Recent techniques for the identification of piecewise affine and hybrid systems

A. Lj. Juloski[1], S. Paoletti[2], and J. Roll[3]

[1] Department of Electrical Engineering, Eindhoven University of Technology
P.O. Box 513, 5600 MB Eindhoven, The Netherlands. A.Juloski@tue.nl
[2] Department of Information Engineering, University of Siena
Via Roma 56, 53100 Siena, Italy. paoletti@dii.unisi.it
[3] Division of Automatic Control, Linköping University
SE-581 83 Linköping, Sweden. roll@isy.liu.se

Summary. The problem of piecewise affine identification is addressed by studying four recently proposed techniques for the identification of PWARX/HHARX models, namely a Bayesian procedure, a bounded-error procedure, a clustering-based procedure and a mixed-integer programming procedure. The four techniques are compared on suitably defined one-dimensional examples, which help to highlight the features of the different approaches with respect to classification, noise and tuning parameters. The procedures are also tested on the experimental identification of the electronic component placement process in pick-and-place machines.

1 Introduction

The focus of this chapter is on the problem of identifying PieceWise Affine (PWA) models of discrete-time nonlinear and hybrid systems from input-output data. PWA systems are obtained by partitioning the state and input space into a finite number of non-overlapping convex polyhedral regions, and by considering linear/affine subsystems sharing the same continuous state variables in each region. The interest in PWA identification techniques is motivated by several reasons. Since PWA maps have universal approximation properties [18, 10], PWA models represent an attractive black-box model structure for nonlinear system identification. In addition, given the equivalence between PWA systems and several classes of hybrid systems [4, 14], the many different analysis, synthesis and verification tools for hybrid systems (see, *e.g.*, [2, 21, 28] and references therein) can be applied to the identified PWA models. PWA systems have indeed many applications in different contexts such as neural networks, electrical networks, time-series analysis and function approximation.

In the extensive literature on nonlinear black-box identification (see, *e.g.*, [27] and references therein), a few techniques can be found that lead to

PWA models of nonlinear dynamical systems. An overview and classification of them is presented in [25]. Recently, novel contributions to this topic have been also proposed in the hybrid systems community [5, 6, 13, 16, 22, 26, 29]. 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 and input space (or the regressor space, for models in regression form). The main difficulty lies in the fact that the identification problem includes a classification problem, in which each data point must be associated to one region and to the corresponding submodel. The problem is even harder when also the number of submodels must be estimated. In this chapter, four recently proposed techniques for the identification of (possibly) discontinuous PWA models are considered, namely the Bayesian procedure [16], the bounded-error procedure [5, 6], the clustering-based procedure [13] and the Mixed-Integer Programming (MIP) procedure [26]. While the MIP procedure formulates the identification problem as a mixed-integer linear or quadratic program that can be solved for the global optimum, the other three procedures can only guarantee suboptimal solutions. On the other hand, the very high worst-case computational complexity of MILP/MIQP problems makes the approach in [26] affordable only when few data are available, or data are clustered together. The four procedures are studied here for what concerns the classification accuracy, and the effects of noise, overestimated model orders and varying the tuning parameters on the identification results. The study of specific cases can indeed shed some light on the properties of the different techniques, and guide the user in their application to practical situations.

This chapter is organized as follows. The considered PWA identification problem is formulated and discussed in Section 2. Section 3 describes the four compared procedures, and introduces several quantitative measures for assessing the quality of the identified models. The different approaches of the four procedures to data classification are addressed in Section 4. The effects of the overestimation of model orders on the identification accuracy are investigated in Section 5, while Section 6 studies the effects of noise. In Section 7 the sensitivity of the identification results to tuning parameters is analyzed for the Bayesian, bounded-error and clustering-based procedures. In Section 8 the four procedures are tested on experimental data from the electronic component placement process in pick-and-place machines [15]. Finally, conclusions are drawn in Section 9.

2 Problem formulation

PieceWise affine AutoRegressive eXogenous (PWARX) models can be seen as collections of affine ARX models equipped with the switching rule determined by a polyhedral partition of the regressor set. Letting $k \in \mathbb{Z}$ be the time index, and $u(k) \in \mathbb{R}$ and $y(k) \in \mathbb{R}$ be the system input and output, respectively,

a PWARX model establishes a relationship between past observations and future outputs in the form

$$y(k) = f(x(k)) + e(k), \tag{1}$$

where $e(k) \in \mathbb{R}$ is the prediction error, $f(\cdot)$ is the PWA map

$$f(x) = \begin{cases} [x'\ 1]\theta_1 \ \text{ if } x \in \mathcal{X}_1 \\ \quad \vdots \\ [x'\ 1]\theta_s \ \text{ if } x \in \mathcal{X}_s, \end{cases} \tag{2}$$

defined over the regressor set $\mathcal{X} = \bigcup_{i=1}^{s} \mathcal{X}_i \subseteq \mathbb{R}^n$ on which the PWARX model is valid, and $x(k) \in \mathbb{R}^n$ is the regression vector with fixed structure depending only on past n_a outputs and n_b inputs:

$$x(k) = [y(k-1)\ \ldots\ y(k-n_a)\ u(k-1)\ \ldots\ u(k-n_b)]' \tag{3}$$

(hence, $n = n_a + n_b$). In (2) s is the number of submodels and $\theta_i \in \mathbb{R}^{n+1}$ are the parameter vectors (PVs) of the affine ARX submodels. The regions $\mathcal{X}_i$ are convex polyhedra which do not overlap, *i.e.*, $\mathcal{X}_i \cap \mathcal{X}_j = \emptyset$ for all $i \neq j$. Hence, for a given data point $\big(y(k), x(k)\big)$, the corresponding active *mode* $\mu(k)$ can be uniquely defined as:

$$\mu(k) = i \ \text{ iff } \ x(k) \in \mathcal{X}_i. \tag{4}$$

Given the data set $\mathcal{D} = \{\big(y(k), x(k)\big)\}_{k=1}^{N}$, the considered identification problem consists of finding the PWARX model that best matches the data according to some specified criterion of fit (*e.g.*, the minimization of the sum of absolute or squared prediction errors, see [27]). For fixed model orders n_a and n_b, this problem involves the estimation of the number of submodels s, the PVs $\{\theta_i\}_{i=1}^{s}$ and the polyhedral partition $\{\mathcal{X}_i\}_{i=1}^{s}$. It also includes a classification problem in which each data point is associated to one region and the corresponding submodel. In general, the simultaneous optimal estimation of all the quantities above leads to very complex, nonconvex optimization problems with potentially many local minima, which complicate the use of local search minimization algorithms. One of the main difficulties concerns the selection of the number of submodels s. Constraints on s must be introduced in order to keep the number of submodels low and to avoid overfit. Heuristic and suboptimal approaches to the identification of PWARX models have been proposed in the literature (see [25] for an overview). Most of these approaches either assume a fixed s, or adjust s iteratively (*e.g.*, by adding one submodel at a time) in order to improve the fit. When s is fixed, the identification of a PWARX model amounts to a PWA regression problem, namely the problem of reconstructing the PWA map $f(\cdot)$ from the finite data set $\mathcal{D}$.

Note that, if the partition of the regressor set is either known or fixed a priori, the problem complexity reduces to that of a linear identification problem, since the data points can be classified to corresponding data clusters $\{\mathcal{D}_i\}_{i=1}^{s}$, and standard linear identification techniques can be applied to estimate the PVs for each submodel [19].

3 The compared procedures

In this section, four recently proposed procedures for the identification of PWA models are briefly introduced and described. These are the Bayesian procedure [16], the bounded-error procedure [5, 6], the clustering-based procedure [13] and the Mixed-Integer Programming (MIP) procedure [26].

In its basic formulation, the MIP procedure considers hinging-hyperplane ARX models, which form a subclass of PWARX models with continuous PWA map $f(\cdot)$ [10]. For this class of models, the identification problem is formulated as a mixed-integer linear or quadratic program that can be solved for the global optimum.

The Bayesian, bounded-error and clustering-based procedures identify models in PWARX form. The basic steps that these procedures perform are data classification and parameter estimation, followed by the reconstruction of the regions. The bounded-error procedure also estimates the number of submodels. The first two steps are performed in a different way by each procedure, as described in the following sections, while the estimation of the regions can be carried out in the same way for all procedures. Basically, given the clusters $\{\mathcal{D}_i\}_{i=1}^{s}$ of data points provided by the data classification phase, the corresponding clusters of regression vectors $\mathcal{R}_i = \{x(k) \mid \big(y(k), x(k)\big) \in \mathcal{D}_i\}$ are constructed. Then, for all $i \neq j$ a separating hyperplane of the clusters $\mathcal{R}_i$ and $\mathcal{R}_j$ is sought, *i.e.*, a hyperplane

$$M_{ij}' x = m_{ij}, \tag{5}$$

with $M_{ij} \in \mathbb{R}^n$ and $m_{ij} \in \mathbb{R}$, such that $M_{ij}' x(k) < m_{ij}$ for all $x(k) \in \mathcal{R}_i$ and $M_{ij}' x(k) > m_{ij}$ for all $x(k) \in \mathcal{R}_j$. If such a hyperplane cannot be found (*i.e.*, the data set is not linearly separable) one is interested in finding a generalized separating hyperplane which minimizes the number of misclassified data points or some misclassification cost. Robust Linear Programming (RLP) [7] and Support Vector Machines (SVM) [11] methods (and their extensions to the multi-class case [8, 9]) can be employed. The minimization of the number of misclassifications is equivalent to solving a MAXimum Feasible Subsystem (MAX FS) problem for a system of linear inequalities (see [24] and references therein). The interested reader is referred to [5, 13, 23] for a detailed overview.

3.1 Mixed-integer programming procedure

The procedure proposed in [25, 26] is, in its basic formulation, an algorithm for optimal identification of Hinging-Hyperplane ARX (HHARX) models [10], which are described by

$$\begin{aligned} y(k) &= f(x(k);\theta) + e(k) \\ f(x(k);\theta) &= \varphi(k)'\theta_0 + \sum_{i=1}^{M} \sigma_i \max\{\varphi(k)'\theta_i, 0\}, \end{aligned} \tag{6}$$

where $\varphi(k)' = [x(k)'\ 1]$, $\theta' = [\theta_0'\ \theta_1' \dots \theta_M']$, and $\sigma_i \in \{-1, 1\}$ are fixed a priori. It is easy to see that HHARX models form a subclass of PWARX models for which the PWA map $f(\cdot)$ is continuous. The number of submodels s is bounded by the quantity $\sum_{j=0}^{n} \binom{M}{j}$, which only depends on the dimension n of the regressor space and the number M of hinge functions.

The identification problem considered in [25, 26] selects the optimal parameter vector θ^* by solving

$$\theta^* = \arg\min_{\theta} \sum_{k=1}^{N} |y(k) - f(x(k); \theta)|^p, \tag{7}$$

where $p = 1$ or 2. Assuming a priori known bounds on θ (which may be taken arbitrarily large), problem (7) can be reformulated as a mixed-integer linear or quadratic program (MILP/MIQP) by introducing binary variables

$$\delta_i(k) = \begin{cases} 0 & \text{if } \varphi(k)'\theta_i \leq 0 \\ 1 & \text{otherwise,} \end{cases} \tag{8}$$

and auxiliary continuous variables $z_i(k) = \max\{\varphi(k)'\theta_i, 0\}$. The MILP/MIQP problems can then be solved for the global optimum.

The optimality of the described algorithm comes at the cost of a theoretically very high worst-case computational complexity, which means that it is mainly suitable for small-scale problems. To be able to handle somewhat larger problems, different suboptimal approximations were proposed in [25]. Various extensions are also possible so as to handle non-fixed σ_i, discontinuities, general PWARX models, etc., again at the cost of increased computational complexity. For more details, see [25, 26].

3.2 Bayesian procedure

The Bayesian procedure [16, 17] is based on the idea of exploiting the available prior knowledge about the modes and the parameters of the hybrid system. The PVs θ_i are treated as random variables, and described through their probability density functions (*pdf*s) $p_{\theta_i}(\cdot)$. A priori knowledge on the parameters can be supplied to the procedure by choosing appropriate a priori parameter *pdf*s $p_{\theta_i}(\cdot; 0)$. Various parameter estimates, such as expectation or maximum a posteriori probability estimate, can be easily obtained from the parameter *pdf*s. 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 in [16, 17].

Data classification and parameter estimation are carried out through sequential processing of the collected data points. At iteration k the data point $\big(y(k), x(k)\big)$ is considered, and attributed to the mode $\hat{\mu}(k)$ using maximum likelihood. Then, the a posteriori *pdf* of $\theta_{\hat{\mu}(k)}$ is computed using as a fact that $\big(y(k), x(k)\big)$ was generated by mode $\hat{\mu}(k)$. To numerically implement

the described procedure, particle filtering algorithms are used (see, *e.g.*, [3]). After the parameter estimation phase, each data point is 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 [16, 17]. For each data point $\big(y(k), x(k)\big)$ attributed to mode i, the price for misclassification into mode j is defined as

$$\nu_{ij}(x(k)) = \log \frac{p(\big(y(k), x(k)\big) \mid \mu(k) = i)}{p(\big(y(k), x(k)\big) \mid \mu(k) = j)}, \tag{9}$$

where $p(\big(y(k), x(k)\big) \mid \mu(k) = \ell)$ is the likelihood that $\big(y(k), x(k)\big)$ was generated by mode ℓ. Prices for misclassification are plugged into MRLP.

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 a priori parameter *pdf*s $p_{\theta_i}(\cdot; 0)$, and the *pdf* $p_e(\cdot)$ of the error term.

3.3 Bounded-error procedure

The main feature of the bounded-error procedure [5, 6, 23] is to impose that the error $e(k)$ in (1) is bounded by a given quantity $\delta > 0$ for all the samples in the estimation data set $\mathcal{D}$.

At *initialization*, the estimation of the number of submodels s, data classification and parameter estimation are performed simultaneously by partitioning the (typically infeasible) set of N linear complementary inequalities

$$|y(k) - \varphi(k)'\theta| \leq \delta, \quad k = 1, \ldots, N, \tag{10}$$

where $\varphi(k)' = [x(k)'\ 1]$, into a minimum number of feasible subsystems (MIN PFS problem). Then, an iterative *refinement* procedure is applied in order to deal with data points $\big(y(k), x(k)\big)$ satisfying $|y(k) - \varphi(k)'\theta_i| \leq \delta$ for more than one θ_i. These data are termed *undecidable*. The refinement procedure alternates between data reassignment and parameter update. If desirable, it enables the reduction of the number of submodels. For given positive thresholds α and β, submodels i and j are merged if $\alpha_{i,j} < \alpha$, where

$$\alpha_{i,j} = \|\theta_i - \theta_j\|_2 / \min\{\|\theta_i\|_2, \|\theta_j\|_2\}. \tag{11}$$

Submodel i is discarded if the cardinality of the corresponding data cluster $\mathcal{D}_i$ is less than βN. Data points that do not satisfy $|y(k) - \varphi(k)'\theta_i| \leq \delta$ for any θ_i are discarded as *infeasible* during the classification process, making it possible to detect outliers. In [5, 6] parameter estimates are computed through the ℓ_∞ projection estimator, but any other projection estimate, such as least squares, can be used [20].

The bounded-error procedure requires that the model orders n_a and n_b are fixed. The main tuning parameter is the bound δ: The larger δ, the smaller the required number of submodels at the price of a worse fit of the data.

The optional parameters α and β, if used, also implicitly determine the final number of submodels returned by the procedure. Another tuning parameter is the number of nearest neighbors c used to attribute undecidable data points to submodels in the refinement step.

3.4 Clustering-based procedure

The clustering-based procedure [13] is based on the rationale that small subsets of regression vectors that lie close to each other could be very likely attributed to the same region and the same submodel. The main steps of the procedure are hence the following:

- *Local regression.* For $k = 1, \ldots, N$, a local data set $\mathcal{C}_k$ is built by collecting $\big(y(k), x(k)\big)$ and the data points $\big(y(j), x(j)\big)$ corresponding to the $c-1$ regression vectors $x(j)$ that are nearest to $x(k)$. Local parameter vectors θ_k^{LS} are then computed for each local data set $\mathcal{C}_k$ through least squares.
- *Construction of feature vectors.* The centers $m_k = \frac{1}{c}\sum_{(y,x)\in\mathcal{C}_k} x$ are computed, and the feature vectors $\xi_k = [(\theta_k^{LS})'\ m_k']'$ are formed.
- *Clustering.* Feature vectors are partitioned into s groups $\{\mathcal{F}_i\}_{i=1}^s$ through clustering. To this aim, a "K-means"-like algorithm exploiting suitably defined confidence measures on the feature vectors is used. The confidence measures allow to reduce the influence of outliers and poor initializations.
- *Parameter estimation.* Since the mapping of the data points onto the feature space is bijective, the data clusters $\{\mathcal{D}_i\}_{i=1}^s$ can be easily built according to the rule: $\big(y(k), x(k)\big) \in \mathcal{D}_i \leftrightarrow \xi_k \in \mathcal{F}_i$. The PVs $\{\theta_i\}_{i=1}^s$ are estimated from data clusters through 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 its main tuning knob. A modification to the clustering-based procedure is proposed in [12] to allow for the simultaneous estimation of the number of submodels.

3.5 Quality measures

In the following sections the four procedures described above will be compared on suitably defined test examples. To this aim, some quantitative measures for assessing the quality of the identification results are introduced here.

When the true system generating the data is known and belongs to the considered model class, the accuracy of the estimated PVs can be evaluated by computing the quantity:

$$\Delta_\theta = \max_{i=1,\ldots,s} \frac{\|\theta_i - \bar{\theta}_i\|_2}{\|\bar{\theta}_i\|_2}\,, \tag{12}$$

where $\bar{\theta}_i$ and θ_i are the true and identified PVs for mode i, respectively. Δ_θ is zero for the perfect estimates, and increases as estimates get worse. In general,

a sensible quality measure for the estimated regions is harder to define. For the one-dimensional bi-modal case ($n = 1$ and $s = 2$) the following measure:

$$\Delta_{\mathcal{X}} = \left| \frac{m_{12}}{M_{12}} - \frac{\bar{m}_{12}}{\bar{M}_{12}} \right| \tag{13}$$

is used, where $\bar{M}_{12}$, $\bar{m}_{12}$, M_{12}, m_{12} are the coefficients of the true and estimated separating hyperplane (5), respectively.

A general quality measure, which is also applicable when the true system is not known, is provided by the averaged sum of the squared residuals:

$$\hat{\sigma}_e^2 = \frac{1}{s} \sum_{i=1}^{s} \frac{\mathrm{SSR}_i}{|\mathcal{D}_i|}, \tag{14}$$

where the set $\mathcal{D}_i$ contains the data points classified to mode i, $|\cdot|$ here denotes the cardinality of a set, and the sum of squared residuals (SSR) for mode i is defined as follows:

$$\mathrm{SSR}_i = \sum_{(y(k),x(k))\in\mathcal{D}_i} (y(k) - [x(k)'\ 1]\theta_i)^2. \tag{15}$$

The quality of the identified model is considered acceptable if $\hat{\sigma}_e^2$ is small and/or close to the expected noise variance of the true system.

Models with good one-step ahead prediction properties may perform poorly in simulation. To evaluate the model performance in simulation, a suitable measure of fit is

$$FIT = 100 \cdot \left(1 - \frac{\|\hat{\mathbf{y}} - \mathbf{y}\|_2}{\|\mathbf{y} - \bar{y}\|_2}\right), \tag{16}$$

where $\mathbf{y} = [y(1) \ldots y(N)]'$ is the vector of system outputs, $\bar{y}$ is the mean value of $\mathbf{y}$, and $\hat{\mathbf{y}} = [\hat{y}(1) \ldots \hat{y}(N)]'$ is the vector of simulated outputs, obtained by building regression vectors $x(k)$ from real inputs and previously simulated outputs. (16) can be interpreted as the percentage of the output variation that is explained by the model.

In experimental identification, (14) and (16) are useful for selecting good models from a set obtained by applying each identification procedure with different values of the tuning parameters and/or of the model orders.

4 Intersecting hyperplanes

From the descriptions in the previous section, it is evident that each identification procedure implements a different approach to parameter estimation and data classification. The aim of this section is to evaluate how the different procedures are able to deal with data points that are consistent with more

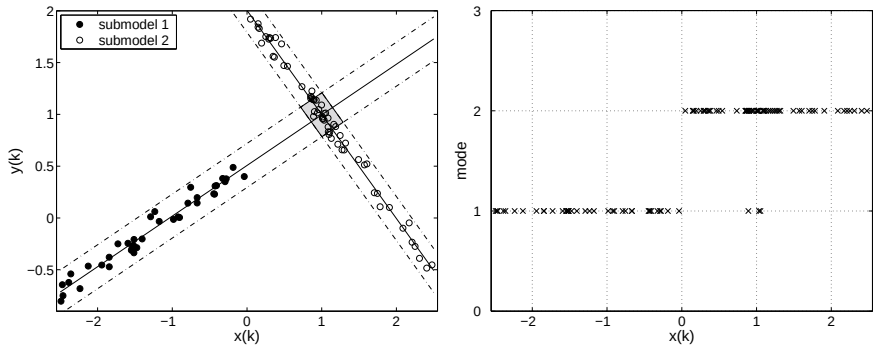

Fig. 1. Left: Results of data classification for the bounded-error, clustering-based and MIP procedures. All the three procedures yield $\Delta_\theta = 0.0186$ (using least squares) and $\Delta_\mathcal{X} = 0.0055$ (using RLP). **Right**: Data classification by attributing each data point to the submodel which generates the smallest prediction error.

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 indeed lead to misclassifications when estimating the polyhedral regions.

In order to illustrate this problem, an example where the submodels of the true system intersect over the regressor set $\mathcal{X}$ is designed. Consider the one-dimensional PWARX system $y(k) = \bar{f}(x(k)) + \eta(k)$, where the additive noise $\eta(k)$ is normally distributed with zero mean and variance $\sigma_\eta^2 = 0.005$, and the PWA map $\bar{f}(\cdot)$ is defined as:

$$\bar{f}(x) = \begin{cases} 0.5x + 0.5 & \text{if } \; x \in [-2.5,\, 0] \\ -x + 2 & \text{if } \; x \in (0,\, 2.5]. \end{cases} \tag{17}$$

$N = 100$ regressors $x(k)$ are generated. The 80% is uniformly distributed over $[-2.5,\, 2.5]$, and the remaining 20% over $[0.85,\, 1.15]$, so that the intersection of the two submodels is excited thoroughly. Results for the bounded-error, clustering-based and MIP procedures are shown in Figure 1, left. Note that an extension of the MIP procedure described in Section 3.1 was applied in order to handle discontinuities in the model.

In this example the three procedures classify correctly all the data points, and both the PVs and the switching threshold are estimated accurately. However, this might not be the case in general. For the clustering-based procedure the quality of the results depends on the choice of the cardinality c of the local data sets (see Section 7). In addition, the clustering may fail when there is a large variance on the centers m_k corresponding to similar θ_k^{LS}.

The bounded-error procedure is applied with $\delta = 3\,\sigma_\eta$ and $c = 10$. The gray area in Figure 1, left, represents the region of all possible *undecidable* data points for the fixed δ. In [5] undecidable data points were discarded during the classification process to avoid errors in the region estimation phase.

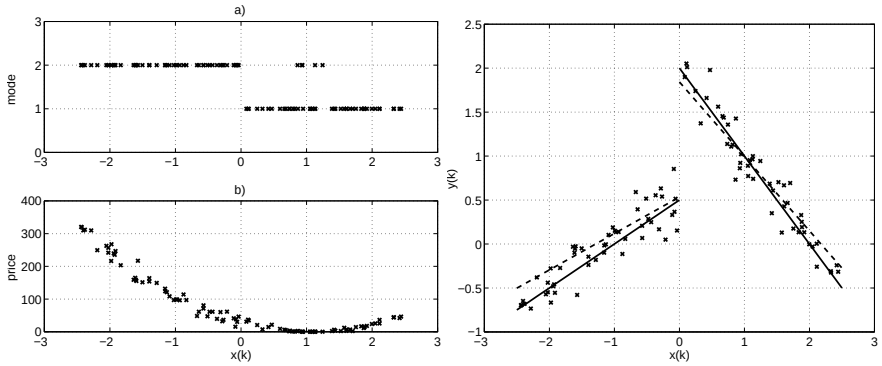

Fig. 2. Left: a) Data points attributed to modes. b) Pricing function for misclassifications. **Right:** Estimation data set (crosses), true system (solid lines) and estimated model (dashed lines). The Bayesian procedure yields $\Delta_\theta = 0.1366$ and $\Delta_\mathcal{X} = 0.0228$.

However, in this way a non-negligible amount of information is lost when a large number of undecidable data points shows up. In [6, 23] undecidable data points are attributed to submodels by considering the assignments of the c nearest neighbors. Also for this method, classification results depend on the choice of c (see again Section 7). It is interesting to point out that, if parameter estimates are computed through the ℓ_∞ projection estimator, one gets $\Delta_\theta = 0.0671$ in this example. As expected, parameter estimates are worse than using least squares, since the noise of the true system is normally distributed.

Classification results of the three procedures are compared to those obtained by attributing each data point to the submodel which generates the smallest prediction error [29]. Results using this approach are shown in Figure 1, right. Three data points around the intersection of the two submodels are misclassified. This leads to non-linearly separable classes which determine a larger error in the estimation of the switching threshold ($\Delta_\mathcal{X} = 0.0693$ in this example using RLP).

The Bayesian procedure is applied on a different data set, shown in Figure 2, right. The procedure is initialized with a priori parameter *pdf*s $p_{\theta_1}(\cdot;0) = p_{\theta_2}(\cdot;0) \sim \mathcal{U}([-2.5, 2.5] \times [-2.5, 2.5])$, where $\mathcal{U}(I)$ denotes the uniform distribution over the set I. Note that a priori parameter *pdf*s overlap. Results of data classification are shown in Figure 2, left(a). There are five wrongly classified data points. These points are close to the intersection of the two submodels. The misclassification pricing function (9) is plotted in Figure 2, left(b). The weight for misclassification of the wrongly attributed data points is small compared to the weight for misclassification of those correctly attributed.

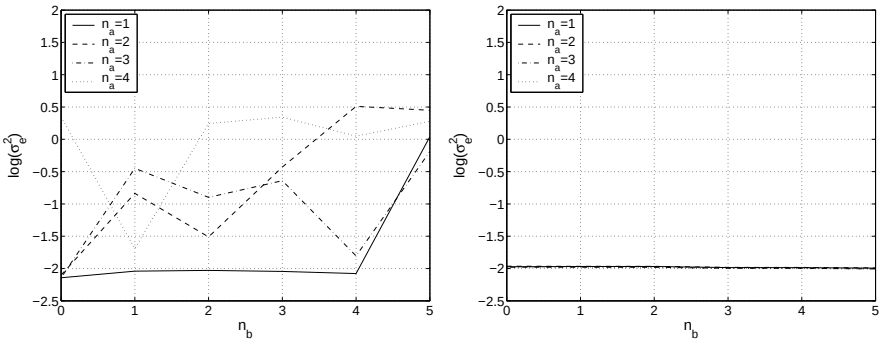

Fig. 3. Left: $\hat{\sigma}_e^2$ for the clustering-based procedure ($s = 2$, $c = 20$). **Right:** $\hat{\sigma}_e^2$ for the MIP procedure (2-norm criterion, $M = 1$).

5 Overestimation of model orders

The four identification procedures require that the model orders n_a and n_b are fixed. In order to investigate the effects of overestimated model orders, the one-dimensional PWAR system $y(k) = \bar{f}(y(k-1)) + \eta(k)$ is considered, where the additive noise $\eta(k)$ is normally distributed with zero mean and variance $\sigma_\eta^2 = 0.01$, and the PWA map $\bar{f}(\cdot)$ is defined as:

$$\bar{f}(x) = \begin{cases} 2x + 10 & \text{if } x \in [-10, 0) \\ -1.5x + 10 & \text{if } x \in [0, 10]. \end{cases} \tag{18}$$

The sequence $y(k)$ is generated with initial condition $y(0) = -10$. A fictitious input sequence is also generated as $u(k) \sim \mathcal{U}([-10, 10])$. The four identification procedures are applied for all combinations of $n_a = 1, \ldots, 4$ and

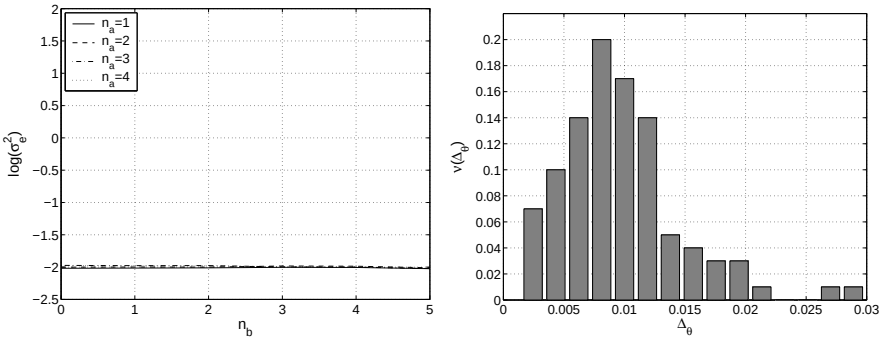

Fig. 4. Left: $\hat{\sigma}_e^2$ for the bounded-error procedure ($\delta = 3\,\sigma_\eta$, $c = 10$, α and β not used). **Right:** Approximate distribution of Δ_θ over $Q = 100$ runs using the bounded-error procedure and different realizations of Gaussian noise with $\sigma_\eta^2 = 0.075$.

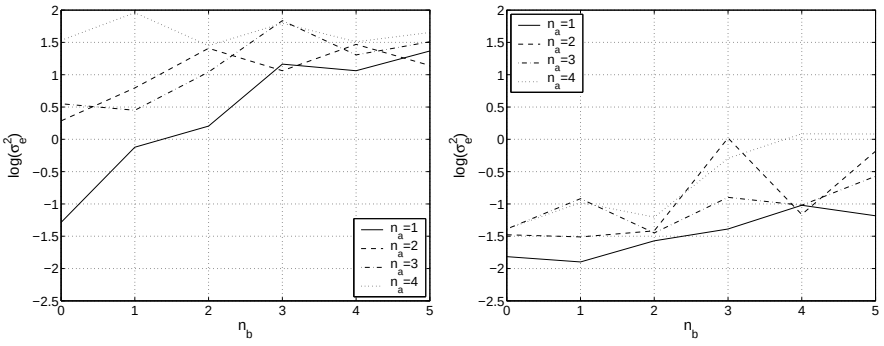

Fig. 5. Left: $\hat{\sigma}_e^2$ for the Bayesian procedure with $s = 2$ and unprecise initial parameter *pdf*s. **Right:** With precise initial parameter *pdf*s.

$n_b = 0, 1, \ldots, 5$, and $\hat{\sigma}_e^2$ is computed for each identified model. Note that the true system orders are $n_a = 1$ and $n_b = 0$.

Figure 3, left, shows the log-values of $\hat{\sigma}_e^2$ for models with different model orders identified by the clustering-based procedure. For true system orders the procedure identifies a model with $\hat{\sigma}_e^2$ close to the noise variance, but the performance significantly deteriorates when the model orders are overestimated. This is due to the adopted rationale that data points close to each other in the regressor space could be very likely attributed to the same submodel. When overestimating the model orders, the regression vector is extended with elements which do not contain relevant information for identification, but alter the distances in the feature space. This may determine misclassifications during clustering, and consequent bad estimates of the PVs and of the regions.

Since the MIP procedure solves the optimization problem (7) at the optimum ($p = 2$ is considered since the noise is normally distributed), from Figure 3, right, it is apparent that the procedure has no difficulties in estimating the over-parameterized models.

Results for the bounded-error procedure are shown in Figure 4, left. For the case $n_a = 1$, $n_b = 0$, a value of δ allowing to obtain $s = 2$ submodels is sought. The procedure is then applied to the estimation of the over-parameterized models using the same δ. When extending the regression vector, the minimum number of feasible subsystems of (10) does not increase, and remains equal in this example. Hence, the minimum partition obtained for $n_a = 1$, $n_b = 0$ is also a solution in the over-parameterized case. This explains the very good results shown in Figure 4, left, which are comparable to those obtained by the MIP procedure. The enhanced version [6, 23] of the greedy algorithm [1] is applied here for solving the MIN PFS problem. For completeness, it is reported that the corresponding values of Δ_θ obtained[1] range between 0.005 and 0.012.

[1]To compute Δ_θ, the entries of the true parameter vectors corresponding to superfluous elements in the regression vector are set to 0.

Values of $\hat{\sigma}_e^2$ for the Bayesian procedure with two different initializations are shown in Figure 5. In the left plot, the a priori parameter *pdf*s for the case $n_a = 1$, $n_b = 0$ are chosen as $p_{\theta_1}(\cdot;0) = p_{\theta_2}(\cdot;0) \sim \mathcal{U}([-5,5] \times [-20,20])$. For increased orders, added elements in the parameter vectors are taken to be uniformly distributed in the interval $[-5,5]$ (while their "true" value should be 0). In the right plot, the a priori parameter *pdf*s for the case $n_a = 1$, $n_b = 0$ are chosen as $p_{\theta_1}(\cdot;0) \sim \mathcal{U}([0,4] \times [8,12])$ and $p_{\theta_2}(\cdot;0) \sim \mathcal{U}([-4,0] \times [8,12])$, and all added elements are taken to be uniformly distributed in the interval $[-0.5,0.5]$. This example shows the importance of proper choices for the initial parameter *pdf*s in the Bayesian procedure. With precise initial *pdf*s the algorithm estimates relatively accurate over-parameterized models. When the a priori information is not adequate, the performance rapidly deteriorates.

6 Effects of noise

This section addresses the effects of noise on the identification accuracy. The first issue of interest is the effect that different noise realizations with the same statistical properties have on the identification results. The second issue is how different statistical properties of the noise affect the identification results.

To shed some light on the above issues, an experiment is designed with the one-dimensional PWARX system $y(k) = \bar{f}(x(k)) + \eta(k)$, and the PWA map $\bar{f}(\cdot)$ defined as in (18). The additive noise $\eta(k)$ is normally distributed with zero mean and variance σ_η^2. A noiseless data set of $N = 100$ data points is generated with $x(k) \sim \mathcal{U}([-10,10])$. For fixed σ_η^2, $Q = 100$ noise realizations are drawn, and added to the noiseless data set. For each noise realization, a model is identified using the different procedures, and the value of Δ_θ is computed for the identified models. In this way an approximate distribution of Δ_θ for each σ_η^2 and each procedure can be constructed, and its mean and variance

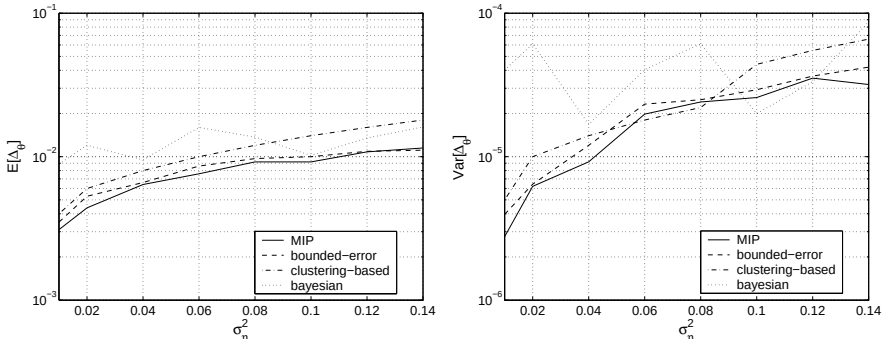

Fig. 6. Left: Estimated means of the Δ_θ distributions for several values of the noise variance σ_η^2. **Right:** Estimated variances.

can be estimated. Figure 4, right, shows one such distribution obtained using the bounded-error procedure.

Figure 6 shows the estimated means and variances of the Δ_θ distributions as functions of σ_η^2 for the four procedures. From the analysis of the plots, it is apparent that the MIP procedure achieves the best performance with respect to noise. The bounded-error procedure performs well when δ is chosen close to $3\sigma_\eta$ and the true PVs are quite different, as in this example. However, in practical situations such a value is unlikely to be available, and several trials are needed to find a suitable value for δ. As the noise level increases, the clustering-based procedure requires to increase the cardinality c of the local data sets in order to reduce the variance of the estimates (see Section 7 and the discussion in [13, Section 3.1]). With precise initialization as in Section 5, the Bayesian procedure achieves comparable performance to the other procedures, while with imprecise initialization the quality measures are the worst of all procedures (not shown in Figure 6).

7 Effects of varying the tuning parameters

The four identification procedures described in Section 3 require that some parameters which directly determine the structure of the identified models are fixed a priori. These are the model orders n_a and n_b for all procedures, the number of modes s for the Bayesian and the clustering-based procedure, and the number of hinge functions M for the MIP procedure. The Bayesian, bounded-error and clustering-based procedures also have several tuning parameters whose influence on identification results is not immediately obvious. In this section, the effects of varying the tuning parameters will be illustrated for these procedures by means of examples.

To investigate the role of the parameter c of the clustering-based procedure, an experiment is designed where the approximate distribution of Δ_θ is

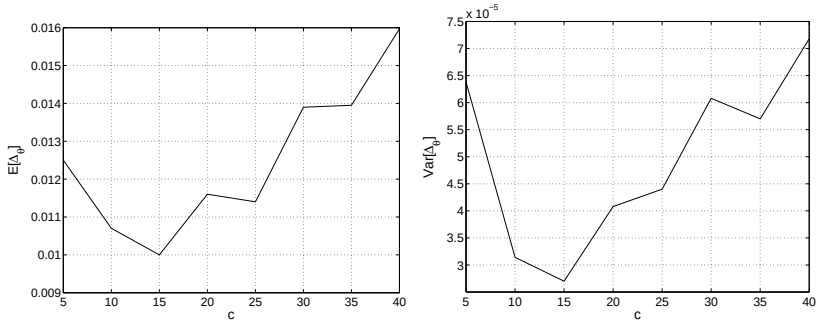

Fig. 7. Left: Estimated means of the Δ_θ distributions for several values of c using the clustering-based procedure, and $\sigma_\eta^2 = 0.075$. **Right:** Estimated variances.

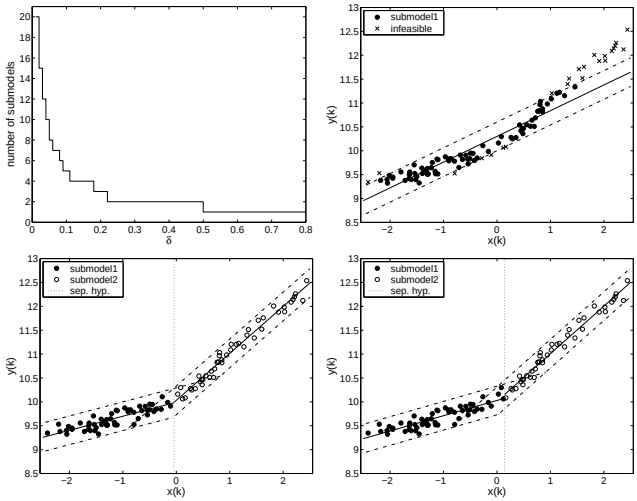

Fig. 8. Upper left: Plot of the estimated minimum number of submodels as a function of δ for the bounded-error procedure. **Upper right:** Results of data classification with $\delta = 0.3$ and $\alpha = 10\%$. **Lower left:** With $\delta = 0.3$, $\alpha = 2\%$ and $c = 5$, yielding $\Delta_\theta = 0.0035$ and $\Delta_\mathcal{X} = 0.0411$. **Lower right:** With $\delta = 0.3$, $\alpha = 2\%$ and $c = 40$, yielding $\Delta_\theta = 0.0035$ and $\Delta_\mathcal{X} = 0.1422$.

computed as described in Section 6 for different values of c and $\sigma_\eta^2 = 0.075$. The estimated means and variances of such distributions are shown in Figure 7. It is apparent that there exists an optimal value of c for which the identified model is most accurate. In the considered example the procedure gives the best results for $c = 15$, corresponding to the minimum of both curves in Figure 7. For the selection of c in practical cases, see [13].

The bounded-error procedure has several tuning parameters, and finding their right combination is not always straightforward. The role of the parameters δ, α and c will be investigated here by considering the one-dimensional PWARX system $y(k) = \bar{f}(x(k)) + \eta(k)$, where the additive noise $\eta(k)$ is normally distributed with zero mean and variance $\sigma_\eta^2 = 0.01$, and the PWA map $\bar{f}(\cdot)$ is defined as:

$$\bar{f}(x) = \begin{cases} 0.3x + 10 & \text{if } \ x \in [-2.5, 0] \\ x + 10 & \text{if } \ x \in (0, 2.5]. \end{cases} \tag{19}$$

$N = 100$ regressors $x(k)$ are generated uniformly distributed over $[-2.5, 2.5]$. Figure 8, upper left, shows a plot of the estimated minimum number of feasible subsystems of (10) as a function of δ. In general, such a plot can be used to select an appropriate value for δ close to knee of the curve. In this example, choosing δ in the suggested way allows to estimate the correct number of modes $s = 2$. Using $\delta = 0.3$, the initialization provides parameter estimates

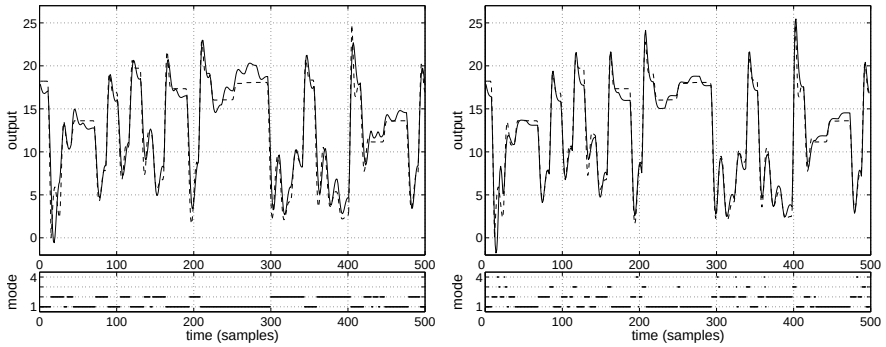

Fig. 9. Left: Simulation results on validation data using the PWARX model identified by the clustering-based procedure with $n_a = 2$, $n_b = 2$, $s = 2$ and $c = 90$ ($FIT = 74.7127\%$). **Right:** Simulation results on validation data using the HHARX model identified by the MIP procedure with $n_a = 2$, $n_b = 1$ and $M = 2$ ($FIT = 81.5531\%$). Solid - simulated output, dashed - system output.

with $\alpha_{1,2} = 5.9\%$. If α is selected greater than $\alpha_{1,2}$, the two submodels are merged into one during the refinement phase. In this case, a high number of infeasible data points shows up (Figure 8, upper right), which indicates poor fit of the data. Hence, the refinement procedure can be repeated using a smaller value of α. The role of c with respect to classification is illustrated in Figure 8, bottom. On the left, perfect classification is obtained using $c = 5$, whereas on the right three data points are misclassified using $c = 40$, which causes a larger error when estimating the switching threshold.

For the Bayesian procedure the most important tuning knobs are the a priori *pdf*s of the parameters. Effects of improper choices are clearly illustrated in Sections 5 and 6. The more precise the a priori knowledge on the system, the better the procedure is expected to perform (see also Section 8).

8 Experimental example

In this section the four procedures are applied to the identification of the electronic component placement process in pick-and-place machines[2]. Pick-and-place machines are used to automatically place electronic components on printed circuit boards. The process consists of a mounting head carrying the electronic component, which is pushed down until it comes in contact with the circuit board, and then is released. The input to the system is the voltage applied to the motor driving the mounting head. The output of the system is the position of the mounting head. A detailed description of the process and of the experimental setup can be found in [15, 17].

[2]The authors would like to thank Hans Niessen for some of the results presented in this section.

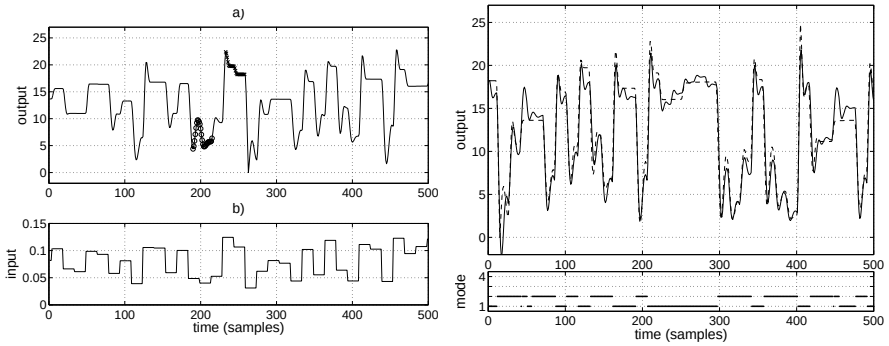

Fig. 10. Left: Data set used for identification with the Bayesian procedure. a) Output signal. Data marked with ∘ and × are those used for initializing the free mode and the impact mode, respectively. b) Input signal. **Right:** Simulation results on validation data using the identified PWARX model with $n_a = 2$, $n_b = 2$ and $s = 2$ ($FIT = 77.1661\%$). Solid - simulated output, dashed - system output.

A data record over an interval of $T = 15$ s is available. The data sets used for identification with the Bayesian, clustering-based and MIP procedures are sampled at 50 Hz, while the data set used for identification with the bounded-error procedure is sampled at 150 Hz. Two modes of operation can be distinguished through physical insight into the process. In the *free mode* the mounting head moves unconstrained, whereas in the *impact mode* the carried component is in contact with the circuit board. Hard nonlinear phenomena due to dry friction are also present.

Results obtained by applying the four identification procedures are shown in Figures 9, 10 and 11. Note that the aim in this section is only to show that all the four procedures are able to estimate sensible models of the experimental process. A fair comparison of models identified by different procedures is not always possible here, mainly because they were obtained using different model structures and/or different data sets.

For identification using the Bayesian and clustering-based procedures, a data set consisting of 750 samples is partitioned into two overlapping sets of 500 points each. The first set is used for estimation, and the second for validation. Figure 9, left, shows the simulation results on validation data for the best model identified by the clustering-based procedure. The best model is obtained for a high value of c. A possible explanation of this, given in [15, 17], is that using large local data sets the effects of dry friction can be "averaged out" as a process noise. Differences between the measured and simulated responses due to unmodeled dry friction are clearly visible on the time interval $(225, 300)$.

Physical insight into the process helps the initialization of the Bayesian procedure. Although the mode switch does not occur at a fixed height of the mounting head, data points below a certain height can be most likely

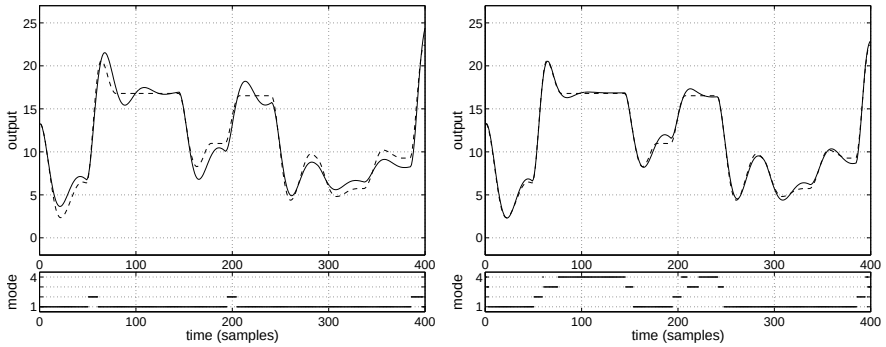

Fig. 11. Left: Simulation results on validation data using the model with $s = 2$ modes identified by the bounded-error procedure ($FIT = 81.3273\%$). **Right:** Using the model with $s = 4$ modes ($FIT = 93.4758\%$). Solid - simulated output, dashed - system output.

attributed to the free mode. A similar consideration holds for the impact mode, see Figure 10, left(a). The a priori information is exploited to obtain rough estimates θ_i^{LS} of the PVs of the two modes through least squares. As in [13], the empirical covariance matrix V_i of θ_i^{LS} is also computed. The a priori parameter pdfs are then taken to be normal distributions with means θ_i^{LS} and covariance matrices V_i. Simulation results on validation data are shown in Figure 10, right. It is apparent that the identified model benefits from the prior knowledge, and yields a higher value of FIT than the model obtained through the clustering-based procedure.

The MIP procedure identifies a model with $n_a = 2$, $n_b = 1$ and $M = 2$ using $N = 150$ estimation data points. The 2-norm criterion is considered, and $s = 4$ submodels are obtained. Simulation results on the same validation data set as for the Bayesian and clustering-based procedures are shown in Figure 9, right. Although a smaller number of data points is used to estimate the model, it is apparent that the use of more than two submodels is favorable to improve the fit.

The bounded-error procedure identifies models with orders $n_a = 2$ and $n_b = 2$ using $N = 1000$ estimation data. Two models with $s = 2$ and $s = 4$ modes are identified for $\delta = 0.06$ and $\delta = 0.04$, respectively. For $s = 4$, multi-category RLP [8] is used for region estimation. Simulation results on validation data are shown in Figure 11. Again, it is apparent that the fit improves as the number of submodels increases, *i.e.*, as δ decreases. The active mode evolution at the bottom of Figure 11, left, clearly shows that one of the two submodels is active in situations of high incoming velocity of the mounting head (*i.e.*, rapid transitions from low to high values of the mounting head position). One submodel modeling the same situation is also present in the identified model with $s = 4$ modes.

9 Conclusions

In this chapter the problem of PWA identification has been addressed by studying four recently proposed techniques for the identification of PWARX or HHARX models. The four techniques have been compared on suitably defined test examples, and tested on the experimental identification of the electronic component placement process in pick-and-place machines.

Identification using the clustering-based procedure is straightforward, as only one parameter has to be tuned. However, poor results are obtained when the model orders are overestimated, since distances in the feature space become corrupted by irrelevant information. The bounded-error procedure gives good results when the right combination of the tuning parameters is found, but several attempts are often needed for finding such a combination. The Bayesian procedure is designed to take advantage of prior knowledge on the system, and has been shown to be very effective in the pick-and-place machine identification. The MIP procedure provides globally optimal results and needs no parameter tuning, but requires the solution of MILP/MIQP problems whose theoretical worst-case computational complexity is very high.

References

1. Amaldi E, Mattavelli M (2002) The MIN PFS problem and piecewise linear model estimation. Discrete Applied Mathematics 118:115–143
2. Antsaklis PJ, Nerode A (eds) (1998) Special issue on hybrid systems. IEEE Transactions on Automatic Control 43(4):457–579
3. Arulampalam MS, Maskell S, Gordon N, Clapp T (2002) A tutorial on particle filters for online nonlinear/non-gaussian Bayesian tracking. IEEE Transactions on Signal Processing 50(2):174–188
4. Bemporad A, Ferrari-Trecate G, Morari M (2000) Observability and controllability of piecewise affine and hybrid systems. IEEE Transactions on Automatic Control 45(10):1864–1876
5. Bemporad A, Garulli A, Paoletti S, Vicino A (2003) A greedy approach to identification of piecewise affine models. In: Maler O, Pnueli A (eds) Hybrid Systems: Computation and Control. Lecture Notes on Computer Science pp. 97–112. Springer Verlag
6. Bemporad A, Garulli A, Paoletti S, Vicino A (2004) Data classification and parameter estimation for the identification of piecewise affine models. In: Proceedings of the 43rd IEEE Conference on Decision and Control
7. Bennett KP, Mangasarian OL (1992) Robust linear programming discrimination of two linearly inseparable sets. Optimization Methods and Software 1:23–34
8. Bennett KP, Mangasarian OL (1994) Multicategory discrimination via linear programming. Optimization Methods and Software 3:27–39
9. Bredensteiner EJ, Bennett KP (1999) Multicategory classification by support vector machines. Computational Optimization and Applications 12:53–79
10. Breiman L (1993) Hinging hyperplanes for regression, classification, and function approximation. IEEE Transactions on Information Theory 39(3):999–1013

11. Cortes C, Vapnik V (1995) Support-vector networks. Machine Learning 20:273–297
12. Ferrari-Trecate G, Muselli M (2003) Single-linkage clustering for optimal classification in piecewise affine regression. In: Engell S, Gueguen H, Zaytoon J (eds) Proceedings IFAC Conference on Analysis and Design of Hybrid Systems
13. Ferrari-Trecate G, Muselli M, Liberati D, Morari M (2003) A clustering technique for the identification of piecewise affine systems. Automatica 39(2):205–217
14. Heemels WPMH, De Schutter B, Bemporad A (2001) Equivalence of hybrid dynamical models. Automatica 37(7):1085–1091
15. Juloski A, Heemels WPMH, Ferrari-Trecate G (2004) Data-based hybrid modelling of the component placement process in pick-and-place machines. Control Engineering Practice 12(10):1241–1252
16. Juloski A, Wieland S, Heemels WPMH (2004) A Bayesian approach to identification of hybrid systems. In: Proceedings 43rd IEEE Conference on Decision and Control
17. Juloski A (2004) Observer design and identification methods for hybrid systems. PhD thesis, Department of Electrical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands
18. Lin JN, Unbehauen R (1992) Canonical piecewise-linear approximations. IEEE Transactions on Circuits and Systems – I: Fundamental Theory and Applications 39(8):697–699
19. Ljung L (1999) System Identification: Theory for the User. 2nd ed., Prentice Hall
20. Milanese M, Vicino A (1991) Optimal estimation theory for dynamic systems with set membership uncertainty: an overview. Automatica 27(6):997–1009
21. Morse AS, Pantelides CC, Sastry SS, Schumacher JM (eds) (1999) Special issue on hybrid systems. Automatica 35(3):347–535
22. Münz E, Krebs V (2002) Identification of hybrid systems using apriori knowledge. In: Proceedings 15th IFAC World Congress
23. Paoletti S (2004) Identification of piecewise affine models. PhD thesis, Department of Information Engineering, University of Siena, Siena, Italy
24. Pfetsch ME (2002) The maximum feasible subsystem problem and vertex-facet incidences of polyhedra. PhD thesis, Technischen Universität Berlin, Berlin, Germany
25. Roll J (2003) Local and piecewise affine approaches to system identification. PhD thesis, Department of Electrical Engineering, Linköping University, Linköping, Sweden
26. Roll J, Bemporad A, Ljung L (2004) Identification of piecewise affine systems via mixed-integer programming. Automatica 40(1):37–50
27. Sjöberg J, Zhang Q, Ljung L, Benveniste A, Delyon B, Glorennec P, Hjalmarsson H, Juditsky A (1995) Nonlinear black-box modeling in system identification: a unified overview. Automatica 31(12):1691–1724
28. Van der Schaft AJ, Schumacher JM (2000) An introduction to hybrid dynamical systems. Vol. 251 of Lecture Notes in Control and Information Sciences. Springer Verlag
29. Vidal R, Soatto S, Ma Y, Sastry S (2003) An algebraic geometric approach to the identification of a class of linear hybrid systems. In: Proceedings 42nd IEEE Conference on Decision and Control