. In particular, the choice of the input signal for identification should be such that not only all the affine dynamics are sufficiently excited, but also accurate shaping of the boundaries of the regions is possible.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Aleksandar Juloski (Department of Electrical Engineering, Eindhoven University of Technology, The Netherlands) for providing the data used for identification in Section VI-B. Data are by courtesy of Assembleon, Eindhoven (www.assembleon.com).

REFERENCES

[1] J. Sjöberg, Q. Zhang, L. Ljung, A. Benveniste, B. Delyon, P. Glorennec, H. Hjalmarsson, and A. Juditsky, "Nonlinear black-box modeling in system identification: a unified overview," *Automatica*, vol. 31, no. 12, pp. 1691–1724, 1995.

[2] A. Juditsky, H. Hjalmarsson, A. Benveniste, B. Delyon, L. Ljung, J. Sjöberg, and Q. Zhang, "Nonlinear black-box models in system identification: mathematical foundations," *Automatica*, vol. 31, no. 12, pp. 1725–1750, 1995.

[3] E. D. Sontag, "Nonlinear regulation: The piecewise linear approach," *IEEE Transactions on Automatic Control*, vol. 26, no. 2, pp. 346–358, 1981.

[4] J.-N. Lin and R. Unbehauen, "Canonical piecewise-linear approximations," *IEEE Transactions on Circuits and Systems – I: Fundamental Theory and Applications*, vol. 39, pp. 697–699, Aug. 1992.

[5] L. Breiman, "Hinging hyperplanes for regression, classification, and function approximation," *IEEE Transactions on Information Theory*, vol. 39, pp. 999–1013, May 1993.

[6] A. Bemporad, G. Ferrari-Trecate, and M. Morari, "Observability and controllability of piecewise affine and hybrid systems," *IEEE Transactions on Automatic Control*, vol. 45, no. 10, pp. 1864–1876, 2000.

[7] W. P. M. H. Heemels, B. De Schutter, and A. Bemporad, "Equivalence of hybrid dynamical models," *Automatica*, vol. 37, pp. 1085–1091, July 2001.

[8] J. Roll, *Local and Piecewise Affine Approaches to System identification*. PhD thesis, Department of Electrical Engineering, Linköping University, Linköping, Sweden, 2003. http://www.control.isy.liu.se/publications/.

[9] G. Ferrari-Trecate, M. Muselli, D. Liberati, and M. Morari, "A clustering technique for the identification of piecewise affine systems," *Automatica*, vol. 39, no. 2, pp. 205–217, 2003.

[10] R. Vidal, S. Soatto, Y. Ma, and S. Sastry, "An algebraic geometric approach to the identification of a class of linear hybrid systems," in *Proceedings of the 42nd IEEE Conference on Decision and Control*, (Maui, Hawaii), pp. 167–172, 2003.

[11] J. Ragot, G. Mourot, and D. Maquin, "Parameter estimation of switching piecewise linear systems," in *Proceedings of the 42nd IEEE Conference on Decision and Control*, (Maui, Hawaii), pp. 5783–5788, 2003.

[12] J. Roll, A. Bemporad, and L. Ljung, "Identification of piecewise affine systems via mixed-integer programming," *Automatica*, vol. 40, pp. 37–50, 2004.

[13] M. Milanese and A. Vicino, "Optimal estimation theory for dynamic systems with set membership uncertainty: an overview," *Automatica*, vol. 27, no. 6, pp. 997–1009, 1991.

[14] M. Milanese, J. P. Norton, H. Piet-Lahanier, and E. Walter, eds., *Bounding Approaches to System Identification*. New York: Plenum Press, 1996.

[15] E. Amaldi and M. Mattavelli, "The MIN PFS problem and piecewise linear model estimation," *Discrete Applied Mathematics*, vol. 118, pp. 115–143, 2002.

[16] K. P. Bennett and O. L. Mangasarian, "Robust linear programming discrimination of two linearly inseparable sets," *Optimization Methods and Software*, vol. 1, pp. 23–34, 1992.

[17] C. Cortes and V. N. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, pp. 273–297, 1995.

[18] K. P. Bennett and O. L. Mangasarian, "Multicategory discrimination via linear programming," *Optimization Methods and Software*, vol. 3, pp. 27–39, 1994.

[19] E. J. Bredensteiner and K. P. Bennett, "Multicategory classification by support vector machines," *Computational Optimization and Applications*, vol. 12, pp. 53–79, 1999.

[20] A. Bemporad, A. Garulli, S. Paoletti, and A. Vicino, "A greedy approach to identification of piecewise affine models," in *Hybrid Systems: Computation and Control* (O. Maler and A. Pnueli, eds.), vol. 2623 of *Lecture Notes in Computer Science*, pp. 97–112, Springer Verlag, 2003.

[21] A. Bemporad, A. Garulli, S. Paoletti, and A. Vicino, "Set membership identification of piecewise affine models," in *Proceedings of the 13th IFAC Symposium on System Identification*, (Rotterdam, The Netherlands), pp. 1826–1831, 2003.

[22] A. Lj. 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*, vol. 12, no. 10, pp. 1241–1252, 2004.

[23] S. Paoletti, *Identification of piecewise affine models*. PhD thesis, Dipartimento di Ingegneria dell'Informazione, Università di Siena, Siena, Italy, 2004. http://www.dii.unisi.it/∼paoletti.

[24] E. Amaldi and V. Kann, "The complexity and approximability of finding maximum feasible subsystems of linear relations," *Theoretical Computer Science*, vol. 147, no. 1-2, pp. 181–210, 1995.

[25] E. Amaldi and R. Hauser, "Randomized relaxation methods for the maximum feasible subsystem problem," Tech. Rep. 2001-90, DEI, Politecnico di Milano, Italy, 2001.

[26] M. E. Pfetsch, *The Maximum Feasible Subsystem Problem and Vertex-Facet Incidences of Polyhedra*. PhD thesis, Technischen Universität Berlin, 2002. http://www.zib.de/pfetsch/publications.en.html.

[27] V. N. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer-Verlag, 1995.

[28] L. Ljung, *System Identification Toolbox User's Guide*. The MathWorks, Inc., 6 ed., 2003.

[29] A. Lj. Juloski, W. P. M. H. Heemels, G. Ferrari-Trecate, R. Vidal, S. Paoletti, and J. H. G. Niessen, "Comparison of four procedures for the identification of hybrid systems," in *Hybrid Systems: Computation and Control* (M. Morari and L. Thiele, eds.), vol. 3414 of *Lecture Notes in Computer Science*, pp. 354–369, Springer Verlag, 2005.

[30] A. Lj. Juloski, S. Paoletti, and J. Roll, "Recent techniques for the identification of piecewise affine and hybrid systems," in *Current trends in nonlinear systems and control* (L. Menini, L. Zaccarian, and C. Abdallah, eds.), Birkäuser, 2005. In press.