A Bayesian Approach to Identification of Hybrid Systems

A. Lj. Juloski, S. Weiland, and W. P. M. H. Heemels

***Abstract*—In this paper, we present a novel procedure for the identification of hybrid systems in the class of piecewise ARX systems. The presented method facilitates the use of available *a priori* knowledge on the system to be identified, but can also be used as a black-box method. We treat the unknown parameters as random variables, described by their probability density functions. The identification problem is posed as the problem of computing the *a posteriori* probability density function of the model parameters, and subsequently relaxed until a practically implementable method is obtained. A particle filtering method is used for a numerical implementation of the proposed procedure. A modified version of the multicategory robust linear programming classification procedure, which uses the information derived in the previous steps of the identification algorithm, is used for estimating the partition of the piecewise ARX map. The proposed procedure is applied for the identification of a component placement process in pick-and-place machines.**

***Index Terms*—Hybrid systems, identification.**

I. Introduction

In this paper, we present a novel procedure for the identification of hybrid systems in the class of Piecewise AutoRegressive systems with eXogenous inputs (PWARX systems). PWARX models are a generalization of the classical ARX models, in the sense that the regressor space is partitioned into a finite number of polyhedral regions, where in each region the input-output relation is defined through an ARX-type model. PWARX models represent a broad class of hybrid systems, and they form a subclass of piecewise affine (PWA) models [1], which are under mild conditions equivalent to other hybrid modeling formalisms, such as mixed logic dynamics (MLD) systems [2] and linear complementarity (LC) models [3]–[5]. In recent years, a number of methods for stability analysis, optimal control design and verification have been developed for the above mentioned classes. In this paper, we will focus on the identification of hybrid systems in this class.

Based on the observed data the identification problem amounts to determining the parameters of the ARX sub-models together with the regions of the regressor space where each of the models is valid. The main problem in the identification of PWARX models is the problem of *data classification*—that is, the problem to assign each data point to one specific sub-model. When the data has been classified, the parameters of the submodels can be determined, and the regions where each of the submodels is valid can be estimated using techniques for pattern classification [6].

The problem of the identification of PWA and PWARX models has been considered before, and to date several approaches exist for the identification of such models (see [7] and the references therein). As pointed out in [7], most of the existing approaches assume that the system dynamics are continuous over the switching surfaces, while the approaches that allow for discontinuities started appearing only recently [6], [8]–[10]. The identification procedure proposed in this paper allows for discontinuous system dynamics as well.

In the *clustering-based procedure* [6], the data classification and the parameter estimation steps are performed simultaneously by solving an optimal clustering problem in the parameter space. In the *greedy procedure* [8] the data classification and the parameter estimation steps are accomplished by partitioning an infeasible set of linear inequalities into a minimal number of feasible subsets. In the *algebraic procedure* [9], [10] the parameter estimation is accomplished by finding the roots of suitably defined polynomials, while the data points are classified to the submodel that gives the smallest prediction error. For a comparison of the mentioned procedures and the procedure presented in this paper, see [11], [12].

In this paper, we take a Bayesian approach to the problem of identifying PWARX models. Specifically, we treat the unknown parameters as random vectors, and describe them in terms of their joint probability density function (pdf). The probability density function contains the complete stochastic information about the parameters, and different parameter estimates, such as expected values or maximum a posteriori probability estimates, can be easily inferred.

We assume that an *a priori* joint parameter pdf is given. After the data has been classified the *a posteriori* joint parameter pdf will be computed, using Bayes' rule. Furthermore, we compute the probability that the observed data is generated by the given classification. One data classification is then considered *better* than another if it has a higher probability. Following this line of reasoning, the identification problem amounts to finding *the best* data classification. This approach to model comparison is similar to the Bayesian framework which was used in [13] for neural networks.

The classification problem is a combinatorial optimization problem, where all possible mode sequences have to be explored, in order to find an optimal solution. To reduce complexity we resort to a sequential approach, where each data point

Manuscript received June 8, 2004; revised June 3, 2005. Recommended by Guest Editor A. Vicino. This work was supported by the STW/PROGRESS Grant EES.5173, and in part by European project Grant SICONOS (IST2001-37172) and HYCON Network of Excellence (FP6-IST-511368).

A. Juloski and S. Weiland are with the Department of Electrical Engineering, Eindhoven University of Technology, 5600 MB Eindhoven, The Netherlands (e-mail: a.juloski@tue.nl; s.weiland@tue.nl).

M. Heemels is with the Embedded Systems Institute, 5600 MB Eindhoven, The Netherlands (e-mail: maurice.heemels@embeddedsystems.nl).

Digital Object Identifier 10.1109/TAC.2005.856649

is classified to the best mode based on the information available so far. In the optimization literature, this strategy is known as the *greedy strategy*, as it makes the best possible local decision in order to approach the global optimum. In order to calculate the resulting pdfs, we propose a method of particle approximations [14], [15].

Region estimations are based on a modification of the multicategory robust linear programming procedure (MRLP) [16]. The modification consists in introducing a suitably defined pricing function, that assigns weights to the misclassification of data points. An advantage of using pricing functions is that more information is preserved from the classification phase. This is illustrated by the example in Section VIII.

By choosing a prior parameter probability density function the user can supply the available a priori knowledge to the identification procedure. This is a major advantage of the framework presented here. Including a priori knowledge is much harder in other identification methods, such as the ones described in [6], [8], and [9].

The need for using *a priori* knowledge was observed in an experimental case study of a component placement process in a pick-and-place machine [17], [18]. The *a priori* knowledge may stem from physical insight in the system or from previous identification experiments. Specifically, if parameters have a physical interpretation then the pdfs can be chosen so as to match the physically meaningful values, and the procedure can be forced to identify a model that can be interpreted in physical terms. The approach presented here can also be used to improve the previously identified models with targeted identification experiments. Also, models of increasing complexity can be built from a series of identification experiments, where in each experiment only a subset of the modes of the physical system is excited and identified. We believe that these are important advantages in any practical identification problem. We also discuss some ways to initialize the procedure without using *a priori* knowledge.

The parameter estimation of switching autoregressive probabilistic models in a probabilistic setting was also examined in time series analysis, in the econometrics literature (see, e.g., [19]–[21]). There, the switching law was modeled as a Markov chain, defined by its transition probability matrix. Both the parameter values and the transition probabilities are subsequently estimated by direct numerical optimization of a suitably defined likelihood function. As this optimization problem is nonconvex and possibly possesses local minima, good initial estimates are needed in this procedure. As an alternative, the optimization procedure can be repeated several times, starting from different random initial conditions. Another approach to find parameters values that (locally) maximize the likelihood function when unobserved states are present (e.g., unmeasured modes of the system) is the expectation-maximization type algorithms [22]. These approaches are, in principle, also applicable to parameter estimation of PWARX models. The key differences to our approach are that, first, we consider models with deterministic switching, and second, the optimization criterion is different, which has consequences in theoretical and numerical aspects. Specifically, our approach allows for including a priori information, decoupling of mode and parameter estimation, and for sequential data processing.

The remainder of the paper is organized as follows. Preliminaries are given in Section II. The class of PWARX models is introduced in Section III. The identification problem is formally stated in Section IV. In Sections V and VI, we derive the suboptimal identification algorithm, and the particle filtering approach, as a way to implement it. In Section VII, we present the modified MRLP procedure. In Section VIII, we give an example that illustrates the presented ideas. In Section IX, we discuss several ways to obtain the *a priori* probability density functions of the model parameters, so as to initialize the procedure. The connection with the clustering procedure, and the improvement that our method can provide are explained in Section IX-B. An experimental example is given in Section X. Conclusions are given in Section XI.

II. Preliminaries

Let a vector of random variables $\theta_i \in \Theta_i \subseteq \mathbb{R}^n$ be described by a probability density function (pdf) p_{θ_i}. If the pdf p_{θ_i} takes the form

$$p_{\theta_i}(\theta) = \delta\left(\theta - \theta_i^0\right) \tag{1}$$

where δ is the Dirac delta distribution, then $\theta_i = \theta_i^0$, with probability one, which will mean that the value of θ_i is known.

Different estimates of θ_i can be easily obtained from the probability density function. For instance, the *expectation* of θ_i is given as

$$\hat{\theta}_i^E = E[\theta_i] = \int_{\Theta_i} \theta\, p_{\theta_i}(\theta)\, d\theta. \tag{2}$$

The covariance matrix V_i, which is a measure of the quality of the estimate $\hat{\theta}_i^E$, is defined as

$$V_i^E = \int_{\Theta_i} \left(\theta - \hat{\theta}_i^E\right)\left(\theta - \hat{\theta}_i^E\right)^\top p_{\theta_i}(\theta) d\theta. \tag{3}$$

We define the *dispersion* of the estimate $\hat{\theta}_i^E$ as the spectral radius of the covariance matrix

$$\rho_i^E = \rho\left(V_i^E\right) = \lambda_{\max}\left(V_i^E\right) \tag{4}$$

where $\lambda_{\max}(\cdot)$ denotes the maximal eigenvalue. Note that $\rho_i^E = 0$ if and only if (1) holds. Dispersion is useful for the comparison of different estimates of θ_i.

III. Model Class

We consider piecewise autoregressive exogenous (PWARX) models of the form

$$y(k) = f(x(k)) + e(k) \tag{5}$$

where $k \geq 0, x(k)$ is a vector of regressors defined as

$$x(k) = [y(k-1) \ldots y(k-n_a) \quad u(k-1) \ldots u(k-n_b)]^\top \tag{6}$$

and

$$f(x) = \begin{cases} \theta_1^\top \begin{bmatrix} x \\ 1 \end{bmatrix} & \text{if } x \in \mathcal{X}_1 \\ \vdots & \vdots \\ \theta_s^\top \begin{bmatrix} x \\ 1 \end{bmatrix} & \text{if } x \in \mathcal{X}_s \end{cases} \tag{8}$$

is a piecewise affine map where y denotes the scalar real valued measured output and u the scalar real valued input signal.

The parameters n_a and n_b in (6) and the number of modes s are assumed to be known. Therefore, $\theta_i \in \Theta_i \subseteq \mathbb{R}^{n+1}$, where $n = n_a + n_b$. The sets $\mathcal{X}_i$ are assumed to be bounded convex polyhedra, described by

$$\mathcal{X}_i = \{x \in \mathbb{R}^n \mid H_i x \leq h_i\} \tag{9}$$

where H_i is a real valued matrix of compatible dimensions, h_i is a real valued vector, and the inequality holds element-wise. The set $\mathcal{X} = \bigcup_{i=1}^s \mathcal{X}_i$ is assumed to be a bounded polyhedron, and we assume that $\{\mathcal{X}_i\}_{i=1}^s$ is a partition of $\mathcal{X}$ (in the sense that the interiors of $\mathcal{X}_i$ and $\mathcal{X}_j$ do not intersect for $i \neq j$)[1]

Assumption III.1: The realization of the additive noise e in the model (5) is a sequence of independent, identically distributed random values, with an *a priori* known probability density function p_e.

Let a data sequence $(x(k), y(k)), k = 1, \ldots, T$ for $T > 0$ be given. We define the *mode function* $\mu : \{1, \ldots, T\} \to \{1, \ldots, s\}$ that assigns mode $\mu(k)$ to the data pair $(x(k), y(k)), k = 1, \ldots, T$ as

$$\mu(k) := i \text{ whenever } x(k) \in \mathcal{X}_i. \tag{10}$$

For a given data set $\{(x(k), y(k))\}_{k=1}^T$ the partitioning $\{\mathcal{X}_i\}_{i=1}^s$ of $\mathcal{X}$ *induces* the mode function μ, as given by (10) (see also footnote 1).

Conversely, given the mode function μ, the problem to find regions $\{\mathcal{X}_i\}_{i=1}^s$ such that whenever $\mu(k) = i$, we have that $x(k) \in \mathcal{X}_i$ is the *region estimation* problem. The region estimation problem can be solved using standard techniques for data classification [6]. The problem of region estimation therefore can, in principle, be replaced by the problem of estimating the mode $\mu(k)$ for each data pair $(x(k), y(k))$. We will refer to the latter problem as the *classification problem*. These problems will be formalized in the next section.

[1]Since regions $\mathcal{X}_i$ are closed sets by definition (9) it may happen that $\mathcal{X}_i$ and $\mathcal{X}_j$ share a common facet. Technically, the point x, lying on the shared facet would belong to both $\mathcal{X}_i$ and $\mathcal{X}_j$. We neglect this issue, as it has no consequence on the presented procedure.

IV. Problem Statement

The identification problem consists of estimating the unknown parameter vectors θ_i, for $i = 1 \ldots s$, and the regions $\{\mathcal{X}_i\}_{i=1}^s$, described by (9), given the data pairs $(x(k), y(k))$, for $k = 1, \ldots, T$. With $\vartheta \in \Theta \subseteq \mathbb{R}^{(n+1)s}$ we will denote a vector $\vartheta = \operatorname{col}(\theta_1, \ldots, \theta_s)$, where the operator $\operatorname{col}(\cdot)$ stacks its operands into a column vector. With $\mathcal{M}$, we will denote the space of all possible mode sequences. The identification problem can then be posed as follows.

Problem IV.1 (Full Identification Problem): Given the joint *a priori* probability density function of the parameters and of the partition $p_{\vartheta, \{\mathcal{X}_i\}_{i=1}^s}$ and the data set $\{(x(k), y(k))\}_{k=1}^T$, determine the conditional joint pdf of the parameters and the partition

$$p_{\vartheta, \{\mathcal{X}_i\}_{i=1}^s}\left(\vartheta, \{\mathcal{X}_i\}_{i=1}^s \mid \{(x(k), y(k))\}_{k=1}^T\right). \tag{11}$$

The pdf (11) contains the complete statistical information about the parameters and the partition that can be inferred from the observed data and prior information. In conjunction with the system definition (5), the pdf (11) can be used to predict the output value $y(k)$ given the regressor $x(k)$ in the sense that we can obtain the probability density function $p_{y(k)}$. In that sense, (5) and (11) form a complete model of the system. Point estimates of the parameters and the partition can be obtained as the maximum likelihood estimates, for instance

$$\begin{aligned} &\{\vartheta^*, \{\mathcal{X}_i^*\}_{i=1}^s\} \\ &= \arg\max p_{\vartheta, \{\mathcal{X}_i\}_{i=1}^s}\left(\vartheta, \{\mathcal{X}_i\}_{i=1}^s \mid \{(x(k), y(k))\}_{k=1}^T\right) \end{aligned} \tag{12}$$

where the maximum is taken over all possible parameters and partitions satisfying the assumptions on the model (5). Another possibility is to compute the expectations of parameters and the partition.

Problem IV.1 can, at least in principle, be solved using Bayes' rule as in (7) provided that a suitable probability space has been defined for the random variable $(\vartheta, \{\mathcal{X}_i\}_{i=1}^s)$.

However, as (12) involves joint optimization over parameters and partitions, which is a hard nonconvex optimization problem with many local maxima, computing (7) and (12) is numerically infeasible. Therefore, we will instead consider a problem in which the computation of the joint probability density function of the parameters and the partition will be replaced by the joint probability density function of the parameters and the mode function.

Problem IV.2 (Mode and Parameter Identification Problem): Given the joint *a priori* probability density function of the parameters and the mode $p_{\vartheta,\mu}$ and the data set

$$\begin{aligned} &p_{\vartheta, \{\mathcal{X}_i\}_{i=1}^s}\left(\vartheta, \{\mathcal{X}_i\}_{i=1}^s \middle| \{(x(k), y(k))\}_{k=1}^T\right) \\ &= \frac{p\left(\{(x(k), y(k))\}_{k=1}^T \mid \vartheta, \{\mathcal{X}_i\}_{i=1}^s\right) p_{\vartheta, \{\mathcal{X}_i\}_{i=1}^s}(\vartheta, \{\mathcal{X}_i\}_{i=1}^s)}{\int\limits_{\Theta, \{\mathcal{X}_i\}_{i=1}^s} p\left(\{(x(k), y(k))\}_{k=1}^T \mid \vartheta, \{\mathcal{X}_i\}_{i=1}^s\right) p_{\vartheta, \{\mathcal{X}_i\}_{i=1}^s}(\vartheta, \{\mathcal{X}_i\}_{i=1}^s)\, d(\vartheta, \{\mathcal{X}_i\}_{i=1}^s)} \end{aligned} \tag{7}$$

$\{(x(k), y(k))\}_{k=1}^{T}$, determine the conditional joint pdf of the parameters and the mode

$$p_{\vartheta,\mu}\left(\vartheta, \mu \,|\, \{(x(k), y(k))\}_{k=1}^{T}\right). \tag{13}$$

Point estimates of the parameters and the mode can be obtained from (13) as:

$$\{\vartheta^*, \mu^*\} = \arg\max p_{\vartheta,\mu}\left(\vartheta, \mu \,|\, \{(x(k), y(k))\}_{k=1}^{T}\right) \tag{14}$$

where the maximum is taken over all possible parameter values Θ and mode sequences $\mathcal{M}$. Once the optimal parameters and the mode function $\{\vartheta^*, \mu^*\}$ are computed, the partition $\{\mathcal{X}_i\}_{i=1}^{s}$ can be subsequently reconstructed as discussed in Section III. The pdf (13) can, in principle, be computed using Bayes' rule in a similar way to (7)

$$p_{\vartheta,\mu}(\vartheta, \mu \,|\, \{(x(k), y(k))\}_{k=1}^{T}) = \frac{p\left(\{(x(k), y(k))\}_{k=1}^{T} \,|\, \vartheta, \mu\right) p_{\vartheta,\mu}(\vartheta, \mu)}{\sum_{\mathcal{M}} \int_{\Theta} p\left(\{(x(k), y(k))\}_{k=1}^{T} \,|\, \vartheta, \mu\right) p_{\vartheta,\mu}(\vartheta, \mu) d\vartheta}. \tag{15}$$

Note that defining the a probability space for the random variable (ϑ, μ) is straightforward.

Computing (15) and (14) is a combinatorial problem, where all possible mode sequences have to be examined, in order to find the optimal mode sequence and the corresponding parameter values. As such, it is computationally infeasible for larger data sets. Therefore, we need to further relax the considered problem. We make the following assumption.

Assumption IV.3: Assume that ϑ and μ are a priori independent random variables. In particular, this means that $p_{\vartheta}(\vartheta \,|\, \mu) = p_{\vartheta}(\vartheta)$. Furthermore, assume that a priori all mode sequences are equally probable, i.e., p_{μ} is constant for all $\mu \in \mathcal{M}$.

We will further relax Problem IV.2 by separating the mode and parameter estimation.

Problem IV.4 (Classification Problem): Given the data and the *a priori* pdf for the parameters $p_{\vartheta}(\vartheta)$ determine

$$p_{\mu}\left(\mu \,|\, \{(x(k), y(k))\}_{k=1}^{T}\right) \tag{16}$$

and the point estimate of μ

$$\mu^* = \arg\max p_{\mu}\left(\mu \,|\, \{(x(k), y(k))\}_{k=1}^{T}\right) \tag{17}$$

where the maximum is taken over all possible sequences $\mu \in \mathcal{M}$.

Under the Assumption IV.3, the mode probability density function $p_{\mu}(\mu \,|\, \{(x(k), y(k))\}_{k=1}^{T})$ can be computed using Bayes' rule as

$$p_{\mu}\left(\mu \,|\, \{(x(k), y(k))\}_{k=1}^{T}\right) = \frac{p\left(\{(x(k), y(k))\}_{k=1}^{T} \,|\, \mu\right)}{\sum_{\mathcal{M}} p\left(\{(x(k), y(k))\}_{k=1}^{T} \,|\, \mu\right)} \tag{18}$$

where

$$p\left(\{(x(k), y(k))\}_{k=1}^{T} \,|\, \mu\right) = \int_{\Theta} p\left(\{(x(k), y(k))\}_{k=1}^{T} \,|\, \vartheta, \mu\right) p_{\vartheta}(\vartheta) d\vartheta \tag{19}$$

and

$$p\left(\{(x(k), y(k))\}_{k=1}^{T} \,|\, \vartheta, \mu\right) = \prod_{k=1}^{T} p_e(y(k) - \theta_{\mu(k)}[x(k)^{\top} \quad 1]^{\top}). \tag{20}$$

Remark IV.5: Consider the case when all values of all parameters θ_i in (5) are known exactly, before the identification commences. Denote the known values of parameters with θ_i^0. The joint *a priori* parameter probability density function then takes the form

$$p_{\theta_1,\theta_2,\ldots,\theta_s}(\theta_1, \theta_2, \ldots, \theta_s) = \prod_{i=1}^{s} \delta\left(\theta_i - \theta_i^0\right)$$

and for a given mode function μ, the integral (19) becomes

$$p\left(\{(x(k), y(k))\}_{k=1}^{T} \,|\, \mu\right) = \prod_{k=1}^{T} p_e\left(y(k) - \theta_{\mu(k)}^0 [x(k)^{\top} \quad 1]^{\top}\right).$$

Assume, now that $\theta_i^0 = \theta_j^0$, for some $i, j, i \neq j$. Denote with μ^* the mode function that optimizes Problem IV.4. Choose now any k^* such that $\mu^*(k^*) = i$. It is easy to see that the function

$$\mu^{**}(k) = \begin{cases} j, & \text{if } k = k^* \\ \mu^*(k), & \text{otherwise} \end{cases}$$

is also an optimal solution of the Problem IV.4, as it attains the the same value of the optimality criterion. In other words, some or all of data points that are generated by the mode i can be assigned to the mode j and *vice versa*, while the classification is considered equally good. This may cause problems when reconstructing the regions on the basis of the estimated mode function.

The methodology that we will pursue in this paper amounts to first finding the optimal mode sequence μ^* by solving the classification problem, and using this mode sequence to solve the parameter estimation problem. The parameter estimation problem can be posed as follows.

Problem IV.6 (Parameter Estimation): Given the a priori joint parameter pdf p_{ϑ} and the mode function μ, compute

$$p_{\vartheta}\left(\vartheta \,|\, \{(x(k), y(k))\}_{k=1}^{T}, \mu\right). \tag{21}$$

Point estimates of the parameter values can be obtained as maximum likelihood estimates, or as expected values

$$\vartheta^E = E\vartheta. \tag{22}$$

Under the Assumption IV.3 we can solve Problem IV.6 using Bayes' rule as

$$p_\vartheta(\vartheta \mid \{(x(k),y(k))\}_{k=1}^T, \mu) = \frac{p\left(\{(x(k),y(k))\}_{k=1}^T \mid \vartheta, \mu\right) p_\vartheta(\vartheta)}{\int_\Theta p\left(\{(x(k),y(k))\}_{k=1}^T \mid \vartheta, \mu\right) p_\vartheta(\vartheta) d\vartheta}. \quad (23)$$

Once the joint parameter pdf (23) is computed (22) can be easily solved numerically. In the sequel, we will focus on the Problem IV.4.

V. Suboptimal Identification Algorithm

The optimization Problem IV.4 is a combinatorial optimization problem, where all possible mode sequences have to be explored in order to obtain an optimal solution. For large data sets such a search quickly becomes computationally intractable. Hence, we have to resort to suboptimal minimization algorithms.

We will consider the data points sequentially, and aim to find the best possible classification of the data pair $(x(k), y(k))$, with data points up to $k-1$ already classified. The described optimization strategy is known in the optimization literature as the *greedy strategy*—the algorithm tries to make the best possible local decision, in order to approach the global optimum. Let $p_\theta(\cdot; k)$ denote the pdf of the parameter θ after k steps of the algorithm. Let $p_\theta(\cdot; 0)$ denote the *a priori* parameter pdf, i.e., $p_\theta(\cdot; 0) = p_\theta(\,\cdot\,)$.

Assumption V.1: Assume that the a priori joint pdf of $\theta_1, \ldots, \theta_s$ takes the form

$$p_{\theta_1,\ldots,\theta_s}(\theta_1, \ldots, \theta_s; 0) = \prod_{i=1}^s p_{\theta_i}(\theta_i; 0) \quad (24)$$

That is, for all $i \neq j$ parameters θ_i and θ_j are assumed to be mutually independent at step 0.

Assume now that the parameters are independent at time step $k-1$, i.e.,

$$p_{\theta_1,\ldots,\theta_s}(\theta_1, \ldots, \theta_s; k-1) = \prod_{i=1}^s p_{\theta_i}(\theta_i; k-1).$$

We consider the following problem.

Problem V.2: For $k = 1, \ldots, T$ find the most likely mode $\mu(k)$ of the data pair $(x(k), y(k))$, given the a priori joint parameter pdf $p_{\theta_1,\ldots,\theta_s}(\theta_1, \ldots, \theta_s; k-1)$ at step $k-1$, i.e.,

$$\mu(k) = \arg\max_i p((x(k),y(k)) \mid \mu(k) = i) \quad (25)$$

where

$$p((x(k),y(k)) \mid \mu(k) = i) = \int_{\Theta_i} p((x(k),y(k)) \mid \theta) p_{\theta_i}(\theta; k-1) d\theta \quad (26)$$

and

$$p((x(k),y(k)) \mid \theta) = p_e(y(k) - \theta[x(k)^\top \quad 1]^\top). \quad (27)$$

Problem V.2 is solved in a straightforward way, by computing (26) for $i = 1, \ldots, s$ and choosing $\mu(k)$, according to (25). If $\mu(k) = i$ the *a posteriori* joint parameter pdf is computed using Bayes' rule as

$$p_{\theta_1,\ldots,\theta_s}(\theta_1, \ldots, \theta_s; k) = p_{\theta_{\mu(k)}}\left(\theta_{\mu(k)}; k\right) \prod_{\substack{i=1 \\ i \neq \mu(k)}}^s p_{\theta_i}(\theta_i; k-1) \quad (28)$$

where

$$p_{\theta_{\mu(k)}}(\theta; k) = \frac{p_e(y(k) - \theta^\top [x(k)^\top \quad 1]^\top) p_{\theta_{\mu(k)}}(\theta; k-1)}{\int_{\Theta_i} p_e(y(k) - \theta^\top [x(k)^\top \quad 1]^\top) p_{\theta_{\mu(k)}}(\theta; k-1) d\theta}. \quad (29)$$

Hence, if the parameters were independent at step $k-1$, from (28) it follows that after classifying the kth data point they will remain independent. From Assumption V.1, it follows that if the parameters were initially independent, they will remain independent throughout the parameter estimation procedure. The *a posteriori* joint parameter pdf is obtained by updating the pdf of the parameter that generated the data pair, while the pdf of the other parameters remains unchanged. Furthermore, if we define the support of $p_\theta(\,\cdot\,)$ as

$$\operatorname{supp} p_\theta = \{\theta \mid p_\theta(\theta) \neq 0\}$$

then from (29) it immediately follows that

$$\operatorname{supp} p_{\theta_{\mu(k)}}(\cdot; k) \subseteq \operatorname{supp} p_{\theta_{\mu(k)}}(\cdot; k-1).$$

That is, with every newly available data sample the support of the pdf of the parameter vector of the classified generating mode is nonexpanding.

Now, we are ready to formally state the algorithm for classification and parameter estimation.

Algorithm V.3 (Classification and Parameter Estimation):

- **Step 1:** Obtain the *a priori* probability density functions $p_{\theta_i}(\cdot; 0)$ for $i = 1, \ldots, s$; set $k = 1$.
- **Step 2:** Assign the data pair $(x(k), y(k))$ to the mode $\mu(k)$ with the highest likelihood using (25).
- **Step 3:** Compute the *a posteriori* pdf of the parameter $\theta_{\mu(k)}, p_{\theta_{\mu(k)}}(\cdot; k)$ using (29). For all $j \neq \mu(k)$, set $p_{\theta_j}(\cdot; k) = p_{\theta_j}(\cdot; k-1)$.
- **Step 4:** $k = k+1$; goto Step 2 until $k > T$ ♢

The schematic representation of the Algorithm V.3, for the case $s = 2$ is given in Fig. 1.

The Algorithm V.3 is derived by considering one data point at a time. It is possible, along the same lines, to derive a family of suboptimal algorithms, that would classify $m \leq T$ data points in each step. Note that, as m increases the complexity of the combinatorial optimization problem that has to be solved in each step increases exponentially. In particular, for $m = T$, the optimization problem becomes the classification Problem IV.4.

VI. Particle Filtering Approximation

Analytical solutions to (25) and (29) are intractable for general noise and parameter probability density functions. To turn Algorithm V.3 into a feasible computational scheme, we propose the particle filtering approach [14]. Here, we present only

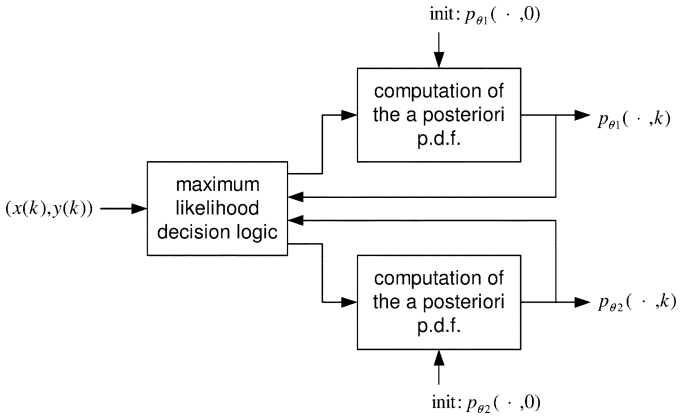

Fig. 1. Schematic representation of algorithm operation for two modes.

the main ideas of this approach. For a detailed exposition on implementing, tuning and convergence results for particle filters, see, e.g., [14], [15], [23], and the references therein.

The underlying idea of particle filtering methods is to approximate the pdf $p_{\theta_i}(\cdot;k)$ defined over a dense set Θ_i with a pdf supported in a finite number of points $\theta_i^{l,k} \in \Theta_i$, $l = 1, \ldots, N$ called particles. The pdf $p_{\theta_i}(\cdot;k)$ is then approximated as

$$p_{\theta_i}(\theta;k) \approx \hat{p}_{\theta_i}(\theta;k) := \sum_{l=1}^{N} w_i^{l,k} \delta\left(\theta - \theta_i^{l,k}\right) \qquad (30)$$

where $w_i^{l,k} > 0$ is a weight associated with the particle $\theta_i^{l,k}$ and $\sum_{l=1}^{N} w_i^{l,k} = 1$.

Algorithms that sample particles $\theta_i^{l,k}$ according to any given probability density function can be found in the literature (e.g., Metropolis–Hastings algorithm, Gibbs sampler, etc. [24]).

Estimates (2) and (3) can be obtained from (30) in a straightforward way. Combining (30) with (26) we obtain the following approximation for (26):

$$p((x(k),y(k)) \,|\, \mu(k) = i) \approx \sum_{l=1}^{N} w_i^{l,k-1} p_e\left(y(k) - \theta_i^{l,k-1} x(k)\right). \qquad (31)$$

To compute the recursion (29) we use a modification of the sample importance resampling (SIR) particle filtering algorithm [14]. This results in the following computational scheme.

Algorithm VI.1 (SIR Particle Filtering):

- FOR $l = 1$ TO N
- — diversify particles: $\theta_i^{l,k} = \theta_i^{l,k-1} + \varepsilon^l$, where $\varepsilon^l \sim \mathcal{N}(0, \Sigma_\varepsilon)$
- — compute weights $w_i^{l,k} = p((x(k),y(k)) \,|\, \theta_i^{l,k})$ using (27)

  END FOR
- normalize:

$$w_i^{l,k} := \frac{w_i^{l,k}}{\sum_{l=1}^{N} w_i^{l,k}}$$

  for $l = 1, \ldots N$
- draw N samples from distribution (30) to obtain the new set of particles $\theta_i^{l,k}$, where $w_i^{l,k} = N^{-1}$. $\diamondsuit$

Algorithms for sampling distributions of the type (30) are standard [see, for instance, [14, Alg. 2]]. Since we are using the SIR algorithm for estimating constant parameters it is necessary to diversify the particles [25]. For this purpose, we add the normally distributed random term ε^l to each particle in the first step of the Algorithm VI.1. The variance matrix Σ_ε is the tuning parameter of the algorithm. This method of particle diversification is simple, but increases the variance of the estimates. Other particle filtering algorithms with better statistical properties but higher computational load, can be found in the literature (see, for instance, [25]).

VII. Partition Estimation

Once the entire data set has been passed through the Algorithm V.3, the final pdfs of the parameters $p_{\theta_i}(\cdot;T)$ are available and all data points can be attributed to the mode with the highest likelihood, using (25). In other words, the mode function μ is re-estimated, using $p_{\theta_1,\ldots,\theta_s}(\cdot;T)$, in order to obtain the most likely μ. After this classification, standard techniques from pattern recognition can be applied to determine the regions $\{\mathcal{X}_i\}_{i=1}^{s}$ (see, e.g., [16]).

However, the method of maximum likelihood classification does not necessarily classify the data points to the correct mode. This problem is especially important when the hyperplanes defined by two parameter vectors θ_i and θ_j intersect over the region $\mathcal{X}_j$. Then, data points near this intersection may be wrongly attributed to the mode i. This issue will be illustrated in the example in Section VIII. Wrongly attributed data points may in turn lead to errors in determining the separating hyperplanes. In this section, we propose a modified version of the MRLP algorithm from [16] that aims to alleviate this problem.

Define the set $\mathcal{D}_i$ as

$$\mathcal{D}_i = \{x(k) \,|\, \mu(k) = i\} \qquad (32)$$

where $\mu(k)$ is computed as in (25), with $p_{\theta_i}(\cdot,T)$. Hence, $\mathcal{D}_i$ consists of all data points that are attributed to the mode i on the basis of the *a posteriori* pdf $p_{\theta_i}(\cdot;T)$.

Definition VII.1 [16]: The sets $\{\mathcal{D}_i\}_{i=1}^{s}$ are *piece-wise-linearly separable* if there exist $w_i \in \mathbb{R}^n$, $\gamma_i \in \mathbb{R}$ for $i = 1, \ldots, s$ such that

$$\left\langle \begin{bmatrix} x \\ 1 \end{bmatrix} \quad \begin{bmatrix} w_i \\ \gamma_i \end{bmatrix} \right\rangle > \left\langle \begin{bmatrix} x \\ 1 \end{bmatrix}, \begin{bmatrix} w_j \\ \gamma_j \end{bmatrix} \right\rangle \qquad (33)$$

for all $x \in \mathcal{D}_i$ and all $j \neq i$. Here, $\langle \cdot, \cdot \rangle$ denotes the standard inner product in $\mathbb{R}^{n+1}$.

Given w_i, γ_i the mode of the data point x can be estimated as

$$\tilde{\mu}(x) = \arg\max_i \left\langle \begin{bmatrix} x \\ 1 \end{bmatrix}, \begin{bmatrix} w_i \\ \gamma_i \end{bmatrix} \right\rangle \qquad (34)$$

and the hyperplane that separates regions $\mathcal{X}_i$ and $\mathcal{X}_j$ is given by

$$\{x \in \mathbb{R}^n \,|\, (w_i - w_j)^\top x = \gamma_i - \gamma_j\}. \qquad (35)$$

If the sets $\mathcal{D}_i$ are piecewise-linearly separable then the matrices H_i, h_i defining the region $\mathcal{X}_i$ as in (9) can be formed as

$$H_i = \mathrm{col}_j((w_i - w_j)^\top) \quad h_i = \mathrm{col}_j(\gamma_i - \gamma_j) \qquad (36)$$

where $j = 1, \ldots, s, j \neq i$, and the operator $\mathrm{col}_j(\cdot)$ evaluates its argument for all admissible values of index j and stacks the results into a column vector. Note that only simply connected and convex regions with up to $s-1$ vertices can be described in this way.

If the sets $\mathcal{D}_i$ are not piecewise-linearly separable some data points are going to violate (33). If the data point $x \in \mathcal{D}_i$ is classified to the region $\mathcal{X}_j$ (i.e., if $\tilde{\mu}(x) = j$) the *violation* $\zeta_{ij}(x) : \mathcal{D}_i \to \mathbb{R}$ can be defined as

$$\zeta_{ij}(x) = (-x(w_i - w_j) + (\gamma_i - \gamma_j) + 1)_+ \tag{37}$$

where $q_+ = \max\{q, 0\}$. Standard MRLP algorithm finds w_i, γ_i by minimizing the sum of averaged violations (37), through a single linear program [16].

In our case, we will weight the violations (37) according to the following principle: If the probability that the regressor $x \in \mathcal{D}_i$ belongs to mode i is approximately equal to the probability that it belongs to mode j, then the corresponding violation $\zeta_{ij}(x)$, if positive, should not be penalized highly. We define the *weighting function* $\xi_{ij} : \mathcal{D}_i \to \mathbb{R}$ as

$$\xi_{ij}(x(k)) = \log \frac{p((x(k), y(k)) \,|\, \mu(k) = i)}{p((x(k), y(k)) \,|\, \mu(k) = j)}. \tag{38}$$

Since for any $j \neq i$

$$p((x(k), y(k)) \,|\, \mu(k) = i) > p((x(k), y(k)) \,|\, \mu(k) = j)$$

the weight (38) is always nonnegative, and is equal to zero when the two likelihoods are exactly equal.

The weighting function (38) takes into account only the relative size of the mode likelihoods. If outliers are present in the data set, mode likelihoods may be negligible, but their ratio, formed as in (38), may still be significant. Another possible choice of the weighting function ξ_{ij}, which uses the absolute sizes of mode likelihoods is

$$\xi_{ij}(x(k)) = p((x(k), y(k)) \,|\, \mu(k) = i) - p((x(k), y(k)) \,|\, \mu(k) = j). \tag{39}$$

The optimization problem can be stated as

$$\min_{w_i, \gamma_i} \sum_{i=1}^{s} \sum_{\substack{j=1 \\ j \neq i}}^{s} \sum_{x \in \mathcal{D}_i} \xi_{ij}(x)\, \zeta_{ij}(x). \tag{40}$$

Problem (40) can be further cast as a linear program, in a same way as in [16].

By introducing pricing functions more information is preserved from the classification phase to the region estimation phase. This is an advantage over the region estimation procedures presented in [8] and [9]. We will illustrate this issue with an academic example in the next section.

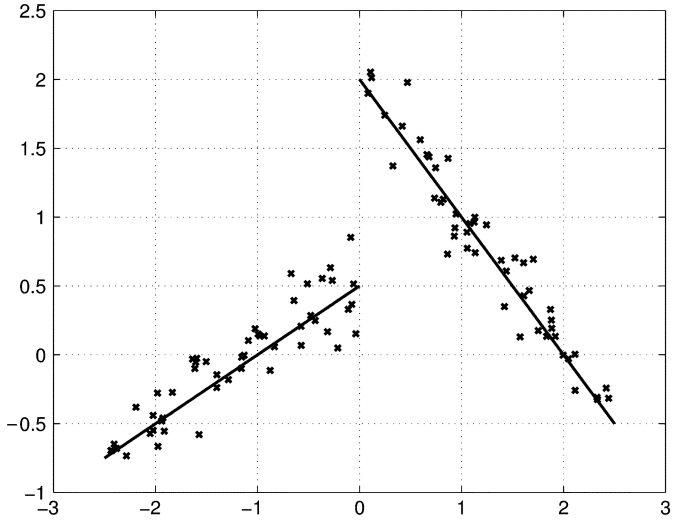

Fig. 2. Data set used for identification together with the true model.

VIII. Example

Let the data $\{(x(k), y(k))\}_{k=1}^{100}$ be generated by a system of type (5) where

$$f(x) = \begin{cases} \begin{bmatrix} 0.5 & 0.5 \end{bmatrix} \begin{bmatrix} x \\ 1 \end{bmatrix}, & \text{if } x \in [-2.5, 0) \\ \begin{bmatrix} -1 & 2 \end{bmatrix} \begin{bmatrix} x \\ 1 \end{bmatrix}, & \text{if } x \in [0, 2.5] \end{cases} \tag{41}$$

and $e(k)$ is a sequence of normally distributed random numbers, with zero mean and variance $\sigma_e^2 = 0.025$. The data set of $T = 100$ data points together with the true model is shown in Fig. 2.

A priori pdfs are chosen as $p_{\theta_1}(\theta_1) = p_{\theta_2}(\theta_2) = \mathcal{U}([-2.5, 2.5] \times [-2.5, 2.5])$, where $\mathcal{U}$ denotes the uniform distribution. A particle approximation to this pdf, with $N = 200$ particles for each pdf, is given in Fig. 3(left). The particle filtering Algorithm VI.1 is applied, with $\Sigma_\varepsilon^2 = diag\{0.001, 0.001\}$ and the final particle distribution at step $k = 100$ is shown in Fig. 3(right). The estimates of the parameter vectors are

$$\theta_1^E = \begin{bmatrix} 0.4619 \\ 0.5279 \end{bmatrix} \quad \theta_2^E = \begin{bmatrix} -0.9350 \\ 1.9006 \end{bmatrix}. \tag{42}$$

Data points are classified using (25), and the results are depicted in Fig. 4(a). Several data points that belong to mode 1 are attributed to mode 2. These points are near the virtual intersection of the two planes defined by the parameter vectors. In Fig. 4(b), the weighting functions (see (38)) $\xi_{1,2}$ and $\xi_{2,1}$ for misclassification of points are shown. The weight for misclassification of wrongly attributed points is small in comparison to the weight for misclassification of the correctly attributed points. The region for mode 1 is estimated as $x \geq 0.0228$ while the region corresponding to mode 2 is estimated as $x < 0.0228$.

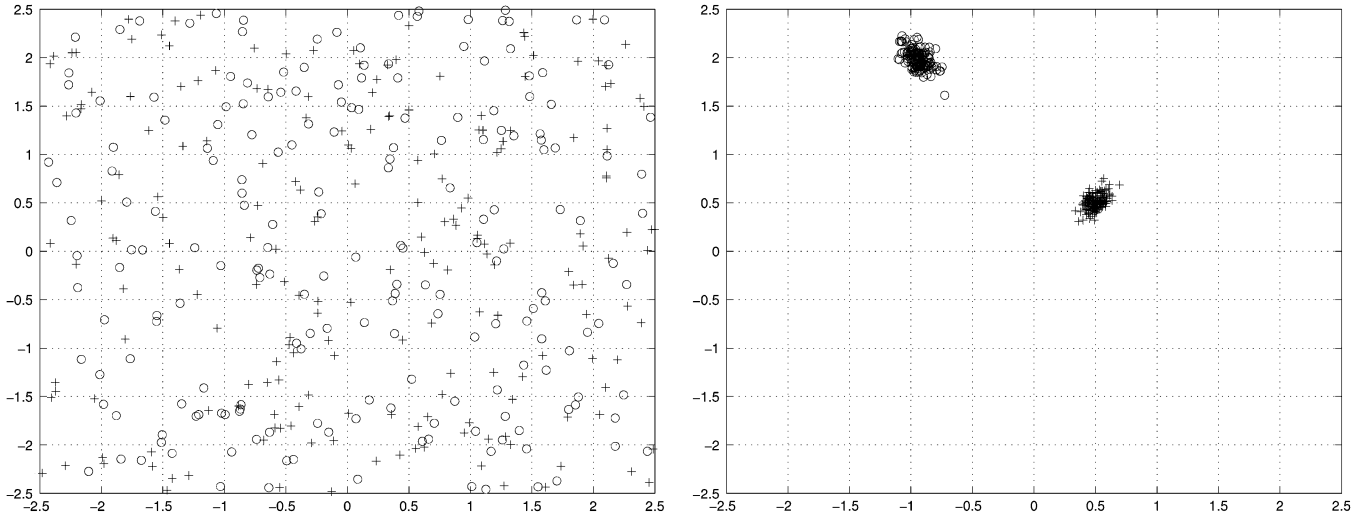

Fig. 3. **Left:** Particle approximation of the initial pdfs of the parameters θ_1, θ_2 ($+$: particles of p_{θ_1}, $\circ$: particles of p_{θ_2}). **Right:** Final particle approximation of pdf of the parameters θ_1 $(+), \theta_2(\circ)$.

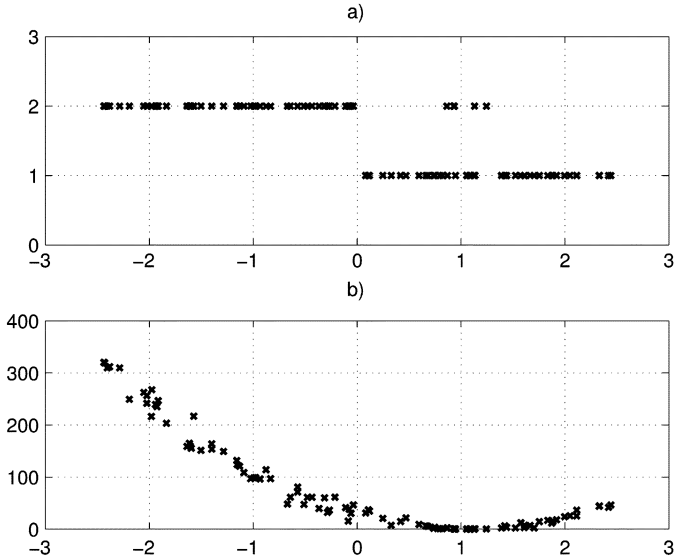

Fig. 4. (a) Data points attributed to the different modes. (b) Pricing functions $\xi_{2,1}(\mathcal{D}_2)$ and $\xi_{1,2}(\mathcal{D}_1)$ for the wrong classification.

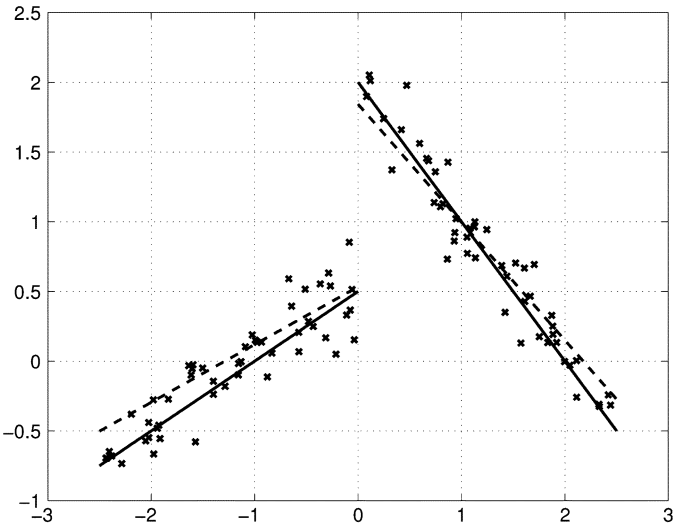

Fig. 5. True model (solid), the identified model (dashed) and the data set used for identification.

The identified model, together with the true model and the data set is depicted in Fig. 5.

IX. Initialization

In this section, we will discuss in more detail three different ways to obtain *a priori* probability density functions $p_{\theta_i}(\cdot;0), i = 1, \ldots, s$.

A. Initialization Using Mode Knowledge

If $m > n+1$ data pairs $(x(k_1), y(k_1)), \ldots (x(k_m), y(k_m))$ are attributed to the mode i, the least squares estimate of the value of the parameter vector θ_i^{LS} may be obtained as

$$\begin{aligned} \theta_i^{\mathrm{LS}} &= (\Phi_i^\top \Phi_i)^{-1} \Phi_i^\top y_i \\ \Phi_i &= \begin{bmatrix} x(k_1) & x(k_2) & \cdots & x(k_m) \\ 1 & 1 & \cdots & 1 \end{bmatrix}^\top \\ y_i &= [y(k_1) \quad y(k_2) \quad \cdots y(k_m)]^\top. \end{aligned} \tag{43}$$

The empirical covariance matrix of θ_i^{LS} can be computed as [26]

$$\begin{aligned} V_i &= \frac{\mathrm{SSR}_i}{m-(n+1)} (\Phi_i^\top \Phi_i)^{-1} \\ \mathrm{SSR}_i &= y_i^\top (I - \Phi_i (\Phi_i^\top \Phi_i)^{-1} \Phi_i^\top) y_i \end{aligned} \tag{44}$$

where SSR_i denotes a sum of squared residuals. This information is sufficient to initialize the parameter θ_i as a normally distributed random variable

$$p_{\theta_i}(\cdot;0) = \mathcal{N}(\theta_i^{\mathrm{LS}}, V_i). \tag{45}$$

Samples from the normal distribution can be easily obtained with some of the mentioned algorithms for sampling from general multidimensional distributions (or using built-in MATLAB functions).

B. Initialization Via Clustering Procedure

In this section, we will show that our procedure can be initialized using the ideas from the clustering procedure [6]. For

the sake of completeness we discuss the relevant steps from the clustering procedure briefly. For a detailed exposition, see [6].

For each regressor $x(l)$ in the data set, we collect $c > n+1$ nearest regressors, and form a *local data set* (LD) $\mathcal{C}_l$. The rationale behind this procedure is that regressors that are close in the regressor space are likely to belong to the same partition; we distinguish two types of LDs—*pure* LDs—when all the regressors collected in one LD indeed belong to the same partition, and *mixed* LDs, when they do not belong to the same partition. For the procedure to work properly the ratio between the number of pure and mixed LDs should be high.

From each LD we can obtain an estimate θ^l, using (43), and the variance V^l of the estimate θ^l, using (44). To each estimate the following confidence measure is assigned:

$$w^l = \frac{1}{\sqrt{(2\pi)^{n+1}\det(V^l)}}. \tag{46}$$

Ideally, pure LDs will produce good estimates, with high values of w^l, while for mixed LDs w^l will be low. Estimates obtained from pure LDs are expected to form groups (clusters) in the parameter space, while estimates from mixed LDs will be isolated points.

The next step is to form s clusters $\{\mathcal{D}_i\}_{i=1}^s$ in the parameter space, by solving the following optimization problem:

$$\min_{\{\mathcal{D}_i\}_{i=1}^s, \{m_i\}_{i=1}^s} J\left(\{\mathcal{D}_i\}_{i=1}^s, \{m_i\}_{i=1}^s\right). \tag{47}$$

The clustering functional J is given as

$$J\left(\{\mathcal{D}_i\}_{i=1}^s, \{m_i\}_{i=1}^s\right) = \sum_{i=1}^{s} \sum_{\theta^l \in \mathcal{D}_i} \|\theta^l - m_i\|^2_{(V^l)^{-1}} \tag{48}$$

where m_i is the center of the cluster $\mathcal{D}_i$. The optimization problem (47) is computationally hard, but there exist efficient algorithms that provide suboptimal solutions, e.g., the K-means algorithm [6]. The weight V^l in (48) is used to minimize the influence of θ^l that correspond to mixed LDs, which in turn may lead to wrong assignment of those parameter vectors.

Points attributed to the ith cluster, $\theta^l \in \mathcal{D}_i$, together with the associated weights w^l can be used to form a probability density function of type (30)

$$p_{\theta_i}(\theta) = q^{-1} \sum_{\theta^l \in \mathcal{D}_i} w^l \delta(\theta - \theta^l) \tag{49}$$

where q is a normalizing constant

$$q = \sum_{\theta^l \in \mathcal{D}_i} w^l. \tag{50}$$

In the clustering procedure, after the clustering step, the bijective relation

$$(x(l), y(l)) \leftrightarrow \mathcal{C}_l \leftrightarrow \theta^l \in \mathcal{D}_i$$

is used to classify data pairs to modes. Data pairs $(x(l), y(l))$ that correspond to mixed LDs may be wrongly classified. In our procedure, mixed LDs yield a point with low weight in the discrete approximation of the parameter pdf. This point will be discarded in the SIR particle filtering Algorithm VI.1, and will have no adverse consequences on the classification of the corresponding data pair.

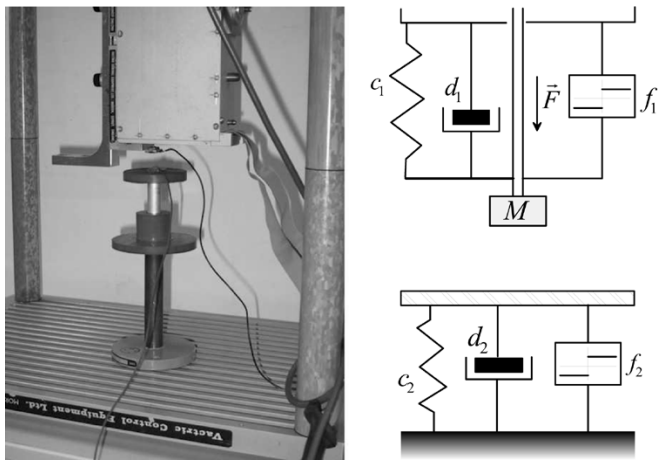

Fig. 6. **Left:** (a) Photo of the experimental setup. **Right:** (b) Schematic representation of the experimental setup.

C. Brute Force Initialization

Parameters θ_i can be estimated in an optimal way as the solution of the following problem:

$$\{\theta_i\}_{i=1}^s = \arg\min_{\theta_i} \sum_{k=1}^{T} \|y(k) - \theta_{\mu(k)}[x^\top \quad 1]\|^2. \tag{51}$$

When the sequence $\mu(k), k = 1, \ldots T$ is known problem (51) is an ordinary least squares problem. In our case, since the mode sequence is not known, problem (51) is a combinatorial optimization problem, where all possible mode sequences must be explored, which can be computationally intractable for larger values of T.

In order to obtain a rough estimate of the parameter values a small enough subset of the complete data set $(x(k), y(k)), k = 1, \ldots T$ can be chosen, and a computationally tractable problem of type (51) can be formulated. The solution of this problem gives estimates of the parameter values $\hat{\theta}_i$, together with the variances of the estimates V_i. This information is sufficient to describe the parameters as $\theta_i \sim \mathcal{N}(\hat{\theta}_i, V_i)$.

X. Experimental Example

In order to demonstrate the proposed identification procedure we applied it to the data collected from an experimental setup made around the mounting head from a pick-and-place machine. The purpose of the setup is to study the component placement process on a printed circuit board (PCB) in the controlled conditions. The same experimental setup was previously successfully identified using the clustering procedure [6], and the greedy procedure [12]. The experimental setup and the identification results with the clustering procedure are described in more detail in [17] and [18].

A photo and the schematic representation of the experimental setup are given in Fig. 6. The setup consists of the mounting

head, from an actual pick-and-place machine, which is fixed above the impacting surface [the small disc in Fig. 6(a)]. The impacting surface is in contact with the ground via the spring [the spring c_2 in Fig. 6(b)], within the outer tube in Fig. 6(a). The mechanical construction under the impacting surface is such that only the movement on the vertical axis is enabled [inner tube, which can slide inside the outer tube in Fig. 6(a)]. This construction exhibits linear and dry friction phenomena, represented in Fig. 6(b) by the damper d_2 and the block f_2, respectively. The chosen design of the impacting surface simulates the elasticity properties of the PCB as well as hard mechanical constraints due to saturations. It also introduces some side effects, such as dry friction.

The mounting head contains: A vacuum pipette which can move on the vertical axis [the mass M in Fig. 6(b)] and which is connected via the spring to the casing [the spring c_1 in Fig. 6(b)]; an electrical motor which enables the movement [represented by force $\vec{F}$ in Fig. 6(b)]; and a position sensor, which measures the position of the pipette, relative to the upper retracted position. The position axis is pointed downwards, i.e., the value of the position increases when the pipette moves downwards. The motion of the pipette is also subject to friction phenomena [the damper d_1 and the dry friction block f_1 in Fig. 6(b)].

The dynamics of the experimental setup exhibits, in a first approximation, four different modes of operation.

- **Upper saturation:** The pipette is in the upper retracted position (i.e., cannot move upwards, due to the physical constraints).
- **Free mode:** The pipette is not in contact with the impacting surface, but is not in the upper saturation.
- **Impact mode:** The pipette is in contact with the impacting surface, but is not in lower saturation.
- **Lower saturation:** The pipette is in the lower extended position, (i.e., cannot move downwards due to the physical constraints).

We stress that the switch between the impact and free modes does not occur at a constant head position, because of the movement of the impacting surface. For the upper and lower saturations, although they occur at a fixed position, they introduce dynamic behaviors due to bouncing when hitting the constraints.

The control input is the voltage applied to the motor, which is converted up to a negligible time constant to the force $\vec{F}$. The input signal for the identification experiment should be chosen in a way that modes of interest are sufficiently excited. To obtain the data for identification, the input signal $u(t)$ is chosen as

$$u(t) = a_k \quad \text{when} \quad t \in [k\tau, (k+1)\tau) \tag{52}$$

where $\tau > 0$ is fixed, and the amplitude a_k is a random variable, with uniform distribution in the interval $[a, b]$. By properly choosing the boundaries of the interval $[a, b]$ only certain modes of the system are excited. For instance, one can choose to excite free and impact modes, without reaching upper and lower saturations. Physical insight into the operation of the setup facilitates the initialization of the 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. Data points that belong to saturations can also be distinguished. This *a priori* information may be exploited in a way described in Section IX-A.

TABLE I
COMPUTATION TIMES FOR BI-MODAL IDENTIFICATION AND IDENTIFICATION WITH LOWER SATURATION

	Bi-modal id.	Id. with lower saturation
Parameter estimation	55 s	72 s
Final assignment	16 s	23 s
Computing pricing functions	16 s	30 s
Number of constraints in MRLP	208	918
Solving MRLP	1.7 s	2.6 s

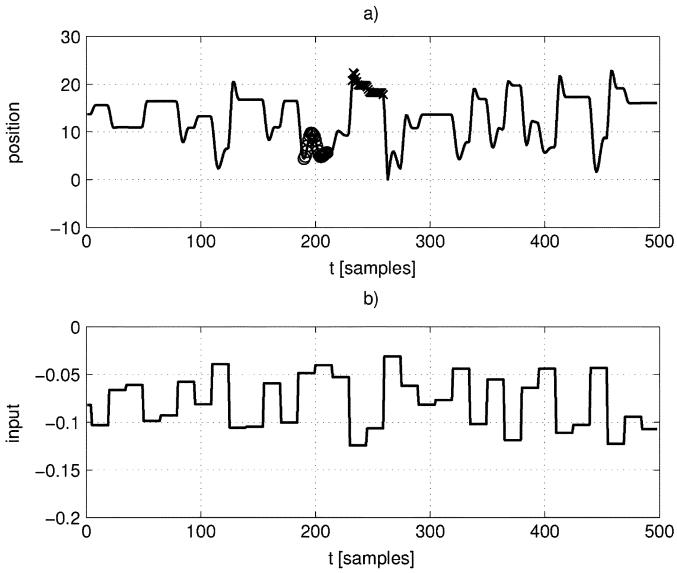

Fig. 7. Bimodal identification. The data set used for identification (a) position (points marked with ∘: data points used for the initialization of the free mode; points marked with ×: data points used for initialization of impact mode (b) input signal

In the sequel, we present two identification experiments: in the first experiment only free and impact modes are excited; in the second experiment free, impact and lower saturation modes are excited. The collected data sets consist of 750 points, and are divided into two overlapping sets of 500 points: One is used for identification, while the second is used for validation of the identified models.

In all examples, the weighting function (39) is used. As a pdf of the noise we used $p_e \sim \mathcal{N}(0, 1)$.

The parameter estimation, the final assignment and the computation of pricing functions were done in Matlab, on a Pentium 4, 2-Ghz computer with 512 Mb of memory under Windows XP. The region estimation (i.e., solving the linear program resulting from MRLP procedure) was done using the CPLEX software on a dual Pentium Xeon 2.8 Ghz computer, with 4 GB of memory under Linux. This computer was simultaneously used by several other users. The computation times are given in Table I.

A. Bimodal Identification

The data set used for identification is depicted in Fig. 7. Portions of the data set that are used for initialization of free and impact mode are marked with × and ∘, respectively. Models

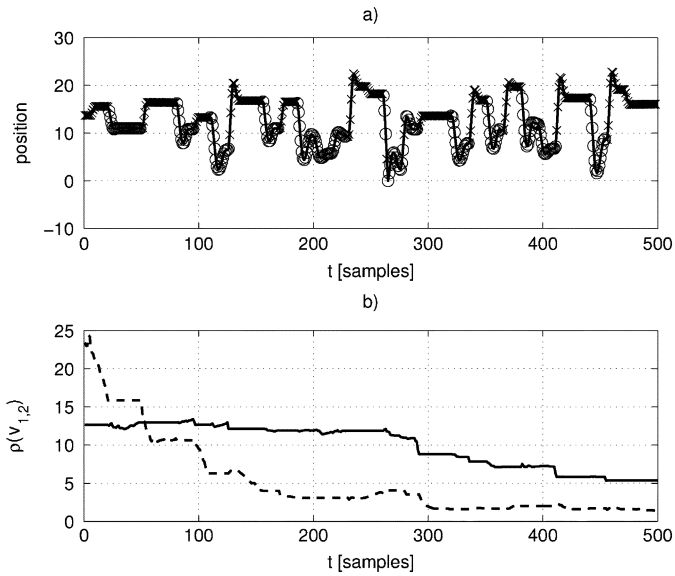

Fig. 8. Bimodal identification. (a) Classified data points (∘: free mode, ×: impact mode). (b) $\rho^E_{1,2}$ (solid line: free mode; dashed line: impact mode).

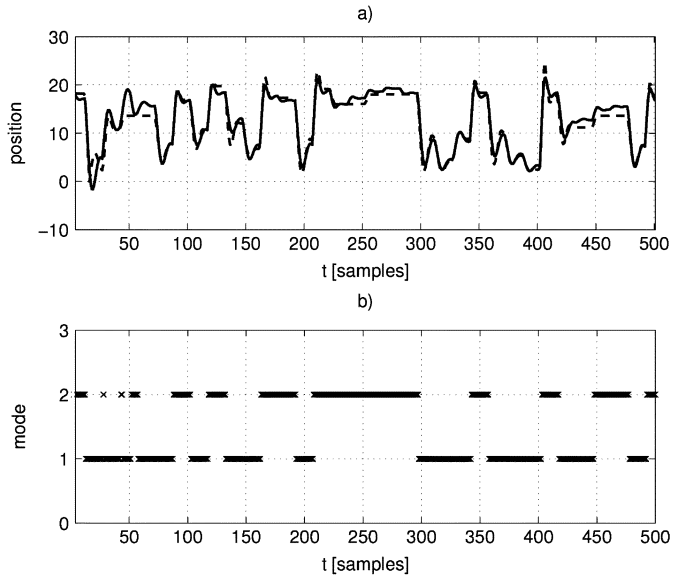

Fig. 9. Bimodal identification. (a) Simulation of the identified model (solid line: simulated response, dashed line: measured response. (b) Modes active during the simulation.

with $s = 2, n_a = 2, n_b = 1$ are identified. The computation times are given in Table I.

The final classification of the data points is depicted in Fig. 8(a). In Fig. 8(b), the spectral radii of the variance matrices $\rho^E_{1,2}$ at each step of the classification are depicted. *Simulation* of the identified model using the validation data, together with the modes active during the simulation is depicted in Fig. 9.

From Fig. 8(a), we see that the identification procedure separated the data points into two groups, that correspond to impact and free modes. From Fig. 8(b), we see that the estimates of the parameters, described by the spectral radii of the covariance matrices (4), improve during the iterations of the algorithm.

From the comparison of the simulated response of the model and the measured response we see that the identified model is satisfactory. However, the system response in both impact and free modes is nonlinear, because of the presence of dry friction. The effects of the dry friction are especially pronounced in the impact mode, and can be observed in Fig. 7, for instance on a time interval around 300, where small changes in the input signal produce no change in the measured output, because the dry friction is in stick phase. Since the impact and the free modes are described by one linear model each, the effects of the dry friction can not be properly described in the identified model. For instance, the discrepancy between the simulated and measured response in Fig. 9 on the time interval around 250 is due to this effect. While the dry friction is in the stick phase in the real system, and no change in the position is visible, the identified model predicts a linear step response.

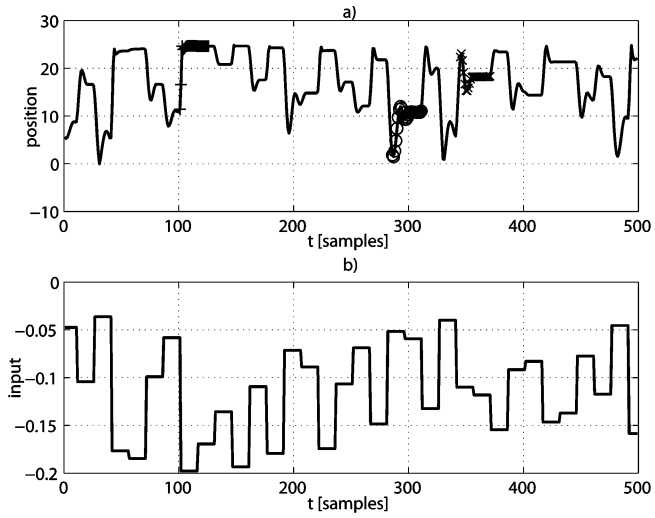

Fig. 10. Identification with saturations. Data set used for identification (a) position (points marked with +: data points used for the initialization of the lower saturation mode; points marked with ∘: data points used for initialization of impact mode; points marked with ×: data points used for the initialization of free mode (b) input signal.

B. Identification With Lower Saturation

The data set used for identification is depicted in Fig. 10. Portions of the dataset that are used for initialization of free, impact and saturation mode are marked with $\times, \circ$, and $+$, respectively. Models with $s = 3, n_a = 2, n_b = 2$ are identified. The computation times are given in Table I.

Final classification of data points is depicted in Fig. 11(a). In Fig. 11(b), spectral radii of variance matrices $\rho^E_{1,2,3}$ at each step of the classification are depicted. Simulation of the identified model, together with the modes active during the simulation is depicted in Fig. 12. The parameters of the identified model are given in the Table II.

From Fig. 11(a), we see that data points are classified into three groups, corresponding to the impact, free, and saturation modes. From 11(b), we see that the estimates of the parameters are improving during the iterations of the algorithm. From Fig. 12, we see that the simulated response is satisfactory, and that the modes active during the simulation correspond well to intuitive classification of data. The response in the free mode does not match the measured response precisely, while the responses in impact and saturation modes are predicted remarkably well.

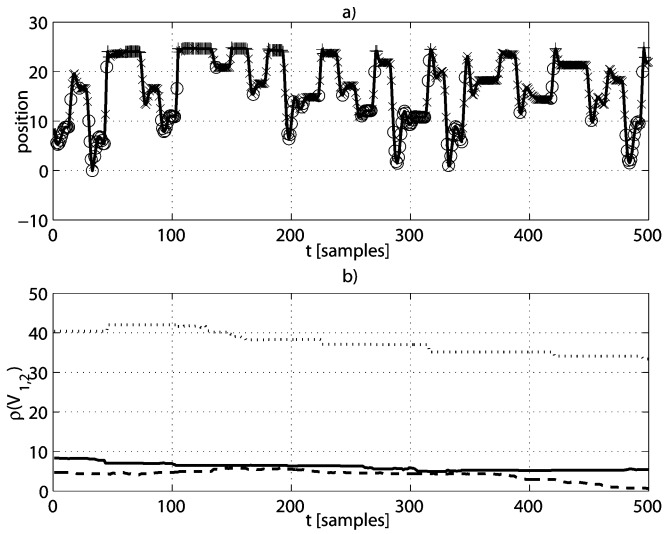

Fig. 11. Identification with saturations. (a) Classified data points (∘: free mode, ×: impact mode; + lower saturation). (b) $\rho^{E}_{1,2,3}$ (solid line: free mode; dashed line: impact mode; dotted line: lower saturation).

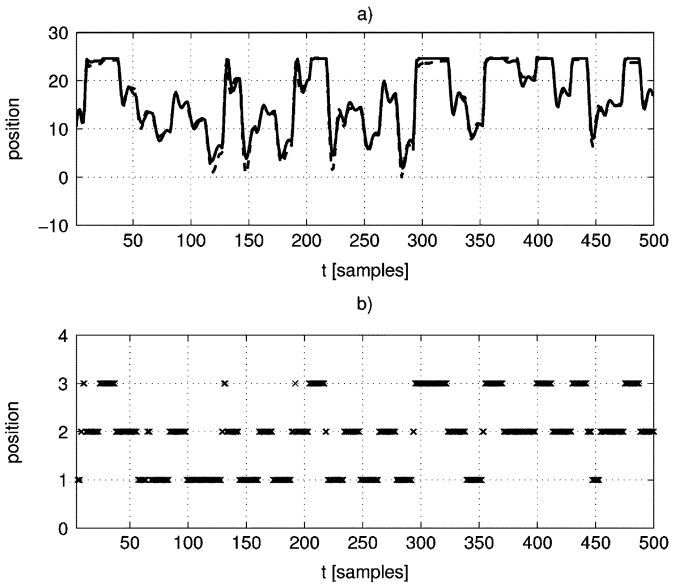

Fig. 12. Identification with saturations. (a) Simulation of the identified model (solid line: simulated response, dashed line: measured response. (b) Modes active during the simulation.

It is interesting to consider further the saturation mode. From the physical considerations we know that the position of the mounting head stays close to the certain value saturation level y_s, as long as the system is in saturation. To gain some insight about the predicted saturation level y_s from the identified model, consider the "steady state" situation in which the system is in saturation at time $k-2$ and $k-1$ (i.e., $y(k-2)=y(k-1)=y_s$) and the value of the input is constant in the time instants $k-2, k-1$ and such that system stays in saturation also at the time instant k (i.e., $y(k)=y_s$). According to the ARX dynamics of mode 3, y_s and u must satisfy the linear constraints

$$y_s(1-\theta_{3,1}-\theta_{3,2}) = (\theta_{3,3}+\theta_{3,4})u+\theta_{3,5} \quad \text{and} \quad H_3[y_s\ y_s\ u\ u]^\top \le h_3. \quad (53)$$

The minimal and maximal values that the output y can have under the previously stated assumptions can be found by solving the linear programs $\min y_s$ and $\max y_s$, respectively, in the unknowns y_s and u subject to the constraints (53). We found $\min y_s = 24.3903, \max y_s = 24.5047$. These values very precisely correspond to the values that the measured output takes while in saturation. In [18], the minimal and maximal values of y_s under the same assumptions were determined from the model identified using the clustering procedure, and the following values were obtained: $\min y_s = 23.2, \max y_s = 24.2$. The computed minimal and maximal values of y_s identified using the Bayesian procedure are tighter than the values identified with the clustering procedure.

XI. Conclusion

In this paper, we have presented a novel method for the identification of hybrid systems in PWARX form. The presented method facilitates incorporation of the available a priori information on the system to be identified, but can also be initialized and used as a black-box method.

Unknown model parameters are treated as random variables described by their pdfs. We pose the identification problem as the problem of computing the a posteriori pdf of the model parameters given the observed data set and the prior information. The identification problem is subsequently relaxed, until the procedure which can be practically implemented is obtained. A modified MRLP procedure, based on pricing functions is used for the estimation of the regions. Pricing functions preserve the valuable information from the classification phase for the region estimation. The applicability and the effectiveness of the proposed algorithm is illustrated by an academic and experimental example.

The suboptimal approach taken in the derivation of the proposed algorithm is to consider one data pair in each step of the algorithm (sequential processing), and to determine the optimal classification of the considered data pair, assuming that all previously considered data is processed optimally. In other words, the proposed algorithm aims to find the best possible local decision, with the purpose to approach the global optimum. In the optimization literature this approach to optimization is known as the greedy approach. Sequential data processing brings about several possibilities. First, the parameter estimation part of the algorithm can be implemented in a *recursive* fashion, where the pdfs of the parameters are updated as the new data measurements become available. New and efficient algorithms for data classification that allow sequential data processing started appearing recently [27], [28]. However, still the complete previous measurement history must be memorized. Another possibility is *incremental* identification. As noticed in [18] in practice it is frequently possible to excite only some of the modes of the physical system. The basic idea would be to reconstruct first the modes visible in the simpler experiments, and then enhance the model with additional behaviors appearing in richer data sets, by using the pdfs of the already identified modes as priors. Again, data classification algorithms that allow for sequential data processing are necessary. Both topics will be the subject of the future research.

TABLE II
PARAMETERS OF THE IDENTIFIED MODEL $s = 3, n_a = 2, n_b = 2$

$$\theta_1 = \begin{bmatrix} 1.5234 & -0.7407 & -35.5370 & -5.9786 & -0.2236 \end{bmatrix}^\top,$$
$$\theta_2 = \begin{bmatrix} 1.3163 & -0.6682 & -31.1894 & -10.5764 & 1.5656 \end{bmatrix}^\top,$$
$$\theta_3 = \begin{bmatrix} -0.1161 & 0.0647 & 2.1464 & -1.2081 & 25.9239 \end{bmatrix}^\top,$$

$$H_1 = \begin{bmatrix} 0.039677 & 0.012028 & -14.109 & -17.308 \\ 99.531 & -57.568 & -3159.7 & -121.09 \\ -1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 1 \end{bmatrix}, H_2 = \begin{bmatrix} -0.039677 & -0.012028 & 14.109 & 17.308 \\ 99.491 & -57.58 & -3145.6 & -103.79 \\ -1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix},$$

$$H_3 = \begin{bmatrix} -99.531 & 57.568 & 3159.7 & 121.09 \\ -99.491 & 57.58 & 3145.6 & 103.79 \\ 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

$$h_1 = \begin{bmatrix} 2.8771 & 1588.4132 & 40 & 40 & 40 & 40 & 0.3 & 0.3 & 0.3 & 0.3 \end{bmatrix}^\top,$$
$$h_2 = \begin{bmatrix} -2.8771 & 1585.5631 & 40 & 40 & 40 & 40 & 0.3 & 0.3 & 0.3 \end{bmatrix}^\top,$$
$$h_3 = \begin{bmatrix} -1588.4132 & -1585.5631 & 40 & 40 & 40 & 0.3 & 0.3 & 0.3 & 0.3 \end{bmatrix}^\top$$

Another possible suboptimal approach for solving the classification problem would be to first classify all of the available data on the basis of the available a priori knowledge (batchwise processing), and after that compute the a posteriori parameter pdfs on the basis of all data points that are classified to the respective mode (or estimate the parameters in some other way, e.g., using least squares, as in Section IX-A). Conceptually speaking, this approach would give good results if the *a priori* knowledge on the parameter values is precise enough to enable good classification. However, if this is the case, sequential processing is expected to perform equally well.

We use particle filters to represent and compute with the general probability density functions. However in some special cases (such as uniform or normal distributions for e and/or parameters) specific properties of the algorithm may be inferred from the explicit expressions of the update rule for determining a posteriori distributions in Algorithm V.3. This will be investigated in future research, as well.

Further research will also focus on the investigation of properties of the presented method: The influence of the quality of the available a priori knowledge, the convergence properties of the proposed algorithm and the relation between the obtained suboptimal solutions and the optimal ones.

REFERENCES

[1] E. D. Sontag, "Nonlinear regulation: The piecewise linear approach," *IEEE Trans. Autom. Control*, vol. AC-26, no. 2, pp. 346–358, Feb. 1981.

[2] A. Bemporad and M. Morari, "Control of systems integrating logic, dynamics and constraints," *Automatica*, vol. 35, no. 3, pp. 407–427, 1999.

[3] A. J. v. d. Schaft and J. M. Schumacher, "The complementary-slackness class of hybrid systems," *Math. Control, Signals, Syst.*, vol. 9, pp. 266–301, 1996.

[4] W. Heemels, J. Schumacher, and S. Weiland, "Linear complementarity systems," *SIAM J. Appl. Math.*, vol. 60, no. 4, pp. 1234–1269, 2000.

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

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

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

[8] A. Bemporad, A. Garulli, S. Paoletti, and A. Vicino, "A greedy approach to identification of piecewise affine models," *Hybrid Systems: Computation and Control 2003*, ser. Lecture Notes in Computer Science, vol. 2623, pp. 97–112, 2003.

[9] R. Vidal, S. Soatto, and S. Sastry, "An algebraic geometric approach for identification of linear hybrid systems," in *Proc. 42nd IEEE Conf. Decision and Control*, 2003, pp. 167–172.

[10] Y. Ma and R. Vidal, "Identification of deterministic switched arx systems via identification of algebraic varieties," in *Proceedings of Hybrid Systems: Computation and Control*. ser. Lecture Notes in Computer Science, M. Morari and L. Thiele, Eds. New York: Springer-Verlag, 2005, vol. 3414, pp. 449–465.

[11] H. Niessen, A. Juloski, G. Ferrari-Trecate, and W. Heemels, "Comparison of three procedures for the identification of hybrid systems," in *Proc. Conf. Control Applications*, Taipei, Taiwan, 2004.

[12] A. Juloski, W. Heemels, G. Ferrari-Trecate, R. Vidal, S. Paoletti, and J. Niessen, "Comparison of four procedurs for the identificaiton of hybrid systems," in *Hybrid Systems: Computation and Control*. ser. Lecture Notes in Computer Science, M. Morari and L. Thiele, Eds. New York: Springer-Verlag, 2005, vol. 3414, pp. 354–400.

[13] D. MacKay, "Bayesian methods for adaptive models," Ph.D. dissertation, California Inst. Technol., Pasadena, CA, 1992.

[14] M. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/nongaussian bayesian tracking," *IEEE Trans. Signal Process.*, vol. 50, no. 2, pp. 174–188, Feb. 2002.

[15] A. Doucet, N. de Freitas, and N. Gordon, Eds., *Sequential Monte Carlo Methods in Practice*. New York: Springer-Verlag, 2001.

[16] K. Bennet and O. Mangasarian, "Multicategory discrimination via linear programming," *Optim. Meth. Software*, vol. 4, pp. 27–39, 1994.

[17] A. Juloski, W. Heemels, and G. Ferrari-Trecate, "Identification of an experimental hybrid system," in *Proc. IFAC Symp. Analysis and Design of Hybrid Systems*, Saint Malo, France, 2003.

[18] ——, "Data-based hybrid modeling of the component placement process in pick-and-place machines," *Control Eng. Pract.*, vol. 12, pp. 1241–1252, 2004.

[19] J. Hamilton, "A new approach to economic analysis of nonstationary time series and the business cycle," *Econometrica*, vol. 57, pp. 357–384, 1989.

[20] J. D. Hamilton and S. R. Susmel, "Autoregressive conditional heteroskedasticity and changes in regime," *J. Econom.*, vol. 64, pp. 307–333, 1994.

[21] B. Hansen, "The likelihood ratio test under nonstandard conditions: Testing the Markov switching model of GNP," *J. Appl. Econom.*, vol. 7, pp. 61–82, 1992.

[22] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Royal Statist. Soc.*, vol. 39, pp. 1–38, 1977.

[23] D. Crisan and A. Doucet, "A survey of convergence results on particle filtering methods for practitioners," *IEEE Trans. Signal Process.*, vol. 50, pp. 736–746, 2002.

[24] G. Fishman, *Concepts, Algorithms, and Applications*. New York: Springer-, 1996.

[25] C. Berzuini and W. Gilks, "RESAMPLE-MOVE filtering with cross-model jumps," in *Sequential Monte-Carlo Methods in Practice*, A. Doucet, N. de Freitas, and N. Gordon, Eds. New York: Springer-Verlag, 2001, ch. 6, pp. 117–138.

[26] L. Ljung, *System Identification—Theory for the User*. Upper Saddle River, NJ: Prentice-Hall, 1999.
[27] G. Fung and O. L. Mangasarian, "Incremental support vector machine classification," in *Proc. 2nd SIAM Int. Conf. Data Mining*, Philadelphia, PA, 2002, pp. 247–260.
[28] G. Fung and O. Mangsarian, "Multicategory proximal support vector machine classifiers," *Mach. Learn.*, vol. 59, no. 1–2, pp. 77–97, 2005.

Aleksandar Juloski was born in Belgrade, Serbia and Montenegro, in 1976. He received the Dipl.Ing. degree in electrical engineering from the Faculty of Electrical Engineering, University of Belgrade, in 1999, and the Ph.D. degree from Eindhoven University of Technology, Eindhoven, The Netherlands, in 2004, with the thesis "Observer Design and Identification Methods for Hybrid Systems: Theory and Experiments."

He spent one year developing software for embedded controllers with the Mihajlo Pupin Institute in Belgrade. He is currently working as a Postdoctoral Researcher in the Control Systems Group, Department of Electrical Engineering, Eindhoven University of Technology. His research interests are modeling, identification and control of hybrid systems, and industrial applications.

Siep Weiland received both the M.Sc. (1986) and Ph.D. degrees in mathematics from the University of Groningen, Groningen, The Netherlands.

He is Associate Professor at the Control Systems Group, Dept. of Electrical Engineering, Eindhoven University of Technology. He was a Postdoctoral Research Associate at the Department of Electrical Engineering and Computer Engineering, Rice University, Houston, TX, from 1991 to 1992. Since 1992, he has been affiliated with Eindhoven University of Technology. His research interests are the general theory of systems and control, robust control, model approximation, modeling and control of hybrid systems, identification, and model predictive control.

Dr. Weiland was an Associate Editor of the IEEE TRANSACTIONS ON AUTOMATIC CONTROL from 1995 to 1999, of the *European Journal of Control* from 1999 to 2003, of the *International Journal of Robust and Nonlinear Control* from 2001 to 2004, and he is currently an Associate Editor for *Automatica*.

Maurice Heemels was born in St. Odiliënberg, The Netherlands, in 1972. He received the M.Sc. degree (with honors) from the Department of Mathematics and the Ph.D. degree (*cum laude*) from the Department of Electrical Engineering of the Technische Universiteit Eindhoven, Eindhoven, The Netherlands, in 1995 and 1999, respectively.

From 2000 to 2004, he was an Assistant Professor in the Control Systems Group of the Department of Electrical Engineering of the Technische Universiteit Eindhoven, Eindhoven, The Netherlands. In June 2004, he moved to the Embedded Systems Institute in Eindhoven, where he is working as a Research Fellow. He spent three months as a Visiting Professor at the ETH in Zürich, Switzerland, in 2001, and a same period with Océ, Venlo, The Netherlands, in 2004. His research interests include modeling, analysis, and control of hybrid and nonsmooth systems and their applications to industrial design problems for embedded systems.

Dr. Heemels was awarded the ASML Prize for the Best Ph.D. Thesis of the Technische Universiteit Eindhoven in 1999–2000 in the area of fundamental research.