Copyright © IFAC Analysis and Design of Hybrid
Systems, Brittany, France, 2003

IFAC
Publications
www.elsevier.com/locate/ifac

SINGLE-LINKAGE CLUSTERING FOR OPTIMAL CLASSIFICATION IN PIECEWISE AFFINE REGRESSION

Giancarlo Ferrari-Trecate * **Marco Muselli** **

* *INRIA, Domaine de Voluceau, Rocquencourt, France.*
Email: Giancarlo.Ferrari-Trecate@inria.fr

** *IEIIT, Italian National Research Council, Genoa, Italy.*
Email: Marco.Muselli@ieiit.cnr.it

Abstract: When performing regression with piecewise affine maps, the most challenging task is to classify the data points, i.e. to correctly attribute a data point to the affine submodel that most likely generated it. In this paper, we consider a regression scheme similar to the one proposed in (Ferrari-Trecate *et al.*, 2001; Ferrari-Trecate *et al.*, 2003) that reduces the classification step to a clustering problem in presence of outliers. However. instead of the K-means procedure adopted in (Ferrari-Trecate *et al.*, 2001; Ferrari-Trecate *et al.*, 2003), we propose the use of single-linkage clustering that estimates automatically the number of submodels composing the piecewise affine map. Moreover we prove that, under mild assumptions on the data set, single-linkage clustering can guarantee optimal classification in presence of bounded noise. ***Copyright © 2003 IFAC***

Keywords: Piecewise affine functions, hybrid systems, identification, clustering.

1. INTRODUCTION

In this paper we consider the problem of reconstructing a Piece-Wise Affine (PWA) map from a finite number of noisy datapoints. A PWA map $f : \mathbb{X} \mapsto \mathbb{R}$ is defined by the equations

$$f(x) = f_q(k) \quad \text{if} \quad x \in \bar{\mathcal{X}}_q \tag{1}$$

$$f_q(x) = \left[\, x^T \ 1 \,\right] \bar{\theta}_q \tag{2}$$

where $\mathbb{X} \subset \mathbb{R}^n$ is a bounded polyhedron, $\{\bar{\mathcal{X}}_q\}_{q=1}^{\bar{s}}$ is a polyhedral partition of $\mathbb{X}$ in $\bar{s}$ regions and $\bar{\theta}_q \in \mathbb{R}^{n+1}$ are Parameter Vectors (PVs). Therefore, a PWA map is composed of $\bar{s}$ affine submodels defined by the pairs $(\bar{\theta}_q, \bar{\mathcal{X}}_q)$. The dataset $\mathcal{N}$ collects the samples $(x(k), y(k))$, $k = 1, \dots, N$ generated by the model

$$y(k) = f(x(k)) + \eta(k) \tag{3}$$

where $\eta(k)$ are noise samples. We assume that all the PVs are different and that the number $\bar{s}$ of submodels is unknown. Then, the aim of PWA regression is to estimate $\bar{s}$, the PVs and the regions by using the information provided by $\mathcal{N}$.

When considering hybrid systems, an input/output description of a PWA system (see (Sontag, 1981) for a definition) with inputs $u(k) \in \mathbb{R}^m$ and outputs $y(k) \in \mathbb{R}$ is provided by Piece-Wise ARX models that are defined by equation (3) where k is now the time index and the vector of regressors $x(k)$ is given by

$$[y(k-1) \dots y(k-n_a) u^T(k-1) \dots u^T(k-n_b)]^T.$$

It is apparent that, if the orders n_a and n_b are known, the identification of a Piece-Wise ARX model amounts to a PWA regression problem.

In order to highlight the main difficulties of PWA regression, consider the partition $\{\bar{\mathcal{F}}_q\}_{q=1}^{\bar{s}}$ of the dataset defined by the rule: $(x(k), y(k)) \in \bar{\mathcal{F}}_q$ if the datapoint is *associated* with the q-th submodel (i.e. if $x(k) \in \bar{\mathcal{X}}_q$). When $\bar{s}$ and the sets $\bar{\mathcal{F}}_q$ are known, the identification problem can be easily solved. In fact, the PVs can be estimated by solving a linear regression problem for each dataset $\bar{\mathcal{F}}_q$. Moreover, the regions can be reconstructed by finding the hyperplanes separating pairwise the sets $\{x : (x, y) \in \bar{\mathcal{F}}_q\}$, $\{x : (x, y) \in \bar{\mathcal{F}}_{q'}\}$ for all indices $q \neq q'$.

We stress that, independently of the algorithms used for estimating the submodels, the sets $\{\bar{\mathcal{F}}_q\}_{q=1}^{\bar{s}}$ partition the dataset in the optimal way. A key problem in PWA regression is that such sets are not known a priori because the regions are unknown. Then, it is apparent that any PWA regression method aims at providing, implicitly or explicitly, an estimate s of $\bar{s}$ and estimates $\mathcal{F}_q$ of the sets $\bar{\mathcal{F}}_q$. We say that an algorithm achieves *optimal classification* if it gives $s = \bar{s}$ and $\mathcal{F}_q \subset \bar{\mathcal{F}}_q$. In the previous relation, we used "$\subset$" instead of "$=$" because, in many practical cases, it is enough to correctly classify sufficiently many data points for obtaining good identification results. The remaining datapoints can be attributed a posteriori to the submodels and used for refining the estimated PVs and the regions (Ferrari-Trecate *et al.*, 2003; Ferrari-Trecate and Schinkel, 2003). We point out that the classification task is also the most challenging step in PWA regression, because, in absence of further information about the submodels, it cannot be decoupled from the estimation of the PVs.

Various methods have been proposed in literature for circumventing the problems associated with classification by exploiting both the knowledge of $\bar{s}$ and some a priori information about the model structure. For instance, in (Groff *et al.*, 2000), only monodimensional PWA functions are considered. In (Heredia and Gonzalo, 2000) and (Johansen and Foss, 1995) a gridding procedure is used to find the regions, that are constrained to have rectangular shape. More classical techniques, like Hinging Hyperplanes (Breiman, 1993) or neural networks with PWA activation functions (Haykin, 1994) focus on the estimation of continuous PWA models (that can represent only a limited number of logic features, in the case of hybrid behaviors). Analogously, in (Bemporad *et al.*, 2001), the attention is restricted to special subclasses of PWA systems.

We consider a PWA regression algorithm (summarized in Section 2) similar to the one proposed in (Ferrari-Trecate *et al.*, 2001; Ferrari-Trecate *et al.*, 2003). The method reported in (Ferrari-Trecate *et al.*, 2001; Ferrari-Trecate *et al.*, 2003), has two main features. First, it is capable to reconstruct general PWA maps without imposing special constraints on the PVs and the regions. Second, it allows to reduce the classification problem to a clustering problem that is solved through a K-means-like procedure. As any other supervised clustering algorithm, K-means requires to specify the number of clusters that must be found (Fritzke, 1997) and this corresponds to assuming the knowledge of $\bar{s}$.

In this paper, we investigate the use of single-linkage clustering (Duda and Hart, 1973) that is capable to find automatically the proper number of clusters. In particular, in Section 3 we prove that single-linkage clustering, coupled with a suitable rule for accepting or rejecting the clusters found, can guarantee optimal classification if the noise level is small enough. Single-linkage algorithms require the specification of a threshold whose choice may be critical. We study this point by showing theoretically (Section 3) and through a benchmark problem (Section 4) that in the small-noise case, the optimality of the results is robust against threshold mis-specification.

2. THE PWA REGRESSION ALGORITHM

In this section we summarize the overall PWA regression method that is structured in three steps. The conceptual skeleton is same of the algorithm reported in (Ferrari-Trecate *et al.*, 2003; Ferrari-Trecate *et al.*, 2001).

1. Local Regression. For $j = 1, \ldots, N$ we build a Local Dataset (LD) $\mathcal{C}_j$ collecting $(x(j), y(j))$ and its $c - 1$ distinct neighboring datapoints, i.e. the pairs $(\tilde{x}, \tilde{y}) \in \mathcal{N}$ that satisfy

$$\|x(j) - \tilde{x}\|^2 \leq \|x(j) - \hat{x}\|^2, \quad \forall (\hat{x}, \hat{y}) \in \mathcal{N} \backslash \mathcal{C}_j. \quad (4)$$

The cardinality c of an LD is a parameter of the algorithm, and we assume $c > n + 1$. We refer to $\mathcal{C}_j$ as a *pure LD* if it collects only datapoints associated with a single submodel (and we say that $\mathcal{C}_j$ is *associated* with this submodel). Otherwise the LD is termed *mixed*. Note that the distinction between pure and mixed LDs is conceptual and cannot be done, in practice, at this stage of the algorithm. For each LD $\mathcal{C}_j$ we compute a Local Parameter Vector (LPV) θ_j through least squares

$$\theta_j = Q_j Y_j, \quad Q_j = (\Phi_j^T \Phi_j)^{-1} \Phi_j^T Y_j \quad (5)$$

$$\Phi_j = \begin{bmatrix} x_1 & \cdots & x_c \\ 1 & \cdots & 1 \end{bmatrix}^T, \quad Y_j = \begin{bmatrix} y_1 & \cdots & y_c \end{bmatrix}^T$$

where x_i and y_i, $i = 1, \ldots, c$, satisfy $(x_i, y_i) \in \mathcal{C}_j$. In view of the bijection between LDs and LPVs, an

LPV is *pure* if it is obtained from a pure LD and *mixed* otherwise. Analogously, and with obvious meaning, each pure LPV is *associated* with a submodel.

An intuitive explanation of step 1 is the following (a more thorough description is provided in (Ferrari-Trecate *et al.*, 2003)). The LD $\mathcal{C}_j$ collects datapoints characterizing the behavior of the PWA map in a neighborhood of $x(j)$. Then, if $\mathcal{C}_j$ is pure and associated with the q-th submodel, θ_j is an estimate of $\bar{\theta}_q$. Moreover, if the noise level is "low" all the LPVs associated to a submodel are close to each other. On the other hand, if $\mathcal{C}_j$ is mixed, then θ_j is likely to be different from all the pure LPVs. The reason is that mixed LPVs account for the model mismatch since they results from fitting, with an affine model, datapoints generated by at least two different submodels. As a consequence, we expect that, in the LPV space, all pure LPVs are concentrated in $\bar{s}$ dense clouds, whereas the mixed LPVs form a pattern of isolated points. For this reason, mixed LPVs will be also referred to as *outliers*. We point out that an "high" noise level can make the clouds scattered and possibly overlapping (examples are reported in (Ferrari-Trecate *et al.*, 2003)), a fact that may compromise the accuracy of the results obtained in the next steps. The only countermeasure against noise is to increase the size c of the LDs, thus reducing the covariance of pure LPVs but

2. Clustering. The aim of this step is to isolate the clouds of pure LPVs through clustering. The clustering procedure we adopt is described in Section 3. Here we provide only a theoretical characterization of the clustering quality. Consider the sets $\{\bar{\mathcal{D}}_q\}_{q=1}^{\bar{s}}$, each one collecting only the pure LPVs associated with the same submodel and let Θ denote the set of all LPVs. Then, we have $\Theta = \cup_{q=1}^{\bar{s}} \bar{\mathcal{D}}_q \cup \mathcal{D}_{mix}$, where $\mathcal{D}_{mix}$ collects all mixed LPVs. The outcome of the clustering algorithm is, in general, a collection of s clusters $\{\mathcal{D}_q\}_{q=1}^{s}$ partitioning Θ. The best result we can obtain is characterized by the following definition.

Definition 1. The clusters $\{\mathcal{D}_q\}_{q=1}^{s}$ are optimal if each one satisfies either $\mathcal{D}_q = \bar{\mathcal{D}}_q$ (type-I clusters) or $\mathcal{D}_q \subset \mathcal{D}_{mix}$ (type-II clusters).

If $\{\mathcal{D}_q\}_{q=1}^{s}$ are optimal and we had a procedure for detecting the type-II clusters, we could discard them and compute the optimal number of submodels $s = \bar{s}$ simply by counting the type-I clusters. Moreover, type-I clusters provide the best possible starting point for the next step.

3. Estimation of the submodels. By using the bijective maps

$$(x(j), y(j)) \longleftrightarrow \mathcal{C}_j \longleftrightarrow \theta_j \qquad (6)$$

we can build the sets $\{\mathcal{F}_q\}_{q=1}^{s}$ according to the rule: $(x(j), y(j)) \in \mathcal{F}_q \Leftrightarrow \theta_j \in \mathcal{D}_q$. The points in each final set $\mathcal{F}_q$ can be used for estimating the PVs of each submodel (through, for instance, least squares). Also the regions $\{\mathcal{X}_q\}_{q=1}^{s}$ can be found on the basis of the final sets by resorting to multicategory pattern recognition algorithms. For further details on the implementation aspects, we defer the reader to (Ferrari-Trecate *et al.*, 2001; Ferrari-Trecate *et al.*, 2003). Note that if only type-I optimal clusters are used for forming the final sets $\mathcal{F}_q$, the inclusions $\mathcal{F}_q \subset \bar{\mathcal{F}}_q$ are verified, so matching the definition of optimality given in Section 1.

3. SINGLE-LINKAGE CLUSTERING

Single-linkage clustering algorithms are hierarchical methods that proceed by aggregating nearest-neighbor clusters. We consider the following distance between (disjoint) clusters: $d(\mathcal{D}, \mathcal{D}') = \min_{\theta \in \mathcal{D}, \theta' \in \mathcal{D}'} \|\theta - \theta'\|$. Single-linkage clustering is summarized in the following procedure (Duda and Hart, 1973).

Algorithm 1.

```
1. Initialize the list of clusters $\mathcal{L}^{(0)} = \{\mathcal{D}_1, \ldots, \mathcal{D}_N\}$ with $\mathcal{D}_q = \{\theta_q\}$. Fix $\ell > 0$. Set $i = 0$.
2. While card $\mathcal{L}^{(i)} > 1$ or
       $$\delta^{(i)} = \min_{\mathcal{D}_q, \mathcal{D}_{q'} \in \mathcal{L}^{(i)}} d(\mathcal{D}_q, \mathcal{D}_{q'}) < \ell \qquad (7)$$
   2.1. $\mathcal{D}_{\bar{q}} = \mathcal{D}_{\tilde{q}} \cup \mathcal{D}_{\hat{q}}$, where $\mathcal{D}_{\tilde{q}}$ and $\mathcal{D}_{\hat{q}}$ are the minimizers of (7).
   2.2. $\mathcal{L}^{(i+1)} = \mathcal{L}^{(i)} \setminus \mathcal{D}_{\hat{q}}$. Set $i = i + 1$.
```

A fundamental property of Algorithm 1 is that the sequence $\delta^{(i)}$ is increasing, thus implying that close LPVs are aggregated before distant LPVs. The parameter ℓ must be provided by the user and represents a guess on the minimal distance between clusters. Obviously, the number of clusters produced depends on ℓ: too small values, produce over-fragmented clusters where too large values lead to over-aggregation.

In order to accomplish step 3 of the PWA regression procedure, we must also provide a meaning for distinguishing between type-I and type-II clusters.

Discrimination rule: A cluster $\mathcal{D}_q$ is rejected if $\mathrm{card}(\mathcal{D}_q) \leq c$.

The idea underlying the previous rule is that, in view of step 1, an LPV can appear in Θ at most c times. Therefore, also the maximal cardinality of a mixed LPV is c. Note also that duplicate LPVs are the first to be aggregated. However, for reasonable choices of ℓ, sooner or later the algorithm will start aggregating different LPVs. Intuitively, if pure LPVs form dense and separated clouds, they will be merged in different clusters before than distinct outliers. If we are able to stop the algorithm (through a proper choice of ℓ) when all pure LPVs are merged, we obtain optimal clusters. Moreover, if type-I clusters have more than c elements (a property that is likely to be true in a non-pathological scenario) the discrimination rule allows us to reject all the type-II clusters before proceeding with the final estimation of the submodel (step 3).

The previous discussion opens the question of the existence of values of ℓ capable to stop the clustering at the right aggregation level. In Theorem 1, we prove that if the noise level is small enough, then such an ℓ always exists. Hereafter, $\beta > 0$ denotes an upper bound on the noise level, i.e. $\|\eta(k)\| < \beta$, $k = 1, \ldots, N$.

We start showing that all pure LPVs associated with the same submodel fall within a ball centered in the true PV whose radius depends on the noise level.

Lemma 1. For each $\theta_j \in \bar{\mathcal{D}}_q$ it holds $\theta_j \in B(\bar{\theta}_q, \rho_q(\beta))$. Moreover $\rho_q(\beta) \to 0$ for $\beta \to 0$.

Proof. The proof is similar to the one of Lemma 1 in (Ferrari-Trecate and Schinkel, 2003). □

Before stating the main Theorem, we need to introduce some assumptions on the spatial localization of the mixed LPVs.

Assumption 1. For a given $\beta > 0$, there exists $\ell_{mix}(\beta) > 0$ verifying, $\forall \theta_j \in \mathcal{D}_{mix}$,

$$\ell_{mix}(\beta) < \min_{\theta_i \in \bigcup_{q=1}^{s} \bar{\mathcal{D}}_q} \|\theta_j - \theta_i\| \tag{8}$$

$$\ell_{mix}(\beta) < \min_{\theta_i, \theta_j \in \mathcal{D}_{mix}, \theta_i \neq \theta_j} \|\theta_j - \theta_i\|. \tag{9}$$

Moreover, it holds $\lim_{\beta \to 0} \ell_{mix}(\beta) = \bar{\ell}_{mix} > 0$. □

Assumption 1 is very mild, if the noise is small enough. In fact, the inequality (8) requires that mixed LPVs are separated from pure LPVs. We point out that this is extremely likely to happen for a small β since, from Lemma 1 pure LPVs collapse in the PVs $\bar{\theta}_q$. Also the inequality (9), requiring that distinct mixed LPVs are not arbitrarily close, intuitively follows from the fact that a mixed LPV θ_q is obtained by fitting, datapoints in $\mathcal{C}_q$ associated with at least two different submodels. By recalling that distinct mixed LPVs are obtained on the basis of different LDs, it is highly probable that they are not concentrated in dense clouds.

Theorem 1. Assume that $\mathrm{card}(\bar{\mathcal{D}}_q) > c$, $q = 1, \ldots, \bar{s}$. Then, there exist $\beta > 0$ and $\ell > 0$ such that, under Assumption 1, the following facts are true

(1) The clusters produced by the single-linkage algorithm are optimal for $\ell = \bar{\ell}$.
(2) By applying the discrimination rule, only type-I clusters are accepted.

Proof. Let $\bar{\rho}(\beta) = \max_{q \in \{1, \ldots, s\}} \rho_q(\beta)$ where $\rho_q(\beta)$ is defined as in Lemma 1. Lemma 1 and the fact that all PVs are different imply that there exists $\beta_1 > 0$ such that all the balls $B(\theta_q, \rho_q(\beta_1))$ are disjoint and distant more than $2\bar{\rho}(\beta_1)$. In this scenario, for all θ_i, $\theta_j \in \bar{\mathcal{D}}_q$ we have that $\|\theta_i - \theta_j\| < 2\bar{\rho}(\beta_1)$. Then, in absence of mixed LPVs the choice $\bar{\ell} = 2\bar{\rho}(\beta_1)$ guarantees that chain clustering stops when all the optimal clusters of type-I are found. If mixed LPVs are present, Assumption 1 guarantees that is possible to find $\beta_2 > 0$, $\beta_2 \leq \beta_1$, such that $\ell_{mix}(\beta_2) > 2\bar{\rho}(\beta_2)$. By choosing $\bar{\ell} = 2\bar{\rho}(\beta_2)$, the algorithm stops with optimal clusters and point 1 follows.

Note that, if $\bar{\ell}$ is chosen as in point 1, the algorithm terminates before aggregating different mixed LPVs in the same cluster. This follows from the fact that Assumption 1 guarantees that the distance between such LPVs is at least $\bar{\ell}$ and the sequence $\delta^{(i)}$ in step 2 of Algorithm 1 is increasing. Therefore at most c identical mixed LPVs can be present in type-II clusters. From the assumption that the cardinality of type-I clusters is bigger than c, point 2 follows. □

Theorem 1 deserves some comments. First, a careful examination of the proof reveals that the threshold $\bar{\ell}$ should be bigger than the diameters of the sets $\bar{\mathcal{D}}_q$, (that are disjoint for β small enough) and smaller than the minimum distances between each $\bar{\mathcal{D}}_q$ and all the mixed LPVs. One may argue that, in practice, it could be difficult to find the proper value of $\bar{\ell}$. However, following the proof, one discovers that, for a sufficiently small β, the corresponding value of $\bar{\ell}$ is not unique. More precisely, there is an interval of $\bar{\ell}$-values guaranteeing optimality. This is important in application since leaves some freedom in choosing the threshold without spoiling the clustering accuracy (see also the example in Section 4). Second, optimal clustering, according to Def. 1, is guaranteed by the single-linkage algorithm even in presence of mixtures of noisy and noiseless data points. This constitute an importante difference with respect

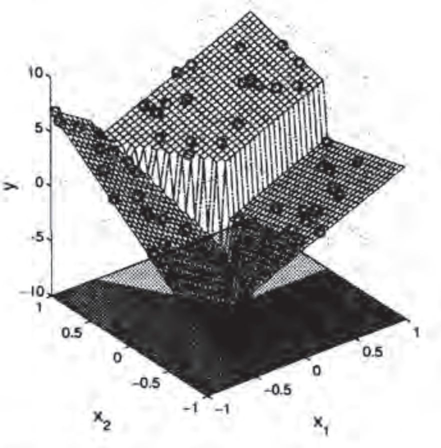


Fig. 1. Identification experiment: The PWA map (10) and the dataset.

to the K-means procedure reported in (Ferrari-Trecate *et al.*, 2003; Ferrari-Trecate *et al.*, 2001) that guarantees good results only if the noise fulfills some uniformity conditions (see Theorem 2 in (Ferrari-Trecate and Schinkel, 2003)).

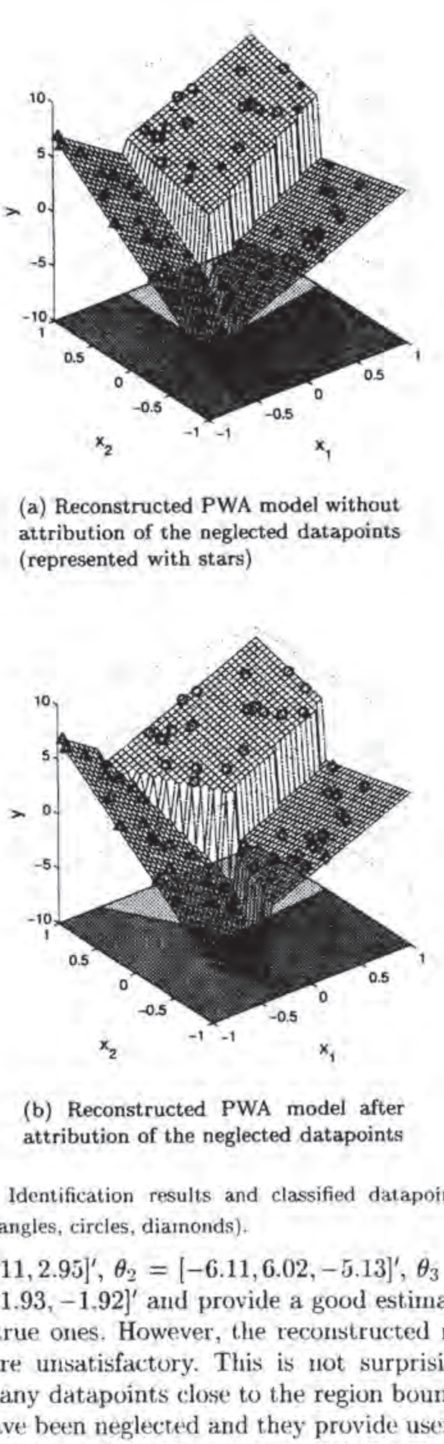
4. EXAMPLE

In this Section we discuss the performance of single-linkage clustering through an artificial example. The PWA map to be reconstructed is

$$f(x_1,x_2)=\begin{cases}4x_1+2x_2+3 \\ \quad \text{if } 5x_1+2.9x_2\geq 0 \text{ and } x_2\geq 0\\ -6x_1+6x_2-5 \\ \quad \text{if } 5x_1+2.9x_2<0 \text{ and}\\ \quad\quad 5x_1-2.9x_2<0\\ 4x_1-2x_2-2 \\ \quad \text{if } 5x_1-2.9x_2\geq 0 \text{ and } x_2<0\end{cases} \tag{10}$$

The dataset, composed of 70 data points corrupted with Gaussian noise of variance 0.01, and the true models are shown in Figure 1.

For identification, an LDs size of $c=6$ was chosen and single linkage clustering was performed with $\ell = 1$. After applying the discrimination rule, only three clusters survived, so providing a correct estimate of the number of submodels. The final datasets $\{\mathcal{F}_q\}_{q=1}^{3}$ are depicted in Figure 2(a). We point out that any value of ℓ in the interval $[0.39, 4.35]$ produces identical results. The data points corresponding to neglected clusters (then not included in the final datasets) are represented with stars. Note that all such points are close to the boundary between different regions. This is expected because mixed LDs must contain points belonging to different submodels. In the last step of the identification procedure, the PVs are estimated via least squares and the regions are reconstructed through Multicategory Robust Linear Programming (Ferrari-Trecate *et al.*, 2003; Ferrari-Trecate *et al.*, 2001). Figure 2(a) shows the results. The PVs are $\theta_1 = [4.03, 2.11, 2.95]'$, $\theta_2 = [-6.11, 6.02, -5.13]'$, $\theta_3 = [3.93, -1.93, -1.92]'$ and provide a good estimate of the true ones. However, the reconstructed regions are unsatisfactory. This is not surprising since many datapoints close to the region boundaries have been neglected and they provide useful information on the regions shape. We highlight that, at this stage, such points can be attributed to the identified submodels by using, for instance, Maximum Likelihood. The corresponding results are plot in Figure 2(b). A visual comparison with Figure 1 reveals that they are satisfactory.



(a) Reconstructed PWA model without attribution of the neglected datapoints (represented with stars)

(b) Reconstructed PWA model after attribution of the neglected datapoints

Fig. 2. Identification results and classified datapoints (triangles, circles, diamonds).

Remark 1. We discuss the clustering performance when ℓ is outside the interval $[0.39, 4.35]$. For $\ell < 0.39$ only few LPVs are aggregated in each cluster. This means that the number of clusters passing the discrimination rule is zero, one or two. However, over-fragmentation can be easily detected by considering the small fraction of points in each cluster. For $\ell > 4.35$, the outliers starts to be merged with the sets $\mathcal{D}_q$. This does not mean that the identification results are necessarily bad since it may happen that the final sets $\mathcal{F}_q$ are still optimal. In fact, for $\ell \in [4.35, 5, 65]$ we still found the results of Figure 2(b), after attribution of the neglected points. For $\ell > 5.65$ the number of clusters decreases progressively from three to one. In this case, the absence of clusters with few points indicates either the absence of outliers or over-aggregation. An examination of the residuals reveals the latter case, so suggesting the use of a smaller threshold.

5. DISCUSSION AND CONCLUDING REMARKS

PWA regression is significantly different from linear regression, due to the fact that regions and parameters of the submodels must be jointly estimated. In this paper, we focused on the most challenging step in PWA regression, namely the task of classifying the datapoints, and discussed the application of single-linkage clustering for solving it. The results of single-linkage clustering depends on the value of a threshold that must be specified a priori and this opens the problem of its tuning. In principle, one may argue that choosing the proper value is no easier than guessing the number of submodels, as required by almost all PWA identification algorithms.

In our opinion the two approaches are complementary and the decision of which one must be adopted is problem-dependent. Here, we discuss an important application of PWA regression to data analysis for which the use of the threshold ℓ is the only possible choice. We consider the problem of detecting submodels switches in long time series. More precisely, assume that input-output data of a PWARX model have been collected over a long time horizon, and that the parameters of the PWARX models are slowly varying in the overall horizon. In order to perform automatic detection of switchings, the usual recipe is to split the dataset in consecutive windows to be processed one after the other. In this scenario, the main difficulty is that the number of modes active in each window cannot be known a priori. On the other hand, the same threshold ℓ (tuned on the basis of experiments performed over few time windows) could be used for each data sequence, so letting the identification algorithm free to choose the optimal number of active submodels.

Finally, we point out that single-linkage clustering is not the only possible alternative for estimating the number of models. In fact, some unsupervised clustering algorithms (for instance, Growing Neural Gas (Fritzke, 1997)) can be used to this purpose. However, it should be noted that the performance of such methods usually depends on many parameters (like learning rates or forgetting factors), whose choice is seldom perspicuous and may be much more difficult than tuning the threshold parameter in single-linkage clustering.

REFERENCES

Bemporad, A., J. Roll and L. Ljung (2001). Identification of hybrid systems via mixed-integer programming. *Proc. IEEE Conference on Decision and Control* pp. 786–792.

Breiman, L. (1993). Hinging hyperplanes for regression, classification, and function approximation. *IEEE Trans. Inform. Theory* **39**(3), 999–1013.

Duda, R.O. and P.E. Hart (1973). *Pattern Classification and Scene Analysis*. Wiley.

Ferrari-Trecate, G. and M. Schinkel (2003). Conditions of optimal classification for piecewise affine regression. In: *Proc. 5th International Workshop on Hybrid Systems: Computation and Control* (A. Pnueli and O. Maler, Eds.). Lecture Notes in Computer Science. Springer-Verlag.

Ferrari-Trecate, G., M. Muselli, D. Liberati and M. Morari (2001). A clustering technique for the identification of piecewise affine systems. In: *Proc. 4th International Workshop on Hybrid Systems: Computation and Control* (M. Di Benedetto and A. Sangiovanni-Vincentelli, Eds.). Vol. 2034 of *Lecture Notes in Computer Science*. pp. 218–231. Springer-Verlag.

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

Fritzke, B. (1997). Some competitive learning methods. Technical report. Institute for Neural Computation. Ruhr-Universit at Bochum.

Groff, R.E., D.E. Koditschek and P.P. Khargonekar (2000). Piecewise linear homeomorphisms: The scalar case.. *Proc. Int. Joint Conf. on Neural Networks* **3**, 259–264.

Haykin, S. (1994). *Neural networks - a comprehensive foundation*. Macmillan, Englewood Cliffs.

Heredia, E.A. and R.A. Gonzalo (2000). Nonlinear filters based on combinations of piecewise polynomials with compact support. *IEEE Trans. on Signal Processing* **48**(10), 2850–2863.

Johansen, T.A. and B.A. Foss (1995). Identification of non-linear system structure and parameters using regime decomposition. *Automatica* **31**(2), 321–326.

Sontag, E. D. (1981). Nonlinear regulation: The piecewise linear approach. *IEEE Trans. Automatic Control* **26**(2), 346–358.