Available online at www.sciencedirect.com
SCIENCE DIRECT®
automatica
Automatica 41 (2005) 905–913
www.elsevier.com/locate/automatica

Brief paper

Identification of piecewise affine systems based on statistical clustering technique☆

Hayato Nakada*, Kiyotsugu Takaba, Tohru Katayama

Department of Applied Mathematics and Physics, Graduate School of Informatics, Kyoto University, Yoshida-Honmachi, Sakyo-ku, Kyoto 606-8501, Japan

Received 30 November 2003; received in revised form 19 November 2004; accepted 13 December 2004

Abstract

This paper is concerned with the identification of a class of piecewise affine systems called a piecewise affine autoregressive exogenous (PWARX) model. The PWARX model is composed of ARX sub-models each of which corresponds to a polyhedral region of the regression space. Under the temporary assumption that the number of sub-models is known a priori, the input–output data are collected into several clusters by using a statistical clustering algorithm. We utilize support vector classifiers to estimate the boundary hyperplane between two adjacent regions in the regression space. In each cluster, the parameter vector of the sub-model is obtained by the least squares method. It turns out that the present statistical clustering approach enables us to estimate the number of sub-models based on the information criteria such as CAIC and MDL. The estimate of the number of sub-models is performed by applying the identification procedure several times to the same data set, after having fixed the number of sub-models to different values. Finally, we verify the applicability of the present identification method through a numerical example of a Hammerstein model.

Keywords: Piecewise affine autoregressive exogenous model; Identification; Statistical clustering; Support vector classifier; Number of sub-models

1. Introduction

Hybrid systems are composed of both continuous dynamics governed by physical laws and discrete-event dynamics driven by logic and rules. Recently, much attention has been paid to hybrid systems from various viewpoints (van der Schaft & Schumacher, 2000).

This paper deals with a system representation of hybrid systems called a piecewise affine (PWA) system. A number of research works on PWA systems (Imura & van der Schaft, 2000; Johansson, 2003; Johansson & Rantzer, 1998; Nakada & Takaba, 2003) have been reported. Recently, it has been proved by Bemporad, Ferrari-Trecate, and Morari (2000) and by Heemels, De Schutter, and Bemporad (2001) that the PWA system is equivalent to the other hybrid system models such as mixed logical dynamical systems, linear complementarity systems, extended linear complementarity systems and max–min-plus-scaling systems.

The identification of PWA systems is quite important, because there are many linear systems with PWA nonlinearities such as saturation or relay elements, and a general nonlinear system can be treated as a PWA system by approximating a nonlinear function by a PWA one with arbitrary accuracy. There are some applications to the identification of real systems, e.g. a nonlinear electrical circuit (Ferrari-Trecate, Muselli, Liberati, & Morari, 2003), a fermentation process (Fantuzzi, Simani, Beghelli, & Rovatti, 2002) and a pick-and-place machine (Juloski, Heemels, & Ferrari-Trecate, 2004).

☆ This paper was presented at IFAC Workshop on Adaptation and Learning in Control and Signal Processing (ALCOSP 04) and IFAC Workshop on Periodic Control Systems (PSYCO 04), Yokohama, Japan, August, 2004. This paper was recommended for publication in revised form by Associate Editor Antonio Vicino under the direction of Editor T. Söderström.

* Corresponding author. Tel.: +81 75 753 9104; fax: +81 75 753 5507.
E-mail addresses: nakada@amp.i.kyoto-u.ac.jp (H. Nakada), takaba@amp.i.kyoto-u.ac.jp (K. Takaba), katayama@amp.i.kyoto-u.ac.jp (T. Katayama).

0005-1098/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.automatica.2004.12.005

In the identification of PWA systems, a piecewise affine autoregressive exogenous (PWARX) model (Amaldi & Mattavelli, 2002; Bemporad, Garulli, Paoletti, & Vicino, 2003; Ferrari-Trecate et al., 2003; Ragot, Mourot, & Maquin, 2003; Roll, Bemporad, & Ljung, 2004; Vidal, Soatto, & Sastry, 2003) is used as a typical model of PWA systems. The identification based on the PWARX model includes the estimation of both polyhedral partition on the regression space and the parameter of the ARX sub-model corresponding to each polyhedral region. In the case where the partition of the regression space is known a priori, the parameters of ARX sub-models can be estimated almost straightforwardly by the least squares method. However, the necessity of estimating the partition of the regression space as well as system parameters in the identification of PWA systems makes the development of identification methods extremely difficult. Jordan and Jacobs (1994) proposed an EM algorithm for hierarchical models that could be exploited for identifying PWARX models. Amaldi and Mattavelli (2002) considered a combinatorial optimization problem called the MIN PFS problem to estimate a piecewise linear model, which was approximately solved by a greedy method. Their result was applied by Bemporad et al. (2003) to the identification of the PWARX models with estimation of the number of sub-models. Ferrari-Trecate et al. (2003) developed an identification method by clustering parameter vectors, each of which is locally estimated from several data neighboring to each data. The optimality of the data classification utilized in this method was shown by Ferrari-Trecate and Schinkel (2003). Recently, Ragot et al. (2003) have proposed a method for identifying the parameter of sub-models when choosing an adapted weighting function, which allows one to select the data for which each sub-model is active. Vidal et al. (2003) have taken an algebraic geometric approach, in which the problem of estimating both the parameters and the number of sub-models is casted as a polynomial factorization problem. Moreover, Roll et al. (2004) have reduced the identification problem of a special class of PWARX models to a mixed-integer programming problem.

In this paper, we present a new method for the identification of a PWARX model based on a statistical clustering of measured data via a Gaussian mixture model and support vector classifiers (SVCs). A major advantage of the statistical clustering technique is that the statistical information such as the log-likelihood function enables us to estimate the number of sub-models of a PWARX model. We show how to estimate the number of sub-models based on the statistical information criteria such as the consistent Akaike's information criterion (CAIC) (Bozdogan, 1987), and the minimum description length (MDL) criterion (Rissanen, 1978), after discussing the case where the number of sub-models is available in the identification procedure. We also verify the applicability of the present identification method to a Hammerstein model, which is a popular model of nonlinear systems and composed of a linear system with a static nonlinearity. If the nonlinearity is a PWA function, the present method can be easily applied to Hammerstein models.

The organization of this paper is as follows. In Section 2, we formulate the identification problem of a PWARX model. The measured data are classified into clusters based on a Gaussian mixture model in Section 3. Sections 4 and 5 are, respectively, devoted to estimating boundary hyperplanes between two adjacent polyhedral regions on the regression space and the parameters of local sub-models. In Section 6, the two criteria CAIC and MDL are introduced to estimate the number of sub-models. We apply the present identification method to a Hammerstein model in Section 7. Finally, we conclude this paper in Section 8.

2. Piecewise affine autoregressive exogenous (PWARX) model

In this paper, we consider the identification problem of a PWA system. We introduce a useful model called a PWARX model (Bemporad et al., 2003; Ferrari-Trecate et al., 2003; Roll et al., 2004) for the identification of a PWA system.

A PWARX model is given by

$$y_k = \begin{cases} \theta_1^{\mathrm{T}} \begin{bmatrix} x_k \\ 1 \end{bmatrix} + e_k, & \text{if } x_k \in \mathscr{X}_1, \\ \vdots & k = 1, 2, \ldots, N. \\ \theta_s^{\mathrm{T}} \begin{bmatrix} x_k \\ 1 \end{bmatrix} + e_k, & \text{if } x_k \in \mathscr{X}_s, \end{cases} \quad (1)$$

The vectors $u_k \in \mathbf{R}^m$, $y_k \in \mathbf{R}^p$ and $e_k \in \mathbf{R}^p$ are the input, the output and the noise at time k, respectively. The regression vector $x_k \in \mathbf{R}^n$ is denoted by

$$x_k = [y_{k-1}^{\mathrm{T}} \; y_{k-2}^{\mathrm{T}} \; \cdots \; y_{k-n_y}^{\mathrm{T}} \; u_{k-1}^{\mathrm{T}} \; u_{k-2}^{\mathrm{T}} \; \cdots \; u_{k-n_u}^{\mathrm{T}}]^{\mathrm{T}},$$

where $n = pn_y + mn_u$ with non-negative integers n_y and n_u. Let $\mathscr{X} \subseteq \mathbf{R}^n$ be the regression space, and $\mathscr{X}_i$, $i = 1, 2, \ldots, s$ represent convex polyhedral subsets of $\mathscr{X}$.[1] The number of sub-models is denoted by s, and $\theta_i \in \mathbf{R}^{(n+1)\times p}$, $i = 1, 2, \ldots, s$ are parameter matrices to be estimated. We assume that N data samples

$$z_k = \begin{bmatrix} x_k \\ y_k \end{bmatrix} \in \mathbf{R}^{n+p}, \quad k = 1, 2, \ldots, N$$

are generated by the PWARX model (1). We also assume that the noise e_k is uncorrelated with the regression vector x_k.

Example 1. Consider the following single-input–single-output (SISO) PWARX model taken form Bemporad et al. (2003).

$$y_k = [-0.4 \;\; 1 \;\; 1.5]\bar{x}_k + e_k, \quad \text{if } x_k \in \mathscr{X}_1 = \{x : \; [4 \;\; -1 \;\; 10]\bar{x} < 0\},$$

[1] Each polyhedron $\mathscr{X}_i$ is assumed to satisfy $\mathscr{X}_i \neq \emptyset \; \forall i \in \{1, 2, \ldots, s\}$, $\mathscr{X}_i \cap \mathscr{X}_j = \emptyset \; \forall i, j \in \{1, 2, \ldots, s\}$, $i \neq j$ and $\bigcup_{i=1}^{s} \mathscr{X}_i = \mathscr{X}$.

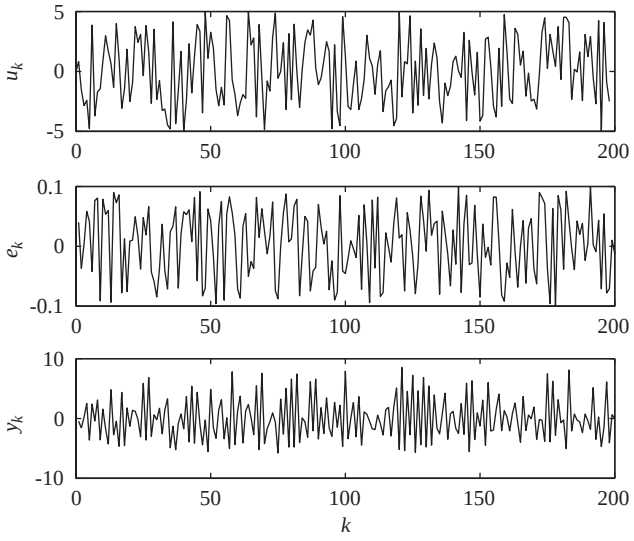

Fig. 1. Time series of input u_k, noise e_k and output y_k of Example 1.

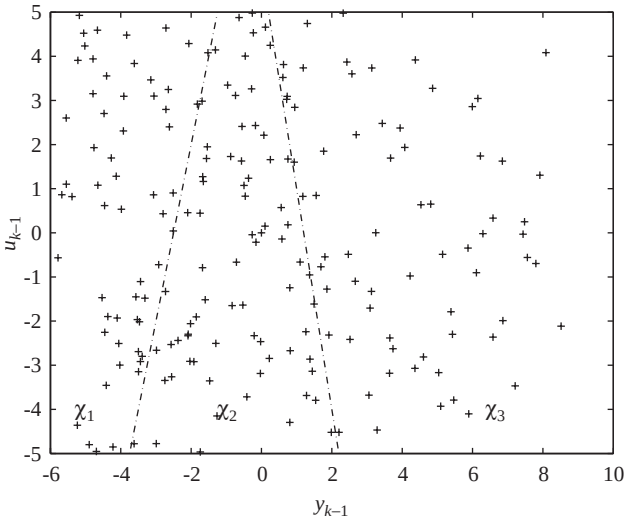

Fig. 2. Data of Example 1 in the regression space $\mathbf{R}^n$.

$$y_k = [0.5 \quad -1 \quad -0.5]\bar{x}_k + e_k,$$
$$\text{if } x_k \in \mathscr{X}_2 = \left\{ x : \begin{bmatrix} -4 & 1 & -10 \\ 5 & 1 & -6 \end{bmatrix} \bar{x} \leqslant 0 \right\},$$
$$y_k = [-0.3 \;\; 0.5 \;\; -1.7]\bar{x}_k + e_k,$$
$$\text{if } x_k \in \mathscr{X}_3 = \{x : \; [-5 \;\; -1 \;\; 6]\bar{x} < 0\},$$

where $\bar{x}_k = [x_k^{\mathrm{T}} \;\; 1]^{\mathrm{T}}$, $x_k = [y_{k-1} \;\; u_{k-1}]^{\mathrm{T}}$, and in this case $s = 3$, $n_y = 1$, $n_u = 1$, $n = 2$, $m = 1$ and $p = 1$. It is easy to see that the boundary hyperplanes $\mathscr{H}_{12}$ between $\mathscr{X}_1$ and $\mathscr{X}_2$, and $\mathscr{H}_{23}$ between $\mathscr{X}_2$ and $\mathscr{X}_3$ are described as

$$\mathscr{H}_{12} = \{x : \; h_{12}^{\mathrm{T}}\bar{x} = 0\}, \quad h_{12} = [0.4 \;\; -0.1 \;\; 1]^{\mathrm{T}},$$
$$\mathscr{H}_{23} = \{x : \; h_{23}^{\mathrm{T}}\bar{x} = 0\}, \quad h_{23} = [-0.8333 \;\; -0.1667 \;\; 1]^{\mathrm{T}},$$

respectively. The input u_k and the noise e_k are white noises generated from uniform distributions over the intervals $[-5, 5]$ and $[-0.1, 0.1]$, respectively. Fig. 1 shows an example of time series of u_k, e_k and y_k with $N = 200$. The data in the regression space $\mathbf{R}^2$ are depicted in Fig. 2. The dash–dotted lines represent the boundary hyperplanes $\mathscr{H}_{12}$ and $\mathscr{H}_{23}$.

For the time being, we make the following temporary assumption.

Assumption 1. The number of sub-models s is given a priori.

In some cases, we are able to know the number of sub-models s in advance, so the above assumption is reasonable. For more practical situations where s cannot be obtained a priori, we need to estimate it. The estimation of the number of sub-models will be discussed in Section 6 based on the statistical information criteria.

In addition, it is assumed that the dynamics of all the sub-models are sufficiently excited by the input sequence u_k, $k = -n_u + 1, -n_u + 2, \ldots, N - 1$ in order to estimate the parameters of all sub-models.

We state the identification problem considered in this paper.

Problem 1. Estimate the parameter matrices θ_i, $i = 1, 2, \ldots, s$, and the boundary hyperplanes between all pairs of two adjacent[2] polyhedral regions $\mathscr{X}_i$, $i = 1, 2, \ldots, s$ from the measured data z_k, $k = 1, 2, \ldots, N$ generated by (1) under Assumption 1.

A general framework for the identification of a PWA system is summarized as follows (Bemporad et al., 2003; Ferrari-Trecate et al., 2003):

Phase 1: *Clustering of the measured data.* In this phase, the measured (input–output) data are classified into several clusters by using a data clustering technique. Among many clustering techniques (Miyamoto, 1999), the *K*-means method is employed by Ferrari-Trecate et al. (2003) for the PWA system identification. Bemporad et al. (2003) also applied a greedy method to the identification problem without a priori assumption on the number of sub-models. As an alternative to those previous approach, we will consider the statistical clustering approach to the PWA system identification.

Phase 2: *Estimation of the boundary hyperplanes on the regression space.* Since the polyhedral regions on the regression space are characterized by boundary hyperplanes between adjacent clusters, we need to estimate these hyperplanes. The well-known SVCs (Boyd & Vandenberghe, 2003; Vapnik, 1998) are typically used for this purpose.

[2] Two polyhedral regions $\mathscr{X}_i$ and $\mathscr{X}_j$ $(i, j \in \{1, 2, \ldots, s\}, \; i \neq j)$ are said to be *adjacent* if their closures share an $(n-1)$-dimensional boundary hyperplane.

Phase 3: *Parameter estimation of each sub-model.* Once the measured data are correctly classified, we can apply the standard least squares (LS) method (Ljung, 1999; Söderström & Stoica, 1989) to the parameter estimation of each sub-model (Bemporad et al., 2003; Ferrari-Trecate et al., 2003; Ragot et al., 2003).

3. Clustering of the measured data

We wish to classify the data $z_k,\ k=1,2,\ldots,N$ in $\mathbf{R}^{n+p}$ into s clusters. For this purpose, we classify the set of the time indices $\{1,2,\ldots,N\}$ into s non-empty disjoint clusters $\mathscr{C}_i,\ i=1,2,\ldots,s$. In this section, we employ a statistical clustering method based on *a Gaussian mixture model* (Alpaydın, 1998; Dempster, Laird, & Rubin, 1977; Jordan & Jacobs, 1994; Mitra, Pal, & Siddiqi, 2003; Miyamoto, 1999; Redner & Walker, 1984).

We assume that the probability density of the data z_k is given by a Gaussian mixture model

$$p(z;\Phi)=\sum_{i=1}^{s}\alpha_i\,p_i(z;\mu_i,\Sigma_i),$$

where $\Phi := (\alpha,\mu,\Sigma)$ with scalar parameters $\alpha := (\alpha_1,\alpha_2,\ldots,\alpha_s)$ satisfying $\sum_{i=1}^{s}\alpha_i=1$, $(n+p)$-dimensional mean vectors $\mu:=(\mu_1,\mu_2,\ldots,\mu_s)$ and $(n+p)\times(n+p)$-dimensional covariance matrices $\Sigma := (\Sigma_1,\Sigma_2,\ldots,\Sigma_s)$. The function $p_i(z;\mu_i,\Sigma_i)$ is a multivariate Gaussian density given by

$$p_i(z;\mu_i,\Sigma_i)=\frac{1}{(2\pi)^{(n+p)/2}[\det(\Sigma_i)]^{1/2}}\times\exp\left\{-\frac{1}{2}(z-\mu_i)^{\mathrm{T}}\Sigma_i^{-1}(z-\mu_i)\right\},\quad i=1,2,\ldots,s.$$

Once we obtain a good parameter Φ that describes the probability density of the data accurately, the probability that the index k is classified into $\mathscr{C}_i$ is

$$P(k\in\mathscr{C}_i)=\frac{\alpha_i\,p_i(z_k;\mu_i,\Sigma_i)}{p(z_k;\Phi)}.$$

This probability serves as a criterion for the data clustering. We can use this probability as a criterion for clustering. For example, if we wish to carry out the clustering in a deterministic way, each index k is classified into $\mathscr{C}_i$ for which $P(k\in\mathscr{C}_i)$ is maximal over $i=1,2,\ldots,s$ (Miyamoto, 1999).

In the remainder of this section, we will show how to find the best parameter $\Phi=(\alpha,\mu,\Sigma)$ based on the maximum-likelihood (ML) estimation. From the data $z_k,\ k=1,2,\ldots,N$, we wish to find the parameter $\Phi=(\alpha,\mu,\Sigma)$ that attains the maximum of the log-likelihood function

$$L(\Phi)=\sum_{k=1}^{N}\ln p(z_k;\Phi)=\sum_{k=1}^{N}\ln\left(\sum_{i=1}^{s}\alpha_i\,p_i(z_k;\mu_i,\Sigma_i)\right)\tag{2}$$

so that the mixture model fits the data as good as possible.

We get the (possibly local) maximum by iteratively updating Φ via the well-known expectation–maximization (EM) algorithm (Alpaydın, 1998; Dempster et al., 1977; Jordan & Jacobs, 1994; Mitra et al., 2003; Miyamoto, 1999; Redner & Walker, 1984). The EM algorithm is basically composed of two steps: the Expectation step (E-step) and the Maximization step (M-step). The M-step involves the maximization of the log-likelihood function that is redefined at the E-step of each iteration.

Algorithm 1 (*EM algorithm*). *Step* 1: *Set the initial values* $\Phi^{(0)}=(\alpha^{(0)},\mu^{(0)},\Sigma^{(0)})$, *and set the iteration counter* l *to* $l:=0$.

Step 2: *For* $\Phi^{(l)}=(\alpha^{(l)},\mu^{(l)},\Sigma^{(l)})$, *execute the following procedures*:

(E-step): *Compute*

$$\psi_{ik}^{(l)}=\frac{\alpha_i^{(l)}\,p_i(z_k;\mu_i^{(l)},\Sigma_i^{(l)})}{p(z_k;\Phi^{(l)})},\quad k=1,2,\ldots,N,\quad i=1,2,\ldots,s,$$

$$\Psi_i^{(l)}=\sum_{k=1}^{N}\psi_{ik}^{(l)},\quad i=1,2,\ldots,s.$$

(M-step): *Update* $\Phi^{(l)}=(\alpha^{(l)},\mu^{(l)},\Sigma^{(l)})$ *by*

$$\alpha_i^{(l+1)}:=\frac{\Psi_i^{(l)}}{N},\quad i=1,2,\ldots,s,$$

$$\mu_i^{(l+1)}:=\frac{1}{\Psi_i^{(l)}}\sum_{k=1}^{N}\psi_{ik}^{(l)}z_k,\quad i=1,2,\ldots,s,$$

$$\Sigma_i^{(l+1)}:=\frac{1}{\Psi_i^{(l)}}\sum_{k=1}^{N}\psi_{ik}^{(l)}(z_k-\mu_i^{(l+1)})(z_k-\mu_i^{(l+1)})^{\mathrm{T}},\quad i=1,2,\ldots,s.$$

Step 3: *If a prescribed convergence condition such as*

$$\max_{i=1,2,\ldots,s}\left\{\frac{\|\alpha_i^{(l+1)}-\alpha_i^{(l)}\|}{\|\alpha_i^{(l)}\|},\frac{\|\mu_i^{(l+1)}-\mu_i^{(l)}\|}{\|\mu_i^{(l)}\|},\frac{\|\Sigma_i^{(l+1)}-\Sigma_i^{(l)}\|}{\|\Sigma_i^{(l)}\|};\right\}\leqslant\varepsilon,\quad \varepsilon\ll 1\tag{3}$$

is satisfied, then set $l^*:=l+1$ *and exit. The optimal ML estimate of* Φ *is obtained by* $\Phi^*=\Phi^{(l^*)}$. *Otherwise, set* $l:=l+1$ *and go back to Step* 2.

Note that we can improve the maximum by starting the algorithm from several initial parameters $\Phi^{(0)}$.

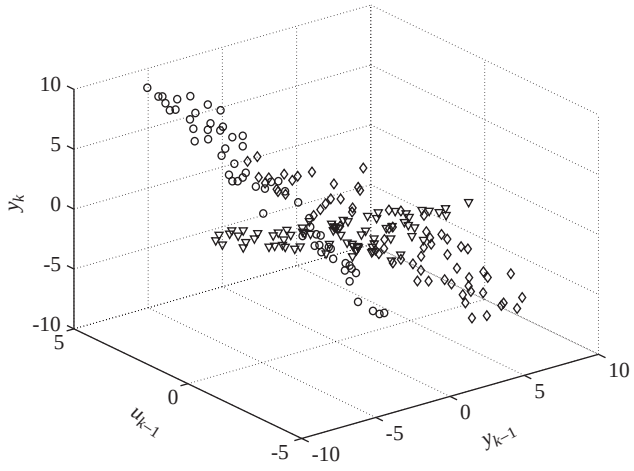

Fig. 3. Result of clustering in $\mathbf{R}^{n+p}$ for Example 1.

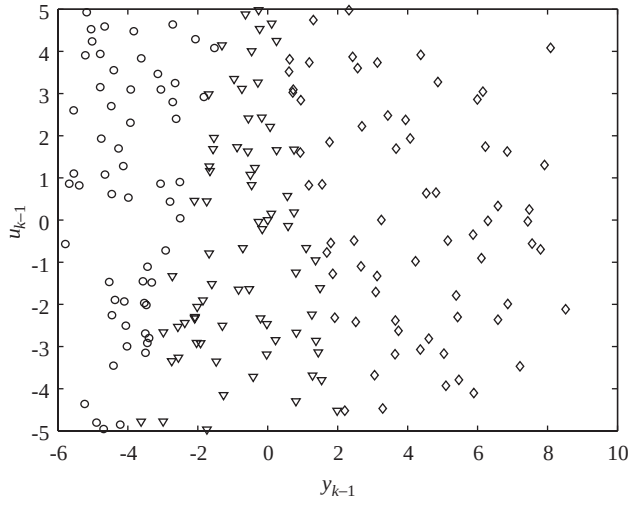

Fig. 4. Projection of clusters onto the regression space $\mathbf{R}^n$ for Example 1.

We apply the EM algorithm to the data in Example 1. We fix the weighting parameters to $\alpha_i^{(0)} = 1/s = 1/3,\ i = 1, 2, 3$, and the initial covariance matrices to $\Sigma_i^{(0)} = 10 I_3$. As for the initial mean vectors $\mu_i^{(0)},\ i = 1, 2, 3$, we set three data arbitrarily chosen from $z_k,\ k = 1, 2, \ldots, N$. The convergence tolerance in (3) is specified by $\varepsilon = 1 \times 10^{-2}$. We compute the local maximum $L(\Phi)$ via the multi-starting technique with 20 initial mean vectors. The results of the data clustering in this example are shown in Figs. 3 and 4, where the data in the same cluster are plotted with the same marks.

4. Estimation of the boundary hyperplanes on the regression space

We describe a method for classifying two adjacent clusters in the regression space $\mathbf{R}^n$ with a hyperplane based on SVCs (Vapnik, 1998; Boyd & Vandenberghe, 2003). The simplest version of SVC is used by Bemporad et al. (2003), while the similar technique is also utilized by Ferrari-Trecate et al. (2003).

Before estimating the partition of the regression space, we need to check the adjacency of the clusters. We employ the so-called Delaunay graph (Edelsbrunner, 1987) for this purpose. This graph describes the adjacency of the Voronoi cells associated with the data. If there exists at least one branch between two regression vectors in different two clusters, then these clusters are regarded as adjacent in the regression space. It should be noted that information on the adjacency of clusters obtained from the Delaunay graph enables us to reduce the number of estimation of boundary hyperplanes, compared to a naive approach where all hyperplanes between all pairs of different clusters are estimated.

The linear separability of the data in two adjacent clusters is not guaranteed in the statistical clustering method described in Section 3 because the observation data are disturbed by the measurement noise e_k, and because the data is classified in the data space $\mathbf{R}^{n+p}$ of higher dimension than that of the regression space $\mathbf{R}^n$. Hence, there is a possibility that some data cannot be linearly classified by a boundary hyperplane, which leads to a misclassified error. We employ the so-called *soft margin SVCs* (Vapnik, 1998; Boyd & Vandenberghe, 2003) in order to overcome this difficulty.

We wish to classify two adjacent clusters $\mathscr{C}_i$ and $\mathscr{C}_j$ by a separating hyperplane $\mathscr{H}_{ij} = \{x :\ a_{ij}^{\mathrm{T}} x + b_{ij} = 0\}$, where $a_{ij} \in \mathbf{R}^n$ and $b_{ij} \in \mathbf{R}$. For this purpose, we solve a quadratic programming (QP) problem

$$\begin{aligned} &\text{minimize } \|a_{ij}\|^2 + \gamma \sum_{h \in \{i,j\},\ k \in \mathscr{C}_h} v_{hk} \\ &\text{subject to } a_{ij}^{\mathrm{T}} x_k + b_{ij} \geqslant 1 - v_{ik},\ v_{ik} \geqslant 0,\ k \in \mathscr{C}_i, \\ &\qquad\qquad a_{ij}^{\mathrm{T}} x_k + b_{ij} \leqslant -(1 - v_{jk}),\ v_{jk} \geqslant 0,\ k \in \mathscr{C}_j. \quad (4) \end{aligned}$$

Here, we think of $\sum_{h \in \{i,j\},\ k \in \mathscr{C}_h} v_{hk}$ as a measure of degree of misclassification, which is desired to be minimized. The quantity $\|a_{ij}\|$ in the objective function represents the inverse of the margin between the two clusters. The scalar $\gamma > 0$ is a prescribed constant.

We solve the QP problem (4) for Example 1. Then, we obtain

$$\begin{aligned} \hat{h}_{12} &= [0.4279 \quad -0.1062 \quad 1]^{\mathrm{T}}, \\ \hat{h}_{23} &= [-0.8297 \quad -0.1645 \quad 1]^{\mathrm{T}} \end{aligned}$$

with the trade-off parameter $\gamma = 100$, where the third elements of both $\hat{h}_{12}$ and $\hat{h}_{23}$ are normalized. Fig. 5 shows the estimated boundary hyperplanes (solid lines) and the true ones (dash–dotted lines).

This section has concentrated on the case where only two clusters are adjacent. In this case, it is not guaranteed that the partition produced by boundary hyperplanes all pairs of different clusters does not leave *holes* in the regression space, as pointed out by Ferrari-Trecate et al. (2003). One remedy for leaving such a hole is to use the multi-category

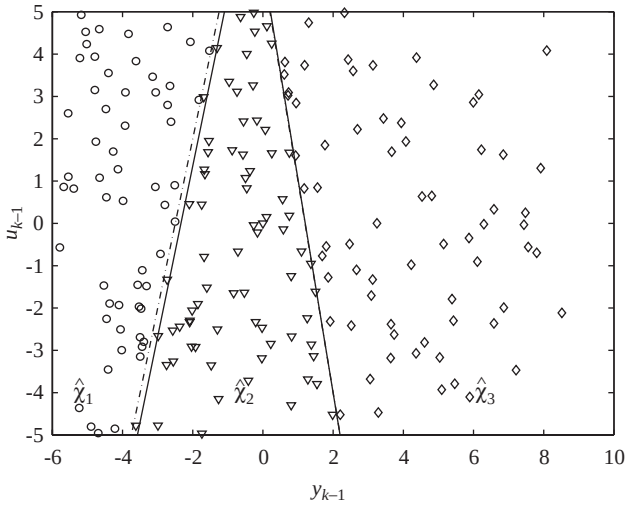

Fig. 5. Estimation of boundary hyperplanes for Example 1.

support vector machine (Bredensteiner & Bennett, 1999). This method provides the estimates of boundary hyperplanes in the case where more than two clusters are mutually adjacent.

5. Parameter estimation of each sub-model

Based on the data classified in the previous section, we estimate the parameter θ_i, $i=1,2,\ldots,s$ by the LS method (Söderström & Stoica, 1989; Ljung, 1999). For each $i=1,2,\ldots,s$, the parameter can be estimated by the formula

$$\begin{aligned}
&\hat{\theta}_i = (X_i^{\mathrm{T}} X_i)^{-1} X_i^{\mathrm{T}} Y_i,\\
&X_i = [\bar{x}_{k_{i1}} \;\; \bar{x}_{k_{i2}} \;\; \cdots \;\; \bar{x}_{k_{iN_i}}]^{\mathrm{T}},\\
&\bar{x}_{k_{il}} = \begin{bmatrix} x_{k_{il}} \\ 1 \end{bmatrix}, \quad l=1,2,\ldots,N_i,\\
&Y_i = [y_{k_{i1}} \;\; y_{k_{i2}} \;\; \cdots \;\; y_{k_{iN_i}}]^{\mathrm{T}},\\
&\mathscr{C}_i = \{k_{i1}, k_{i2}, \ldots, k_{iN_i}\}. \qquad (5)
\end{aligned}$$

Here, the quantity N_i denotes the cardinality of $\mathscr{C}_i$ which is assumed to satisfy $N_i \geqslant n+1$ so that we can estimate θ_i. Obviously, $\sum_{i=1}^{s} N_i = N$ holds.

We compute the parameter estimates $\hat{\theta}_i$, $i=1,2,3$ for Example 1. From (5), we obtain[3]

$$\begin{aligned}
&\hat{\theta}_1 = [-0.3994 \;\; 1.0010 \;\; 1.5147]^{\mathrm{T}},\\
&\hat{\theta}_2 = [0.5025 \;\; -0.9991 \;\; -0.4858]^{\mathrm{T}},\\
&\hat{\theta}_3 = [-0.2992 \;\; 0.4979 \;\; -1.7126]^{\mathrm{T}}.
\end{aligned}$$

Moreover, we repeat the estimation for 50 different sets of data in order to examine the statistical characteristic of the present method. The input u_k and the noise e_k are distributed uniformly in the intervals $[-5,5]$ and $[-0.1,0.1]$ with $N=200$, respectively. In Table 1, we show the mean and the standard deviation of the 50 estimates obtained. We see from this table that it is possible to estimate the parameter vectors and the boundary hyperplanes with small standard deviations.

[3] The total computation time required to get the estimates $\hat{h}_{12}$, $\hat{h}_{23}$ and $\hat{\theta}_i$, $i=1,2,3$ is 34.22 s on a computer with a 3.4 GB Intel Pentium 4 processor, a 2.0 GB memory and Matlab 6.0, which contains the time for the data clustering shown in Section 3.

Table 1
Estimation result for Example 1 (50 times)

	True	Mean	Standard deviation
h_{12}	$\begin{bmatrix} 0.4 \\ -0.1 \\ 1 \end{bmatrix}$	$\begin{bmatrix} 0.4059 \\ -0.0942 \\ 1 \end{bmatrix}$	$\begin{bmatrix} 0.0215 \\ 0.0155 \\ 0 \end{bmatrix}$
h_{23}	$\begin{bmatrix} -0.8333 \\ -0.1667 \\ 1 \end{bmatrix}$	$\begin{bmatrix} -0.8262 \\ -0.1679 \\ 1 \end{bmatrix}$	$\begin{bmatrix} 0.0660 \\ 0.0260 \\ 0 \end{bmatrix}$
θ_1	$\begin{bmatrix} -0.4 \\ 1 \\ 1.5 \end{bmatrix}$	$\begin{bmatrix} -0.3993 \\ 0.9997 \\ 1.5047 \end{bmatrix}$	$\begin{bmatrix} 0.0675 \\ 0.0028 \\ 0.0326 \end{bmatrix}$
θ_2	$\begin{bmatrix} 0.5 \\ -1 \\ -0.5 \end{bmatrix}$	$\begin{bmatrix} 0.4985 \\ -1.0001 \\ -0.4991 \end{bmatrix}$	$\begin{bmatrix} 0.0061 \\ 0.0023 \\ 0.0090 \end{bmatrix}$
θ_3	$\begin{bmatrix} -0.3 \\ 0.5 \\ -1.7 \end{bmatrix}$	$\begin{bmatrix} -0.3000 \\ 0.5001 \\ -1.7012 \end{bmatrix}$	$\begin{bmatrix} 0.0042 \\ 0.0025 \\ 0.0172 \end{bmatrix}$

If the estimates of the parameters $\hat{\theta}_i$ and $\hat{\theta}_j$ $(i \neq j)$, respectively, corresponding to two adjacent estimated polyhedra $\hat{\mathscr{X}}_i$ and $\hat{\mathscr{X}}_j$ are close to each other and if $\hat{\mathscr{X}}_i \cup \hat{\mathscr{X}}_j$ is connected and convex, then these polyhedra can be merged into a new polyhedron. In this case, we can obtain a refined parameter estimate in the new polyhedron by applying the LS method again to the corresponding new cluster.

It might be noted that the weighted LS method can be used for the parameter estimation, where one perhaps could use $\psi_{ik}^{(l^*)}$ in the EM algorithm as weight, to cope with the misclassified data. For Example 1, the weighted LS estimates are almost the same as the LS estimates, since the data are attributed to the correct modes. The usage of the weighted LS method might reduce the consequences of the possible misclassification for more complicated systems.

6. Estimation of the number of sub-models

In this section, we discuss how to estimate the number of sub-models. Since the number of clusters is equal to that of sub-models, we show how to estimate the number of clusters based on the information criteria associated with the maximum-likelihood estimation in Section 3 (Hu & Xu, 2003; Jain, Dulin, & Mao, 2000).

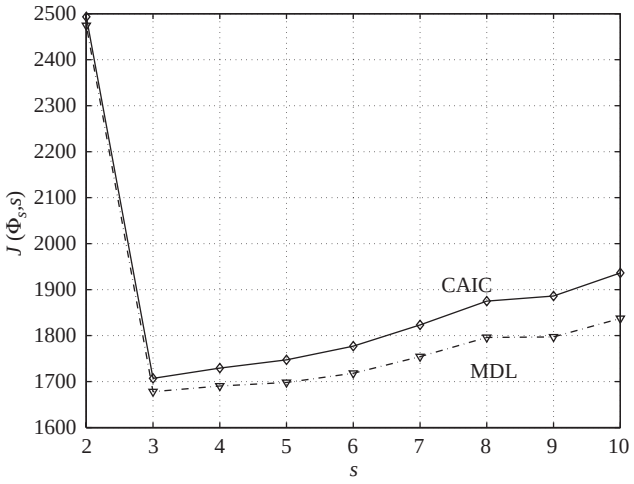

Fig. 6. Estimation of the number of sub-models s based on CAIC and MDL for Example 1.

Firstly, we fix two positive integers $s_{\min}$ and $s_{\max}$ so that the true number of sub-models s is assumed to be in the interval $[s_{\min}, s_{\max}]$. Next, for all $s = s_{\min}, \ldots, s_{\max}$, we compute the parameter estimate Φ_s, where Φ_s denotes the estimate of Φ for a fixed s. Then, the estimate of s is given by

$$\hat{s} = \underset{s=s_{\min},\ldots,s_{\max}}{\operatorname{argmin}} \; J(\Phi_s, s),$$

where $J(\Phi_s, s)$ is a criterion specified below. Among existing information criteria for model selection (Hu & Xu, 2003; Jain et al., 2000), we employ the consistent Akaike's information criterion (CAIC) (Bozdogan, 1987) and the MDL criterion (Rissanen, 1978). These criteria have the form

$$J(\Phi_s, s) = -2L(\Phi_s) + A(N)D(s), \tag{6}$$

where $L(\Phi_s)$ is the log-likelihood function of Φ_s defined by (2), and the second term penalizes the numbers of data and clusters. In (6), $D(s)$ represents the number of independent parameters in Φ_s, so that in this case,

$$D(s) = (s-1) + s(n+p) + \tfrac{1}{2}s(n+p)(n+p+1),$$

while $A(N)$, a function of the number of samples N, is, respectively, given by

$$A(N) = \begin{cases} \ln N + 1 & \text{(CAIC)}, \\ \ln N & \text{(MDL)}. \end{cases}$$

There is a possibility that these criteria might overestimate the number of clusters. In such a case, we can reduce the overestimated number of clusters by merging two adjacent clusters $\mathscr{C}_i$ and $\mathscr{C}_j$ $(i \neq j)$ if the corresponding parameter estimates $\hat{\theta}_i$ and $\hat{\theta}_j$ computed by (5) are close to each other.

For Example 1, we set $(s_{\min}, s_{\max}) = (2, 10)$. Fig. 6 shows the values of $J(\Phi_s, s)$ for CAIC and MDL. We see from the figure that we obtain the correct estimate $\hat{s} = 3$ from these criteria.

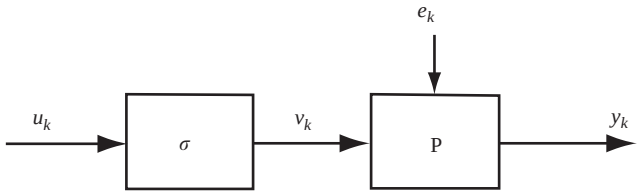

Fig. 7. A Hammerstein model with a saturation nonlinearity.

Finally, we repeat 50 simulation runs with $(s_{\min}, s_{\max}) = (2, 10)$. The result is shown in Table 2, where MDL sometimes overestimates the number of sub-models, but CAIC always returns the correct estimation $\hat{s} = 3$ for this example.

7. Application to Hammerstein model

A Hammerstein model in Fig. 7 is a popular model used in the nonlinear system identification, which is composed of a linear time-invariant (LTI) system and a static nonlinearity. If the static nonlinearity is a static PWA function, we can easily apply the present identification method to the Hammerstein model as shown below.

Example 2. We consider an SISO Hammerstein model. The plant $\boldsymbol{P}$ is an LTI system given by

$$y_k = -a_1 y_{k-1} - a_2 y_{k-2} + b_1 v_{k-1} + e_k,$$

where a_1, a_2 and b_1 are scalar constants, and σ is a saturation function described as

$$v_k = \sigma(u_k) = \begin{cases} u_{\max}, & \text{if } u_k > u_{\max}, \\ u_k, & \text{if } u_{\min} \leqslant u_k \leqslant u_{\max}, \\ u_{\min}, & \text{if } u_k < u_{\min}, \end{cases}$$

where $u_{\max}$ and $u_{\min}$ are scalar constants giving the upper and lower saturation bounds, respectively. It is easy to verify that the Hammerstein model of Fig. 7 can be expressed as the following PWARX model:

$$y_k = [-a_1 \;\; -a_2 \;\; 0 \;\; b_1 u_{\max}]\bar{x}_k + e_k, \quad \text{if } x_k \in \mathscr{X}_1 = \{x : [0 \;\; 0 \;\; 1 \;\; -u_{\max}]\bar{x}_k > 0\},$$

$$y_k = [-a_1 \;\; -a_2 \;\; b_1 \;\; 0]\bar{x}_k + e_k, \quad \text{if } x_k \in \mathscr{X}_2 = \left\{x : \begin{bmatrix} 0 & 0 & -1 & u_{\max} \\ 0 & 0 & 1 & -u_{\min} \end{bmatrix}\bar{x}_k \geqslant 0\right\},$$

$$y_k = [-a_1 \;\; -a_2 \;\; 0 \;\; b_1 u_{\min}]\bar{x}_k + e_k, \quad \text{if } x_k \in \mathscr{X}_3 = \{x : [0 \;\; 0 \;\; -1 \;\; u_{\min}]\bar{x}_k > 0\},$$

where $\bar{x}_k = [x_k^{\mathrm{T}} \;\; 1]^{\mathrm{T}}$ and $x_k = [y_{k-1} \;\; y_{k-2} \;\; u_{k-1}]^{\mathrm{T}}$.

Note that, unlike Example 1, the dynamics of this Hammerstein model is continuous on the boundary hyperplanes between any pairs of adjacent regions.

We fix $a_1 = 0.5$, $a_2 = 0.1$, $b_1 = 1$, $u_{\max} = 2$ and $u_{\min} = -1$. The number of data is $N = 250$. The input u_k and the measurement noise e_k are normally distributed with means 0

Table 2
Estimation of the number of sub-models based on CAIC and MDL for Example 1 (true: $s = 3$)

$\hat{s}$	2	3	4	5	···	10	Total times
CAIC	0	50	0	0	···	0	50
MDL	0	42	8	0	···	0	50

both and variances 2^2 and $(0.2)^2$, respectively. Other parameters in the identification procedure are similarly fixed to Example 1.

Firstly, we estimate the number of sub-models s with $(s_{\min}, s_{\max}) = (2, 6)$. Then, the number of sub-models is estimated as $\hat{s} = 3$, where the minimal values of CAIC and MDL are 2910.6 and 2866.6, respectively.

Under $\hat{s} = 3$, the two boundary hyperplanes are estimated by

$$\hat{h}_{12} = [0.0084 \;\; 0.0128 \;\; -0.5032 \;\; 1]^{\mathrm{T}},$$
$$\hat{h}_{23} = [0.0060 \;\; -0.0372 \;\; 1.0050 \;\; 1]^{\mathrm{T}}$$

at $\gamma = 0.18$, where the third elements of both $\hat{h}_{12}$ and $\hat{h}_{23}$ are normalized. We also obtain[4] the estimated parameter vectors

$$\hat{\theta}_1 = [-0.4974 \;\; -0.1206 \;\; 0.0487 \;\; 1.8127]^{\mathrm{T}},$$
$$\hat{\theta}_2 = [-0.4584 \;\; -0.0462 \;\; 0.9519 \;\; -0.0023]^{\mathrm{T}},$$
$$\hat{\theta}_3 = [-0.5303 \;\; -0.1259 \;\; 0.0473 \;\; -0.8561]^{\mathrm{T}}.$$

This example indicates that the present identification method of a PWARX model is applicable to a Hammerstein model with PWA nonlinearities.

8. Conclusion

In this paper, we have developed a new identification method of PWARX models. More specifically, our method consists of the following three techniques: the statistical clustering based on a Gaussian mixture model with the EM algorithm, the estimation of the regression space partition via soft margin support vector classifiers, and the least squares estimation of the parameter of the ARX sub-models. The advantage of the present identification method is that the information criteria such as CAIC and MDL enable us to estimate the number of sub-models. We have also shown the applicability of the present identification method to a Hammerstein model through a numerical example.

Finally, one of future topics is to design the inputs u_k for the PWARX model identification. Juloski et al. (2004) have produced the uniformly distributed inputs for identification in an experimental setting by adjusting the bounds of the distribution. However, the theoretical characterization of a suitable input sequence for the PWARX model identification is still open. Moreover, other possible future topics are to use another mixture model such as latent variable models (Bishop & Tipping, 1998) in the data clustering, to apply the present method to more realistic systems, and to extend the present identification method to the identification of a PWARX model plus moving-average terms.

[4] The computation time is 161.92 seconds taken to obtain the above $\hat{h}_{12}$, $\hat{h}_{23}$ and $\hat{\theta}_i, i = 1, 2, 3$ on the same computational environment as Example 1, including the time for the data clustering under $\hat{s} = 3$.

Acknowledgements

The authors would like to thank Dr. Hideyuki Tanaka, Kyoto University, and the anonymous referees for their helpful advice on this work.

References

Alpaydın, E. (1998). Soft vector quantization and the EM algorithm. *Neural Networks*, *11*(3), 467–477.

Amaldi, E., & Mattavelli, M. (2002). The MIN PFS problem and piecewise linear model estimation. *Discrete Applied Mathematics*, *118*, 115–143.

Bemporad, A., Ferrari-Trecate, G., & Morari, M. (2000). Observability and controllability of piecewise affine and hybrid systems. *IEEE Transactions on Automatic Control*, *45*, 1864–1876.

Bemporad, A., Garulli, A., Paoletti, S., & Vicino, A. (2003). A greedy approach to identification of piecewise affine models. *Hybrid systems: computation and control* (HSCC 2003), Lecture notes in computer science, vol. 2623 (pp. 97–112). Berlin: Springer.

Bishop, C. M., & Tipping, M. E. (1998). A hierarchical latent variable model for data visualization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(3), 281–293.

Boyd, S., & Vandenberghe, L. (2003). *Convex optimization*. Cambridge: Cambridge University Press.

Bozdogan, H. (1987). Model selection and akaike's information criterion (AIC): the general theory and its analytical extensions. *Psychometrika*, *52*, 345–370.

Bredensteiner, E. J., & Bennett, K. P. (1999). Multicategory classification by support vector machines. *Computational Optimization and Applications*, *12*(1–3), 53–79.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, *39*, 1–37.

Edelsbrunner, H. (1987). *Algorithms in combinatorial geometry*. Springer-Verlag.

Fantuzzi, C., Simani, S., Beghelli, S., & Rovatti, R. (2002). Identification of piecewise affine models in noisy environment. *International Journal of Control*, *75*(18), 1472–1485.

Ferrari-Trecate, G., Muselli, M., Liberati, D., & Morari, M. (2003). A clustering technique for the identification of piecewise affine systems. *Automatica*, *39*(2), 205–217.

Ferrari-Trecate, G., & Schinkel, M. (2003). Conditions of optimal classification for piecewise affine regression. *Hybrid systems: computation and control* (HSCC 2003), Lecture notes in computer science, vol. 2623 (pp. 188–202). Berlin: Springer.

Heemels, W. P. M. H., De Schutter, B., & Bemporad, A. (2001). Equivalence of hybrid dynamical models. *Automatica*, *37*(7), 1085–1091.

Hu, X., & Xu, L. (2003). A comparative study of several cluster number selection criteria. *Intelligent data engineering and automated learning* (*IDEAL 2003*), Lecture notes in computer science, vol. 2690 (pp. 195–202). Berlin: Springer.

Imura, J., & van der Schaft, A. (2000). Characterization of well-posedness of piecewise linear systems. *IEEE Transactions on Automatic Control*, *45*(9), 1600–1619.

Jain, A. K., Dulin, R. P. W., & Mao, J. (2000). Statistical pattern recognition: a review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(1), 4–37.

Johansson, M. (2003). *Piecewise linear control systems: A computational approach*. Berlin: Springer.

Johansson, M., & Rantzer, A. (1998). Computation of piecewise quadratic lyapunov functions for hybrid systems. *IEEE Transactions on Automatic Control*, *43*(4), 555–559.

Jordan, M. I., & Jacobs, R. A. (1994). Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, *6*(2), 181–214.

Juloski, A.L., Heemels, W.P.M.H., & Ferrari-Trecate, G. (2004). Data based hybrid modelling of the component placement process in pick-and-place machines. *Control Engineering Practice*, to appear.

Ljung, L. (1999). *System identification—theory for the user*. (2nd ed)., Englewood Cliffs, NJ: Prentice-Hall.

Mitra, P., Pal, S. K., & Siddiqi, M. A. (2003). Non-convex clustering using expectation maximization algorithm with rough set initialization. *Pattern Recognition Letters*, *24*(6), 863–873.

Miyamoto, S. (1999). *Introduction to cluster analysis*. Morikita Shuppan Co., Ltd.

Nakada, H., & Takaba, K. (2003). Local stability analysis of piecewise affine systems. *Proceedings of the 6th European control conference*.

Ragot, J., Mourot, G., & Maquin, D. (2003). Parameter estimation of switching piecewise linear system. *Proceedings of the 42nd IEEE conference on decision and control* (pp. 5783–5788).

Redner, R. A., & Walker, H. F. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review*, *26*(2), 195–239.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, *14*(5), 465–471.

Roll, J., Bemporad, A., & Ljung, L. (2004). Identification of piecewise affine systems via mixed-integer programming. *Automatica*, *40*(1), 37–50.

Söderström, T., & Stoica, P. (1989). *System identification*. Englewood Cliffs, NJ: Prentice-Hall.

van der Schaft, A., & Schumacher, H. (2000). *An introduction to hybrid dynamical systems*. Berlin: Springer.

Vapnik, V. (1998). *Statistical learning theory*. New York: Wiley.

Vidal, R., Soatto, S., & Sastry, S. (2003). An algebraic geometric approach to the identification of linear hybrid systems. *Proceedings of the 42nd IEEE conference on decision and control* (pp. 167–172).

Hayato Nakada received B.Eng. degree in informatics and mathematical science and M.Inf. (Master of Informatics) degree in applied mathematics and physics from Kyoto University, Japan in 2000 and 2002, respectively. He is currently a doctoral student at the Department of Applied Mathematics and Physics, Graduate School of Informatics, Kyoto University. His current research interests include analysis, control and identification of hybrid systems and saturating systems.

Kiyotsugu Takaba received B.Eng. degree in applied mathematics and physics, M.Eng. degree in applied systems science and Dr.Eng. degree in applied mathematics and physics all from Kyoto University in 1989, 1991 and 1996, respectively. Since 1991, he has been with the Department of Applied Mathematics and Physics in Kyoto University, where he is presently an Associate Professor. His research interest includes robust and optimal control for multi-variable dynamical systems. He is a member of IEEE, ISCIE and SICE.

Tohru Katayama received B.E., M.E. and Ph.D. degrees all in applied mathematics and physics from Kyoto University, Kyoto, in 1964, 1966 and 1969, respectively. Since 1986, he has been in the Department of Applied Mathematics and Physics, Kyoto University. He had visiting positions at UCLA from 1974 to 1975, and at University of Padova in 1997. He was an Associate Editor of IEEE Transactions on Automatic Control from 1996 to 1998, and is now a Subject Editor of International Journal of Robust and Nonlinear Control, and the Chair of IFAC Technical Committee of Stochastic Systems for 1999–2002, and the Chair of IFAC Coordinating Committee of Signals and Systems for 2002–2005. His research interests include estimation theory, stochastic realization, subspace method of identification, blind identification, and control of industrial processes.