

# Identification of hybrid systems: a tutorial

Simone Paoletti<sup>1</sup>, Aleksandar Lj. Juloski<sup>2,\*</sup>,  
Giancarlo Ferrari-Trecate<sup>3,\*\*</sup> and René Vidal<sup>4,\*\*\*</sup>

<sup>1</sup> Dipartimento di Ingegneria dell'Informazione, Università di Siena, Via Roma 56, 53100 Siena, Italy;

<sup>2</sup> Department of Electrical Engineering, Eindhoven University of Technology,  
P.O. Box 513, 5600 MB Eindhoven, The Netherlands;

<sup>3</sup> Dipartimento di Informatica e Sistemistica, Università di Pavia, Via Ferrata 1, 27100 Pavia, Italy;

<sup>4</sup> Center for Imaging Science, Department of Biomedical Engineering, Johns Hopkins University,  
3400 N. Charles St., Baltimore MD 21218, USA

*This tutorial paper is concerned with the identification of hybrid models, i.e. dynamical models whose behavior is determined by interacting continuous and discrete dynamics. Methods specifically aimed at the identification of models with a hybrid structure are of very recent date. After discussing the main issues and difficulties connected with hybrid system identification, and giving an overview of the related literature, this paper focuses on four different approaches for the identification of switched affine and piecewise affine models, namely an algebraic procedure, a Bayesian procedure, a clustering-based procedure, and a bounded-error procedure. The main features of the selected procedures are presented, and possible interactions to still enhance their effectiveness are suggested.*

**Keywords:** Hybrid System Identification; Switched Affine and Piecewise Affine Models; Classification; Parameter Estimation; Linear Separation

## 1 Introduction

Hybrid systems are heterogeneous dynamical systems whose behavior is determined by interacting continuous and discrete dynamics. The continuous dynamics is described by variables taking values from a continuous set, while the discrete dynamics is described by variables taking values from a discrete, typically finite, set. The continuous or discrete-valued variables may depend on independent variables such as time, which in turn may be continuous or discrete. Some of the variables can also be discrete-event driven in an asynchronous manner.

Hybrid systems arise not only from the interaction of

logic devices and continuous processes. They can be used to describe real phenomena that exhibit discontinuous behaviors. For instance, the trajectory of a bouncing ball results from the alternation between free fall and elastic contact. Moreover, hybrid models can be used to approximate continuous phenomena by concatenating different models from a simple class. For instance, a nonlinear dynamical system can be approximated by switching among various linear models.

Due to their many potential applications, hybrid systems have attracted increasing attention in the control community during the last decade. Numerous results on analysis, verification, computation, stability and control of hybrid systems have appeared in the literature. However, most of the theoretical developments hinge on the assumption that a hybrid model of the process at hand is available. In some situations it is possible to obtain such a model starting from first principles. On the other hand, first principles modelling is too complicated or even impossible to apply in most practical situations, and the model needs to be identified on the basis of experimental data.

### 1.1 Paper contribution

In the first part, this paper provides an introduction to the topic of hybrid system identification. The main issues and difficulties connected with hybrid system identification are discussed, and an overview of the related literature is given. It is stressed that most effort in this area has been devoted to the identification of switched affine and piecewise affine (PWA) models.

PWA systems are obtained by partitioning the state-input domain into a finite number of non-overlapping convex polyhedral regions, and by considering linear/affine subsystems in each region [66]. The interest in PWA system identification is motivated by several rea-

\*E-mail: A.Juloski@tue.nl

\*\*E-mail: giancarlo.ferrari@unipv.it

\*\*\*E-mail: rvidal@cis.jhu.edu

Correspondence to: S. Paoletti. E-mail: paoletti@dii.unisi.it

sons. The equivalence between PWA models and several classes of hybrid models [4, 34, 67] makes PWA system identification techniques suitable to obtain hybrid models from data. Moreover, the universal approximation properties of PWA maps [14, 49] make PWA models attractive also for nonlinear system identification [64].

Identification of PWA models is a challenging problem that involves the estimation of both the parameters of the affine submodels, and the coefficients of the hyperplanes defining the partition of the state-input domain (or the regressors domain, for models in input-output form). The main difficulty lies in the fact that the identification problem includes a classification problem where each data point must be associated to the most suitable submodel. Concerning the partitioning, two alternative approaches can be distinguished:

1. the partition is fixed a priori;
2. the partition is estimated along 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 must be 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 harder when also the number of submodels must be estimated.

Different techniques leading to PWA models of smooth dynamical systems can be found in the extensive literature on nonlinear black-box identification. A nice overview is presented in [61]. However, most of these approaches assume that the system dynamics is continuous. Recently, novel contributions allowing for discontinuities have been proposed in both the hybrid systems and the nonlinear identification communities. An iterative algorithm that sequentially estimates the parameters of the model and classifies the data through the use of adapted weights is described in [60]. A method based on statistical clustering of measured data via a Gaussian mixture model and support vector classifiers is presented in [56]. Several optimization problem formulations of the identification problem are proposed in [54, 55]. In [62] 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. Subspace identification of piecewise linear systems is addressed in [10, 71], while recursive identification of switched hybrid system is addressed in [32, 75].

Among the variety of the proposed approaches, the authors of this paper contributed by developing four different procedures for the identification of switched affine and piecewise affine models, namely the algebraic procedure [78], the clustering-based procedure [27], the Bayesian procedure [47], and the bounded-error procedure [5]. These techniques have been successfully applied in real problems, such as identification of the electronic component placement process in pick-and-place machines [5, 43, 47], modelling of a current transformer [27], traction control [11], and motion segmentation in computer vision [76, 77]. The main features of the selected techniques are presented in the second part of the paper, where possible interactions to still enhance their effectiveness are also suggested.

## 1.2 Paper outline

This paper is organized as follows. Section 2 introduces the classes of switched affine and piecewise affine models, both in state space and input-output form. Section 3 reports several formulations of the identification problem for these model classes, and gives an overview of the related literature. Different identification approaches are classified along the lines proposed in [61]. The problems of data classification and region estimation are addressed in Section 4 for those approaches that firstly classify the data, then estimate the affine dynamics, and finally reconstruct the polyhedral partition. Most recent contributions for the identification of models with hybrid and discontinuous characteristics belong to this category. Four procedures falling into the category analyzed in Section 4, are finally described and discussed in Section 5. Section 6 draws the conclusions, and foreshadows interesting topics for future research.

## 2 Switched affine and piecewise affine models

In this section, discrete-time switched affine models both in state space and in input-output form are introduced. Switched affine models are defined as collections of linear/affine models, connected by switches that are indexed by a discrete-valued additional variable, called the discrete state. Models for which the discrete state is determined by a polyhedral partition of the state-input domain, are called piecewise affine models. They can be used to model a large number of physical processes (see, e.g. [3, 17, 18, 48, 69]), and are suitable to approximate nonlinear dynamics, e.g., via multiple linearizations at different operating points. Moreover, piecewise affine models are equivalent to several classes of hybrid models, and can

therefore be used to describe systems that exhibit hybrid structure.

## 2.1 Models in state space form

A general discrete-time switched affine model in *state space* form is described by the equations

$$\begin{aligned} \mathbf{x}_{k+1} &= A_{\sigma(k)} \mathbf{x}_k + B_{\sigma(k)} \mathbf{u}_k + f_{\sigma(k)} + \mathbf{w}_k \\ \mathbf{y}_k &= C_{\sigma(k)} \mathbf{x}_k + D_{\sigma(k)} \mathbf{u}_k + g_{\sigma(k)} + \mathbf{v}_k, \end{aligned} \quad (1)$$

where  $\mathbf{x}_k \in \mathbb{R}^n$ ,  $\mathbf{u}_k \in \mathbb{R}^p$  and  $\mathbf{y}_k \in \mathbb{R}^q$  are, respectively, the *continuous* state, the input and the output of the system at time  $k \in \mathbb{Z}$ , and  $\mathbf{w}_k \in \mathbb{R}^n$  and  $\mathbf{v}_k \in \mathbb{R}^q$  are noise/error terms. The *discrete* state  $\sigma(k)$ , describing in what affine dynamics the system is at time  $k$ , is assumed to take only a finite number of values, i.e.  $\sigma(k) \in \{1, \dots, s\}$ , where  $s$  is the number of affine submodels. In general,  $\sigma(k)$  can be a function of  $k$ ,  $\mathbf{x}_k$ ,  $\mathbf{u}_k$ , or some other external input. The real matrices/vectors  $A_i$ ,  $B_i$ ,  $f_i$ ,  $C_i$ ,  $D_i$  and  $g_i$ ,  $i = 1, \dots, s$ , having appropriate dimensions, describe each affine dynamics. Hence, model (1) can be seen as a collection of affine models with continuous state  $\mathbf{x}_k$ , connected by switches that are indexed by the discrete state  $\sigma(k)$ .

The evolution of the discrete state can be described in a variety of ways. In *Jump Linear* (JL) models,  $\sigma(k)$  is an unknown, deterministic and finite-valued input. In *Jump-Markov Linear* (JML) models, the dynamics of  $\sigma(k)$  is modelled as an irreducible Markov chain governed by the transition probabilities  $\pi(i, j) \triangleq P(\sigma(k+1) = j | \sigma(k) = i)$ . In *PieceWise Affine* (PWA) models [66],  $\sigma(k)$  is given by the rule

$$\sigma(k) = i \quad \text{iff} \quad \begin{bmatrix} \mathbf{x}_k \\ \mathbf{u}_k \end{bmatrix} \in \Omega_i, \quad i = 1, \dots, s, \quad (2)$$

where  $\{\Omega_i\}_{i=1}^s$  is a complete partition<sup>1</sup> of the state-input domain  $\Omega \subseteq \mathbb{R}^{n+p}$ . The regions  $\Omega_i$  are assumed to be convex polyhedra described by

$$\Omega_i = \left\{ \begin{bmatrix} \mathbf{x} \\ \mathbf{u} \end{bmatrix} \in \mathbb{R}^{n+p} : \bar{H}_i \begin{bmatrix} \mathbf{x} \\ \mathbf{u} \\ 1 \end{bmatrix} \preceq_{[i]} \mathbf{0} \right\}, \quad (3)$$

where  $\bar{H}_i \in \mathbb{R}^{\bar{\mu}_i \times (n+p+1)}$ ,  $i = 1, \dots, s$ , and  $\bar{\mu}_i$  is the number of linear inequalities defining the  $i$ th polyhedral region  $\Omega_i$ . With abuse of notation, in (3) the symbol  $\preceq_{[i]}$  denotes a  $\bar{\mu}_i$ -dimensional vector whose elements can be the symbols  $\leq$  and  $<$  in order to avoid that the regions  $\Omega_i$  overlap over common boundaries.

<sup>1</sup>A collection  $\{\mathcal{A}_i\}_{i=1}^s$  is said to be a (complete) partition of  $\mathcal{A} \subseteq \mathbb{R}^m$  if  $\cup_{i=1}^s \mathcal{A}_i = \mathcal{A}$  and  $\mathcal{A}_i \cap \mathcal{A}_j = \emptyset, \forall i \neq j$ .

**Remark 2.1** PWA models form a special class of hybrid models. Other descriptions for hybrid systems include *Mixed Logical Dynamical* (MLD) models [6], *Linear Complementarity* (LC) models [33, 70], *Extended Linear Complementarity* (ELC) models [20], and *Max-Min-Plus-Scaling* (MMPS) models [21]. Equivalences among these five classes of systems are shown in [4, 34]. Such results are very important for transferring theoretical properties and tools (e.g., control and identification techniques) from one class to another, as they imply that one can choose the most convenient hybrid modelling framework for the study of a particular hybrid system.

## 2.2 Models in input-output form

For fixed model orders  $n_a$  and  $n_b$ , a *Switched AutoRegressive eXogenous* (SARX) model is defined by introducing the regression vector

$$\mathbf{r}_k = \left[ \mathbf{y}_{k-1}^\top \cdots \mathbf{y}_{k-n_a}^\top \mathbf{u}_k^\top \mathbf{u}_{k-1}^\top \cdots \mathbf{u}_{k-n_b}^\top \right]^\top, \quad (4)$$

and then by expressing the output  $\mathbf{y}_k$  as a piecewise affine function of  $\mathbf{r}_k$ , namely

$$\mathbf{y}_k = \theta_{\sigma(k)}^\top \begin{bmatrix} \mathbf{r}_k \\ 1 \end{bmatrix} + \mathbf{e}_k, \quad (5)$$

where  $\sigma(k) \in \{1, \dots, s\}$  is the discrete state,  $s$  is the number of submodels,  $\theta_i$ ,  $i = 1, \dots, s$ , are the matrices of parameters defining each submodel, and  $\mathbf{e}_k \in \mathbb{R}^q$  is a noise/error term. In the following, the vector  $\varphi_k = \begin{bmatrix} \mathbf{r}_k \\ 1 \end{bmatrix}$  will be called the *extended* regression vector.

SARX models represent a subclass of the switched affine models (1), and can be easily transformed into that form by defining the continuous state as

$$\mathbf{x}_k = \left[ \mathbf{y}_{k-1}^\top \cdots \mathbf{y}_{k-n_a}^\top \mathbf{u}_{k-1}^\top \cdots \mathbf{u}_{k-n_b}^\top \right]^\top. \quad (6)$$

As for the models in state space form, the evolution of the discrete mode  $\sigma(k)$  can be described in a variety of ways. In *PieceWise AutoRegressive eXogenous* (PWARX) models the switching mechanism is determined by a polyhedral partition of the regressors domain  $\mathcal{R} \subseteq \mathbb{R}^d$ , where  $d = q \cdot n_a + p \cdot (n_b + 1)$ . This means that for these models the discrete state  $\sigma(k)$  is given by

$$\sigma(k) = i \quad \text{iff} \quad \mathbf{r}_k \in \mathcal{R}_i, \quad i = 1, \dots, s, \quad (7)$$

where  $\{\mathcal{R}_i\}_{i=1}^s$  is a complete partition of  $\mathcal{R}$ . Each region  $\mathcal{R}_i$  is a convex polyhedron described by

$$\mathcal{R}_i = \left\{ \mathbf{r} \in \mathbb{R}^d : H_i \begin{bmatrix} \mathbf{r} \\ 1 \end{bmatrix} \preceq_{[i]} \mathbf{0} \right\} \quad (8)$$

where  $H_i \in \mathbb{R}^{\mu_i \times (d+1)}$ ,  $i = 1, \dots, s$ ,  $\mu_i$  is the number of linear inequalities defining the  $i$ th polyhedral region  $\mathcal{R}_i$

and, as in (3), the symbol  $\preceq_{[i]}$  denotes a  $\mu_i$ -dimensional vector whose elements can be the symbols  $\leq$  and  $<$ . In general, the shape of  $\mathcal{R}$  reflects the physical constraints on the inputs and the outputs of the system. For instance, typical constraints on the output can be  $\|\mathbf{y}_k\|_\infty \leq y_{max}$  or  $\|\mathbf{y}_k - \mathbf{y}_{k-1}\|_\infty \leq \Delta y_{max}$ , where  $\|\cdot\|_\infty$  is the infinity norm of a vector.

By introducing the piecewise affine map

$$f(\mathbf{r}) = \begin{cases} \theta_1^\top \varphi & \text{if } H_1 \varphi \preceq_{[1]} \mathbf{0} \\ \vdots & \vdots \\ \theta_s^\top \varphi & \text{if } H_s \varphi \preceq_{[s]} \mathbf{0}, \end{cases} \quad (9)$$

with  $\varphi = [\mathbf{r}_1^\top]$ , it will be useful to rewrite the model defined by (5), (7) and (8) as

$$\mathbf{y}_k = f(\mathbf{r}_k) + \mathbf{e}_k. \quad (10)$$

**Remark 2.2** The PWA map (9) can be discontinuous along the boundaries defined by the polyhedra (8), as shown in Figure 1. Though, for the sake of simplicity, in the following the subscript  $[i]$  will be removed from the notation  $\preceq_{[i]}$ , one must always take care of the definition of the regions, to avoid that the PWA map is multiply defined over common boundaries of the regions  $\mathcal{R}_i$ .

### 3 Hybrid system identification

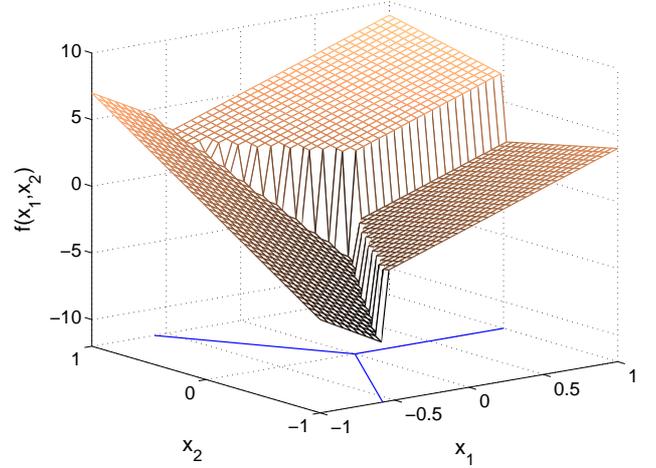
In this section, the identification problem will be firstly addressed for input-output models, and then for state space models. An overview of the related literature is finally presented. For the sake of clarity, single input-single output systems (i.e.  $p = q = 1$ ) are considered. To this aim, notations  $y_k$ ,  $u_k$  and  $e_k$  will be used instead of  $\mathbf{y}_k$ ,  $\mathbf{u}_k$  and  $\mathbf{e}_k$ . The discussion can be straightforwardly extended to multi input-single output systems (i.e.  $p > 1$  and  $q = 1$ ). Multi input-multi output systems (i.e.  $p > 1$  and  $q > 1$ ) are also handled by state-space techniques, while in the input-output case one can identify a model for each output by considering the other outputs as additional inputs<sup>2</sup>.

#### 3.1 Identification problem for SARX models

For SARX models (5), the general identification problem reads as follows.

**Problem 3.1** Given a collection of  $N$  input-output pairs  $(y_k, u_k)$ ,  $k = 1, \dots, N$ , estimate the model orders  $n_a$  and

<sup>2</sup>Though this approach may lead in general to a larger number of regions than necessary, since the overall partition is obtained by intersecting the partitions of the single models.



**Figure 1.** Discontinuous PWA map of two variables with  $s = 3$  regions.

$n_b$ , the number of submodels  $s$ , and the parameter vectors  $\theta_i$ ,  $i = 1, \dots, s$ . Moreover, estimate the discrete state  $\sigma(k)$  for  $k > \max\{n_a, n_b\}$ .

If the system generating the data has the structure (5), an exact algebraic solution to Problem 3.1 is presented in [51, 74, 78] for the case of noiseless data (though the approach can be amended to work also with noisy data). The algorithm only requires to fix upper bounds  $\bar{n}_a$ ,  $\bar{n}_b$ , and  $\bar{s}$  on the model orders and the number of submodels, respectively. A description of the algorithm will be given in Section 5.

If the model orders are fixed, the problem is to fit the data to  $s$  hyperplanes. This problem is addressed in the field of data analysis, and several approaches are proposed where  $s$  is either estimated from data or fixed a priori. One way to estimate  $s$  is by solving the following problem.

**Problem 3.2** Given  $\delta > 0$ , find the smallest number  $s$  of vectors  $\theta_i$ ,  $i = 1, \dots, s$ , and a mapping  $k \mapsto \sigma(k)$  such that

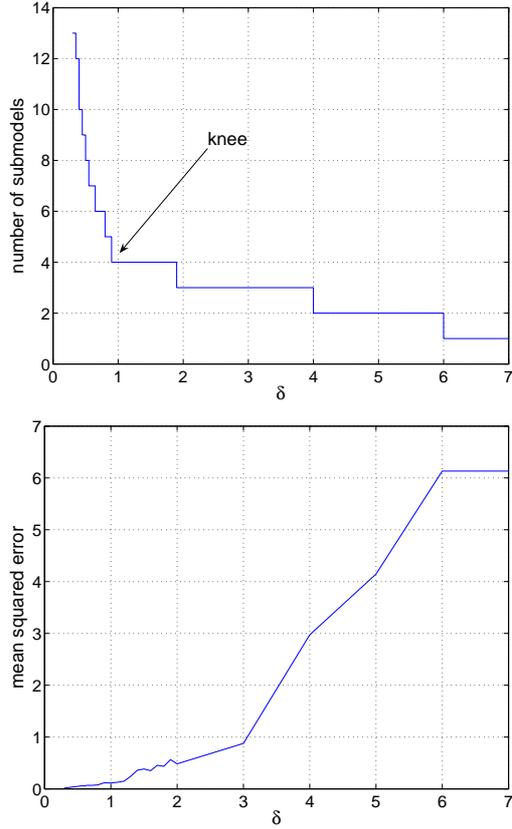
$$|y_k - \varphi_k^\top \theta_{\sigma(k)}| \leq \delta \quad (11)$$

for all  $k = \bar{n}, \dots, N$ , where  $\bar{n} = \max\{n_a, n_b\} + 1$ .

Problem 3.2 consists in finding a *Partition* of the system of inequalities

$$|y_k - \varphi_k^\top \theta| \leq \delta, \quad k = \bar{n}, \dots, N, \quad (12)$$

into a *Minimum* number of *Feasible Subsystems* (MIN PFS problem). The bound  $\delta$  in (12) is not necessarily given a priori (e.g., if the noise is bounded, and the bound



**Figure 2.** Number of submodels and mean squared error as a function of  $\delta$  for a data set generated by a SARX system with four discrete states and Gaussian additive noise with zero mean and variance  $\sigma^2 = 0.1$ .

is known), rather it can be adjusted in order to find the desired trade off between fit and accuracy. In fact, the smaller  $\delta$ , the larger is typically the number of submodels needed to fit the data<sup>3</sup>, while on the other hand, the larger  $\delta$ , the worse is the fit, since larger errors are allowed. Figure 2 shows two typical plots of the number of submodels and the Mean Squared Error (MSE) as a function of  $\delta$  when solving Problem 3.2 for a given data set. The choice of a suitable  $\delta$  is typically made at the knee of the  $s$ -curve, where also the MSE is kept low. The MIN PFS problem is NP-hard, and a suboptimal greedy randomized algorithm to tackle its solution is proposed in [1].

If  $s$  is fixed, the well-known optimization approach used in linear system identification (i.e. choose the parameters of a linear model such that they minimize some prediction error norm) can be generalized to the identification of SARX models. Given a nonnegative function

<sup>3</sup>In this case overfit may occur, i.e. the model adjusts to the particular noise realization.

$\ell(\cdot)$ , such as  $\ell(\varepsilon) = \varepsilon^2$  or  $\ell(\varepsilon) = |\varepsilon|$ , the estimation of the parameter vectors  $\theta_i$ ,  $i = 1, \dots, s$ , and of the discrete state  $\sigma(k)$  can be in fact formulated as the following optimization problem:

$$\begin{cases} \min_{\theta_i, \chi_{k,i}} & \sum_{k=\bar{n}}^N \sum_{i=1}^s \ell(y_k - \varphi_k^\top \theta_i) \chi_{k,i} \\ \text{s.t.} & \sum_{i=1}^s \chi_{k,i} = 1 \quad \forall k \\ & \chi_{k,i} \in \{0, 1\} \quad \forall k, i. \end{cases} \quad (13)$$

In (13), each binary variable  $\chi_{k,i}$  describes whether the data point  $(y_k, \mathbf{r}_k)$  is associated to the  $i$ th submodel, under the constraint that each data point must be associated to only one submodel. The discrete state  $\sigma(k)$  can be finally reconstructed according to the rule:

$$\sigma(k) = i \quad \text{iff} \quad \chi_{k,i} = 1. \quad (14)$$

The optimization problem in (13) is a *mixed integer* program that is computationally intractable, except for small instances. In principle, branch and bound algorithms could be applied, but the search tree increases exponentially with the number of data  $N$  and the number of submodels  $s$ . It is shown in [55] that (13) can be transformed into a smooth constrained optimization problem by relaxing the integer constraints, i.e. by requiring  $\chi_{k,i} \in [0, 1]$ ,  $\forall k, i$ . The global optimum of the relaxed problem coincides with the global optimum of (13). Moreover, an integer solution can be readily obtained from the solution of the relaxed problem. By the same reasoning, it is also shown that (13) can be transformed into the following non-smooth unconstrained optimization problem:

$$\min_{\theta_i} \sum_{k=\bar{n}}^N \min_{i=1, \dots, s} \ell(y_k - \varphi_k^\top \theta_i). \quad (15)$$

In order to not get trapped in a local minimum, suitable optimization techniques must be used to tackle the solution of the equivalent problems. It is reported in [55] that state-of-the-art solvers, such as [38], are able to solve (15) in reasonable time at least for sample problems.

An alternative to the formulation (13) is the clustering algorithm proposed in [12], which groups the given data points into  $s$  clusters by generating  $s$  planes that represent a local solution to the non-convex problem of minimizing the sum of squares of the 2-norm distances between each point and a nearest plane.

### 3.2 Identification problem for PWARX models

For PWARX models defined by (5), (7) and (8), the general identification problem reads as follows.

**Problem 3.3** Given a collection of  $N$  input-output pairs  $(y_k, u_k)$ ,  $k = 1, \dots, N$ , estimate the model orders  $n_a$  and  $n_b$ , the number of submodels  $s$ , the parameter vectors  $\theta_i$  and the regions  $\mathcal{R}_i$ ,  $i = 1, \dots, s$ .

Note that, in the case of piecewise affine models, the partition of the regressors domain automatically implies the estimation of the discrete state according to (7).

All techniques specifically developed for the identification of PWARX models, assume fixed orders  $n_a$  and  $n_b$ . The estimation of the model orders can be based on preliminary data analysis, and carried out by algebraic techniques such as [51, 74], or classical model order selection techniques (see [50]). Hence, in the following the orders  $n_a$  and  $n_b$  are given, and  $\bar{n} = \max\{n_a, n_b\} + 1$ .

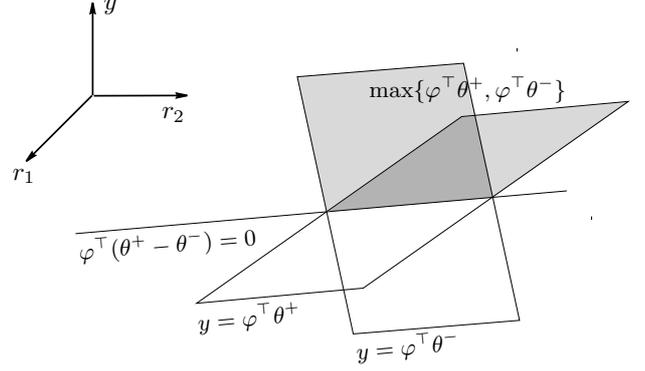
The considered identification problem consists in finding the PWARX model that best matches the given data according to a specified criterion of fit. It involves the estimation of:

- The number of discrete states  $s$ .
- The parameters  $\theta_i$ ,  $i = 1, \dots, s$ , of the affine submodels.
- The coefficients  $H_i$ ,  $i = 1, \dots, s$ , of the hyperplanes defining the partition of the regressors set.

This issue also underlies a classification problem such that each data point is associated to one region, and to the corresponding submodel. The simultaneous optimal estimation of all the quantities mentioned above is a very hard, computationally intractable problem. To the best of the authors' knowledge, no satisfactory formulation in the form of a single optimization problem has been even provided for it. One of the main concerns is how to choose  $s$  in a sensible way. For instance, perfect fit is obtained by letting  $s = N$ , i.e. one submodel per each data point, which is clearly an inadequate solution. Penalties on increasing  $s$  should be therefore introduced in order to keep the number of submodels reasonably low, and to avoid overfit because the model is given too many degrees of freedom. An additional difficulty is how to express efficiently the constraint that the collection  $\{\mathcal{R}_i\}_{i=1}^s$  must form a complete partition of the regressors domain  $\mathcal{R}$ .

The problem becomes easy if the number of discrete states  $s$  is fixed, and the regions (8) are either known or fixed *a priori*. In that case each regression vector  $\mathbf{r}_k$  can be associated to one submodel according to (7). Hence, by introducing the quantities

$$\chi_{k,i} = \begin{cases} 1 & \text{if } \mathbf{r}_k \in \mathcal{R}_i \\ 0 & \text{otherwise} \end{cases} \quad \forall k, i, \quad (16)$$



**Figure 3.** Two hinging hyperplanes  $y = \varphi^\top \theta^-$  and  $y = \varphi^\top \theta^+$ , and the corresponding hinge function  $y = \max\{\varphi^\top \theta^+, \varphi^\top \theta^-\}$ , where  $\varphi = [r_1 \ r_2 \ 1]^\top$ .

the identification problem reduces to the following optimization problem:

$$\min_{\theta_i} \frac{1}{N} \sum_{k=\bar{n}}^N \sum_{i=1}^s \ell(y_k - \varphi_k^\top \theta_i) \chi_{k,i}, \quad (17)$$

where  $\ell(\cdot)$  is a given nonnegative function. If  $\ell(\varepsilon) = \varepsilon^2$ , (17) is an ordinary least-squares problem in the unknowns  $\theta_i$ .

In [61, 62] the identification problem is reformulated in the form of mixed integer linear or quadratic programs for the class of Hinging-Hyperplane ARX (HHARX) models [14], which are described by

$$y_k = f(\mathbf{r}_k; \theta) + e_k$$

$$f(\mathbf{r}_k; \theta) = \varphi_k^\top \theta_0 + \sum_{i=1}^M \sigma_i \max\{\varphi_k^\top \theta_i, 0\}, \quad (18)$$

where  $\theta = [\theta_0^\top \ \theta_1^\top \ \dots \ \theta_M^\top]^\top$ , and  $\sigma_i \in \{-1, 1\}$  are fixed a priori. It is easy to see that HHARX models are a subclass of PWARX models for which the PWA map (9) is continuous. The number of submodels  $s$  is bounded by the quantity  $\sum_{j=0}^d \binom{M}{j}$ , which only depends on the length  $d$  of the regression vector, and the number  $M$  of hinge functions (see Fig. 3). The identification problem considered in [62] selects the optimal parameter vector  $\theta^*$  by solving

$$\theta^* = \arg \min_{\theta} \sum_{k=\bar{n}}^N |y_k - f(\mathbf{r}_k; \theta)|^p, \quad (19)$$

where  $p = 1$  or  $2$ . Assuming a priori known bounds on  $\theta$  (which can be taken arbitrarily large), (19) can be reformulated as a mixed-integer linear or quadratic program

(MILP/MIQP) by introducing auxiliary continuous variables  $z_i(k) = \max\{\varphi_k^\top \theta_i, 0\}$ , and binary variables

$$\delta_i(k) = \begin{cases} 0 & \text{if } \varphi_k^\top \theta_i \leq 0 \\ 1 & \text{otherwise.} \end{cases} \quad (20)$$

The MILP/MIQP problems can then be solved for the global optimum. The optimality of the described approach comes at the cost of a theoretically very high worst-case computational complexity, which means that it is mainly suitable for small-scale problems (e.g., when it is very costly to obtain data). To be able to handle somewhat larger problems, different suboptimal approximations are proposed in [61]. Various extensions are also possible for handling non-fixed  $\sigma_i$ , discontinuities, general PWARX models, etc., again at the cost of increased computational complexity.

Most of the heuristic and suboptimal approaches that are applicable, or at least related, to the identification of PWARX models, either assume a fixed  $s$ , or adjust  $s$  iteratively (e.g., by adding one submodel at a time) in order to improve the fit. A few techniques allow for the automatic estimation of  $s$  from data. An overview of the related literature is presented in Section 3.4.

### 3.3 Identification problem for state space models

For switched affine models defined by (1), or piecewise affine models defined by (1), (2) and (3), the general identification problem reads as follows.

**Problem 3.4** Given a collection of  $N$  input-output pairs  $(y_k, u_k)$ ,  $k = 1, \dots, N$ , estimate the model order  $n$ , the number of submodels  $s$ , and the 6-tuples  $(A_i, B_i, f_i, C_i, D_i, g_i)$ ,  $i = 1, \dots, s$ . Moreover, estimate the discrete state  $\sigma(k)$ ,  $k = 1, \dots, N$ , and, if the model is piecewise affine, the regions  $\Omega_i$ ,  $i = 1, \dots, s$ .

As for the models in input-output form, the difficulty of Problem 3.4 depends on which quantities are assumed to be known. Nevertheless, while for SARX/PWARX models the identification problem is easy if all the quantities (including the switching sequence) are known, and only the parameters of the submodels must be estimated, an additional difficulty arises when dealing with the identification of state space models. If the switching sequence is known, the matrices of each submodel can still be estimated by classical techniques such as subspace identification methods. However, as pointed out in [71], the matrices of the submodels are obtained up to a linear state transformation. This state transformation is different, in

general, for each of the submodels. To combine the submodels they need to be transformed into the same state basis. In [71] it is discussed how the transitions between the submodels can be used to this aim. The algorithm requires a sufficiently large number of transitions for which the states at the transition are linearly independent.

Heuristics and suboptimal techniques for the identification of switched and piecewise affine state space models are summarized in the next subsection.

## 3.4 Literature overview

In this subsection, an overview of different approaches to the identification of switched affine and piecewise affine models is presented. The description is not intended to be exhaustive, and the interested reader is referred to [61] for additional details. The list of references in [61] is completed here with most recent contributions.

### 3.4.1 Switched affine models

Emphasis on the identification of SARX models is put in the contributions [51, 74, 78], where an algebraic procedure for the estimation of the model orders, the number of discrete state and the model parameters, is proposed. The identification of SARX models is also considered in [58], where it is assumed that switchings occur with a certain probability at each time step, and [72, 73], where identification schemes for multi-mode and Markov models are developed. Switched affine models in state space form are considered in [10, 36, 71]. While in [71] the discrete state is assumed to be known, and the focus is mainly on determining the state transformations to express all the submodels in the same state basis (see Section 3.3), in [10] the number of discrete states and the switching times are estimated from data. In both contributions, subspace identification techniques are used to identify the individual submodels. In [36], the estimation of the model orders, the number of submodels and the switching times is carried out by embedding the input-output data into a higher dimensional space, where the problem becomes the one of segmenting the data into distinct subspaces.

### 3.4.2 Piecewise affine models

Work on regression with PWA maps can be found in many fields, such as neural networks, electrical networks, time-series analysis, function approximation. Most of the related approaches assume that the system dynamics is continuous. Indeed, enabling the estimation of discontinuous models is a key feature of algorithms specifically

designed for hybrid system identification. This is motivated by the fact that logic conditions can be represented through discontinuities in the state-update and output maps of the identified PWA model.

**Remark 3.1** If the PWA map is assumed to be continuous, the model parameters and the partition of the domain are not independent. For instance, consider the PWA map (9) with  $s = 2$ . If (9) is continuous, at the switching surface between the two modes it must hold that  $\theta_1^\top [\begin{smallmatrix} r \\ 1 \end{smallmatrix}] = \theta_2^\top [\begin{smallmatrix} r \\ 1 \end{smallmatrix}]$ , and hence  $r$  must satisfy

$$(\theta_1 - \theta_2)^\top [\begin{smallmatrix} r \\ 1 \end{smallmatrix}] = 0. \quad (21)$$

Equation (21) defines a hyperplane which divides the domain into two regions. Each mode of the PWA map is valid on one side of the hyperplane. Exploiting constraints of the type of (21) can be helpful to the identification process.

Different categories of approaches to PWA system identification can be distinguished depending on how the partitioning into regions is done. It follows from the discussion in Section 3.2 that there are mainly two alternative approaches: either the partition is defined a priori, or it is estimated along with the different submodels.

The first approach requires to define a priori the gridding of the domain. For instance, rectangular regions with sides parallel to the coordinate axes are used in [9], while simplices (i.e. polytopes with  $d + 1$  corners, where  $d$  is the dimension of the domain) are considered in [23] and [40]. This approach drastically simplifies the estimation of the linear/affine submodels, since standard linear identification techniques can be used to estimate the submodels, given enough data points in each region. On the other hand, it has the drawback that the number of regions and the need for experimental data, grow exponentially with  $d$ . This approach is therefore impracticable for high-dimensional systems.

The second approach consists in estimating the submodels and the partition of the domain either simultaneously or iteratively. This should allow for the use of fewer regions, since the regions are shaped according to the available data. Depending on how the partition is determined, Roll [61] further distinguishes among four different categories of approaches.

1. The first category relies on the direct formulation of a suitable criterion function to be minimized, such as (19). The parameters of the affine submodels and the coefficients of the hyperplanes defining the partition of the domain are therefore estimated simultaneously by minimizing the criterion function through

numerical methods (e.g., Gauss-Newton search). The algorithms proposed in [3, 15, 29, 41, 59] fall into this category. This way of tackling the identification problem is straightforward, but has the drawback that the optimization algorithm may get trapped in a local minimum. Techniques for reducing the risk of getting stuck in a local minimum can be used, at the cost of increased computational complexity.

2. The second category of approaches is an extension of the first one, and gives more flexibility with respect to the number of submodels. All parameters are identified simultaneously for a model with a very simple partition. If the resulting model is not satisfactory, new submodels/regions are added, in order to improve the value of a criterion function. In other words, instead to be solved at once, the overall identification problem is divided into several steps, each consisting in an easier problem to solve. The algorithms proposed in [14, 22, 35, 37, 39] fall into this category. The algorithm [14] has been analyzed in [59]. The paper [41] also describes an iterative method for introducing new partitions on the domain, when the error obtained is not satisfactory. As for the first category of approaches, there is still a risk to get stuck in a local minimum. When adding new submodels, one should also take into consideration the risk of overfit.
3. The third category contains a variety of approaches, sharing the characteristic that the parameters of the submodels and the partition of the domain are identified iteratively or in different steps, each step considering either the submodels or the regions. The algorithms proposed in [5, 27, 47, 56, 60] start by classifying the data points and estimating the linear/affine submodels simultaneously. Then, region estimation is carried out by resorting to standard linear separation techniques. In [54], the position of rectangular regions is optimized one by one iteratively. Then, each rectangular region is divided into simplices, in which affine submodels are finally identified. In [52], a greedy randomized adaptive search procedure is used to iteratively and heuristically find good partitions of the domain. Other approaches can be found in [30] and [31].
4. The last category of approaches estimates the partition using only information concerning the distribution of the regression vectors, and not the corresponding output values. This means that the domain

is partitioned in such a way that each region contains a suitable number of experimental data to estimate an affine submodel. The algorithms proposed in [16, 68] fall into this category. The major drawback of this category of approaches is that, without considering the output values, a set of data which really should be associated to the same submodel might be split arbitrarily.

It is stressed that most of the aforementioned approaches (e.g. [3, 14, 16, 22, 29, 35, 37, 41, 59]) assume that the system dynamics is continuous, while, e.g., [5, 27, 47, 56, 60] allow for discontinuities. Moreover, only few approaches (e.g., those in the second category, [5, 56], and [26], which is an extension of [27]) estimate also the number of submodels from data.

### 3.4.3 Other hybrid model classes

Recently, some contributions have focused on the class of *PieceWise Output Error* (PWOE) models, which are defined by the equations

$$\begin{aligned} y_k &= w_k + e_k \\ w_k &= f(\mathbf{r}_k), \end{aligned} \quad (22)$$

where  $f(\cdot)$  is the PWA map (9), and the regression vector  $\mathbf{r}_k$  is built as

$$\mathbf{r}_k = [w_{k-1} \dots w_{k-n_a} \ u_k \ u_{k-1} \dots u_{k-n_b}]^\top. \quad (23)$$

In [63] a prediction-error minimization method for piecewise linear output-error predictors is derived under the assumption that the discrete state is known at each time step. Estimation of the discrete state is made possible in [46], where a Bayesian method for identification of PWOE models is proposed.

### 3.4.4 Recursive identification approaches

All the aforementioned algorithms operate in a batch mode, i.e. the model is identified after all the input-output data have been collected. Since the computational complexity of batch algorithms depends on the number of data points, such algorithms may not be suitable for real time applications. An online algorithm for the identification of SARX/PWARX models is proposed in [65]. It exploits a mixture of recursive identification and pattern recognition techniques in order to identify the current parameter values. A different approach is pursued in the recent contributions [32, 75]. A standard recursive identification algorithm is used to estimate the parameters of a “lifted” ARX

model which is independent of the switching sequence, and is built by applying a polynomial embedding to the input-output data. Then, estimates of the ARX submodel parameters are obtained by differentiation. This approach also enables for the estimation of the model orders and the number of submodels.

## 4 Data classification and region estimation

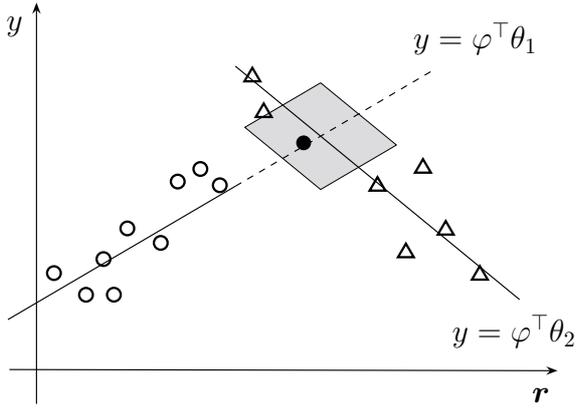
As pointed out in Section 3.4, identification methods allowing for discontinuities in the PWA map (9) are best suited in the context of hybrid systems, since they allow logic conditions to be represented by abrupt changes in the system dynamics. Most recent contributions, such as [5, 27, 47, 56, 60], have thus focused on regression with discontinuous PWA maps. It is interesting to note that all the above mentioned approaches share the idea to tackle the identification problem by firstly classifying the data and estimating the affine submodels, and then estimating the partition of the regressors domain. In this section, the data classification step is discussed in view of the subsequent step of region estimation. Moreover, a brief overview of linear separation techniques is given, and issues related to the estimation of the partition from a finite number of points are highlighted.

### 4.1 Data classification

Methods for the identification of PWARX models that firstly classify the data points and estimate the affine submodels, and then estimate the partition of the regressors domain, split in practice the identification problem into the identification of a SARX model, followed by the shaping of the regions to the clusters of data. In this respect, such methods can be also considered as methods for the identification of SARX models, if the final region estimation step is not addressed. Vice versa, methods developed for the identification of SARX models, such as [51, 74, 78], can be used to initialize the procedures for the identification of PWARX models.

However, in view of the subsequent step of region estimation, data classification for the identification of PWARX models needs to be carefully addressed. The main problem to deal with is represented by data points that are consistent with more than one submodel, namely data points lying in the proximity of the intersection of two or more submodels. Wrong attribution of these data points may lead to misclassifications when estimating the polyhedral regions.

In order to clarify this point, Fig. 4 shows a data set obtained from a one-dimensional PWA model with  $s = 2$



**Figure 4.** Example showing the problem of intersecting submodels. The data point denoted by the black circle could be in principle attributed to both submodels. Wrong attribution yields two non-linearly separable clusters of points.

discrete modes. It assumed that the parameter vectors  $\theta_1$  and  $\theta_2$  have been previously estimated, no matter which method has been used. If each data point  $(y_k, \mathbf{r}_k)$  is associated to the submodel  $i^*$  such that the prediction error is minimized, i.e. according to the rule

$$i^* = \arg \min_{i=1, \dots, s} |y_k - \varphi_k^\top \theta_i|, \quad (24)$$

the point denoted by the black circle is attributed to the first submodel. This yields two non-linearly separable clusters of points. It is stressed that the issue addressed in this example does not depend on the particular choice of (24) for associating each data point to one submodel. If data classification and parameter estimation are performed by solving Problem 3.2 for a given  $\delta > 0$ , the point denoted by the black circle is still attributed to the first submodel in this case. The gray area in Fig. 4 represents the region of all data points satisfying

$$|y_k - \varphi_k^\top \theta_i| \leq \delta \quad (25)$$

for both  $i = 1$  and  $i = 2$ . These data points are termed *undecidable*, because they could be in principle attributed to both submodels.

The identification procedures [5, 27, 47, 56, 60] deal with the problem of intersecting submodels in different ways. For instance, an ad-hoc refinement procedure based on the certainly attributed closest neighbors is proposed in [5], weights for misclassification are introduced in [47], and clustering in a feature space is pursued in [27]. These three approaches will be described in Section 5.

## 4.2 Region estimation

After the data classification step, providing the estimates of the discrete state  $\sigma(k) \in \{1, \dots, s\}$ , it is possible to form  $s$  clusters of regression vectors as

$$\mathcal{A}_i = \{\mathbf{r}_k : \sigma(k) = i\}, \quad i = 1, \dots, s. \quad (26)$$

The problem of region estimation consists in finding a complete polyhedral partition  $\{\mathcal{R}_i\}_{i=1}^s$  of the regressors domain  $\mathcal{R}$  such that  $\mathcal{A}_i \subseteq \mathcal{R}_i$  for all  $i = 1, \dots, s$ . The polyhedral regions (8) are defined by hyperplanes. Hence, the considered problem is equivalent to that of separating  $s$  sets of points by means of linear classifiers (hyperplanes). This problem can be tackled in two different ways:

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

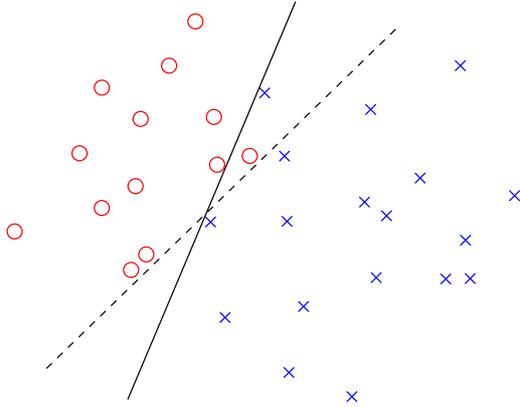
In the first approach, a separating hyperplane is constructed for each pair  $(\mathcal{A}_i, \mathcal{A}_j)$ ,  $i \neq j$ . This amounts to solve  $s(s-1)/2$  *two-class* linear separation problems. Given two sets  $\mathcal{A}_i$  and  $\mathcal{A}_j$ ,  $i \neq j$ , the linear separation problem is to find  $w \in \mathbb{R}^d$  and  $\gamma \in \mathbb{R}$  such that

$$\begin{aligned} w^\top \mathbf{r}_k + \gamma &> 0 \quad \forall \mathbf{r}_k \in \mathcal{A}_i \\ w^\top \mathbf{r}_k + \gamma &< 0 \quad \forall \mathbf{r}_k \in \mathcal{A}_j. \end{aligned} \quad (27)$$

This problem can be easily rewritten as a feasibility problem with linear inequality constraints by introducing the quantities

$$z_k = \begin{cases} 1 & \text{if } \mathbf{r}_k \in \mathcal{A}_i \\ -1 & \text{if } \mathbf{r}_k \in \mathcal{A}_j. \end{cases} \quad (28)$$

If a hyperplane separating without errors the points in  $\mathcal{A}_i$  from those in  $\mathcal{A}_j$  does not exist (this may happen because the sets  $\mathcal{A}_i$  and  $\mathcal{A}_j$  have intersecting convex hulls), a first reasonable approach is to look for a hyperplane that maximizes the number of well-separated points (equivalently, that minimizes the number of misclassified points). This problem amounts to find a pair  $(w, \gamma)$  such that the number of satisfied inequalities in (27) is maximized, and is known in the literature as MAXimum Feasible Subsystem (MAX FS) problem. Although the MAX FS problem is known to be NP-hard, several heuristics have been developed which work well in practice (see [1]). One drawback of the MAX FS approach is that it may not have a unique solution, as shown in Fig. 5.



**Figure 5.** The two sets are not linearly separable. The continuous line and the dashed line represent two different solutions to the problem of minimizing the number of misclassifications. One point is incorrectly classified in both cases.

An alternative approach for linear separation in the inseparable case is the minimization of a suitable cost function associated with errors. In the simplest case, this idea leads to the following linear program:

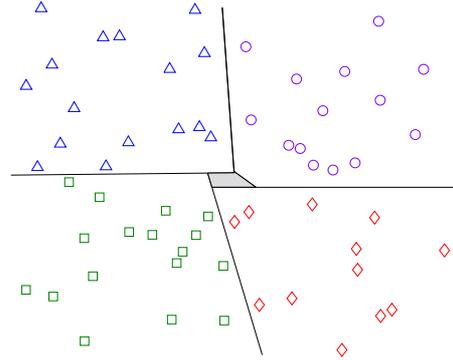
$$\begin{cases} \min_{w, \gamma, v_k} & \sum c_k v_k \\ \text{s.t.} & z_k [w^\top \mathbf{r}_k + \gamma] \geq 1 - v_k \\ & v_k \geq 0 \quad \forall \mathbf{r}_k \in \mathcal{A}_i \cup \mathcal{A}_j, \end{cases} \quad (29)$$

where  $c_k > 0$  are misclassification weights. If the data set is linearly separable, and therefore there exist  $w$  and  $\gamma$  such that the constraints

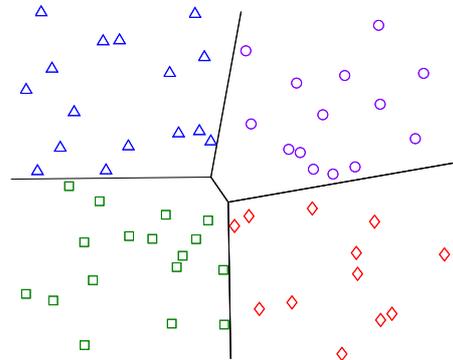
$$z_k [w^\top \mathbf{r}_k + \gamma] \geq 1 \quad (30)$$

are satisfied for all  $\mathbf{r}_k \in \mathcal{A}_i \cup \mathcal{A}_j$ , all the auxiliary variables  $v_k$  can be taken equal to zero. If the data set is not linearly separable, the auxiliary variables  $v_k$  allow the constraints (30) to be violated. Since, at the optimum of (29),  $v_k = \max \{0, 1 - z_k [w^\top \mathbf{r}_k + \gamma]\}$  for all  $\mathbf{r}_k \in \mathcal{A}_i \cup \mathcal{A}_j$ , each variable  $v_k$  can be interpreted as a *misclassification error*. The original *Robust Linear Programming* (RLP) method proposed in [7] is a particular case of (29), while the *Support Vector Machines* (SVM) method [19] solves a quadratic program under the same constraints as (29).

**Remark 4.1** When the set  $\mathcal{A}_i$  has been linearly separated from all the other sets  $\mathcal{A}_j, j \neq i$ , redundant hyperplanes (i.e. not contributing to the boundary of the region  $\mathcal{R}_i$ ) can be eliminated through standard linear programming techniques, so that the number of linear inequalities defining the  $i$ th region is  $\mu_i \leq s - 1$ .



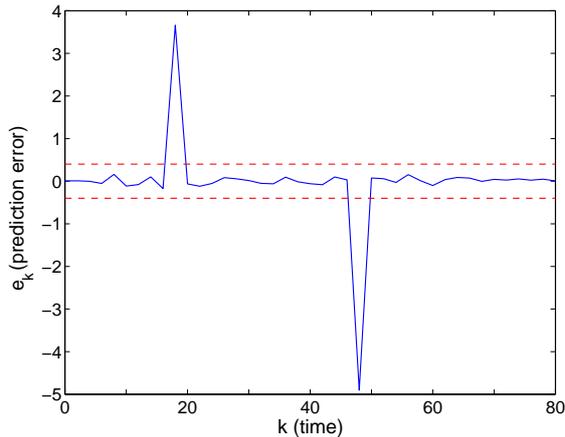
**Figure 6.** Pairwise linear separation of a data set formed by four linearly separable sets. The partition is not complete (the gray area is not covered).



**Figure 7.** Multi-class linear separation of the same data set as Fig. 6. The partition is complete.

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 regressors domain when  $d > 1$ , as shown in Fig. 6. This drawback is quite important, since it causes the model to be not completely defined over the whole regressors domain. If the presence of “holes” in the partition is not acceptable, one can resort to approach *b*). *Multi-class* linear separation techniques 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 [8, 13]. 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 (see Fig. 7). Multi-class problems involve all the available data, and therefore approach *b*) is computationally more demanding than approach *a*).

**Remark 4.2** Even if all data points are correctly classified, it is not possible in general to reconstruct exactly



**Figure 8.** Distinct spikes show up in the plot of the prediction errors due to discontinuity of the PWA map, and wrong assignment of the regression vectors because of errors in estimating the switching surfaces.

the regions from a finite data set. If the true system is characterized by continuous dynamics, small differences in hyperplane orientations are not expected to alter significantly the quality of the model. On the other hand, even small errors in shaping the surfaces along which the true system is discontinuous, may determine large prediction errors, if a regression vector falling close to a discontinuity, is associated to the wrong submodel. Such errors can be typically detected and corrected a posteriori during the validation or the operation of the model, as shown in Fig. 8. If distinct spikes show up in the prediction error plot, the corresponding data points can be re-attributed, and the augmented data set used to re-estimate the regions.

## 5 Four procedures for the identification of SARX/PWARX models

In this section, four procedures for the identification of SARX/PWARX models are briefly discussed, namely the algebraic procedure [51, 74, 78], the clustering-based procedure [27], the Bayesian procedure [47], and the bounded-error procedure [5]. It is stressed that other techniques are available (see Section 3.4). The ones considered here are more closely related to the activities of the authors of this paper, and have been successfully exploited in several real applications, such as identification of the electronic component placement process in pick-and-place machines [5, 43, 47], modelling of a current transformer [27], traction control [11], and motion segmentation in computer vision [76, 77].

While the algebraic procedure focuses on the identifi-

cation of SARX models, the other three procedures are designed for the identification of PWARX models, and are able to deal with discontinuous dynamics. The basic steps that each method performs are the estimation of the discrete state  $\sigma(k)$ , and the estimation of the parameter vectors  $\{\theta_i\}_{i=1}^s$ . Estimation of the polyhedral partition  $\{\mathcal{R}_i\}_{i=1}^s$  of the regressors domain, if needed, can be carried out in the same way for all methods by resorting to the techniques described in Section 4.2.

### 5.1 Algebraic procedure

The method proposed in [51, 74, 78] approaches the identification of SARX models as an algebraic geometric problem. It provides a closed-form solution to the identification problem that is provably correct in the absence of noise.

The key idea behind the algebraic approach is to view the identification of multiple ARX models as the identification of a single, “lifted” ARX model that simultaneously encodes all the ARX submodels and does not depend on the switching sequence. The parameters of the “lifted” ARX model can be identified through standard linear identification techniques after applying a polynomial embedding to the regression vectors. The parameters of the original ARX submodels are then given by the derivatives of this polynomial.

Assuming for simplicity that the number of submodels  $s$  and the model orders  $n_a$  and  $n_b$  are known (these assumptions will be subsequently removed), the algebraic procedure works as follows. If the data are generated by model (5) with  $e_k = 0$  (noiseless case), each data pair  $(y_k, \mathbf{r}_k)$  satisfies

$$y_k - \theta_i^\top \varphi_k = 0 \quad (31)$$

for some  $\theta_i, i = 1, \dots, s$ . Hence, the following equality holds for all  $k$ :

$$\prod_{i=1}^s (b_i^\top \mathbf{z}_k) \quad (32)$$

where  $b_i = [1 \ \theta_i^\top]^\top$  and  $\mathbf{z}_k = [-y_k \ \varphi_k^\top]^\top$ . Equation (32) is called the *hybrid decoupling constraint*, since it is independent of the switching sequence and the mechanism generating the transitions. In view of (32), the *hybrid decoupling polynomial* is defined as

$$p_s(\mathbf{z}) = \prod_{i=1}^s (b_i^\top \mathbf{z}) = h^\top \nu_s(\mathbf{z}), \quad (33)$$

which is a homogeneous polynomial of degree  $s$  in  $\mathbf{z} = [z_1 \ \dots \ z_K]^\top$ ,  $K = n_a + n_b + 3$ . Note that (33) can

be written as a linear combination of all the  $M_s(K) \doteq \binom{s+K-1}{s}$  monomials of degree  $s$  in  $K$  variables. Such monomials are stacked in the vector  $\nu_s(\mathbf{z})$  according to the degree-lexicographic order. The vector  $h \in \mathbb{R}^{M_s(K)}$  contains the so-called *hybrid model parameters*, and encodes the parameter vectors of the  $s$  submodels. Since (32) holds for all  $k$ , the vector  $h$  can be estimated by solving the linear system<sup>4</sup>

$$L_s(K)h = 0, \quad (34)$$

where  $L_s(K) = [\nu_s(\mathbf{z}_{\bar{n}}) \nu_s(\mathbf{z}_{\bar{n}+1}) \dots \nu_s(\mathbf{z}_N)]^\top$ . Once  $h$  has been computed, the vectors  $b_i$  can be reconstructed as

$$b_i = \frac{Dp_s(\mathbf{z}_{k_i})}{[1 \ 0 \ \dots \ 0]Dp_s(\mathbf{z}_{k_i})}, \quad (35)$$

where  $Dp_s(\mathbf{z}) = \frac{\partial p_s(\mathbf{z})}{\partial \mathbf{z}}$ , and  $\mathbf{z}_{k_i}$  is a data point generated by the  $i$ th ARX submodel, which can be chosen automatically once  $p_s(\cdot)$  is known [78]. Given the  $b_i$ 's (and consequently the  $\theta_i$ 's), the discrete state is estimated according to the rule  $\sigma(k) = i^*$ , with  $i^*$  given by (24). As discussed in Section 4.1, enhanced classification rules can be used by incorporating additional knowledge about the switching mechanism (e.g., PWARX models), when available.

The linear system (34) has a unique solution (requiring that the first component of  $h$  is equal to 1) when the data are sufficiently exciting, and  $s$ ,  $n_a$  and  $n_b$  are known exactly. If  $s$  is not known, it is shown in [78] that it can be estimated as<sup>5</sup>

$$s = \arg \min\{i : \text{rank}(L_i(K)) = M_i(K) - 1\}. \quad (36)$$

The algebraic procedure described above can be amended when only upper bounds  $\bar{s}$ ,  $\bar{n}_a$  and  $\bar{n}_b$  for  $s$ ,  $n_a$  and  $n_b$ , respectively, are available. In those cases, the procedure allows for the estimation of all the unknown quantities. More details can be found in [51, 74].

## 5.2 Clustering-based procedure

The clustering-based procedure [27] exploits the fact that the PWA map (9) is locally linear. If the data are generated by (10), there likely exist groups of neighbor regression vectors belonging to the same region (and the same submodel). Parameter vectors computed for these small local data sets should resemble the parameter vector of the corresponding submodel. Hence, information about

<sup>4</sup>The solution must be intended in a least-squares sense in the noisy case.

<sup>5</sup>Evaluation of (36) in the noisy case requires to introduce a threshold for estimating the rank of  $L_i(K)$ .

the submodels can be obtained by clustering the local parameter vectors.

The clustering-based procedure works as follows. The positive integer  $c$  is a fixed parameter.

- *Local regression.* For  $k = \bar{n}, \dots, N$ , a local data set  $\mathcal{C}_k$  is built by collecting  $(y_k, \mathbf{r}_k)$  and the data points  $(y_j, \mathbf{r}_j)$  corresponding to the  $c - 1$  regression vectors  $\mathbf{r}_j$  that are closest<sup>6</sup> to  $\mathbf{r}_k$ . Local parameter vectors  $\theta_k^{LS}$  are then computed for each local data set  $\mathcal{C}_k$  by least squares. For analysis purposes, local data sets containing only data points generated by the same submodel are referred to as *pure*, otherwise they are called *mixed*.

- *Construction of feature vectors.* Each data point  $(y_k, \mathbf{r}_k)$  is mapped onto the feature vector

$$\xi_k = [(\theta_k^{LS})^\top \mathbf{m}_k^\top]^\top, \quad (37)$$

where  $\mathbf{m}_k = \frac{1}{c} \sum_{(y, \mathbf{r}) \in \mathcal{C}_k} \mathbf{r}$  is the center of  $\mathcal{C}_k$ .

- *Clustering.* Feature vectors are partitioned into  $s$  groups  $\{\mathcal{F}_i\}_{i=1}^s$  by applying a “K-means”-like algorithm exploiting suitably defined confidence measures on the feature vectors. The confidence measures make it possible to reduce the influence of outliers and poor initializations.
- *Parameter estimation.* Since the mapping of the data points onto the feature space is bijective, data points are classified into clusters  $\{\mathcal{D}_i\}_{i=1}^s$  according to the rule

$$(y_k, \mathbf{r}_k) \in \mathcal{D}_i \quad \text{iff} \quad \xi_k \in \mathcal{F}_i. \quad (38)$$

A parameter vector  $\theta_i$  is estimated for each data cluster  $\mathcal{D}_i$  by weighted least squares.

The clustering-based procedure requires that the model orders  $n_a$  and  $n_b$ , and the number of submodels  $s$  are fixed. The parameter  $c$ , defining the cardinality of the local data sets, is the main tuning knob. In practical use, the method is expected to perform poorly if the ratio between the number of mixed and pure local data sets is high. The number of mixed local data sets increases with  $c$ . Hence, it is desirable to keep  $c$  as small as possible. On the other hand, when the noise level is high, large values of  $c$  may be needed in order to filter out the effects of noise.

An important feature of the clustering-based procedure is its ability to distinguish submodels characterized by the same parameter vector, but defined in different regions. This is possible because the feature vectors contain also

<sup>6</sup>According to the Euclidean norm.

information on the location of the local data sets. A modification to the clustering-based procedure is proposed in [26] to allow for the simultaneous estimation of the number of submodels. The clustering-based procedure is analyzed in [28], where it is shown that optimal classification can be guaranteed under suitable assumptions in the presence of bounded noise. A software implementation of the clustering-based procedure is also available [24].

### 5.3 Bayesian procedure

The Bayesian procedure [42, 47] is based on the idea of exploiting the available prior knowledge about the modes and the parameters of the hybrid system. The parameter vectors  $\theta_i$  are treated as random variables, and described through their probability density functions (*pdfs*)  $p_{\theta_i}(\cdot)$ . A priori knowledge on the parameters can be supplied to the procedure by choosing appropriate prior parameter *pdfs*. Various parameter estimates, such as expectation or maximum a posteriori probability estimate, can be easily obtained from the parameter *pdfs*. The data classification problem is posed as the problem of finding the data classification with the highest probability. Since this problem is combinatorial, an iterative suboptimal algorithm is derived. It is assumed that the probability density function  $p_e(\cdot)$  of the additive noise term  $e_k$  is given.

Data classification and parameter estimation are carried out by sequential processing of the collected data points. In each iteration, the *pdf* of one of the parameter vectors is updated. Let  $p_{\theta_i}(\cdot; k)$  denote the *pdf* of  $\theta_i$  at iteration  $k$ , when the data point  $(y_k, \mathbf{r}_k)$  is considered. The conditional *pdf*  $p((y_k, \mathbf{r}_k) | \sigma(k) = i)$  is given by

$$\begin{aligned} p((y_k, \mathbf{r}_k) | \sigma(k) = i) &= \\ &= \int_{\Theta_i} p((y_k, \mathbf{r}_k) | \tilde{\theta}) p_{\theta_i}(\tilde{\theta}; k-1) d\tilde{\theta}, \end{aligned} \quad (39)$$

where  $\Theta_i$  is the set of possible values for  $\theta_i$ , and

$$p((y_k, \mathbf{r}_k) | \theta) = p_e(y_k - \theta^\top \varphi_k). \quad (40)$$

The discrete state corresponding to  $(y_k, \mathbf{r}_k)$  is estimated as  $\sigma(k) = i^*$ , where

$$i^* = \arg \max_{i=1, \dots, s} p((y_k, \mathbf{r}_k) | \sigma(k) = i). \quad (41)$$

Then, the assignment of  $(y_k, \mathbf{r}_k)$  to mode  $i^*$  is used to update the *pdf* of  $\theta_{i^*}$  by the Bayes rule, i.e.

$$\begin{aligned} p_{\theta_{i^*}}(\theta; k) &= \\ &= \frac{p((y_k, \mathbf{r}_k) | \theta) p_{\theta_{i^*}}(\theta; k-1)}{\int_{\Theta_{i^*}} p((y_k, \mathbf{r}_k) | \tilde{\theta}) p_{\theta_{i^*}}(\tilde{\theta}; k-1) d\tilde{\theta}}. \end{aligned} \quad (42)$$

*Pdfs* of the other parameter vectors remain unchanged, i.e.  $p_{\theta_i}(\cdot; k) = p_{\theta_i}(\cdot; k-1)$  for  $i \neq i^*$ . For the numerical implementation of the described algorithm, particle filtering algorithms are used [2]. After the parameter estimation phase, each data point is finally attributed to the mode that most likely generated it.

To estimate the regions, a modification of the standard Multicategory RLP (MRLP) method [8] is proposed in [47]. If a regression vector attributed to mode  $i$  ends up in the region  $\mathcal{R}_j$  (this may happen, e.g., in the case of intersecting submodels, see Section 4.1 and Fig. 4), and the probabilities that the corresponding data point is generated by mode  $i$  and mode  $j$  are similar, misclassification should not be penalized highly. To this aim, for each data point  $(y_k, \mathbf{r}_k)$  attributed to mode  $i$ , the price for misclassification into mode  $j$  is defined as

$$\nu_{i,j}(\mathbf{r}_k) = \log \frac{p((y_k, \mathbf{r}_k) | \sigma(k) = i)}{p((y_k, \mathbf{r}_k) | \sigma(k) = j)}, \quad (43)$$

where  $p((y_k, \mathbf{r}_k) | \sigma(k) = \ell)$  is the likelihood that  $(y_k, \mathbf{r}_k)$  was generated by mode  $\ell$ . Note that the price for misclassification is zero if the probabilities are exactly equal. Prices for misclassification are plugged into the MRLP method.

The Bayesian procedure requires that the model orders  $n_a$  and  $n_b$ , and the number of submodels  $s$  are fixed. The most important tuning parameters are the prior parameter *pdfs*  $p_{\theta_i}(\cdot; 0)$ , and the *pdf*  $p_e(\cdot)$  of the error term. In [46] the Bayesian approach has been extended to the identification of piecewise output error models.

### 5.4 Bounded-error procedure

Inspired by ideas from set-membership identification (see, e.g., [53] and references therein), the main feature of the bounded-error procedure [5, 57] is to impose that the error  $e(k)$  in (10) is bounded by a given quantity  $\delta > 0$  for all the samples  $(y_k, \mathbf{r}_k)$  in the estimation data set, i.e.

$$|y_k - f(\mathbf{r}_k)| \leq \delta, \quad \forall k = \bar{n}, \dots, N. \quad (44)$$

Hence, the bounded-error procedure fits a PWARX model satisfying (44) to the data, without any assumption on the system generating the data.

Since any PWARX model satisfying the bounded-error condition (44) is feasible, an initial guess of the number of submodels  $s$  is obtained by addressing Problem 3.2. The solution<sup>7</sup> of Problem 3.2 provides also a raw data classification that suffers two drawbacks. The first one is related

<sup>7</sup>In [5] a method for the solution of Problem 3.2 is proposed to enhance the performance of the greedy randomized algorithm [1].

to the suboptimality of the method used to tackle Problem 3.2, implying that it is not guaranteed to yield the minimum number of submodels. The second one is related to the problem of *undecidable* data points (see Section 4.1), implying that the cardinality and the composition of the feasible subsystems may depend on the order in which they are extracted from (12).

To deal with the aforementioned drawbacks, an iterative *refinement* procedure is applied. The refinement procedure alternates between data reassignment and parameter update. If needed, it enables the reduction of the number of submodels. For given positive thresholds  $\alpha$  and  $\beta$ , submodels  $i$  and  $j$  are merged if  $\alpha_{i,j} < \alpha$ , where

$$\alpha_{i,j} = \frac{\|\theta_i - \theta_j\|}{\min\{\|\theta_i\|, \|\theta_j\|\}}, \quad (45)$$

and  $\|\cdot\|$  is the Euclidean norm. Submodel  $i$  is discarded if the cardinality of the set  $\mathcal{D}_i$  of data points classified to mode  $i$  is less than  $\beta N$ . Data points that do not satisfy (44) are discarded as *infeasible* during the classification process, making it possible to detect outliers. In [5] parameter estimates are computed by the  $\ell_\infty$  projection estimator, i.e.

$$\theta_i = \arg \min_{\theta} \max_{(y_k, \mathbf{r}_k) \in \mathcal{D}_i} |y_k - \varphi_k^\top \theta|, \quad (46)$$

but any other projection estimate, such as least squares, can be used [53].

The bounded-error procedure requires that the model orders  $n_a$  and  $n_b$  are fixed. The main tuning parameter is the bound  $\delta$ . As discussed in Section 3.1, the larger  $\delta$ , the smaller the required number of submodels at the price of a worse fit of the data. The optional parameters  $\alpha$  and  $\beta$ , if used, also implicitly determine the final number of submodels returned by the procedure. Another tuning parameter is the number  $c$  of closest neighbors used to attribute undecidable data points to submodels in the refinement step.

**Remark 5.1** The bounded-error formulation (44) is easily extended to multi-output models. In this case, the output of the system is  $\mathbf{y}_k \in \mathbb{R}^q$ , the PWA map (9) is a  $q$ -valued function, and (44) is replaced by

$$\|\mathbf{y}_k - f(\mathbf{r}_k)\|_\infty \leq \delta, \quad \forall k = \bar{n}, \dots, N, \quad (47)$$

where  $\|\cdot\|_\infty$  is the infinity norm of a vector. The bounded-error procedure [5, 57] is then applicable also to the case  $q > 1$ , provided that Problem 3.2 is reformulated and solved accordingly. The interested reader is referred to [57].

## 5.5 Discussion

The four identification procedures described in this section are compared and discussed in [44] (see also [45]). Specific behaviors of the procedures with respect to classification accuracy, noise level, and tuning parameters are pointed out using simple one-dimensional examples. The procedures are also tested on the experimental identification of the electronic component placement process in pick-and-place machines.

From the comparison, it comes out that the algebraic procedure is well suited when the system generating the data can be accurately described as a switched affine model, and moderate noise is present. The main features of the algebraic procedure are that it can handle the cases with unknown model orders and unknown number of submodels, and it does not require any form of initialization. However, noise and/or nonlinear disturbances affecting the data may cause poor identification results. When trying to identify a PWARX model using the data classification obtained from the algebraic procedure, one must be aware that the minimum prediction error classification rule may lead to wrong data association. In such cases, it is advisable to use one of the classification methods employed by other procedures.

The clustering-based procedure is well suited when there is no prior knowledge on the physical system, and one needs to identify a model with a prescribed structure (i.e. the number of submodels and the model orders are given). Identification using the clustering-based procedure is straightforward, as only one parameter has to be tuned. However, poor results can be obtained when the model orders are overestimated, since distances in the feature space become corrupted by irrelevant information.

The Bayesian procedure is designed to take advantage of prior knowledge and physical insight into the operation modes of the system (like in the pick-and-place machine identification [47]). Another interesting feature is the automatic computation of misclassification weights to be plugged into the linear separation techniques used for region estimation. As a major drawback, poor initialization may lead to poor identification results.

The bounded error procedure is well suited when no prior knowledge on the physical system is available, and one needs to identify a model with a prescribed accuracy (e.g., to approximate nonlinear dynamics in each operation mode). Tuning parameters allow to trade-off the model accuracy with the model complexity, expressed in terms of the mean squared error and the number of submodels, respectively. However, finding the right combi-

nation of the tuning parameters is seldom straightforward, and several attempts are often needed to get a satisfactory model.

It is stressed that mixing the features of the four procedures could still enhance their effectiveness. In particular:

- The algebraic procedure can be used to initialize the other three procedures by providing estimates of the model orders and the number of submodels.
- By exploiting the idea of clustering the feature vectors, the clustering-based procedure is able to distinguish submodels characterized by the same parameter vector, but defined in different regions. This is a pitfall of both the Bayesian and the bounded-error classification procedures. Since these procedures do not exploit the spatial location of the submodels, data points generated by the same parameter vector in different regions are classified as a whole. This may lead to non-linearly separable clusters. Clustering ideas contained in the clustering-based procedure can be extended to the other two procedures in order to detect and split the clusters corresponding to such situations.
- The Bayesian procedure includes the computation of misclassification weights to be plugged into the linear separation techniques used for region estimation. This feature can be extended to the clustering-based and the bounded-error procedures. In the latter case, for each data point  $(y_k, \mathbf{r}_k)$  attributed to mode  $i$ , the price for misclassification into mode  $j$  could be defined as

$$\nu_{i,j}(\mathbf{r}_k) = \lambda \log \max\left\{1, \frac{|y_k - \theta_j^\top \mathbf{r}_k|}{\delta}\right\}, \quad (48)$$

where  $\lambda > 0$  is a scale factor.

- The bounded-error procedure can be used to guess the number of submodels, especially when the dynamics in each operation mode of the true system is nonlinear, and more modes than the true system are thus required to accurately approximate all the nonlinear dynamics.

## 6 Conclusions

Hybrid system identification is an emerging field whose importance grows with the potential new applications of hybrid systems in real life. In order to use the numerous tools for analysis, verification, computation, stability, and control of hybrid systems, a hybrid model of the

process at hand is needed. Hence, techniques for obtaining accurate models are of paramount importance. Most effort in the area of hybrid system identification has recently focused on identifying switched affine and PWA models. In the first part of the paper, different formulations of the identification problem for these model classes have been reported, and an overview of the related literature has been proposed. Although work on regression with continuous PWA maps can be found in the extensive literature on nonlinear black-box identification, most recent contributions aimed at the identification of models with a hybrid and discontinuous structure. Among these, an algebraic procedure, a Bayesian procedure, a clustering-based procedure, and a bounded-error procedure have been successfully applied in several real problems. The four procedures have been the topic of the second part of the paper.

It has emerged that a fundamental issue in PWA system identification is how to keep the computational complexity and the model complexity (number of model parameters) low. The computational complexity for a given algorithm is a function of the number of regions/submodels, the number of experimental data, and the model orders. Hence, there are many trade-off situations when comparing different methods and tuning parameters for given algorithms. For instance, it has been pointed out that prior gridding of the domain drastically simplifies the identification problem, but the numbers of regions and data needed to get good results increase exponentially with the model orders. Hence, this approach may be good for low-dimensional systems.

Another trade-off issue concerns model complexity and quality. The more degrees of freedom are allowed in the model structure, the closer the model can approximate the experimental data. However, since data are typically corrupted by noise, too much flexibility might cause the model to adjust to the noise realization, thereby causing overfit. This is indeed a general problem of system identification, occurring not only for piecewise affine systems.

Several issues still remain open. Though, given the equivalence between PWA models and other classes of hybrid models, PWA system identification techniques can be regarded as general hybrid system identification techniques, the development of specific identification tools for different hybrid model classes would be advisable. In this way, the identification process could take advantage of available prior knowledge that cannot be easily expressed in the PWA formalism.

Including prior knowledge and system physics in the identification process leads to the broader perspective

of application-based hybrid system identification. New frontiers for hybrid system identification are opening, e.g., along with the emerging field of systems biology (see [25] and references therein).

Finally, the choice of persistently exciting input signals for identification (i.e. allowing for the correct identification of all the affine dynamics) is another important topic to be addressed. Preliminary results have been proposed in [75], but the derived persistence of excitation conditions involve both the input and the output data. Moreover, when dealing with discontinuous PWARX models, the choice of the input signal 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.

## References

- [1] Amaldi E, Mattavelli M. The MIN PFS problem and piecewise linear model estimation. *Discrete Appl Math* 2002; 118(1-2):115–143
- [2] Arulampalam MS, Maskell S, Gordon N, Clapp T. A tutorial on particle filters for on-line nonlinear/non-Gaussian Bayesian tracking. *IEEE Trans Signal Process* 2002; 50(2):174–188
- [3] Batruni R. A multilayer neural network with piecewise-linear structure and back-propagation learning. *IEEE Trans Neural Netw* 1991; 2(3):395–403
- [4] Bemporad A, Ferrari-Trecate G, Morari M. Observability and controllability of piecewise affine and hybrid systems. *IEEE Trans Autom Control* 2000; 45(10):1864–1876
- [5] Bemporad A, Garulli A, Paoletti S, Vicino A. A bounded-error approach to piecewise affine system identification. *IEEE Trans Autom Control* 2005; 50(10):1567–1580
- [6] Bemporad A, Morari M. Control of systems integrating logic, dynamics, and constraints. *Automatica* 1999; 35(3):407–427
- [7] Bennett KP, Mangasarian OL. Robust linear programming discrimination of two linearly inseparable sets. *Optim Meth Software* 1992; 1:23–34
- [8] Bennett KP, Mangasarian OL. Multicategory discrimination via linear programming. *Optim Meth Software* 1994; 3:27–39
- [9] Billings SA, Voon WSF. Piecewise linear identification of nonlinear systems. *Int J Control* 1987; 46(1):215–235
- [10] Borges J, Verdult V, Verhaegen M. Iterative subspace identification of piecewise linear systems. In: *Proceedings of the 14th IFAC Symposium on System Identification, Newcastle, Australia. 2006*, pp 368–373
- [11] Borrelli F, Bemporad A, Fodor M, Hrovat D. An MPC/hybrid system approach to traction control. *IEEE Trans Contr Syst Technol* 2006; 14(3):541–552
- [12] Bradley PS, Mangasarian OL.  $k$ -plane clustering. *J Global Optim* 2000; 16:23–32
- [13] Bredensteiner EJ, Bennett KP. Multicategory classification by support vector machines. *Comput Optim Appl* 1999; 12:53–79
- [14] Breiman L. Hinging hyperplanes for regression, classification, and function approximation. *IEEE Trans Inf Theory* 1993; 39(3):999–1013
- [15] Chan KS, Tong H. On estimating thresholds in autoregressive models. *J Time Ser Anal* 1986; 7(3):179–190
- [16] Choi C-H, Choi JY. Constructive neural networks with piecewise interpolation capabilities for function approximations. *IEEE Trans Neural Netw* 1994; 5(6):936–944
- [17] Chua LO, Hasler M, Neiryneck J, Verburgh P. Dynamics of a piecewise-linear resonant circuit. *IEEE Trans Circuits Syst* 1982; 29(8):535–547
- [18] Chua LO, Ying RLP. Canonical piecewise-linear analysis. *IEEE Trans Circuits Syst* 1983; 30(3):125–140
- [19] Cortes C, Vapnik V. Support-vector networks. *Mach Learn* 1995; 20:273–297
- [20] De Schutter B. Optimal control of a class of linear hybrid systems with saturation. *SIAM J Control Optim* 2000; 39(3):835–851
- [21] De Schutter B, Van den Boom TJJ. Model predictive control for max-min-plus-scaling systems. In: *Proceedings of the American Control Conference, Arlington, VA. 2001*, pp 319–324
- [22] Ernst S. Hinging hyperplane trees for approximation and identification. In: *Proceedings of the 37th IEEE Conference on Decision and Control, Tampa, FL. 1998*, pp 1266–1271
- [23] Fantuzzi C, Simani S, Beghelli S, Rovatti R. Identification of piecewise affine models in noisy environment. *Int J Control* 2002; 75(18):1472–1485
- [24] Ferrari-Trecate G. Hybrid Identification Toolbox (HIT). [http://www-rocq.inria.fr/who/Giancarlo.-Ferrari-Trecate/HIT\\_toolbox.html](http://www-rocq.inria.fr/who/Giancarlo.-Ferrari-Trecate/HIT_toolbox.html). 2005
- [25] Ferrari-Trecate G. Hybrid identification methods for the reconstruction of genetic regulatory networks. In: *Proceedings of the European Control Conference, Kos, Greece. 2007*

- [26] Ferrari-Trecate G, Muselli M. Single-linkage clustering for optimal classification in piecewise affine regression. In: Proceedings of the IFAC Conference on Analysis and Design of Hybrid Systems, Saint-Malo, France. 2003
- [27] Ferrari-Trecate G, Muselli M, Liberati D, Morari M. A clustering technique for the identification of piecewise affine systems. *Automatica* 2003; 39(2):205–217
- [28] Ferrari-Trecate G, Schinkel M. Conditions of optimal classification for piecewise affine regression. In: Maler O, Pnueli A (eds). *Hybrid Systems: Computation and Control*, vol. 2623 of Lecture Notes in Computer Science. Springer-Verlag, Berlin/Heidelberg. 2003, pp 188–202
- [29] Gad EF, Atiya AF, Shaheen S, El-Dessouki A. A new algorithm for learning in piecewise-linear neural networks. *Neural Networks* 2000; 13(4-5):485–505
- [30] Gelfand SB, Ravishankar CS. A tree-structured piecewise linear adaptive filter. *IEEE Trans Inf Theory* 1993; 39(6):1907–1922
- [31] Groff RE, Koditschek DE, Khargonekar PP. Piecewise linear homeomorphisms: the scalar case. In: Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks, Como, Italy. 2000, pp 259–264
- [32] Hashambhoy Y, Vidal R. Recursive identification of switched ARX models with unknown number of models and unknown orders. In: Proceedings of the Joint 44th IEEE Conference on Decision and Control and European Control Conference, Sevilla, Spain. 2005, pp 6115–6121
- [33] Heemels WPMH, Schumacher JM, Weiland S. Linear complementarity systems. *SIAM J Appl Math* 2000; 60(4):1234–1269
- [34] Heemels WPMH, De Schutter B, Bemporad A. Equivalence of hybrid dynamical models. *Automatica* 2001; 37(7):1085–1091
- [35] Heredia EA, Arce GR. Piecewise linear system modeling based on a continuous threshold decomposition. *IEEE Trans Signal Process* 1996; 44(6):1440–1453
- [36] Huang K, Wagner A, Ma Y. Identification of hybrid linear time-invariant systems via subspace embedding and segmentation (SES). In: Proceedings of the 43rd IEEE Conference on Decision and Control, Paradise Island, Bahamas. 2004, pp 3227–3234
- [37] Hush DR, Horne B. Efficient algorithms for function approximation with piecewise linear sigmoidal networks. *IEEE Trans Neural Netw* 1998; 9(6):1129–1141
- [38] Huyer W, Neumaier A. Global optimization by multilevel coordinate search. *J Global Optim* 1999; 14:331–355
- [39] Johansen TA, Foss BA. Identification of non-linear system structure and parameters using regime decomposition. *Automatica* 1995; 31(2):321–326
- [40] Julián P, Desages A, Agamennoni O. High level canonical piecewise linear representation using a simplicial partition. *IEEE Trans Circuits Syst I: Fundam Theory Appl* 1999; 46(4):463–480
- [41] Julian P, Jordan M, Desages A. Canonical piecewise-linear approximation of smooth functions. *IEEE Trans Circuits Syst I: Fundam Theory Appl* 1998; 45(5):567–571
- [42] Juloski A. Observer design and identification methods for hybrid systems. PhD thesis, Department of Electrical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands. 2004
- [43] Juloski A, Heemels WPMH, Ferrari-Trecate G. Data-based hybrid modelling of the component placement process in pick-and-place machines. *Control Eng Prac* 2004; 12(10):1241–1252
- [44] Juloski A, Heemels WPMH, Ferrari-Trecate G, Vidal R, Paoletti S, Niessen JHG. Comparison of four procedures for the identification of hybrid systems. In: Morari M, Thiele L (eds). *Hybrid Systems: Computation and Control*, vol. 3414 of Lecture Notes in Computer Science. Springer Verlag, Berlin/Heidelberg. 2005, pp. 354–369
- [45] Juloski A, Paoletti S, Roll J. Recent techniques for the identification of piecewise affine and hybrid systems. In: Menini L, Zaccarian L, Abdallah CT (eds). *Current trends in nonlinear systems and control*, Birkäuser, Boston, MA. 2006, pp. 77–97
- [46] Juloski A, Weiland S. A Bayesian approach to the identification of piecewise linear output error models. In: Proceedings of the 14th IFAC Symposium on System Identification, Newcastle, Australia. 2006, pp 374–379
- [47] Juloski A, Wieland S, Heemels WPMH. A Bayesian approach to identification of hybrid systems. *IEEE Trans Autom Control* 2005; 50(10):1520–1533
- [48] Leenaerts DMW, Van Bokhoven WMG. Piecewise linear modeling and analysis. Kluwer Academic Publishers, Dordrecht, The Netherlands. 1998
- [49] Lin J-N, Unbehauen R. Canonical piecewise-linear approximations. *IEEE Trans Circuits Syst I: Fundam Theory Appl* 1992; 39(8):697–699

- [50] Ljung L. System identification: theory for the user. Prentice-Hall, Upper Saddle River, NJ. 1999
- [51] Ma Y, Vidal R. Identification of deterministic switched ARX systems via identification of algebraic varieties. In: Morari M, Thiele L, Rossi F (eds). Hybrid Systems: Computation and Control, vol. 3414 of Lecture Notes in Computer Science. Springer-Verlag, Berlin/Heidelberg. 2005, pp 449–465
- [52] Medeiros MC, Veiga A, Resende MGC. A combinatorial approach to piecewise linear time series analysis. *J Comput Graph Stat* 2002; 11(1):236–258
- [53] Milanese M, Vicino A. Optimal estimation theory for dynamic systems with set membership uncertainty: an overview. *Automatica* 1991; 27(6):997–1009
- [54] Münz E, Krebs V. Identification of hybrid systems using apriori knowledge. In: Proceedings of the 15th IFAC World Congress, Barcelona, Spain. 2002
- [55] Münz E, Krebs V. Continuous optimization approaches to the identification of piecewise affine systems. In: Proceedings of the 16th IFAC World Congress, Prague, Czech Republic. 2005
- [56] Nakada H, Takaba K, Katayama T. Identification of piecewise affine systems based on statistical clustering technique. *Automatica* 2005; 41(5):905–913
- [57] Paoletti S. Identification of piecewise affine models. PhD thesis, Department of Information Engineering, University of Siena, Siena, Italy. 2004. <http://www.dii.unisi.it/~paoletti>
- [58] Petridis V, Kehagias A. Identification of switching dynamical systems using multiple models. In: Proceedings of the 37th IEEE Conference on Decision and Control, Tampa, FL. 1998, pp 199–204
- [59] Pucar P, Sjöberg J. On the hinge-finding algorithm for hinging hyperplanes. *IEEE Trans Inf Theory* 1998; 44(3):1310–1319
- [60] Ragot J, Mourot G, Maquin D. Parameter estimation of switching piecewise linear systems. In: Proceedings of the 42nd IEEE Conference on Decision and Control, Maui, HI. 2003, pp 5783–5788
- [61] Roll J. 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/>
- [62] Roll J, Bemporad A, Ljung L. Identification of piecewise affine systems via mixed-integer programming. *Automatica* 2004; 40(1):37–50
- [63] Rosenqvist F, Karlström A. Realisation and estimation of piecewise-linear output-error models. *Automatica* 2005; 41(3):545–551
- [64] Sjöberg J, Zhang Q, Ljung L, et al. Nonlinear black-box modeling in system identification: a unified overview. *Automatica* 1995; 31(12):1691–1724
- [65] Skeppstedt A, Ljung L, Millnert M. Construction of composite models from observed data. *Int J Control* 1992; 55(1):141–152
- [66] Sontag ED. Nonlinear regulation: the piecewise linear approach. *IEEE Trans Autom Control* 1981; 26(2):346–358
- [67] Sontag ED. Interconnected automata and linear systems: a theoretical framework in discrete-time. In: Alur R, Henzinger TA, Sontag ED (eds). Hybrid Systems III - Verification and Control, vol. 1066 of Lecture Notes in Computer Science. Springer-Verlag, Berlin/Heidelberg. 1996, pp 436–448
- [68] Strömberg J-E, Gustafsson F, Ljung L. Trees as black-box model structures for dynamical systems. In: Proceedings of the European Control Conference, Grenoble, France. 1991, pp 1175–1180
- [69] Tong H. Threshold models in non-linear time series analysis, vol. 21 of Lecture Notes in Statistics. Springer Verlag, New York. 1983
- [70] Van der Schaft AJ, Schumacher JM. Complementarity modelling of hybrid systems. *IEEE Trans Autom Control* 1998; 43(4):483–490
- [71] Verdult V, Verhaegen M. Subspace identification of piecewise linear systems. In: Proceedings of the 43rd IEEE Conference on Decision and Control, Paradise Island, Bahamas. 2004, pp 3838–3843
- [72] Verriest EI. Multi-mode system identification In: Tao G, Lewis FL (eds). Adaptive Control of Nonsmooth Dynamic Systems. Springer Verlag, Berlin/Heidelberg. 2001, pp 157–190
- [73] Verriest EI, De Moor B. Multi-mode system identification. In: Proceedings of the European Control Conference, Karlsruhe, Germany. 1999
- [74] Vidal R. Identification of PWARX hybrid models with unknown and possibly different orders. In: Proceedings of the IEEE American Control Conference, Boston, MA. 2004, pp 547–552
- [75] Vidal R, Anderson BDO. Recursive identification of switched ARX hybrid models: exponential convergence and persistence of excitation. In: Proceedings of the 43rd IEEE Conference on Decision and Control, Paradise Island, Bahamas. 2004, pp 32–37
- [76] Vidal R, Chiuso A, Soatto S. Applications of hybrid system identification in computer vision. In:

Proceedings of the European Control Conference,  
Kos, Greece. 2007

- [77] Vidal R, Ma Y. A unified algebraic approach to 2-D and 3-D motion segmentation and estimation. *J Math Imaging Vis* 2006; 25(3):403–421
- [78] Vidal R, Soatto S, Ma Y, Sastry S. 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, HI. 2003, pp 167–172