A perceptron-based approach to piecewise linear modeling with an application to time series

Edoardo Amaldi[1], Marco Mattavelli[2] and Jean-Marc Vesin[2]

[1] School of OR and Theory Center, Cornell University, Ithaca, NY 14853, USA
[2] Signal Processing Laboratory, EPFL, CH-1015 Lausanne, Switzerland

Abstract. Piecewise linear models are attractive when modeling a wide range of nonlinear phenomena but determining simultaneously the data partition and the parameter values of the corresponding submodels is a challenging problem. We show that this problem can be formulated as that of partitioning an inconsistent linear system into a minimum number of consistent subsystems and we describe a greedy algorithm, based on a simple variant of the perceptron procedure, which provides good approximate solutions in a short amount of time. The general approach is applied to piecewise linear modeling of time series.

1 Introduction

Although linear models play an important role in the study of a wide range of physical phenomena and technical problems, nonlinear models are often required due to the presence of non-linearity and non-stationarity. In many situations, piecewise linear models are attractive since they allow approximation of complex phenomena but are still simple enough due to local linearity. However, piecewise linear model estimation turns out to be a challenging problem because it involves partitioning the data as well as determining the parameter values.

Conventional approaches in which the data is first partitioned using clustering methods and the parameters of each submodel are then estimated using regression techniques [6] have some major limitations. In particular, they require the number of submodels to be guessed in advance and, since the type of submodel used is not taken into account during the clustering phase, they may lead to meaningless results.

By extending some ideas we developed in [4] for motion analysis, we present a combinatorial optimization approach to piecewise linear model estimation. We describe an algorithm based on a simple variant of the perceptron procedure and we discuss an application to piecewise linear modeling of time series.

2 A combinatorial optimization formulation

As is well known, linear models lead to formulations in terms of linear systems such as $A\ \mathbf{w} = \mathbf{b}$, where A is a $p \times n$ matrix whose rows correspond to the

data and $\mathbf{b}$, $\mathbf{w}$ are n-dimensional vectors. In the presence of non-linearity, the resulting linear systems are obviously inconsistent and classical approximate solutions that minimize least-mean-square error criteria [6] are not meaningful.

The following combinatorial optimization problem provides a natural way of addressing simultaneously the two fundamental issues underlying piecewise linear model estimation.

MIN PCS: Given an inconsistent $A\mathbf{w} = \mathbf{b}$, find a Partition of the set of equations into a MINimum number of Consistent Subsystems.

Given any solution of MIN PCS, the partition into consistent subsystems indicates the data partition and the solution associated with each consistent subsystem provides the parameter values of the corresponding linear submodel. Note that, according to Occam's razor principle, we look for the "simplest" piecewise linear model consistent with the data, which is most likely to be the correct one. Here the focus is on the number of submodels.

In order to cope with noisy data, each equation $\mathbf{a}^k\mathbf{w} = b_k$, where $\mathbf{a}^k$ is the kth row of A and b_k is the kth component of $\mathbf{b}$, can be replaced by the two complementary inequalities:

$$\mathbf{a}^k\mathbf{w} \leq b_k + \varepsilon \qquad \mathbf{a}^k\mathbf{w} \geq b_k - \varepsilon \tag{1}$$

where ε is the maximum tolerable error. If different equations are expected to be affected by different noise levels, different thresholds can of course be used.

Estimating minimum piecewise linear models turns out to be harder than finding least-mean-square solutions.

Theorem *It is* NP*-hard to approximate* MIN PCS *within a factor* $c < 3/2$.

In other words, assuming P $\neq$ NP, no polynomial-time algorithm is guaranteed to provide a partition whose number of consistent subsystems exceeds the minimum one by a multiplicative factor strictly smaller than 3/2. The proof is omitted due to the lack of space. According to this worst-case complexity result, we have to look for efficient heuristics with a good average-case behavior.

3 A greedy algorithm based on perceptron-like procedures

Since in practical applications we are interested in finding approximate solutions of MIN PCS rapidly, we adopt a greedy strategy and subdivide the original problem into a sequence of subproblems. Starting with the original inconsistent system composed of pairs of inequalities (1), close-to-maximum consistent subsystems are extracted iteratively. In other words, at each step we look for a solution satisfying as many pairs of inequalities as possible and then remove them from the current system until the remaining ones are consistent. This clearly yields a partition into consistent subsystems. Although the subproblem is NP-hard [1], it is not necessary and even desirable to look for the largest consistent subsystems. Indeed, it is easily verified that, even if such subsystems were available at each step, the greedy strategy is not guaranteed to yield minimum size partitions.

As we shall see, a simple variant of the perceptron procedure can be adapted to provide good approximate solutions of the subproblem in a short amount of time. Recall that training a perceptron amounts to finding a weight vector $\mathbf{w}$ satisfying the linear system:

$$\begin{array}{ll} \mathbf{a}^k \mathbf{w} \geq 0 & \forall k \text{ such that } b_k = 1 \\ \mathbf{a}^k \mathbf{w} < 0 & \forall k \text{ such that } b_k = -1 \end{array} \tag{2}$$

where $\mathbf{a}^k \in \mathbb{R}^n$, $1 \leq k \leq p$, are the training vectors and $b_k \in \{-1, 1\}$ the corresponding desired outputs. For nonlinearly separable training sets, one looks for an optimal weight vector that correctly classifies as many input vectors $\mathbf{a}^k$ as possible. Several variants of the perceptron procedure have been proposed to search for such maximum consistent subsystems of (2), see [1]. In the thermal perceptron [2], the basic idea is to pay decreasing attention to misclassified input vectors with large errors. Starting with an arbitrary initial weight vector $\mathbf{w}_0$, input vectors $\mathbf{a}^k$ are selected in a random order and the current $\mathbf{w}_i$ is updated as follows:

$$\mathbf{w}_{i+1} := \begin{cases} \mathbf{w}_i + \frac{t}{t_0} \exp\left(-|\mathbf{a}^k \mathbf{w}_i|/t\right) b_k \mathbf{a}^k & \text{if } \mathbf{a}^k \text{ is misclassified by } \mathbf{w}_i \\ \mathbf{w}_i & \text{otherwise,} \end{cases} \tag{3}$$

where the *temperature* t is linearly decreased from an initial t_0 to 0 in a predefined maximum number of cycles $C_{\max}$ through the training set. This annealing process guarantees that corrections will terminate after $C_{\max}$ cycles. See [2, 1] for appropriate choices of t_0 and of the decreasing schedule.

The thermal perceptron procedure can be easily adapted to find close-to-maximum consistent subsystems of pairs of inhomogeneous inequalities such as (1). At each iteration, one picks a pair of inequalities and one checks whether the current solution $\mathbf{w}_i$ violates one of them. If $\mathbf{a}^k \mathbf{w}_i \leq b_k - \varepsilon$, then $\mathbf{w}_{i+1}$ is obtained by adding

$$\frac{t}{t_0} \exp\left(-E_i^k/t\right) \ \mathbf{a}^k \tag{4}$$

to $\mathbf{w}_i$, where $E_i^k = b_k - \varepsilon - \mathbf{a}^k \mathbf{w}_i$ denotes the deviation from satisfiability. Similarly, if $\mathbf{a}^k \mathbf{w}_i \geq b_k + \varepsilon$, then (4) with now $E_i^k = \mathbf{a}^k \mathbf{w} - b_k - \varepsilon$ is substracted from $\mathbf{w}_i$. The next cycle is started when all pairs of inequalities have been considered in a random order.

Although our simple greedy strategy is not guaranteed to yield minimum partitions, it turns out to be very effective experimentally. Note that, unlike in the usual perceptron applications, the weight vectors determined during the training phase are not used to classify new patterns. Instead, they reveal the structure and they provide the parameter values of the piecewise linear model.

To achieve higher robustness for noisy data, a least-mean-square solution can be determined for each subsystem of the partition provided by our algorithm. Thus, piecewise linearity is taken into account in the partition process and robustness is guaranteed in submodel parameter estimation.

4 Piecewise linear modeling of time series

Nonlinear signal models have been the object of an increasing interest over the past few years. A number of applications can be found in different fields such as, for instance, economic system modeling and biomedical signal analysis. The idea of breaking a global linear model into a number of submodels, gave rise to the so-called threshold autoregressive (TAR) models [7, 8] in which the choice of the model at time t is based on the comparison of the signal value at time $t-1$ with pre-defined thresholds. This scheme allows one to model and reproduce phenomena such as jumps and limit cycles, but its inherent limitation derives from the selection of the thresholds, for which no general method exists. The principle, called piecewise linear modeling [9], consists of partitioning the state-space into a certain number of regions and estimating a linear autoregressive submodel for each one of them. A piecewise linear autoregressive model can be described as follows [9]:

$$x_t = \sum_{j=1}^{m} w_{ij} x_{t-j} + u_t \tag{5}$$

where the set of coefficients $(w_{i1}, \ldots, w_{im})$ depends on the position of the vector $(x_{t-1}, \ldots, x_{t-m})$ with respect to a given partition of the state-space $\mathbb{R}^m$, and $\{u_t\}$ is an i.i.d. sequence. Here the x's are the observations and the coefficients w_{ij}'s as well as the partition of the state-space have to be estimated.

Although piecewise autoregressive models are more attractive than the TAR ones, the number of submodels must be selected a priori and an appropriate state-space partition needs to be determined. In [9] a two-stage strategy is suggested. First one determines a state-space decomposition using a Kohonen feature map [3] and then one estimates the parameters of each submodel using robust regression techniques. The purpose of such an unsupervised clustering phase is to avoid guessing the number of submodels in advance. However, besides the delicate convergence issue, it is unclear how a state-space partition can be actually derived from a given feature map. Even more importantly, the clustering process does not take into account the type of submodel used.

According to our general approach, we consider the problem of estimating such piecewise linear models in terms of MIN PCS. Clearly, any sequence of observations $\{x_1, \ldots, x_L\}$ which cannot be modeled (approximated) by a simple autoregressive model, leads to an over-determined inconsistent system of pairs of inequalities such as (1). Given any partition into consistent subsystems, each consistent subsystem defines a group of vectors in the state-space and the associated solution indicates the parameter values of the corresponding submodel.

A number of experiments have been carried out with various well-known time series which have been extensively used as benchmark in the literature. Due to the lack of space, we only mention a few typical results obtained for the chaotic time series generated with the Henon maps and the Sunspot time series [8]. The Henon map is defined by: $x_t = 1 - 1.4x_{t-1}^2 + 0.3x_{t-2}$. The points with coordinates x_{t-1}, x_{t-2} are located on a fractal strange attractor (see Figure 1). As an example, we report the results obtained for a run with a sequence of 2000

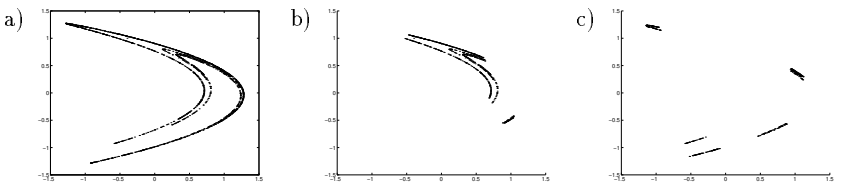

Fig. 1. a) State-space representation of the data of the Henon time series. Regions corresponding to the largest (b) and the second largest (c) submodel.

consecutive values, with a modeling order of the autoregression $m = 2$ and a maximum tolerable error $\epsilon = 0.3$. The resulting partition contains 8 consistent subsystems. Interestingly, the regions corresponding to each submodel are well defined in the state-space, which indicates that they are valuable for modeling purposes. Depending on the region to which the point (x_{t-1}, x_{t-2}) belongs, the corresponding set of autoregressive coefficients provides the estimate of x_t within the predefined ϵ. Figures 1 b) and 1 c) show the regions corresponding to the two largest subsystems.

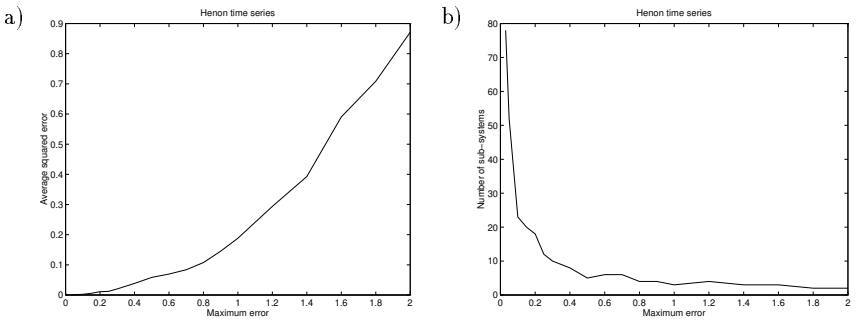

Fig. 2. a) Average squared estimation error of the piecewise linear model and b) number of components in the piecewise state-space decomposition versus ϵ.

An interesting feature of our algorithm is the possibility of selecting the maximum tolerable error ϵ. We have investigated the trade-off between ϵ and model complexity. Figure 2 a) and b) display the total average squared error and, respectively, the number of subsystems as a function of ϵ. While the average squared error clearly increases with ϵ, the number of subsystems remains almost constant for a wide range of values after a sharp decrease. Intuitively, the values of ϵ corresponding to the knee of this curve yield the best compromise between model complexity and average estimation error. We are currently investigating how the minimum description length criterion [5] which is used to estimate this trade-off for linear models can be extended to the case of piecewise linear ones. In this case, not only the set of coefficients but also the description of the state-space partition has to be taken into account.

The same type of tests have been carried out with the well-known Sunspot time series [8]. Comparing our results with those obtained with the classical TAR models [8], we achieve significantly lower average squared errors at the expense of a reasonably small increase in model complexity. We also observe the same behavior of the average squared error and the number of subsystems with respect to ϵ. Finally, it is worth pointing out that in all tests a small number of cycles (100-200) of the perceptron-like procedure was sufficient.

5 Concluding remarks

We have presented a general combinatorial optimization approach for estimating piecewise linear models. Our greedy algorithm based on a variant of the thermal perceptron procedure has been applied to the challenging problem of piecewise linear modeling of time series. The experimental results obtained for various time series, including cahotic ones, indicate that it yields natural partitions of the state-space that are valuable for modeling purposes. To use such piecewise autoregressive models for prediction, signal classification and event detection, state-space partitions can be derived from the resulting linear system partitions using distance measures that take into account the distribution of the state-vectors in each component.

References

1. E. Amaldi. *From finding maximum feasible subsystems of linear systems to feedforward neural network design.* PhD thesis, Department of Mathematics, Swiss Federal Institute of Technology, Lausanne, 1994.
2. M. Frean. A "thermal" perceptron learning rule. *Neural Computation*, 4(6):946–957, 1992.
3. T. Kohonen. The self-organizing map. *Proc. IEEE*, 78:1464–1480, 1990.
4. M. Mattavelli and E. Amaldi. Using perceptron-like algorithms for the analysis and the parameterization of object motion. In F. Girosi et al., editor, *Neural Networks for Signal Processing V, Proceedings of the 1995 IEEE workshop*, pages 303–312, 1995.
5. J. Rissanen. *Stochastic Complexity in Statistical Inquiry.* World Scientific, Singapore, 1989.
6. P. J. Rousseeuw and A. M. Leroy. *Robust regression and outlier detection.* John Wiley & Sons, Inc, New York, 1987.
7. H. Tong. *Threshold Models in Nonlinear Time Series Analysis.* Springer, New York, 1983.
8. H. Tong. *Nonlinear Time Series.* Oxford University Press, Oxford, 1990.
9. J.M. Vesin. A common generalization framework for two classical nonlinear models. In *Proc. IEEE Int. Workshop on Nonlinear Digital Signal Processing*, pages 6–2.3, Tampere, Finland, 1993.

This article was processed using the LaTeX macro package with LLNCS style