ESANN'1999 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 21-23 April 1999, D-Facto public., ISBN 2-600049-9-X, pp. 455-462

Hybrid HMM/MLP models for time series prediction

Joseph Rynkiewicz

SAMOS, Université Paris I - Panthéon Sorbonne
Paris, France
rynkiewi@univ-paris1.fr

Abstract

We present a hybrid model consisting of an hidden Markov chain and MLPs to model piecewise stationary series. We compare our results with the model of gating networks (A.S. Weigend et al. [6]) and we show than, at least on the classical laser time series, our model is more parcimonious and give better segmentation of the series.

1 Introduction

A hard problem in time series analysis is often the non-stationarity of the series in the real world. However an important sub-class of nonstationarity is piecewise stationarity, where the series switch between different regimes with finite number of regimes. A motivation to use this model is that each regime can be represented by a state in a finite set and each state match one expert i.e. a multilayer perceptron (MLP). Addressing this problem, we present a class of models consisting of a mixture of experts, so we have to find which expert does the best prediction for the time series. For example A.S. Weigend et al. [6] introduce a gating network to split the input space.

However in this study, we use instead a hidden Markov chain, because it is a powerfull instrument to find a good segmentation, and is therefore usefull in speech recognition. The potential advantage of hidden Markov chains over gating networks is that the segmentation is only local with gating networks (it decides the probability of a state only with it's inputs), but is global with a hidden Markov chain (the probability of the states at each moment depends on all the observations). So we will use this model for the time series forecasting, which has never been done when the model functions are non-linear functions represented by different MLPs.

2 The model

We write y_{t-p+1}^{t} for the vector $(y_{t-p+1}, ..., y_t)$. Let (X_t), $t \in \mathbb{N}$ be an homogeneous, discrete-time Markov chain in $E = \{e_1, ..., e_N\}$, and (Y_t) the series observations in the set of real numbers. At each time the value of X_t determines the distribution of Y_t .
We consider the model at each time t : $Y_{t+1} = F_{X_{t+1}}(Y_{t-p+1}^{t}) + \sigma_{X_{t+1}}\varepsilon_{t+1}$ where $F_{X_{t+1}} \in \{F_{e_1}, ..., F_{e_N}\}$ is a p-order function represented by a MLP with p entries , $\sigma_{X_{t+1}} \in \{\sigma_{e_1}, ..., \sigma_{e_N}\}$ is a strictly positive real number, σ_{e_i} is the standard deviation for the regime defined by X_t, and ε_t an i.i.d normally distributed $\mathcal{N}(0, 1)$ random variable. The probability density of $\sigma_{e_i}\varepsilon$ will be denoted by Φ_i.

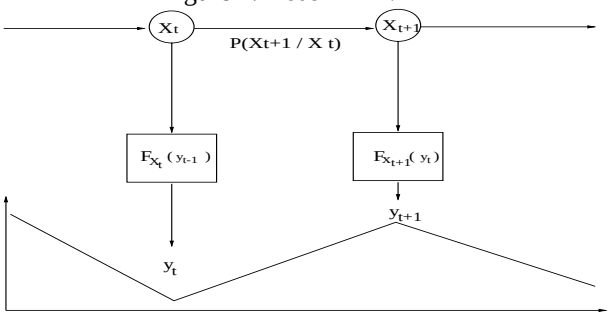
ESANN'1999 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 21-23 April 1999, D-Facto public., ISBN 2-600049-9-X, pp. 455-462

If the state space of (X_t) has N elements it can be identified without loss of generality with the simplex, where e_i are unit vector in $\mathbb{R}^N$ with unity as the ith element and zeros elsewhere. The dynamics of the hidden Markov chain X_t is characterized by the transition matrix $A = (a_{ij})$ with $P(X_{t+1} = e_i / X_t = e_j) = p(e_i/e_j) = a_{ij}$[1]. So if we define : $V_{k+1} := X_{t+1} - AX_t$, we have the following equations for the model :

$$\left\{ \begin{array}{c} X_{t+1} = AX_t + V_{t+1} \\ Y_{t+1} = F_{X_{t+1}}(Y_{t-p+1}^t) + \sigma_{X_{t+1}} \varepsilon_{t+1} \end{array} \right. \quad (1)$$

Moreover, to estimate this model (1) we assume that the initial distribution of the state X_0 is π_0 the uniform distribution on E. Note that conditioning by the initial observations y_{-p+1}^0 will always be implicit.

Figure 1: model HMM/MLP



3 Estimation of the model

3.1 The likelihood

The parameter θ of the model is the vector of weights of the N experts $(w_i)_{1 \leq i \leq N}$, the N standard deviations $(\sigma_i)_{1 \leq i \leq N}$ of each models, the coefficients of the transition matrix $(a_{ij})_{1 \leq i,\, j \leq N}$.
The likelihood of the series $y := (y_1^T)$, for a given path of the hidden Markov chain $x := \{x_t,\ t = 1, ..., T\}$ is :

$$L_\theta(y, x) = \prod_{t=1}^{T} \prod_{i=1}^{N} \left[\Phi_i(y_t - F_{w_i}(y_{t-p}^{t-1})) \right]^{1_{\{e_i\}}(x_t)}$$

$$\times \prod_{t=1}^{T} \prod_{i,j=1}^{N} P_\theta(X_t / X_{t-1})^{1_{\{e_j, e_i\}}(x_{t-1}, x_t)} \times \pi_0(X_0)$$

The likelihood of the series is then :
$E[L_\theta(y, X)] = \sum_x L_\theta(y, x)$, where $\sum_x$ is the sum over all the possibleš paths of the hidden Markov chain.

[1] The traditional notation for a transition matrix is rather $a_{ij} = P(X_{t+1} = e_j / X_t = ei)$ however the transposed notation used here as in Elliott [2] yields us a more convenient notation for the model.

ESANN'1999 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 21-23 April 1999, D-Facto public., ISBN 2-600049-9-X, pp. 455-462

3.2 Maximization of the likelihood

It is easy to show that the exact calculus of the log-likelihood with this method have a complexity of $O(N^T)$ operations, but the E.M. algorithm (Demster et al. [1]) is well suited to find a sequence of parameters which increase the likelihood at each step, and so converge to a local maximum for a very wide class of models and for our model in particular. First we recall the definition of the E.M. algorithm.

3.2.1 E.M. (Expectation/Maximization) algorithm

1. Initialization : Set $k = 0$ and choose θ_0
2. E-Step : Set $\theta^* = \theta_k$ and compute $Q(., \theta^*)$ with
   $Q(\theta, \theta^*) = E_{\theta^*}\left[\ln\left(\frac{L_\theta(y,X)}{L_{\theta^*}(y,X)}\right)\right]$
3. M-Step : Find :
   $\hat{\theta} = \arg\max Q(\theta, \theta^*)$
4. Replace θ_{k+1} with $\hat{\theta}$, and repeat beginning with step 2) until a stopping criterion is satisfied.

The sequence (θ_k) gives nondecreasing values of the likelihood function to a local maximum of the likelihood function . We call $Q(\theta, \theta^*)$ a conditional pseudo-log-likelihood.

3.2.2 Maximization of the conditional pseudo-log-likelihood

Calculus of $Q(\theta, \theta^*)$ (E-Step) for fixed θ^* we have :

$$E_{\theta^*}\left[logL_\theta(y,X) - logL_{\theta^*}(y,X)\right]$$

$$= E_{\theta^*}\left[\sum_{t=1}^{T}\sum_{i,j=1}^{N} 1_{\{e_j,e_i\}}(x_{t-1},x_t) logP_\theta(x_t\,|x_{t-1}\,)\right.$$

$$\left. + \sum_{t=1}^{T}\sum_{i=1}^{N} 1_{\{e_i\}}(x_t)\left[log\Phi_i(y_t - F_{w_i}(y_{t-p}^{t-1}))\right]\right] + Cte$$

So, let : $\omega_t(e_i) = P_{\theta^*}(X_t = e_i\,|y)$ and $\omega_t(e_j,e_i) = P_{\theta^*}(X_{t-1} = e_j,\ X_t = e_i\,|y)$.
We have:
$E_{\theta^*}\left[logL_\theta(y,X) - logL_{\theta^*}(y,X)\right]$

$$= \sum_{t=1}^{T}\sum_{i,j=1}^{N} \omega_t(e_j,e_i) logP_\theta(x_t\,|x_{t-1}\,) + \sum_{t=1}^{T}\sum_{i=1}^{N} \omega_t(e_i)\left[log\Phi_i(y(t) - F_{w_i}(y_{t-p}^{t-1}))\right] + Cte$$

The conditional pseudo-log-likelihood is the sum of two terms U_θ and V_θ, with

$$U_\theta = \sum_{t=1}^{T}\sum_{i,j=1}^{N} \omega_t(e_j,e_i) logP_\theta(X_t\,|X_{t-1}\,)$$

ESANN'1999 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 21-23 April 1999, D-Facto public., ISBN 2-600049-9-X, pp. 455-462

$$V_\theta = \sum_{t=1}^{T} \sum_{i=1}^{N} \omega_t(e_i) \left[log \Phi_i(y(t) - F_{\theta_i}(y_{t-p}^{t-1})) \right]$$

where U_θ depends only on $(a_{ij})_{1 \leq i,\, j \leq N}$, and V_θ depends only on $(w_i)_{1 \leq i \leq N}$ and $(\sigma_i)_{1 \leq i \leq N}$.

To calculate U_θ and V_θ, we compute $\omega_t(e_i)$ and $\omega_t(e_j, e_i)$ with the forward-backward algorithm of Baum et Welch (Rabiner [5]).

Forward dynamic Let : $\alpha_t(e_i) = L_{\theta^*}(X_t = e_i, y_1^t)$ be the propability density of the state e_i and of the observations y_1^t. Then the forward recurrence is the following :

$$\alpha_{t+1}(e_i) = \left(\sum_{j=1}^{N} \alpha_t(e_j) \times P_{\theta^*}(X_{t+1} = e_i \,|X_t = e_j\,) \right) \times \Phi_i^*(y_{t+1} - F_{w_i^*}(y_{t-p+1}^t))$$

Backward dynamic Let : $\beta_t(e_j) = L(y_{t+1}^T \,|X_t = e_j\,)$ the backward recurrence is the following :

$$\beta_t(e_j) = \sum_{i=1}^{N} \Phi_i^*(y_{t+1} - F_{w_i^*}(y_{t-p+1}^t))\beta_{t+1}(e_i) \times P_{\theta^*}(X_{t+1} = e_i \,|X_t = e_j\,)$$

Then we get results :

$$\omega_t(e_i) = \frac{\alpha_t(e_i)\beta_t(e_i)}{\sum_{i=1}^{N} \alpha_t(e_i)\beta_t(e_i)}$$

and

$$\omega_t(e_j, e_i) = \frac{\alpha_t(e_j) P_{\theta^*}(X_{t+1} = e_i \,|X_t = e_j) \Phi_i^*(y_{t+1} - F_{w_i^*}(y_{t-p+1}^t))\beta_{t+1}(e_i)}{\sum_{i,j=1}^{N} \alpha_t(e_j) P_{\theta^*}(X_{t+1} = e_i \,|X_t = e_j) \Phi_i^*(y_{t+1} - F_{w_i^*}(y_{t-p+1}^t))\beta_{t+1}(e_i)}$$

Maximization of the conditional pseudo-log-likelihood (M-step) To maximize the pseudo-log-likelihood, we have to separately maximize U_θ and V_θ .

Maximum of U_θ We find :

$$\boxed{\hat{a}_{ij} = \frac{\sum_{t=1}^{T} \omega_t(e_j, e_i)}{\sum_{t=0}^{T-1} \omega_t(e_j)}}$$

ESANN'1999 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 21-23 April 1999, D-Facto public., ISBN 2-600049-9-X, pp. 455-462

Maximum of V_θ Since the noise is Gaussian we have :

$$V_\theta = \sum_{t=1}^{T} \sum_{i=1}^{N} \omega_t(e_i) \left[\frac{(y_t - F_{w_i}(y_{t-p}^{t-1}))^2}{2\sigma_i^2} + log\left(\sqrt{2\pi}\sigma_i\right) \right]$$

And it is easy to optimize each expert F_{e_i} , by minimizing the cost function weighted by the probability at each time t of the state e_i, so we get

$$\hat{w}_i = \arg\min \sum_{t=1}^{T} \omega_t(e_i) \left[(y_t - F_{w_i}(y_{t-p}^{t-1}))^2 \right]$$

and

$$\hat{\sigma}_i^2 = \frac{1}{\sum_{t=1}^{T} \omega_t(e_i)} \sum_{t=1}^{T} \omega_t(e_i) \left[(y_t - F_{\hat{w}_i}(y_{t-p}^{t-1}))^2 \right]$$

We can then apply the E.M. algorithm, using the calculation and the maximization of the conditional pseudo-log-likelihood.

4 Application to the laser time series

We use here the complete laser time series of "Santa Fe time series prediction and analysis competition". The level of noise is very low, the main source of noise are errors of measurement. We use 11500 patterns for the learning and 1000 patterns for out-of sample data set to validate the estimation. The transition matrix will always be initialized with equal coefficients at the beginning of the learning and we use 10 iterations of the Levenberg-Marquart algorithm to optimize the ponderate cost function.

4.1 Estimation with two experts

We choose to use 2 experts with 10 entries, 5 hidden units, one linear output, and hyperbolic tangent activation functions . Therefore we assume that the hidden Markov chain has two states e_1 and e_2 . The initial transition matrix is :

$$A_0 = \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{pmatrix}$$

We proceed with 200 iterations of the E.M. algorithm. The goal of the learning is to discover the regimes of the series to predict the collapses and the next value prediction of the time series.

ESANN'1999 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 21-23 April 1999, D-Facto public., ISBN 2-600049-9-X, pp. 455-462

4.1.1 Estimation of the conditional probability of the state :

After learning the estimated transition matrix is :

$$\hat{A} = \begin{pmatrix} 0.994 & 0.025 \\ 0.006 & 0.975 \end{pmatrix}$$

Figure 2: The validation series and the conditional probability of states

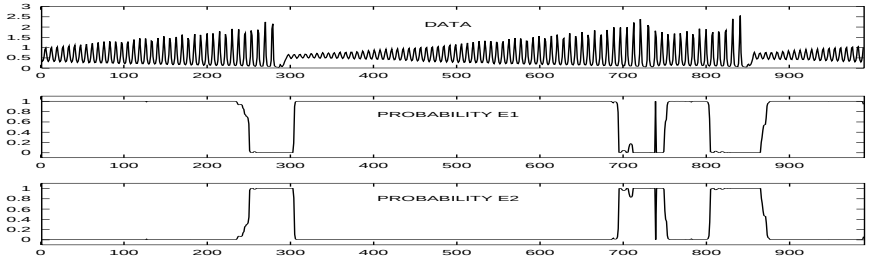


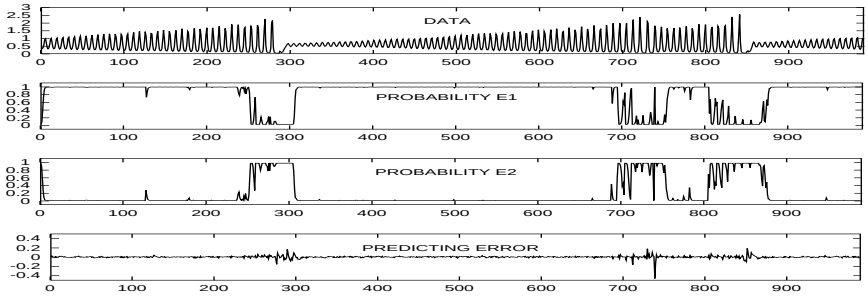
Figure 2 shows the series with the probability of the states conditionally to the observations. We can see that the segmentation is very obvious. The first state matches to the general regime of the series, but the second matches with the collapse regime. Figure 3 deals with the previsions of the state at each time

$$\hat{Q}_{k+1} = A(Q_k)$$

where Q_k is the forward estimation of the state is (with notation of section 3) :

$$Q_k(i) = \frac{\alpha_t(i)}{\sum_{i=1}^{N} \alpha_t(i)}$$

Figure 3: The validation series and the forward prediction of states probability



This prevision is not the same as the conditional probability because here we only use the data until time t to predict the states probability at time t+1. The forecast of the

ESANN'1999 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 21-23 April 1999, D-Facto public., ISBN 2-600049-9-X, pp. 455-462

state probability is clearly not as good as the conditional probability, however we can still predict that when the second state becomes more probable the series will collapse. The forecast of the next point y_{t+1} (single step prediction) is given by :

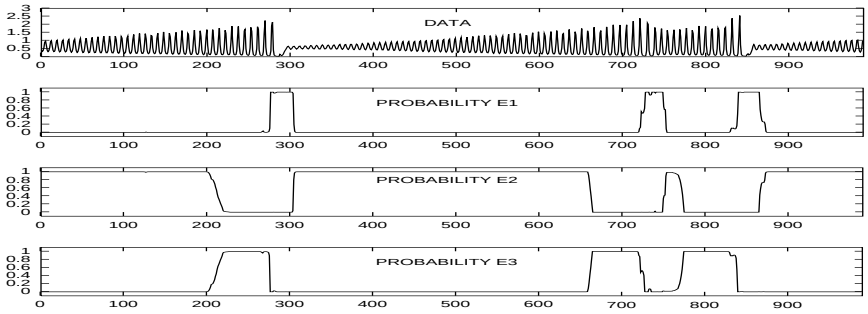
$$\hat{Y}_{t+1} = \sum_{i=1}^{N} \hat{Q}_{k+1}(i) F_i(y_{t-p+1}^{t})$$

The normalized mean square error (E.N.M.S) is then 0.0033.

4.2 Estimation with 3 experts

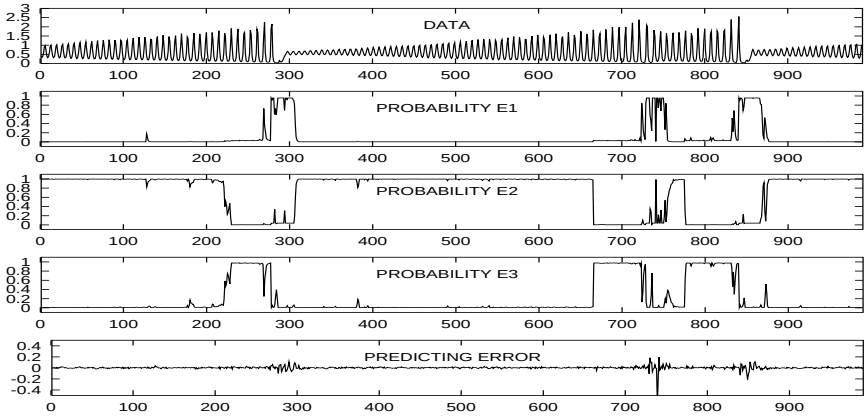
The architecture of experts remains the same. After learning we have below the following results on the validation set.

Figure 4: The validation series and the conditional probability of states



Here the segmentation is still obvious : the second state gives the general regime , the third matches to the pre-collapses, and the first to the collapses.

Figure 5: The validation series with the forward prediction of states probability



The estimated transition matrix is now :

$$\hat{A} = \begin{pmatrix} 0.9548 & 0.0002 & 0.0204 \\ 0.0356 & 0.9955 & 0.0000 \\ 0.0096 & 0.0043 & 0.9796 \end{pmatrix}$$

The E.N.M.S. of the model on the validation set is now 0.0046, it is a little bit more than with two experts. so it is useless to use more experts for the estimation.

5 Conclusion

If we compare with the results of the gating expert of Weigend et al.[6] applied to the laser time series, we can see that we obtain a comparable E.N.M.S., but with many fewer parameters. Indeed the best E.N.M.S. on the validation set is 0.0035 in their case with 6 experts and 0.0033 with 2 experts in our case. Moreover our model gives a much better segmentation of the series. That is to say, the segmentation with the gated expert oscillates always from one state to another (see Weigend et al. [6]), but is very obvious with our model. Finally this model seems to be very promising not only for forecasting but also to predict the change of trends in a wide class of processes like financial crack or avalanche, since the prediction of the next state gives an insight to the future behavior of the process.

References

[1] Demster, N.P., Lair, N.M., Rubin,D.B. Maximum likelihood from incomplete data via the E.M. algorithm. Journal of the Royal statistical society of London, Series B 39:1-38. RSSL 1977

[2] Elliott, R., Aggoun, L., Moore, J. Hidden Markov models : estimation and control, Springer 1997

[3] Morgan, N., Bourlard, H. Connectionist speech recognition : a hybrid approach. Kluwer academic publ., 1994.

[4] Poritz, A.B. Linear predictive hidden Markov models and the speech signals. IEEE transaction on signal processing, Volume 41:N 8:2557-2573. Pergamon, 1982.

[5] Rabiner. L.R. A tutorial on hidden Markov models and selected application in speech application. proceedings of the IEEE, volume 77:257-287. Pergamon, 1989.

[6] Weigend, A., Mangeas, M., Srivastava, A. Nonlinear gated expert for time series : Discovering regimes and avoiding overfitting. International journal of Neural Systems, Volume 6:N 4:373-399, World Scientific publishing, 1995

[7] Weigend, A., Gershenfeld, N. Times series prediction : Forecasting the future and understanding the past. Addison-Wesley 1994

[8] Wu, C. On the convergence property of the E.M. algorithm. Annals of Stat. Volume 11 : 95-103. AMS, 1983