Comparison of four procedures for the identification of hybrid systems

A.Lj. Juloski[1], W.P.M.H. Heemels[2], G. Ferrari-Trecate[3], R. Vidal[4], S. Paoletti[5], and J.H.G. Niessen[6]

[1] Department of Electrical Engineering, Eindhoven University of Technology PO Box 513, 5600MB Eindhoven, The Netherlands, e-mail: a.juloski@tue.nl
[2] Embedded Systems Institute, PO Box 513, 5600 MB Eindhoven, The Netherlands e-mail: maurice.heemels@embeddedsystems.nl
[3] INRIA, Domaine de Voluceau, Rocquencourt - B.P.105, 78153, Le Chesnay Cedex, France, e-mail: Giancarlo.Ferrari-Trecate@inria.fr
[4] Center for Imaging Science, Johns Hopkins University, 308B Clark Hall, 3400 N Charles St, Baltimore, MD 21218, USA, e-mail: rvidal@cis.jhu.edu
[5] Dipartimento di Ingegneria dell'Informazione, Universita' di Siena Via Roma 56, 53100 Siena, Italy, e-mail: paoletti@dii.unisi.it
[6] Nyquist, Industrial Control, P.O. Box 7170, 5605 JD Eindhoven, The Netherlands e-mail: h.niessen@nyquist.com

Abstract. In this paper we compare four recently proposed procedures for the identification of PieceWise AutoRegressive eXogenous (PWARX) and switched ARX models. We consider the clustering-based procedure, the bounded-error procedure, and the Bayesian procedure which all identify PWARX models. We also study the algebraic procedure, which identifies switched linear models. We introduce quantitative measures for assessing the quality of the obtained models. Specific behaviors of the procedures are pointed out, using suitably constructed one dimensional examples. The methods are also applied to the experimental identification of the electronic component placement process in pick-and-place machines.

1 Introduction

In this paper we study four recently proposed procedures for the identification of discrete time piecewise affine (PWA) models. The identification procedures that we compare are the clustering-based procedure [1], the bounded-error procedure [2, 3], the Bayesian procedure [4] and the algebraic procedure [5, 6] (see section 2 for brief descriptions). Of course, there are other methods available in literature, for instance the work given in [7] and [8]. However, due to the specific knowledge of the authors and space limitations, the attention is restricted to the four procedures mentioned before.

There is not much known how the procedures compare in particular situations. Some features of the clustering-based procedure have been analyzed theoretically in [9], but a formal analysis of the properties of the bounded-error,

algebraic and Bayesian procedures for noisy data is currently not available. Therefore, we will study specific examples of PWA models that can help us better understand properties of the methods in practical situations.

To be precise, the PWA models that the clustering-based, bounded-error and the Bayesian procedures identify are PieceWise ARX (PWARX) models of the form:

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

where $e(k)$ is the error term and the PWA map $f(\cdot)$ is defined as:

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

In (2) $x(k)$ is a vector of regressors defined as

$$\begin{aligned} x(k) \triangleq [\, & y(k-1)\, y(k-2)\, \ldots\, y(k-n_a) \\ & u'(k-1)\, u'(k-2)\, \ldots\, u'(k-n_b)\,]', \end{aligned} \tag{3}$$

where k is the time index and $y \in \mathbb{R}$, $u \in \mathbb{R}^m$ are the outputs and the inputs of the system, respectively. For $i = 1, \ldots s$, $\theta_i \in \mathbb{R}^{n+1}$ is a parameter vector (PV) with $n = n_a + n_b$.

The bounded regressor space $\mathbb{X}$ is partitioned into s convex polyhedral regions $\{\mathcal{X}_i\}_{i=1}^s$, i.e.

$$\bigcup_{i=1}^{s} \mathcal{X}_i = \mathbb{X} \subset \mathbb{R}^n \quad \text{and} \quad \mathcal{X}_i \cap \mathcal{X}_j = \varnothing \quad i \neq j. \tag{4}$$

When the partition $\{\mathcal{X}_i\}_{i=1}^s$ is known we can define the mode $\mu(k)$ of the data pair $(x(k), y(k)),\ k = 1, \ldots, N$ uniquely as:

$$\mu(k) := i \text{ if } x(k) \in \mathcal{X}_i. \tag{5}$$

The algebraic procedure identifies switched linear models of the form (1), where

$$f(x) = [x' \quad 1]\theta_i,$$

and $i \in \{1, \ldots s\}$ is arbitrary for each time index k. The problems of estimation of parameters $\theta_1, \ldots, \theta_s$ for switched linear and PWARX models are closely related, and in the sequel we will treat them in parallel. In addition, the identification of PWARX models requires also the estimation of the regions $\mathcal{X}_i$, which would form an extension of the algebraic procedure.

The general identification problem reads as follows: given the data set $\mathcal{N} = \{(x(k), y(k))\}_{k=1}^N$ reconstruct the PWA map $f(\cdot)$, i.e. determine the PVs $\{\theta_i\}_{i=1}^s$ and the polyhedral partition $\{\mathcal{X}_i\}_{i=1}^s$.

Identification of PWARX models is a challenging problem since it involves the estimation of both the PVs $\{\theta_i\}_{i=1}^s$ and the regions of the regressor space $\{\mathcal{X}_i\}_{i=1}^s$ on the basis of the available data set $\mathcal{N}$. In case that regions of the

regressor space are known a priori the problem complexity reduces to that of s linear system identification problems [1].

In order to compare the procedures and asses the quality of the obtained models we propose several quantitative measures in section 2. These measures are "common sense" criteria (not the ones optimized by the methods themselves) and reflect practical needs for identification. In section 3 we will address different approaches to data classification of each of the procedures, and consequences on the accuracy of the identified model. In section 4 we will investigate the effects of the overestimation of model orders. In section 5 we will study the effects of noise. In section 6 we will apply the procedures for the experimental identification of the component placement process in pick-and-place machines. Finally, summary and conclusions are presented in section 7.

2 The compared procedures

In this section we briefly discuss the four procedures we compare. The basic steps that each method performs are: the estimation of the PVs $\{\theta_i\}_{i=1}^s$, the classification of the data points (grouping data points attributed to the i-th mode to the set $\mathcal{F}_i$, $i = 1, \ldots, s$) and the estimation of the corresponding regions $\{\mathcal{X}_i\}_{i=1}^s$, for PWARX models.

The first two steps are performed in a different way by each procedure, as discussed in the sequel, while the estimation of the regions can be done in the same way for all methods. The basic idea is as follows. Having the data points that are attributed to sets $\mathcal{F}_i$ and $\mathcal{F}_j$, we are looking for a separating hyperplane in the regressor space $\mathbb{X}$ described by:

$$M_{ij}' x = m_{ij}, \tag{6}$$

where M_{ij} is a vector, and m_{ij} is a scalar, so that for each $x(k) \in \mathcal{X}_i$, $M_{ij}' x(k) \leq m_{ij}$, and for each $x(k) \in \mathcal{X}_j$ $M_{ij}' x(k) > m_{ij}$. If such a hyperplane can not be found (i.e. the data set is not linearly separable) we are interested in a generalized separating hyperplane which minimizes the number of misclassified data points. The method we use for estimating the separating hyperplanes in this paper is Multicategory Robust Linear Programming (MRLP). This method can solve the classification problem with more than two data classes. For a detailed discussion on MRLP see [10].

2.1 Clustering-based procedure

The clustering-based procedure [1] is based on the rationale that regressors that lie close together are likely to belong to the same partition and the same ARX model. The main steps of the procedure are:

- For each data pair $(x(k), y(k))$ a local data set (LD) $\mathcal{C}_k$ is built containing its $c-1$ nearest datapoints[7] in the regressor space $\mathbb{X}$. LDs that only contain

[7] according to the Euclidean distance.

data pairs belonging to a single subsystem are referred to as *pure* LDs, while LDs containing data generated by different subsystems are called *mixed* LDs.
- Calculate θ_k^{LS} for each LD using least squares on $\mathcal{C}_k$ and compute the mean m_k of $\mathcal{C}_k$. Each datapoint $(x(k), y(k))$ is thereby mapped onto the feature vectors $\xi_k = [(\theta_k^{LS})', m_k']'$.
- Cluster the points $\{\xi_k\}_{k=1}^N$ in s clusters $\mathcal{D}_i$ by minimizing a suitable cost function.
- Since the mapping of the datapoints onto the feature space is bijective, the data subsets $\{\mathcal{F}_i\}_{i=1}^s$ can be built using the clusters $\{\mathcal{D}_i\}_{i=1}^s$. The PVs $\{\theta_i\}_{i=1}^s$ are estimated from data subsets $\mathcal{F}_i$ by least squares.

The clustering procedure requires the model orders n_a, n_b, and the number of models s. The parameter c is the tuning knob of this procedure.

2.2 Bounded-error procedure

The main feature of the bounded-error procedure [2, 3] 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 $\mathcal{N}$. 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

$$|y(k) - \varphi(k)'\theta| \leq \delta, \quad k = 1, \ldots, N, \tag{7}$$

where $\varphi(k)' = [x(k)'\ 1]$, into a minimum number of feasible subsystems (MIN PFS problem). MIN PFS problem is $\mathcal{NP}$-hard, and the suboptimal algorithm based on thermal relaxations is used. Then, an iterative *refinement* procedure is applied in order to deal with data points $(y(k), x(k))$ 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, and, if desirable, enables the reduction of the number of submodels. For given positive thresholds α and β, submodels i and j are merged if $\alpha_{i,j} < \alpha$, with

$$\alpha_{i,j} = \|\theta_i - \theta_j\|_2 / \min\{\|\theta_i\|_2, \|\theta_j\|_2\}, \tag{8}$$

whereas submodel i is discarded if the cardinality of the corresponding data cluster $\mathcal{F}_i$ is less than βN. In [2, 3] parameter estimates are computed through the ℓ_∞ projection estimator, but any other projection estimate, such as least squares, can be used [11].

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.

2.3 Bayesian procedure

The Bayesian procedure [4] is based on the idea of refining the available a priori knowledge about the modes and parameters of the hybrid system. Parameters θ_i of the piece-wise affine map (2) are treated as random variables, and described with their probability density functions (pdfs) $p_{\theta_i}(\cdot)$. A priori knowledge on the parameters can be supplied to the procedure by choosing appropriate a priori 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 in [4], based on sequential processing of data points in the collected data set. It is assumed that the probability density function of the additive noise term e, $p_e(\cdot)$ is given.

The parameter estimation algorithm has N iterations, and in each iteration the pdf of one of the parameters is refined. In the k-th iteration of the algorithm the most probable mode $\mu(k)$ of the data pair $(x(k), y(k))$ is computed, using the available pdfs of the parameter vectors from step $k-1$. Subsequently, the data pair $(x(k), y(k))$ is assigned to the mode i that most likely generated it, and the a posteriori pdf of parameter vector θ_i is computed, using as a fact that the pair $(x(k), y(k))$ was generated by mode i. To numerically implement the Bayesian procedure particle filtering algorithms are used (see e.g. [12]). In order to have a good representation of the pdf a large number of particles may be needed. This accounts for the majority of the computational burden.

After the parameter estimation phase, data points are attributed to the mode that most likely generated them. For the estimation of regions a modification of the standard MRLP procedure is proposed in [4]. Assume that the data point attributed to the mode i ends up in the region $\mathcal{X}_j$. If the probabilities that the data point is generated by both modes are approximately equal, this misclassification should not be penalized highly. Following this idea we introduce the non-negative valued *pricing functions*, which assign price to misclassification of data points. Pricing functions are plugged into the MRLP procedure.

The Bayesian procedure requires model orders n_a and n_b, and the number of modes s. The most important tuning parameters of the procedure are the a priori parameter pdfs $p_{\theta_i}(\cdot, 0)$, and the pdf of the additive noise p_e. Also, the particle filtering algorithm has several tuning parameters.

2.4 Algebraic procedure

The method proposed in [5, 6] approaches the problem of identifying the class of Switched ARX (SARX) models in an algebraic fashion. For deterministic models, it provides a global solution that is provably correct in the noiseless case, even when the number of models and the model orders are unknown and different. For stochastic models, it provides a sub-optimal solution that can be used to initialize any of the iterative approaches. The algebraic method exploits the fact that in the noiseless case ($e = 0$), the data pair $(x(k), y(k))$ satisfies $\boldsymbol{z}'(k)[1 \;\; \theta_i']' \doteq [y(k) \;\; -\varphi'(k)][1 \;\; \theta_i']' = y(k) - \varphi'(k)\theta_i = 0$ for a suitable PV θ_i.

Hence the following homogeneous polynomial of degree s holds for all k[8]

$$p_s(\boldsymbol{z}(k)) = \prod_{i=1}^{s} (\boldsymbol{z}'(k)[1 \;\; \theta_i']') = \nu_s(\boldsymbol{z}(k))' \boldsymbol{h}_s = 0, \tag{9}$$

where $\nu_n(\boldsymbol{z}(k))$ contains all $M_s(n_a, n_b) \doteq \binom{n_a+n_b+s+1}{s}$ monomials of degree s in $\boldsymbol{z}(k)$ and $\boldsymbol{h}_s \in \mathbb{R}^{M_s(n_a,n_b)}$ contains the coefficients of p_s. Therefore, the identification of multiple ARX models can be viewed as the identification of a single, though more complex, hybrid ARX model $\nu_s(\boldsymbol{z}(k))' \boldsymbol{h}_s = 0$ whose hybrid PV $\boldsymbol{h}_s$ depends on the parameters of the ARX models $\{\theta_i\}_{i=1}^s$, but not on the switching sequence or the switching mechanism. Since the polynomial $\boldsymbol{h}_s' \nu_s(\boldsymbol{z}(k)) = 0$ holds for all k, the hybrid PV can be identified by solving the following linear system (using least squares with noisy data)

$$[\nu_s(\boldsymbol{z}(1)) \cdots \nu_s(\boldsymbol{z}(k)) \cdots]' \boldsymbol{h}_s = 0 \quad \text{and} \quad \boldsymbol{h}_s(1) = 1. \tag{10}$$

This linear system has a unique solution when the data are sufficiently exciting and s, n_a and n_b are known perfectly. When only upper bounds $\bar{s}$, $\bar{n}_a$ and $\bar{n}_b$ for s, n_a and n_b, respectively, are available, one can still obtain a unique solution by noticing that the last entries of each θ_i are zero, hence the last entries of $\boldsymbol{h}_{\bar{s}}$ must also be zero. Determining the number of zero entries requires a tuning parameter in the case of noisy data. Given $\boldsymbol{h}_{\bar{s}}$, the number of models s is the number of non-repeated factors in $p_{\bar{s}}$ and the PVs of the original ARX models correspond to the last $\bar{n}_a + \bar{n}_b + 1$ entries of the vector of partial derivatives of $p_{\bar{s}}$, $\frac{\partial p_{\bar{s}}(\boldsymbol{z})}{\partial \boldsymbol{z}} \in \mathbb{R}^{\bar{n}_a+\bar{n}_b+2}$, evaluated at a point $\boldsymbol{z}_i \in \mathbb{R}^{\bar{n}_a+\bar{n}_b+2}$ that is generated by the ith ARX model and can be chosen automatically once $p_{\bar{s}}$ is known. Given the PVs, data pairs $(x(k), y(k)$ are attributed to the model λ satisfying the rule

$$\lambda(k) = \arg \min_{1 \le i \le s} (y(k) - \varphi(k)' \theta_i)^2. \tag{11}$$

This rule is applicable to SARX models, and by extension to all switching mechanisms. However, if additional knowledge about the switching mechanism (e.g. PWARX models) is available, more appropriate classification rules can be used.

2.5 Quality measures

Since our aim is to compare the procedures, some quantitative measures for the quality of the identification results are introduced. These measures will capture the accuracy of the estimated PVs $\{\hat{\theta}_i\}_{i=1}^s$ and the accuracy of the estimated partitions $\{\hat{\mathcal{X}}_i\}_{i=1}^s$.

When the model that generated the data is known, one can measure the accuracy of the identified PV through the quantity:

$$\Delta_\theta = \max_{1 \le i \le s} \left(\min_{1 \le j \le s} \frac{\|\hat{\theta}_i - \theta_j\|_2}{\|\theta_j\|_2} \right), \tag{12}$$

[8] This product equation was introduced independently in [13] in the particular case of $s = 2$ models.

where $\hat{\theta}_i$ are the reconstructed PVs and θ_j are the PVs of the generating model. This measure is only applicable for the cases where the number of submodels is the same for the generating and identified model. Δ_θ is zero for the perfect estimates, and increases as the estimates worsen.

A sensible quality measure for the estimated regions is much harder to define. For the case where $n = 1$ and $s = 2$ we propose the following index:

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

where M_{12}, m_{12}, $\hat{M}_{12}$, $\hat{m}_{12}$ are the coefficients of the separating hyperplanes, defined in (6), of the original and reconstructed model, respectively.

An overall quality measure which is also applicable when the generating model is not known is provided by the sum of squared residuals (one step ahead prediction errors):

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

where the set $\mathcal{F}_i$ contains the datapoints classified to submodel i and the sum of squared residuals (SSR) of submodel i is defined as:

$$\mathrm{SSR}_{\mathcal{F}_i} = \sum_{x(k) \in \mathcal{F}_i} (y(k) - [x(k)' \; 1]\theta_i)^2.$$

The value of the estimated model is considered acceptable if $\hat{\sigma}_\varepsilon^2$ is small and/or near the expected noise of the identified system.

Models with good one-step ahead prediction properties may perform poorly in simulation. To measure the model performance in simulation we propose to use the averaged Sum of the Squared simulation Errors ($\mathrm{SSE}_{\mathrm{sim}}$),

$$\mathrm{SSE}_{\mathrm{sim}} = \frac{1}{N-n} \sum_{k=n+1}^{N} \left(y(k) - \hat{y}(k) \right)^2, \tag{15}$$

where $\hat{y}(k)$ is the output of the simulation obtained by building $x(k)$ from the real inputs and previously estimated outputs. The idea behind (15) is that poorly estimated regions may increase the simulation error, since these poor estimates may lead to wrong choices of the next submodel.

When doing experimental identification $\hat{\sigma}_\varepsilon^2$ and $\mathrm{SSE}_{\mathrm{sim}}$ are useful for selecting acceptable models from a set of identified models obtained by using the procedures with different tuning parameters and estimates of the system orders.

3 Intersecting hyperplanes

If the hyperplanes over the regressor space defined by PVs θ_i and θ_j intersect over $\mathcal{X}_j$, datapoints may be wrongly attributed to the data subset $\mathcal{F}_i$. To shed

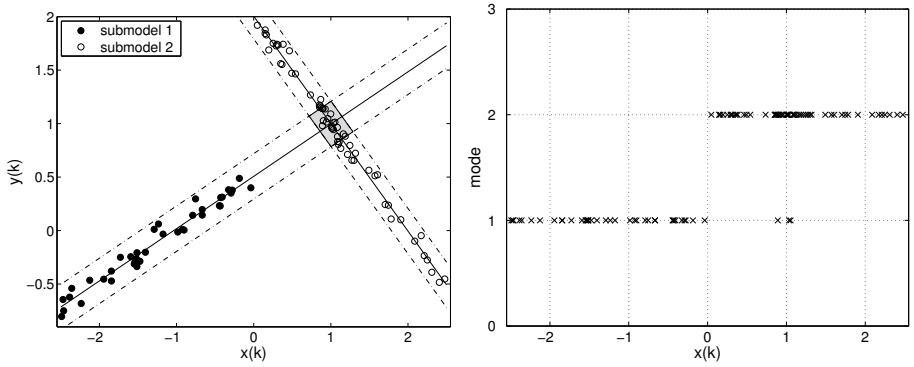

Fig. 1. left: Classification with clustering-based and the bounded-error procedures. Both procedures yield $\Delta_\theta = 0.0186$ and $\Delta_\mathcal{X} = 0.0055$ **right**: Data classification obtained by using the algebraic procedure (yielding $\Delta_\theta = 0.0276$) and attributing each data point to the submodel which generates the smallest prediction error.

some light on this issue, consider the PWARX model $y(k) = f(x(k)) + e(k)$ where f is defined as:

$$f(x) = \begin{cases} \begin{bmatrix} x & 1 \end{bmatrix} \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix} \text{ if } x \in [-2.5,\, 0] \\ \\ \begin{bmatrix} x & 1 \end{bmatrix} \begin{bmatrix} -1 \\ 2 \end{bmatrix} \text{ if } x \in (0,\, 2.5] \end{cases} . \tag{16}$$

The data set used for identification is depicted in figure 1, together with the data classification obtained from the clustering-based and bounded error procedures. It is seen that the clustering-based and the bounded-error procedures do not experience problems with the intersecting PVs in this particular example. The data classification using the algebraic procedure and the minimum prediction error rule (11) is given in figure 1, right. It is seen that the minimum error prediction rule can lead to misclassifications, and hence it is not the most appropriate rule for the case of PWARX models.

The data classification and the price function for misclassification using the Bayesian procedure is depicted in figure 2, left. The price for misclassification of wrongly attributed points is small in comparison to the weight for misclassification of the correctly attributed points. The identified model with the Bayesian procedure, together with the true model is depicted in the figure 2, right.

We stress that the classification methods employed by the clustering-based, bounded-error and the Bayesian methods are based on heuristics. Theoretical analysis of this issue is needed.

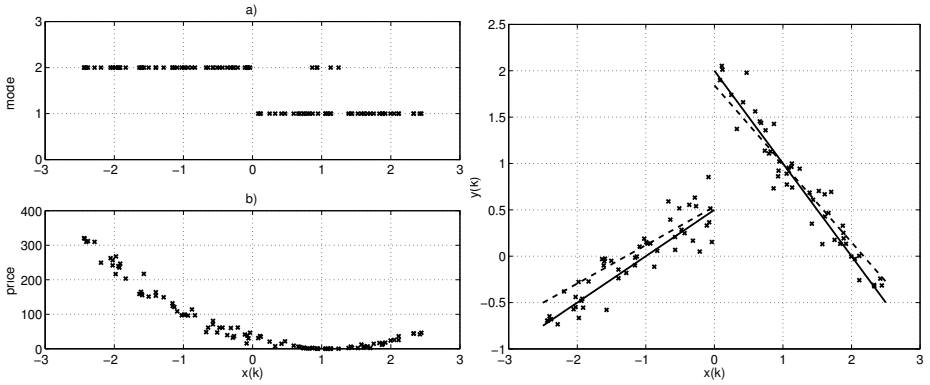

Fig. 2. Identification results for Bayesian procedure, initialized with a priori parameter pdfs $p_{\theta_1}(\cdot;0) = p_{\theta_2}(\cdot;0) \sim \mathcal{U}[-2.5, 2.5] \times [-2.5, 2.5]$, yielding $\Delta_\theta = 0.1366$ and $\Delta_\mathcal{X} = 0.0228$ **left:** a) Data points attributed to modes b) Price function for the wrong classification **right:** Data set used for identification, the true model (solid) and the identified model (dashed)

4 Overestimation of model orders

The clustering based, bounded-error and the Bayesian approach assume that the system orders n_a and n_b are known exactly, but in practice this is seldom the case. The algebraic procedure is able to estimate the model orders directly from the data set.

In order to investigate the effects of overestimating model orders we will consider a 1-dimensional autoregressive autonomous system of the form

$$y(k+1) = \begin{cases} 2y(k) + 10 + e(k), & \text{if} \quad y(k) \in [-10, 0) \\ -1.5y(k) + 10 + e(k), & \text{if} \quad y(k) \in [0, 10]. \end{cases} \tag{17}$$

The additive noise term $e(k)$ is normally distributed, with zero mean and variance $\sigma_e^2 = 0.01$. The sequence $y(k)$ was generated with $y(0) = -10$, and the input was generated as $u(k) \sim \mathcal{U}[-10, 10]$.

The true model orders are $n_a = 1, n_b = 0$. Identification procedures were applied for all combinations of $n_a = 1, \ldots, 4$ and $n_b = 1, \ldots, 5$. Note that for overestimated model orders, the correct model is obtained by setting to zeroes the entries in θ_i, M_{ij}, m_{ij} on positions corresponding to superfluous elements in the regressor.

Figure 3 shows the values of the criterion $\hat{\sigma}_\varepsilon^2$ on the logarithmic scale, for models with different model orders identified by the clustering-based procedure. From figure 3 it is seen that the clustering procedure identifies the model with $\hat{\sigma}_\varepsilon^2$ value close to the noise in the system for true system orders, but that the performance rapidly deteriorates when the model order is overestimated. The problem with the overestimated order lies in the assumption that datapoints

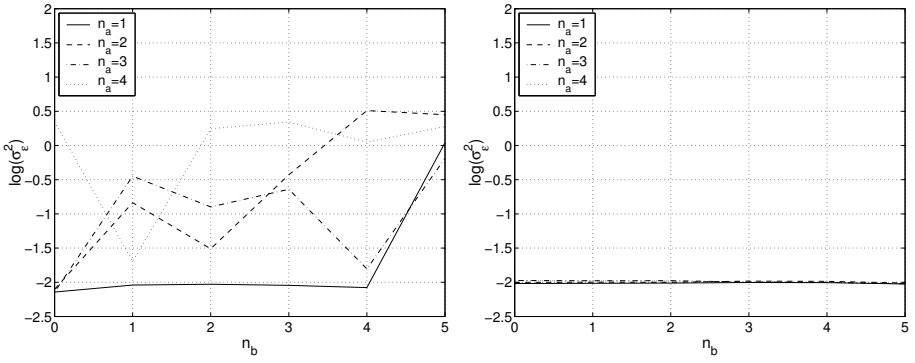

Fig. 3. left: $\hat{\sigma}_\varepsilon^2$ for the clustering procedure with $s = 2$ and $c = 20$ **right:** $\hat{\sigma}_\varepsilon^2$ for the bounded error procedure

close to each other in the regressor space belong to the same subsystem. When overestimating the order of the model regressor is extended with elements which do not contain relevant information for the estimation of the subsystems, but change the distance between the regressors. If the true distance is denoted by d_0, the distance between the extended regressors is $d_e^2 = d_0^2 + d_*^2$, where d_*^2 is due to the added elements, and contains no useful information. Depending on the true and overestimated model orders d_* can easily be of the same or higher order of magnitude as d_0.

The results for the bounded-error procedure are shown in Figure 3, 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 (7) 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. The enhanced version [3] of the greedy algorithm [14] is applied here for solving the MIN PFS problem.

The results for the Bayesian procedure for two different initializations are depicted in the figure 4. In figure 4, left the a priori parameter pdfs for the case $n_a = 1, n_b = 0$ are chosen as $p_{\theta_1}(\cdot; 0) = p_{\theta_2}(\cdot; 0) = \mathcal{U}([-5, 5] \times [-20, 20])$. For increased orders, added elements in the parameter vector are taken to be uniformly distributed in the interval $[-5, 5]$ (while the true value is 0). In figure (4), right for the case $n_a = 1, n_b = 0$ the a priori parameter pdfs are chosen as $p_{\theta_1}(\cdot; 0) = \mathcal{U}([0, 4] \times [8, 12])$, $p_{\theta_2}(\cdot; 0) = \mathcal{U}([-4, 0] \times [8, 12])$, and all added elements are taken to be uniformly distributed in the interval $[-0.5, 0.5]$. This example shows the importance of proper choice of initial parameter pdfs for the Bayesian procedure. With precise initial pdfs the algorithm manages to estimate relatively accurate over-parameterized models. In the case when the a priori information is not adequate the performance of the algorithm deteriorates rapidly.

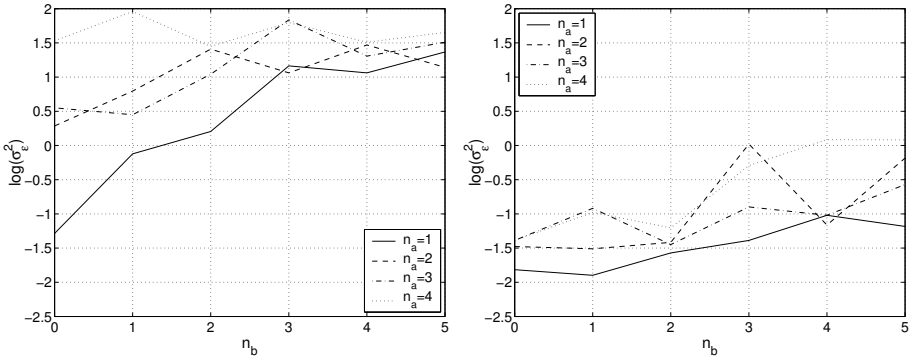

Fig. 4. $\hat{\sigma}_\varepsilon^2$ for the Bayesian procedure **left:** with unprecise initial parameter pdfs **right:** with precise initial parameter pdfs

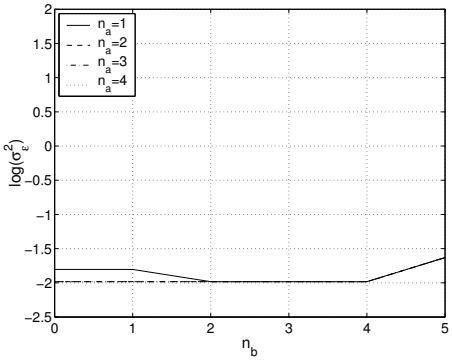

Fig. 5. $\hat{\sigma}_\varepsilon^2$ for the algebraic procedure

The algebraic procedure is applied to the data set with $s = 2$, but unknown model orders. The results are depicted in the figure 5. From 5 we see that the procedure has no difficulties in estimating the over-parameterized model.

5 Effects of noise

In this section we study effects of noise e on the identification procedures. The first issue of interest is the effect that different realizations of noise with the same statistical properties have on the identification results. The second issue is how statistical properties of noise influence identification results.

To shed some light on these issues we designed an experiment with the PWARX model of section 4 (see (17)). For this model we generated a noise-less data set of 100 datapoints. The procedures are applied 100 times on this

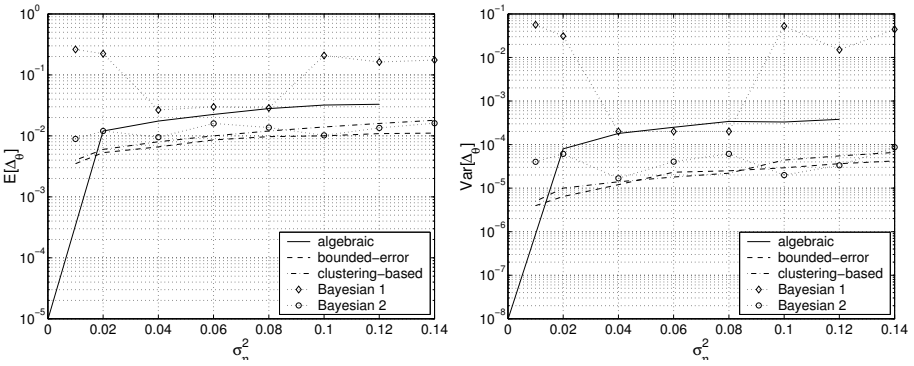

Fig. 6. Means (left) and variances (right) of the Δ_θ distributions for several variances of noise σ_η^2

data set, after adding a different realization of normally distributed noise with zero mean and variance σ_e^2 to the outputs $y(k)$. For each identified model the index Δ_θ is computed. In this way an approximate distribution of Δ_θ for each σ_e^2 can be constructed. For each such distribution we computed its mean and variance. For more details see [15]

Figure 6 depicts means and variances of Δ_θ distributions as functions of σ_e^2 for all four procedures. Again, we have two different initializations for the Bayesian procedure, denoted in figure as "Bayesian 1" and "Bayesian 2". For "Bayesian 1" we used $p_{\theta_1}(\cdot;0) = p_{\theta_2}(\cdot;0) = \mathcal{U}([-5,5]\times[-20,20])$, and for "Bayesian 2" we used $p_{\theta_1}(\cdot;0) = \mathcal{U}([0,4]\times[8,12])$, $p_{\theta_1}(\cdot;0) = \mathcal{U}([-4,0]\times[8,12])$.

From figure 6 we can conclude that the clustering-based procedure and the bounded-error procedure achieve similar performance with respect to noise. The algebraic procedure is more sensitive to noise, as compared to the clustering-based and bounded-error procedures. With precise initialization ("Bayesian 1") the Bayesian procedure achieves performance comparable to clustering-based and bounded-error, while with imprecise initialization ("Bayesian 2") the quality measures are the worst of all procedures.

6 Experimental example

In this section we show the results of the identification of the component placement process in pick-and-place machines. The pick-and-place machine is used for automatically placing electronic components on a Printed Circuit Board (PCB). To study the placement process, an experimental setup was made. The photo and the schematic of the setup are shown in figure 7. A detailed description of the process and the experimental setup can be found in [16].

A data set consisting of 750 samples is collected. The data set is divided into two overlapping sets of 500 points, the first set is used for identification,

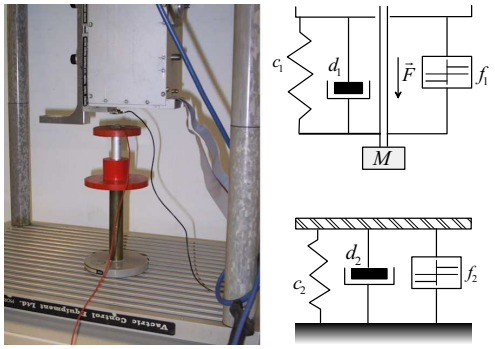

Fig. 7. Photo and the schematic representation of the experimental setup

and the second for validation. All four procedures were applied for several order estimates and with different tuning parameters. The procedures were executed for all the combinations of these orders and tuning parameters. The proposed quality measures $\hat{\sigma}_\varepsilon^2$ and $\mathrm{SSE}_{\mathrm{sim}}$ were used to choose acceptable identified models for which the simulations were plotted. The best identified model was then chosen by visual inspection.

For the clustering-based procedure figure 8, left shows the *simulation* based on the validation data set for the best model obtained. In the upper panel of the figure measured output y_{id} and the simulated output y_{sim} are depicted. The lower panel shows which of the identified submodels is active at each time instant. It turns out that the best models are obtained for high values of c. The same was observed in [16]. A possible explanation is the following: because of the presence of dry friction neither the free nor the impact mode are linear, but with large LD's the effects of dry friction can be 'averaged out' as a process noise. Note that the difference between the measured and simulated responses, which is due to unmodeled dry friction, is clearly visible, e.g. on the time interval $[225, 300]$.

As the number of modes s for the bounded-error procedure is not fixed, in order to identify two modes, the right combination of the parameters α, γ and δ has to be found. For the initial error bound δ we used $3\hat{\sigma}_\varepsilon \approx 1$, obtained from the clustering-based procedure, assuming that this value would be a good estimate for the variance of the measurement noise. Executing the bounded-error procedure with δ's in the vicinity of this $3\hat{\sigma}_\varepsilon^2$ resulted in identified models with only one parameter vector, and a large number of infeasible points. Therefore, we had to lower the error bound to $\delta = 0.30$. For this value of δ the procedure identified a model that distinguishes two submodels. Model identified with this δ had a smaller values of both $\hat{\sigma}_\varepsilon^2$ for the identification data set and SSE_{sim} for the validation data set than the model identified with the clustering-based procedure. The simulation of the validation data set for the best identified model is shown in the figure 8, right.

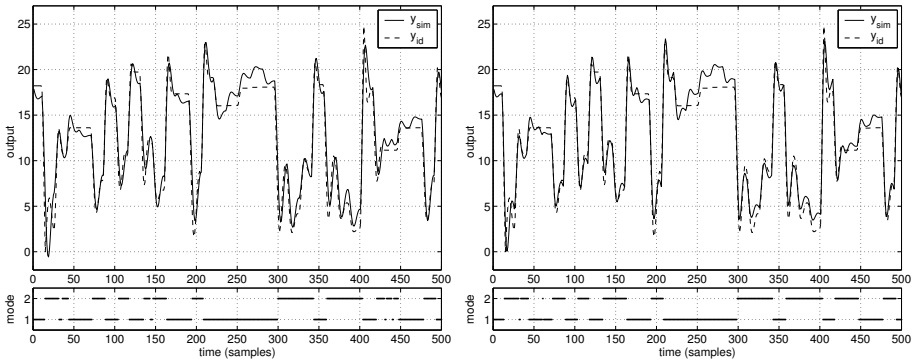

Fig. 8. left: Simulation of the PWARX model generated by the clustering procedure with $n_a = 2$, $n_b = 2$, $s = 2$ and $c = 90$ for the validation data set with $\text{SSE}_{sim} = 1.98$ **right:** Simulation of the PWARX model generated by the bounded-error procedure with $n_a = 2$, $n_b = 2$, $\delta = 0.3$, $\alpha = 0.10$, $\beta = 0.01$ and $c = 40$ for the validation data set with $\text{SSE}_{sim} = 1.72$ **upper fig.:** solid line: predicted response, dashed line: measured response **lower fig.:** active mode.

Physical insight into the operation of the setup facilitates the initialization of the Bayesian procedure. For instance, although the mode switch does not occur at a fixed height of the head, with a degree of certainty data points below certain height may be attributed to the free mode, and, analogously data points above certain height may be attributed to the impact mode. This a priori information may be exploited to obtain the rough estimate of each of the parameters through least squares, θ_i^{LS}. Also, the variance $\tilde{V}_i$ of such estimate may be obtained. This information is sufficient to describe the parameter θ_i as a normally distributed random variable, with a mean θ_i^{LS} and variance $\tilde{V}_i$.

Portions of the identification data set that are used to initialize the procedure are depicted in the figure 9, left, together with the input signal. Results of simulation of the identified model are given in figure 9, right. The model yields a lower value of SSE_{sim} than the two models obtained with the clustering-based and bounded-error procedures.

The algebraic procedure identified the parameters of the model, with $\hat{\sigma}_\varepsilon^2 = 0.0803$. However, the data classification is not satisfactory, as the procedure predicts rapidly oscillating mode values, while in the physical system such oscillations are impossible. It remains for the future work to check if estimated parameters can be used to obtain the satisfactory PWARX model.

7 Conclusions

We conclude the paper by summarizing features and drawbacks of each identification procedure, based on the insights obtained from the considered examples.

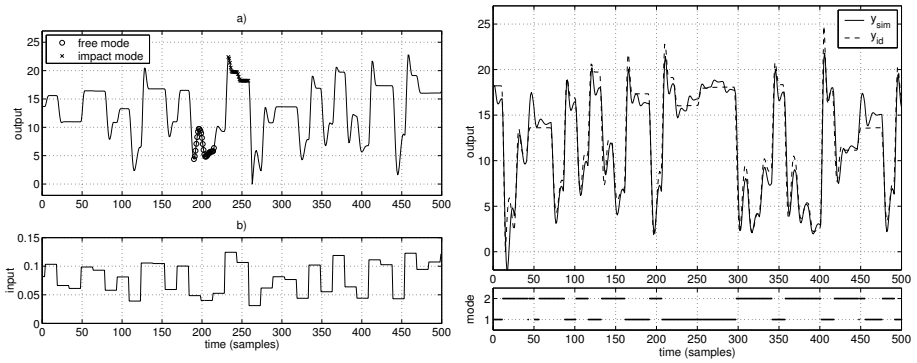

Fig. 9. Bayesian procedure. **left:** Data set used for identification a) position (portion marked with ∘: data points used for the initialization of the free mode; portion marked with ×: data points used for initialization of impact mode b) input signal **right: upper fig.:** Simulation of the identified model (solid line: simulated response, dashed line: measured response), $SSE_{sim} = 1.56$ **lower fig.:** modes active during the simulation

The algebraic procedure is well suited for the cases when the system that generated the data can be accurately described with a switched linear system, and no or little noise is present. It can also handle the cases with unknown model orders. Noise and/or nonlinear disturbances in 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 might lead to inaccurate classification. In such cases, it is better to use one of the classification methods employed by other procedures.

The Bayesian procedure is well suited for the cases where the sufficient physical insight into the underlying data generating process is available. By appropriate choice of the initial parameter pdfs the user might steer the procedure towards identifying the model where the modes of the identified model represent different modes of the physical system. On the other hand, poor initialization may lead to poor identification results.

The bounded error procedure is well suited for the cases when there is no a priori knowledge on the physical system and one needs to identify a model with a prescribed bounded prediction error (e.g. approximation of nonlinear systems). Tuning parameters allow for the tradeoff between the model complexity and accuracy. However, finding the right combination of tuning parameters to get the model with the prescribed structure (number of modes) may be difficult.

The clustering-based procedure is well suited for the cases when there is no a priori knowledge on the physical system, and one needs to identify a model with a prescribed structure. When using the clustering-based procedure one must be

aware of the possible erratic behavior (as described in section 4) in the cases when the model orders are not known exactly.

References

1. Ferrari-Trecate, G., Muselli, M., Liberati, D., Morari, M.: A clustering technique for the identification of piecewise affine and hybrid systems. Automatica **39** (2003) 205–217
2. Bemporad, A., Garulli, A., Paoletti, S., Vicino, A.: 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. Springer Verlag (2003) 97–112
3. Bemporad, A., Garulli, A., Paoletti, S., Vicino, A.: Data classification and parameter estimation for the identification of piecewise affine models. In: Proceedings of the 43rd IEEE Conference on Decision and Control, Paradise Island, Bahamas (2004) 20–25
4. Juloski, A., Weiland, S., Heemels, W.: A Bayesian approach to identification of hybrid systems. In: Proceedings of the 43rd Conference on Decision and Control, Paradise Island, Bahamas (2004) 13–19
5. Vidal, R., Soatto, S., Ma, Y., Sastry, S.: An algebraic geometric approach to the identification of a class of linear hybrid systems. In: Proc. of IEEE Conference on Decision and Control. (2003)
6. Vidal, R.: Identification of PWARX hybrid models with unknown and possibly different orders. In: Proc. of IEEE American Control Conference. (2004)
7. Roll, J., Bemporad, A., Ljung, L.: Identification of piecewise affine systems via mixed-integer programming. Automatica **40** (2004) 37–50
8. Munz, E., Krebs, V.: Identification of hybrid systems using a priori knowledge. In: Preprints of the 15th IFAC world congress, Barcelona, Spain (2002)
9. Ferrari-Trecate, G., Schinkel, M.: Conditions of optimal classification for piecewise affine regression. In Maler, O., Pnueli, A., eds.: Proc. 6th International Workshop on Hybrid Systems: Computation and Control. Volume 2623 of Lecture Notes in Computer Science. Springer-Verlag (2003) 188–202
10. Bennett, K., Mangasarian, O.: Multicategory discrimination via linear programming. Optimization Methods and Software **3** (1993) 27–39
11. Milanese, M., Vicino, A.: Optimal estimation theory for dynamic systems with set membership uncertainty: an overview. Automatica **27** (1991) 997–1009
12. Arulampalam, M., Maskell, S., Gordon, N., Clapp, T.: A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. IEEE Transactions on Signal Processing **50** (2002) 174–188
13. Verriest, E., Moor, B.D.: Multi-mode system identification. In: Proc. of European Conference on Control. (1999)
14. Amaldi, E., Mattavelli, M.: The MIN PFS problem and piecewise linear model estimation. Discrete Applied Mathematics **118** (2002) 115–143
15. Niessen, H., Juloski, A., Ferrari-Trecate, G., Heemels, W.: Comparison of three procedures for the identification of hybrid systems. In: Proceedings of the Conference on Control Applications, Taipei, Taiwan (2004)
16. Juloski, A., Heemels, W., Ferrari-Trecate, G.: Data-based hybrid modelling of the component placement process in pick-and-place machines. Control Engineering Practice **12** (2004) 1241–1252