TREE STRUCTURED NON-LINEAR SIGNAL MODELING AND PREDICTION

Olivier Michel

Ecole Normale Supérieure de Lyon
Laboratoire de Physique, URA 1325 CNRS
46, allée d'Italie
69364 Lyon, France

Alfred Hero

Department of Electrical Engineering
and Computer Science
University of Michigan
Ann Arbor, MI 48109-2122

ABSTRACT

We develop a non-parametric method of nonlinear prediction based on adaptive partitioning of the phase space associated with the process. The partitioning method is implemented with a recursive tree-structured vector quantization algorithm which successively refines the partition by binary splitting where the splitting threshold is determined by a penalized maximum entropy criterion. A complexity penalty is derived and applied to protect against high statistical variability of the predictor structure. We establish an important relation between our tree-structured model for the process and generalized non-linear thresholded AR model (ART). We illustrate our method for two cases where classical linear prediction is ineffective: a chaotic "double-scroll" signal measured at the output of a Chua-type electronic circuit, and a simulated second order ART model.

1. INTRODUCTION

Tree-based models were first introduced as a non-parametric exploratory data analysis technique for non-additive statistical models [1]. The tree-based model represents the data in a hierarchical structure where the leaves of the tree induce a non-uniform partition of the data space. Each leaf can be labeled by a scalar or vector value of a one-step predictor, a non-linear response variable, or a multi-variable quantizer output. Once a cost-complexity metric is specified, the tree can then be recursively grown to efficiently perform particular tasks such as non-linear prediction, pattern classification, and vector quantization (VQ) [2, 3]. The tree-based approach has the several attractive features. Unlike likelihood approaches no parametric model is required, however if one is available it can be incorporated into the tree structure as a constraint. Furthermore, unlike higher order moment approaches a tree model is stable even in the case of heavy tailed densities. Finally, unlike moment-based methods the performance of the optimal decision tree is invariant with respect to monotonic non-linear transforms of the data.

This paper presents an approach to tree structured signal modeling and prediction based on a maximum entropy recursive partitioning of the signal phase space and the local singular value decomposition (SVD). We use the Takens [4] time delay embedding method to construct a phase space for the signal which captures the linear or non-linear dynamics of any finite dimensional state model. We apply recursive tree growing techniques to specify an optimal tiling of the phase space which represents the best piecewise constant approximation to the joint probability density function under a complexity constraint. The partitioning is accomplished by adding or deleting branches (nodes) of the tree according to a maximum entropy principle: we test that the joint distribution is approximately uniform within any candidate partition by comparing the conditional entropy of the data points in the candidate partition to the maximum achievable conditional entropy. By implementing a local SVD over each node of the tree prior to performing the uniformity test we obtain a hierarchical signal model which is very similar to the non-linear auto-regressive (AR) threshold model, referred to as SETAR in [7]. This threshold model has been proposed for many physical signal models involving stochastic resonance and bistable/multi-stable trajectories such as ECG cardiac signals, EEG brain signals, turbulent flow, economic time series and chaotic signals.

This work was partially supported by a NATO postdoctoral fellowship (21B93-France, 01/94-09/94) awarded to Olivier Michel while visiting Dept. of EECS, University of Michigan, Ann Arbor.

2. DESCRIPTION OF TREE-BASED APPROACH

Let $\{\mathbf{X}(k)\}$ be a stationary random process and let P denote the underlying probability measure. For p a positive integer and τ a positive real number define $\underline{\mathbf{X}}(k) = [\mathbf{X}(k), \ldots, \mathbf{X}(k-\tau(p-1))]^T$. $\{\underline{\mathbf{X}}(k)\}_k$ is called the phase trajectory through p-dimensional phase space $I\!R^p$ with embedding parameter τ. Let $\Pi = \{\pi_1, \ldots, \pi_L\}$ be a partition of $I\!R^p$ into L cells, let $\{\underline{q}_1, \ldots, \underline{q}_L\}$ be representative points from each of these cells, and define the function Q: $Q(\underline{x}) = q_l$ if $\underline{x} \in \pi_l$. The random vector $\mathbf{X}_q(k) \stackrel{def}{=} Q(\underline{\mathbf{X}}(k))$ is a quantization of $\underline{\mathbf{X}}(k)$ and the discrete probability distribution function $P_{\underline{X}_q}(\underline{q}_l)$, $l = 1, \ldots, L$, is equal to the theoretical histogram $P(\underline{\mathbf{X}}(k) \in \pi_l)$, $l = 1, \ldots, L$, of $\underline{\mathbf{X}}$. This theoretical histogram is the most complete statistical model of the quantized phase trajectories and can be used to perform optimal non-linear prediction, process classification, and other statistical tasks. For example, the well known minimum mean-squared error predictor of the quantized sample $\mathbf{X}_q(k)$ given the values of $p-1$ past quantized samples $\mathbf{X}_q(k-\tau) = x_1, \ldots, \mathbf{X}_q(k-\tau(p-1)) = x_{p-1}$ is given by the following function of the histogram:

$$\widehat{\mathbf{X}}_q(k) = \sum_{l=1}^{p-1} \underline{e}_1^T \underline{q}_l \frac{P_{\underline{X}_q}(\underline{e}_1^T \underline{q}_l, x_1, \ldots, x_{p-1})}{\sum_{\forall \underline{e}_1^T \underline{q}_l} P_{\underline{X}_q}(\underline{e}_1^T \underline{q}_l, x_1, \ldots, x_{p-1})} \qquad (1)$$

where $\underline{e}_1 = [1, 0, \ldots, 0]^T$. The mean square error improves monotonically as the number L of quantization levels, equivalently the number of partition cells, becomes large and the continuous distribution $P_{\underline{X}}(\underline{x})$ of $\underline{\mathbf{X}}(k)$ becomes well approximated by the staircase function $P_{\underline{X}_q}(\underline{q}_l) \cdot I(\underline{x} \in \pi_l)$, where $I(A)$ denotes the 0-1 indicator function of an event A.

Now in a practical setting only a finite set of realizations of the phase trajectory $\underline{x}(k),\ k = 1, \ldots, N$ is available and the theoretical histogram must be estimated from the data. In this case performance will not improve monotonically in L, in particular $L \ll N$ is necessary to stabilize the histogram estimate. A recursive tree growing procedure can be used to find an increasingly dense sequence of partitions $\Pi^l = \{\pi_1^l, \ldots, \pi_{L_l}^l\}$ of $I\!R^p$ which iteratively minimize a measure of distortion between $\underline{x}(k)$ and its quantization $\underline{x}_q^l(k)$. In this paper we restrict the cells π_j^l to be rectangles in $I\!R^p$.

Assume that at depth l of the tree we have created a partition Π^l and consider the partition cells π_i^l, which we call the i-th parent nodes at level l. We refine the partition Π^l by using a maximum entropy binary splitting rule (described below) to split each partition cell π_i^l into 2^p smaller cells which we call children-nodes of the i-th parent. The sample distributions of data points over the set of children-nodes are each tested against the uniform distribution (null hypothesis H_0) via the Chi-square goodness of fit test. If for a cell π_i^l the null hypothesis H_0 is rejected, the 2^p-ary split is memorized, along with the resulting sub-cells of π_i^l, and 2^p parent nodes at level $l+1$ are created. Otherwise, the splitting procedure is stopped and the parent node π_i^l at level l is declared a terminal node. The set of terminal nodes are called the leaves of the tree. See Fig. 1 for an illustration of the tree growing procedure.

Binary splitting rule: For each parent node the splitting rule is implemented by applying a single threshold to each of the p coordinate axes of the phase space. More specifically, for splitting the parent node π_i^l, the p thresholds $\widehat{T}_1, \ldots, \widehat{T}_p$ are selected as the sample median of the projection of the inscribed data cloud onto each of the coordinate axes $\underline{e}_1, \ldots, \underline{e}_p$:

$$\widehat{T}_j = \text{median}\{\underline{e}_j^T \underline{x}(k) : \underline{x}(k) \in \pi_i^l, k = 1, \ldots, N\}.$$

where, for a scalar sequence $x_1, \ldots, x_n$, the sample median is a point such that roughly half of the iterates fall to the left and half to the right:

$$\text{median}\{x_i\} = \begin{cases} x_{(n/2)}, & n\text{even} \\ x_{([n+1]/2)}, & n\text{odd} \end{cases}$$

and $x_{(1)}, \ldots, x_{(n)}$ denotes the rank ordered sequence.

Under the assumption that the n scalars $\{\underline{e}_j^T \underline{X}(k) : \underline{X}(k) \in \pi_j\}_{k=1}^N$ are conditionally i.i.d. with common continuous pdf $f_{x|\pi_j}$, the sample median $\widehat{T}_j$ is an asymptotically unbiased and consistent estimator of the theoretical median T_j, which is the half mass point of the marginal cumulative distribution function, and it has an asymptotic normal distribution [6]:

$$\widehat{T}_j \sim \mathcal{N}\left(T_j, \frac{1}{4n[f_{x_j|\pi_j}(T_j)]^2}\right).$$

It can be shown [8] that the median splitting rule has a strong optimality property among binary splitting rules: for large N it maximizes the conditional entropy of the quantized phase space. Furthermore, the rule is optimal in the

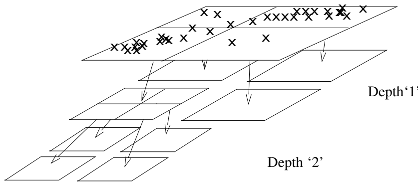

Figure 1: *For a $p = 2$ dimensional phase space embedding, the root-node is split into 4 subsets and the distribution is found to be non-uniform. Among the derived subsets, only the one depicted by the lower left corner square was found to be non-uniform and split further.*

sense of minimizing a lower bound on the average distortion over π_j: $E[g(|\underline{bX}_q - \underline{\mathbf{X}}|)|\pi_j]$ where g is any non-decreasing distortion function.

We use the approximate variance expression

$$var(\widehat{T}_j|\mathbf{n}) = \frac{1}{4\mathbf{n}[f_{x_j|S}(T_j)]^2},$$

where $\mathbf{n}$ is number of points falling into the parent node, to obtain a stopping rule for terminating a node: the size of each partition element must be greater than $c\sqrt{var(\widehat{T}_j|\mathbf{n})}$ where $c > 1$. This stopping rule ensures that the variance of the binary threshold $\widehat{T}_j$ does not exceed three times the width of each quantization cell and thus constrains the complexity of the tree. It is simple to show that this leads to the following simple stopping rule: terminate the node if the number of points $\mathbf{n}$ in its partition is less than $c\ 2^p$. This guarantees that the number of terminal nodes of the final tree will be significantly less than the total number of data points N. In the simulations below we used $c = 2$.

The final tree determines a partition of the phase-space which is described by the set of leaves (terminal cells) of the tree $\pi_1, \ldots, \pi_L$, together with the empirical histogram (cell occupancy rate): $\widehat{P}_{\underline{X}_q}(\pi_j) = N_{\pi_j}/N$, where N_{π_j} is the number of samples $\{\underline{x}(k)\}_{k=1}^N$ which fall into leaf π_j.

3. RELATION TO ART MODELS VIA THE SVD

Let $X_{n+1} = F_{X_n}(X_n) + \varepsilon_n$ be the sampled form of a p-dimensional dynamical system equation. F_{X_n} depicts the dynamical behavior of the system when the state vector has value X_n, that is F may be state-dependent. ε can be regarded as a realization of an observation noise or as a deterministic state perturbation. We first construct a tree based on the "training set" $X_n, 1 \leq n \leq N$ which creates cells in the phase space $\mathcal{R}^p$ for which the distribution of the realizations of the state vector have been estimated to be uniform.

As described in the previous section the tree growing procedure is based on partitioning the phase space into rectangles until the distributions of points within each cell are close to uniformly distributed, i.e. separable into p piecewise constant marginal distributions. We perform a local recursive orthogonalization of the data prior to node splitting in order to produce trees with fewer leaves. Define the locally orthogonalized vector $\underline{\mathbf{X}}^{\pi^l_j}(k)$ produced at depth l for some parent node π^l_j and define $\Lambda(\pi^l_j)$ as the $p \times p$ local covariance matrix:

$$\Lambda(\pi^l_j) = E_{\pi^l_j}\left[(\underline{\mathbf{X}}^{\pi^l_j} - E_{\pi^l_j}[\underline{\mathbf{X}}^{\pi^l_j}])(\underline{\mathbf{X}}^{\pi^l_j} - E_{\pi^l_j}[\underline{\mathbf{X}}^{\pi^l_j}])^T\right]$$

where $E_{\pi^l_j}[\bullet] \stackrel{def}{=} E[\bullet | \underline{\mathbf{X}}^{\pi^l_j} \in \pi^l_j]$. Let $M_j^T diag(\lambda_j) M_j$ denote the eigendecomposition of $\Lambda(\pi^l_j)$. The equation for propagation of the local orthogonalized vector from a cell π^l at depth l to a cell π^{l+1} at depth $l+1$ is:

$$\underline{\mathbf{X}}^{\pi^{l+1}} = M_l(\underline{\mathbf{X}}^{\pi^l} - \underline{C}^{\pi^l}),$$

where subtraction of $\underline{C}^{\pi^l} = E_{\pi^l}[\underline{\mathbf{X}}]$ ensures zero mean. By induction this gives the closed form expression for $\underline{\mathbf{X}}^{\pi^l}$ in terms of the original data $\underline{\mathbf{X}}$ at the root node:

$$\underline{\mathbf{X}}^{\pi^l} = \mathcal{M}_{l-1}\underline{\mathbf{X}} - \underline{\mathcal{C}}_{l-1}, \qquad (2)$$

where $\mathcal{M}_d = \prod_{i=0}^d M_i$ and $\underline{\mathcal{C}}_d = \sum_{i=0}^d \left[\prod_{j=i}^d M_j\right] \underline{C}^{\pi^i}$.

Note that for any parent node π^l the covariance matrix of the rotated data $\underline{\mathbf{X}}^{\pi^l}(k)$ is diagonal, which means that the components of $\underline{\mathbf{X}}^{\pi^l}(k)$ are separable (in the mean squared sense) but not necessarily uniform. On this rotated data the Chi-square test for uniformity can easily be implemented on a coordinate-by-coordinate basis. When the tree growing procedure terminates we will have found a set of partition cells $\pi_1^{l_1}, \ldots, \pi_L^{l_L}$ such that each $\pi^l = \pi^l_j$ contains points $\{\underline{\mathbf{X}}^{\pi^l}(k)\}_k$ which are (approximately) vectors of white noises. Thus, in view of relation (2), we obtain a set of piecewise state-conditioned (and possibly unstable) AR models for $\underline{\mathbf{X}}(k)$. This specifies an AR-threshold (ART) model, which is very similar to the SETAR model [7], for which the transition from one AR model to another is controlled by a values of all coordinates of the phase space.

4. EXAMPLES

First we illustrate a tree-based one step forward quantized prediction (as described in section 3) for chaotic time series. A voltage signal was digitized from the output of a double scroll electronic circuit from the Chua family. We chose an embedding dimension $p = 4$ to generate the phase trajectory $\underline{x} = [x(k), x(k-\tau), x(x-2\tau), x(k-3\tau)]^T$. The reconstruction delay τ was chosen in such a way to minimize the mutual information between the coordinates (see [5] and [4]). The prediction was constructed by growing a tree to generate the quantization intervals, using the midpoint of each terminal cell π for the quantization level $\underline{q}$, and using Eq. (1) with the theoretical histogram $P_{\underline{X}_q}(\underline{q}_l) = P(\underline{X} \in \pi_l)$ replaced by the empirical histogram $\widehat{P}_{\underline{X}}(\pi_l)$. A training set of $N = 4096$ points was used to grow the tree and to estimate the histogram. Figure 2 shows results which indicate that our tree based method performs as well as the popular but more complicated nearest neighbor prediction methods.

Second we consider doing non-linear prediction for the thresholded AR (ART) model given by the equation

$$\mathbf{X}(k) = \begin{cases} 1.71\mathbf{X}(k-1) - .81\mathbf{X}(k-2) + .356 + \varepsilon_k, & \mathbf{X}(k-1) > 0 \\ -.562\mathbf{X}(k-2) - 3.91 + \varepsilon_k, & \mathbf{X}(k-1) \leq 0 \end{cases}$$

A tree was grown in a 3-dimensional reconstructed phase space, from a training series of 4000 points. Figure 3 shows a representation of the final tree. Notice that the tree only contains two leaves and one internal node which almost perfectly separates the phase space into its two constituent linear AR process models. In figure 4, we show results of using the tree of Figure 3 for one-step forward prediction. A classification procedure was performed to determine to which leaf a given time sample $\underline{x}(k)$ of the phase trajectory belongs and we then performed optimal prediction using the estimated AR(3) model for this leaf. By concatenating the coefficients of the estimated AR model into a 3-element vector A and plotting the vectors of coefficient estimates in $I\!R^p$ we see from Figure 5 that the two models are well identified by the prediction tree. In Figure 5 the amplitude of each vector is plotted proportionally to the estimated cell occupancy probability associated with each leaf.

5. REFERENCES

[1] J.N. Sonquist and J.N. Morgan, "The detection of interaction effects," Monograph 35, Survey Research Center, Institute for Social Research, University of Michigan, 1964.

[2] L. Breiman, J.H.Friedman, R.A. Olshen, C.J. Stone, "Classification and Regression Trees," Wadsworth Advanced Books and Software, 1984.

[3] P.A. Chou, T. Lookabaugh, R.M. Gray, "Optimal Pruning with Applications to Tree-Structures Source Coding and Modeling," *IEEE Trans on Inf. theory*, vol.35, No.2, 1989, pp 299-315.

[4] J.P. Eckmann and D. Ruelle, "Ergodic Theory of Chaos and Strange Attractors," *Rev. Mod. Phys.*, Vol. 57, No. 3, pp. 617–656, 1985.

[5] A.M. Fraser, "Information and Entropy in Strange Attractors," *IEEE Trans on Inf. theory*, vol.35, No.2, 1989, pp 245-262.

[6] A.M. Mood, F.A. Graybill, D.C.Boes, "Introduction to the Theory of Statistics," Mc Graw Hill International Editions, Statistics Series, 3rd ed. 1974.

[7] H. Tong, "Non Linear Time Series : a Dynamical system Approach,"Oxford Science Publication, Oxford University Press, NY 1990.

[8] O. Michel and A. O. Hero, "Tree-based modeling of non-linear processes for prediction, detection, and classification," technical report in preparation, 1995.

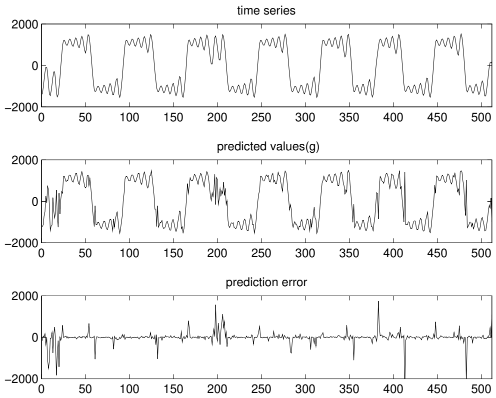

Figure 2: *quantized one step forward prediction for the experimental double scroll system*

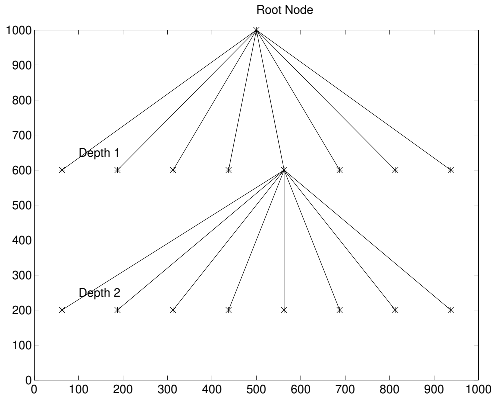

Figure 3: *Tree estimated from a 3-d representation of the system*

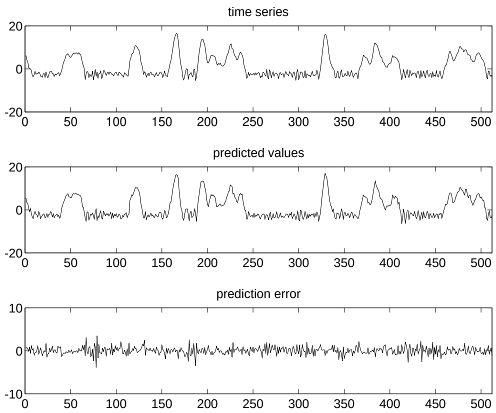

Figure 4: *Tree-based 1 step forward prediction and prediction errors*

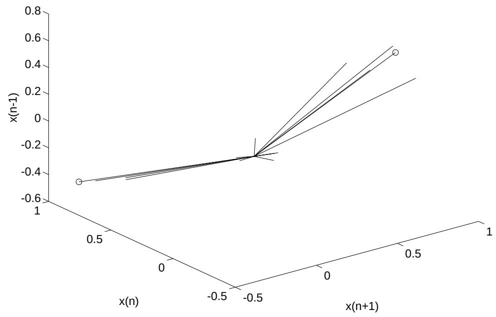

Figure 5: *Vector representation of the signal models obtained in the leaves of the tree. The true models are indicated by a 'o'*