ESANN'2006 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 26-28 April 2006, d-side publi., ISBN 2-930307-06-4.

Modelling switching dynamics using prediction experts operating on distinct wavelet scales

Alexandre Aussem[1] and Pierre Chainais[2] *

COMAD Team, PRISMa Lab., University Lyon 1,
43 boulevard du 11 novembre 1918, 69622 Villeurbanne cedex
aaausem@univ-lyon1.fr

UMR CNRS 6158 LIMOS, Université Blaise Pascal,
Campus des Cézeaux, 63173 Aubière cedex
pchainai@isima.fr

Abstract. We present a framework for modelling the switching dynamics of a time series with correlation structures spanning distinct time scales, based on a neural-based multi-expert prediction model. First, an orthogonal wavelet transform is used to decompose the time series into varying levels of temporal resolution so that the underlying temporal structures of the original time series become more tractable. The transitions between the resolution scales are assumed to be governed by a hidden Markov model (HMM). The best state sequence is obtained by the Viterbi algorithm assuming some prior knowledge on the state transition probabilities and energy-dependent observation probabilities. The model achieves a hard segmentation of the time series into distinct dynamical modes and the simultaneous specialization of the prediction experts on the segments. The predictive ability of this strategy is assessed on a synthetic time series.

1 Introduction

The idea of combining both wavelets and neural networks has attracted much attention over the last decade. Wavelet networks [1] are feedforward neural networks with one hidden layer of nodes, whose basis functions are drawn from a family of orthonormal wavelets. They were introduced for neural network dynamic modelling in presence of varying sampling times, sparse and dense data in different regions and the inherent presence of both large and small dynamics. A wavelet network is first trained to learn the mapping at the coarsest resolution level. In subsequent stages, the network is trained to incorporate elements of the mapping at higher and higher resolutions [2, 3]. Another strategy, aimed at combining neural network forecasts on wavelet-transformed time series, has been proposed in [4, 5, 6]. Contrary to the wavelet networks, the wavelet part is essentially decoupled from learning. A neural network is trained on each time-scale to approximate the underlying law governing the wavelet coefficients. The individual wavelet scale forecasts are afterward recombined to form the current estimate. In this paper, we discuss a new framework for modelling the switching dynamics of a time series with correlation structures spanning distinct time scale.

*Supported by the Alliance Franco-British research partnership programme.

The experts are selected according to the energy at the different scales assuming markovian dynamic transitions. The method is illustrated on artificial data.

2 Multiresolution

A multiresolution [7, 8, 9] of $L^2(\mathbb{R})$ is characterized by a set of subspaces V_j and W_j, $j \in \mathbb{Z}$, of $L^2(\mathbb{R})$. The scaling spaces V_j are increasing and the wavelet space W_j is the difference between V_j and V_{j+1}. The space V_{j+1} is the direct sum of V_j and W_j which intersect only at the zero vector. A function $f(t)$ in the whole space has a piece in each subspace. The piece in V_j is $f_j(t)$. On requirement on the sequence of subspaces is *completeness*: $f_j(t) \to f(t)$ as $j \to \infty$. The completeness condition can be restated as $V_0 \oplus \sum_{j=0}^{\infty} W_j = L^2$. In the orthogonal case, any function $f(t) \in L^2(\mathbb{R})$ can be decomposed as $f(t) = f_0(t) + \sum_{j=0}^{+\infty} \Delta f_j(t)$ where $\Delta f_j = f_{j+1} - f_j$ belongs to W_j. Let $\{\psi_{j,k}(t) := 2^{-j}\psi(2^{-j}t - k), k \in \mathbb{Z}\}$ a basis for W_j, in the orthogonal case we may write

$$\Delta f_j(t) = \sum_{k=-\infty}^{+\infty} d_j^k \psi_{j,k}(t) \tag{1}$$

where d_j^k are the detail coefficients at scale j. The energy in this piece is $\sum_{k=-\infty}^{+\infty} |d_j^k|^2$. In our case, $f(t)$ is unknown, only samples are observed at different instants. These observations are gathered in a time series $\{x_t\}$. The task is to predict the future value $y_t = x_{t+\tau}$ of the time series $\{x_t\}$. τ is the delay parameter. Now, at each time t, we compute an orthogonal wavelet transform on a sliding time window of size *wsize* over the past values $t, t-1, \ldots, t-wsize$. The output is a sequence of coefficients $\{d_0^t, d_1^t, \ldots, d_J^t, a_J^t\}$ up to resolution level J, where a_J^t are the ultimate approximations. The proposed prediction method is based on the assumption that the contribution of scale j to the future value $y_t = x_{t+\tau}$ is dependent upon the fraction of the energy at that scale. In our case, the energy estimate at scale j, for $t > wsize$, noted by E_j^t, is given by

$$E_j^t = \sum_{j=t-K}^{t} {d_j^k}^2 \tag{2}$$

where K is a constant and $wsize = 2^{J+K}$. In our experiments, $K = 2$ to speed up the detection of change points.

3 Prediction experts

In the following we assume that the reader is familiar with the basic principles of HMM. For a thorough introduction, we refer to the tutorial by Rabiner [10]. Our presentation closely derives from [11, 12]. Consider an HMM where each state $i = 1, \ldots, M$ is associated to a prediction expert. The prediction expert

i predicts the future value $y_t = x_{t+\tau}$ of the time series $\{x_t\}$ or some exogene variable, given a vector of past values

$$z_j^t = (x_{t-2^{j-1}d}, \ldots, x_{t-2^{j-1}}, x_t). \tag{3}$$

where d is the embedding dimension. As may be seen, the past sequence at time t is summarized by a vector of time delayed coordinates z_j^t whose sampling rate is divided by two as we move from scale j to scale $j+1$. This comes implicitly from the two scale difference in the dilation equation [8], the reason is to restrain the analysis to the frequencies lying the *octave band* of interest. We suppose the target variables y_t, at each time t, are given by some deterministic function, $f_i(z_i^t)$, where i is current dynamic mode, corrupted with an additive gaussian noise ϵ_t. The noise variance, σ_i^2, does not depend on z_i^t or on t. The observation probability distribution for each state j is supposed to be given by

$$P(y_t|s_t = j, x_1^t) = K \cdot \mathrm{e}^{-\alpha E_j^t} \tag{4}$$

where K is the normalizing factor, $x_1^t = x_1, \ldots, x_t$ and α is some constant and E_i^t is the estimate of the energy at time t and at time-scale i. Note that E_i^t are implicit functions of x_1^t. The transition matrix $A = \{a_{ij}\}$ determines the probability to switch from a state i to a state j. In principle, this matrix can be found by the Baum and Welch method [10]. However, since we focus on problems with only relatively few switching events, the matrix is used to incorporate this prior knowledge in such a way that remaining in the current state is as likely than switching to another state (low switching rate assumption [12]) :

$$a_{ij} = \begin{cases} \frac{k}{k+(M-1)} & \text{if } i = j \\ \frac{1}{k+(M-1)} & \text{if } i \neq j \end{cases} \tag{5}$$

With thus get a constant transition matrix which depend on a single parameter k ($k = M-1$ in our experiments). The restriction on models with a low switching rate reduces the degrees of freedom in model space and effectively prevents from overfitting the data. Note that this model we arrive at differ from the so-called the mixture-of-experts architecture [13]. In such model, the input space is divided into a nested set of regions. Training the gating models for $P(s_t|x_1^t)$ is however difficult especially when too few transitions are observed in the data. Instead, we assume here that the state probability at time t can directly be computed from the energy estimates E_i^τ, for $\tau \leq t$ and $i = 1, \ldots, J$, which are calculated other sliding windows. The number of freedom parameters is significantly reduced and the training is greatly facilitated.

4 Expert training

We seek to model the functions $f_i()$ by prediction experts. The goal is to find the maximum likelihood estimator of expert parameters given the observed data $\{x_t\}$. In the architecture used in [11, 12], the time series $\{x_t\}$ has been observed but not the hidden states $\{s_t\}$. The architecture is interpreted as a statistical

model and the training is performed by the Expectation-Maximization (GEM) algorithm [14] algorithm to find the maximum likelihood estimator of the system. In our case, the dynamic states are readily obtained from the input data $\{x_t\}$. It suffices to calculate the energies E_i^t in order to fix all the parameters of the HMM. The optimal state sequence is then obtained by the well-known Viterbi algorithm [15], (i.e., $argmax_{s_1,\ldots,s_T}\ P(s_1,\ldots,s_T|x_1^T,y_1^T)$) by

$$\delta_t^j = K \cdot \mathrm{e}^{-\alpha E_j^t} \cdot \max_i[\delta_{t-1}^i a_{ij}] \tag{6}$$

where $\delta_t^j = \max_{s_1,\ldots,s_{t-1}} P(s_1,\ldots,s_t = j, x_1^t)$ is the best score along a single path, at time t, which account for the first t observations and ends in state j. We finally obtain a *complete* data set in which the observations are labeled with the optimal sequence of states given our assumptions.

4.1 Detection of switching points

Once the predictors are trained, it suffices to detect the current dynamical mode based on the new incoming data and run the corresponding expert to output the forecast. Clearly, mode changes are not detected instantaneously. According to our prior (with $k = 2$), the switching rate is such that the model remains in its current state with probability $1/2$. However, given that the wavelet energies are estimated from the last two wavelet coefficients, the energy profile varies rapidly as a new dynamic takes place. In our experiments, the current mode at time t is selected according to $argmax_j[\delta_{t-1}^j]$ since E_j^t is not yet available.

5 Experiments on synthetic data

To illustrate the predictive ability of the method, we compare our multi-scale approach to the direct approach using a single prediction model fed with a time delay vector. It is important to note that the comparison is made on the basis on the same number of parameters and the same number of inputs. The number of inputs is deliberately small to increase the difficulty of the prediction task. The prediction models used are standard multilayer neural networks (MLP) trained by conjugate gradient techniques. The same MLP of size $3 \times 4 \times 1$ is used to model each prediction expert up to time-scale $J = 6$. This makes a total of 126 adjustable parameters including bias. The MLP used in the direct approach has size $3 \times 25 \times 1$. We generate a synthetic data set made up from a succession of 5 noisy sine functions with exponentially distributed duration periods and distinct frequencies $f_1 = 3.1$, $f_2 = 5.17$, $f_3 = 9.11$, $f_4 = 24.15$ and $f_5 = 39.77$. These frequencies were chosen rather at random so as to spread the signal over the 6 scales. The signal is corrupted by an additive gaussian noise of variance $\sigma^2 = 0.03$. The noise level was taken sufficiently large to enforce the overlapping of time delay coordinates. The signals were generated over a period spanning 20 seconds with a sampling rate of $500Hz$. This makes a total of 10000 values for x_t, half of the data is used for training and the rest for validation. The

multiresolution is performed on 6 octaves over the past 1024 data points. The target is the average of the next 4 data points.

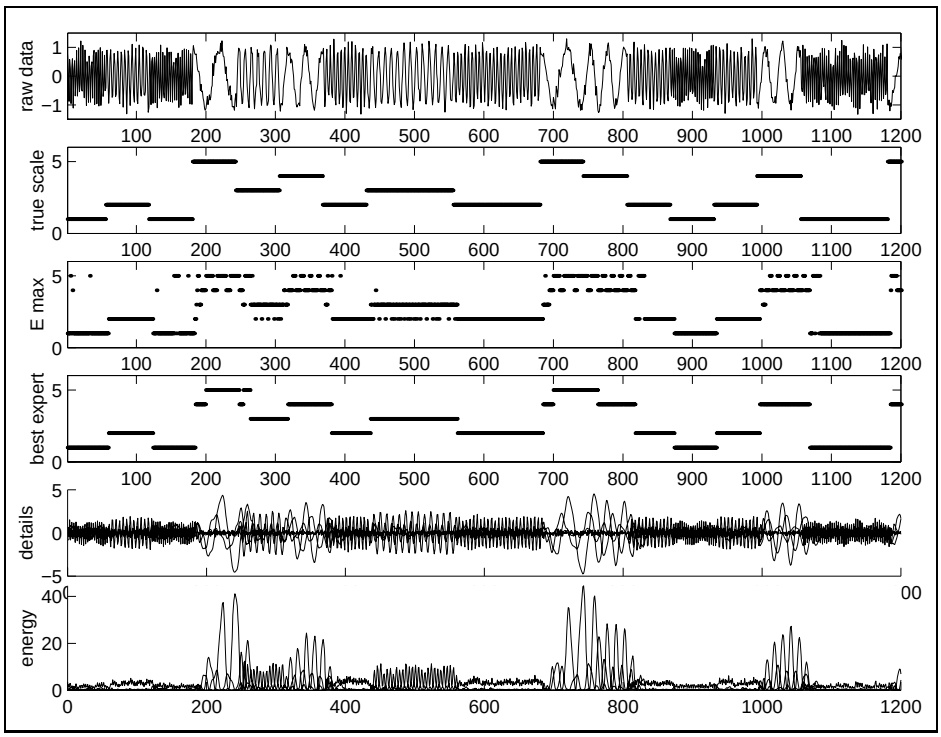

Fig. 1: Random succession of five noisy sine functions. From top to bottom : the raw signal, the true dynamical mode, the scale having maximum energy, the optimal sequence of experts, the wavelet coefficients d_j^t up to scale 6, and the energy estimates.

The results are illustrated in Figure 1. From top to bottom, : the raw signal, the true dynamical mode, the scale having maximum energy, the optimal sequence of experts obtained by the Viterbi algorithm, the wavelet coefficients d_j^t for $j = 1, \ldots, 6$ and the current energy estimate for each scale approximated by the two last wavelet details. As may be observed, the wavelet segmentation approach achieves a very accurate segmentation of the dynamics. Not plotted here, the neural nets achieve accurate forecasts when they are selected as experts and give erroneous predictions outside their own dynamic. The normalized mean squared error (NMSE) on the test set for the reference model is 0.42 and 0.21 for the neuro-wavelet approach. These preliminary results on this toy problem are encouraging given the limited number of inputs and neurons for the experts. Experiments conducted at present on real world teletraffic data will be reported on in due course. Further substantiation through more analysis and experiments are needed to ascertain the data best suited for this approach.

6 Conclusion

We have discussed a new method based on a multiresolution for modelling the switching dynamics of a time series with correlation structures spanning distinct time scales, based on multi-expert prediction models. The best state sequence is obtained by the Viterbi algorithm assuming some prior knowledge on the state transition probabilities and energy-dependent observation probabilities. The model achieves a hard segmentation of the time series into distinct dynamical modes and the simultaneous specialization of the prediction experts on the segments. The predictive ability of this strategy was illustrated on a synthetic time series.

References

[1] Q. Zhang and A. Benveniste. Wavelet networks. *IEEE Trans. on Neural Networks*, 3(6), 1992.

[2] S. Mallat Ch. Bernard and J-J. Slotine. Wavelet interpolation networks. In *European Symposium on Artificial Neural Networks (ESANN'98), Bruges, Belgium*, 1998.

[3] B.L. Zhang and Z.Y. Dong. An adaptive neural-wavelet model for short term load forecasting. *Electric Power Systems Research*, 1(59):121–129, 2001.

[4] A. Aussem, F. Murtagh, and M. Sarazin. Combining neural network forecasts on wavelet-transformed time series. *Connection Science*, 9(1):113–121, 1997.

[5] A. Aussem and F. Murtagh. A neuro-wavelet strategy for web traffic forecasting. *Research in Official Statistics*, 1(1):65–87, 1998.

[6] A. Aussem, J. Campbell, and F. Murtagh. Wavelet-based feature extraction and decomposition strategies for financial forecasting. *Journal of Computational Intelligence in Finance*, 6(2):5–12, 1998.

[7] S. Mallat. A theory for multiresolution signal decomposition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11:674–693, 1989.

[8] G. Strang and T. Nguyen. *Wavelets and Filter Banks*. Wellesley-Cambridge Press, 1996.

[9] J.-L. Starck, F. Murtagh, and A. Bijaoui. *Image Processing and Data Analysis*. Cambridge University Press, 1996.

[10] L.R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceeding of the IEEE*, 77(2):257–285, 1989.

[11] A. Aussem and C. Boutevin. Segmentation of switching dynamics with a hidden Markov model of neural prediction experts. In *Proc. of the European Symposium on Artificial Neural Networks, Bruges, Belgium*, pages 251–256, April 2001.

[12] J. Kohlmorgen, Lemm S., Müller K.-R, Liehr S., and Pawelzik K. Fast change point detection in switching dynamics using a hidden Markov model of prediction experts. In *Proc. of the International Conference on Artificial Neural Networks*, pages 204–208, 1999.

[13] M.I. Jordan and R. A. Jacobs. Hierarchical mixtures of experts and the em algorithm. *Neural Computation*, 6:181–214, 1994.

[14] A. Dempster, N. Laird, and O. D. RubinAgazzi. Maximum likelihood from incomplete data via the em algorithm. *J. Royal Stat. Soc. Series B*, 39:1–38, 1977.

[15] G.D. Forney. The viterbi algorithm. In *Proceedings of the IEEE*, volume 61, 1973.