ANALYSIS OF NONSTATIONARY TIME SERIES BY MIXTURES OF SELF-ORGANIZING PREDICTORS

Jens Kohlmorgen, Steven Lemm, Gunnar Rätsch, Klaus–Robert Müller

GMD FIRST
German National Research Center for Information Technology
Institute for Computer Architecture and Software Technology
Kekuléstr. 7, D-12489 Berlin, Germany
E-mail: {jek, lemm, raetsch, klaus}@first.gmd.de
Web: http://www.first.gmd.de

Abstract. We present a method for the analysis of time series from drifting or switching dynamics. In extension to existing approaches that identify switches or drifts between stationary dynamical modes, the method allows to analyze even continuously varying dynamics and can identify mixtures of more than two dynamical modes. The architecture is based on a mixture of self-organizing Nadaraya-Watson kernel estimators. The mixture model is trained by barrier optimization, a technique for constrained optimization problems. We apply the proposed method to artificially generated data and EEG recordings from the wake/sleep transition.

INTRODUCTION

Time series from alternating dynamics are ubiquitous in real-world systems like, for example, speech, climatological data, physiological recordings (EEG, MEG), and financial markets. It is therefore important to find methods that can deal with time-varying dynamical systems, which possibly might also be nonlinear.

In [9, 14], we introduced the annealed competition of experts (ACE) method for time series from nonlinear *switching* dynamics, where an ensemble of neural network predictors specializes on different dynamical regimes by increasing the competition among the predictors using a deterministic annealing scheme. Related approaches for switching dynamics can be found, e.g., in [1, 3, 5, 8, 12, 15]. In [10], we extended the ability to describe a mode change not only as a switching, but, if appropriate, also as a continuous drift from one predictor to another, and found that physiological signals can be modeled more appropriately by a drifting dynamics model [11].

In this paper we present a different and, compared to [10, 11], rather simple and straightforward approach for the analysis of switching and drifting dynamics. Furthermore, the method is even able to analyze continuously varying dynamics that do not contain any stationary segments, and the mixture dynamics may consist of more than two components. In the following, we present the new method, which is based on a mixture of self-organizing kernel estimators, and apply this approach to artificially generated data and EEG recorded from the wake/sleep transition of a human subject.

THE ALGORITHM

Consider a time series from a nonstationary dynamical system that consists of pairs of input and target data, $\{(\vec{x}_t, y_t)\}$, $1 \leq t \leq T$. In particular, the target data might be a future value of a scalar time series $\{x_t\}$, that is $y_t = x_{t+\tau}$, and the input data might be a d-dimensional vector of past values $\vec{x}_t = (x_t, x_{t-\tau}, \ldots, x_{t-(d-1)\tau})$. This is the typical formulation of a time series prediction problem. The parameter d is called the embedding dimension and τ is called the delay parameter. Note that the extension to multivariate time series is straightforward. For simplicity, however, we restrict ourselves to the scalar notation.

The model

The basic idea of this approach is to model the dynamical system by a time-varying mixture of potentially nonlinear predictors f_s,

$$\hat{y}(t) = \sum_{s=1}^{N} p_{s,t}\, f_s(\vec{x}_t), \tag{1}$$

where $\hat{y}(t)$ is an estimate for the target y_t. Without any constraints regarding the mixing coefficients $p_{s,t}$ and the predictors f_s, there are infinitely many, qualitatively different solutions for fitting the data. For example, given arbitrary predictors f_s and $N-1$ arbitrarily chosen values for all except one $p_{s,t}$ at time t, one can still find a perfect fit for y_t simply by solving (1) for the single remaining parameter $p_{s,t}$. In all previous approaches to this problem, the $p_{s,t}$ were therefore not simply *parameters* to be estimated but time-independent, parameterized *functions* g_s (called gating functions) of either the input $\vec{x}_t$ [7], the input $\vec{x}_t$ and some internal state s_t [1, 3, 15], the prediction performance $(y_t - f_s(\vec{x}_t))^2$ of the individual predictors [5, 8, 14], or even all of these quantities [12]. The above methods consider a *switching* model and assume that only a single predictor is responsible for generating the data at each time step. Moreover, the functions g_s have interpretations as probability functions, namely the conditional probability that expert s

exclusively has generated the data at time t, given the respective quantities. In [10, 11], the actual mixing coefficients are therefore determined separately in a second stage, in order to model mixing dynamics.

We now present a more straightforward, one-stage approach for estimating mixing dynamics: for a given time series $\{(\vec{x}_t, y_t)\}$, $1 \leq t \leq T$, and a number of experts $s = 1 \ldots N$, the mixing proportions $p_{s,t}$ are simply *parameters* to be estimated by an optimization procedure, subject to the following constraints:

$$\sum_{s=1}^{N} p_{s,t} = 1, \quad \forall t \qquad \text{and} \qquad 0 \leq p_{s,t} \leq 1, \quad \forall s,t. \tag{2}$$

These constraints are applied in order to restrict the space of possible solutions, as already discussed. Furthermore, they permit to model the *convex hull* of the underlying multi-modal dynamical system. It includes the cases of switching and drifting dynamics considered in previous work, and, in extension to [10, 11], it allows us to represent mixtures of more than two predictors and is not restricted to a fixed number of discrete mixture states. It even allows us to analyze continuously drifting dynamics without any stationary periods. Note that in the context of mixtures, the $p_{s,t}$ are mixing factors and not probabilities of the individual experts. However, this framework can also be used for merely a switching model, and then the $p_{s,t}$ again have a probabilistic interpretation. Due to the limited space, we will not consider this variant here.

Before we discuss the optimization technique, let us first consider the second part of the model, the function approximators f_s. They represent the set of base dynamics of the model. In general, function approximators contain parameters that need to be adapted to the data. In our case, these parameters would add to the typically already large number of parameters $p_{s,t}$, making the optimization problem much harder. We found, however, an elegant way to introduce function approximators without introducing any new adaptive parameters: we use Nadaraya-Watson kernel estimators [2],

$$f(\vec{x}) = \frac{\sum_{t=1}^{T} y_t \, K_\sigma(\vec{x}, \vec{x}_t)}{\sum_{t=1}^{T} K_\sigma(\vec{x}, \vec{x}_t)}, \tag{3}$$

with a Gaussian kernel for each data point in the training set $\{(\vec{x}_t, y_t)\}$,

$$K_\sigma(\vec{x}, \vec{x}_t) = \exp\left(-\frac{(\vec{x} - \vec{x}_t)^2}{2\sigma^2}\right). \tag{4}$$

The kernel width σ determines the smoothness of the estimator. It is the only free parameter of the estimator and we use it in the following as a fixed smoothness prior. In principle, it might also be adapted during training. In that case, however, care has to be taken to prevent σ from getting too small, which clearly would lead to overfitting.

Since we do not want to estimate a single global predictor but individual prediction experts for different dynamical modes in the data set, we obtain individual experts by weighting each data point in the kernel estimator with the respective mixing proportion $p_{s,t}$,

$$f_s(\vec{x}) = \frac{\sum_{t=1}^{T} y_t \; K_\sigma(\vec{x}, \vec{x}_t) p_{s,t}}{\sum_{t=1}^{T} K_\sigma(\vec{x}, \vec{x}_t) p_{s,t}}. \tag{5}$$

This corresponds to a self-organization of the experts during the optimization of the parameters $p_{s,t}$. In the case of switching dynamics, the estimators would contain exactly the subset of data points they are assigned to, and therefore would simultaneously represent the prediction functions for the respective modes. The more a data point represents a mixture of two or more dynamics, however, the less it is suited to contribute to any of the predictors, since it contains “noise” from the other components. In fact, it turned out that the linear weighting of the data points in (5) is not sufficient to suppress the contribution of mixed (“noisy”) data in the estimators. Therefore, we introduced a nonlinearly weighted estimator,

$$f_s(\vec{x}) = \frac{\sum_{t=1}^{T} y_t \; K_\sigma(\vec{x}, \vec{x}_t) (p_{s,t})^\alpha}{\sum_{t=1}^{T} K_\sigma(\vec{x}, \vec{x}_t) (p_{s,t})^\alpha}. \tag{6}$$

Hence, the contribution of data points from mixtures, $p_{s,t} < 1$, becomes smaller, the larger α is chosen. In our experiments, $\alpha = 2$ turned out to be a good choice for the analysis of mixture dynamics. Note that for switching dynamics, $\alpha = 1$ is already sufficient.

Optimization

Fitting the set of parameters $\theta = \{p_{s,t} : s = 1, \ldots, N; t = 1, \ldots, T\}$ of the above mixture model for a given data set, can be formulated as a constrained optimization problem. The objective function to be minimized is given by

$$E(\theta) = \sum_{t=1}^{T} \left(y_t - \sum_{s=1}^{N} p_{s,t} \, f_s(\vec{x}_t) \right)^2 + C \sum_{t=1}^{T-1} \sum_{s=1}^{N} (p_{s,t+1} - p_{s,t})^2. \tag{7}$$

It is the sum of squared prediction errors plus an additional regularization term, weighted by the constant C, that penalizes changes of the mixing coefficients in time. This is necessary to avoid local minima of the objective function and imposes another smoothness prior: solutions with a simple temporal structure are more likely than those with frequent changes. In fact, our goal is not only to minimize the prediction error of the training data, but also to find a simple model with respect to the dynamical structure.

Minimizing $E(\theta)$ is subject to the constraints in (2). This constrained optimization problem can be solved by a technique called barrier optimization [4, 6]. To this end, the constraints need to be transformed into a set of inequalities of the form $c_i(\theta) \leq 0$, $i = 1, \ldots, m$,

$$\begin{array}{lll} -1 + \sum_{s=1}^{N} p_{s,t} \leq 0, & 1 - \sum_{s=1}^{N} p_{s,t} \leq 0, & \forall t \\ -p_{s,t} \leq 0, & p_{s,t} - 1 \leq 0, & \forall s, t. \end{array}$$

The constrained optimization problem can now be solved by using the so-called barrier (or penalty) error function

$$E_\beta(\theta) = E(\theta) + \sum_{i=1}^{m} \kappa_\beta(-c_i(\theta)), \tag{8}$$

where κ_β is a suitable barrier function and $\beta > 0$ is the penalty parameter. Typical choices for κ_β are $\kappa_\beta(t) = -\beta \log(t)$ [6] or $\kappa_\beta(t) = \beta \exp(-t/\beta)$ [4]. Note that by using the log-barrier, the optimization has to start with a feasible θ, i.e. with all inequalities already being satisfied, while the exp-penalty does not need this condition [4]. Although it is not a problem to find a feasible θ to start with in our case, we nevertheless prefer the exp-penalty.

For a given starting value of β, the function $E_\beta(\theta)$ is minimized using an unconstrained optimization technique. We use the conjugate gradient (CG) method with line-search [2]. After the optimization step, β is decreased according to $\beta := \beta^r$, $r > 1$, and the optimization procedure restarts with the decreased β and the solution θ found in the previous step until a stopping criterion, e.g. an error threshold or a final value of β, has been reached.

In the case of our mixture model, the resulting $\theta = \{p_{s,t}\}$ represents the drift/switch/mix-segmentation of the time series, and, at the same time, the set of Nadaraya-Watson prediction experts for the extracted dynamical modes.

APPLICATIONS

To illustrate our approach, two examples of artificially generated drifting dynamics are discussed first. We then present an application to real-world data: an EEG recording of the wake/sleep transition of a human subject.

Drifting Dynamics of a Mackey-Glass System

We generated time series from drifting dynamics using the Mackey-Glass delay differential equation,

$$\frac{dx(t)}{dt} = \gamma_{t_d} = -0.1x(t) + \frac{0.2x(t-t_d)}{1 + x(t-t_d)^{10}}. \tag{9}$$

It is a high-dimensional chaotic system that was originally introduced as a model of blood cell regulation [13]. In the first example, three stationary operating modes, A, B and C, are established by using different delays, $t_d =$ 17, 23, and 30, respectively. After operating 100 time steps in mode A (with respect to a subsampling step size $\Delta = 6$), the dynamics switches to a mixture of modes A, B, and C. The mixture dynamics is generated for the next 100 time steps by mixing the equations for $t_d = 17$, 23, and 30,

$$\frac{dx(t)}{dt} = a\,\gamma_{17} + b\,\gamma_{23} + c\,\gamma_{30}, \tag{10}$$

using $a = 0.6$, $b = 0.3$, and $c = 0.1$. Thereafter, the system runs stationary in mode B for the following 100 time steps ($t = 201, \ldots, 300$), whereupon it switches to a new mixture, $a = 0.2$, $b = 0.3$, and $c = 0.5$, until it reaches $t = 400$. Finally, from $t = 401, \ldots, 500$, it runs stationary in mode C.

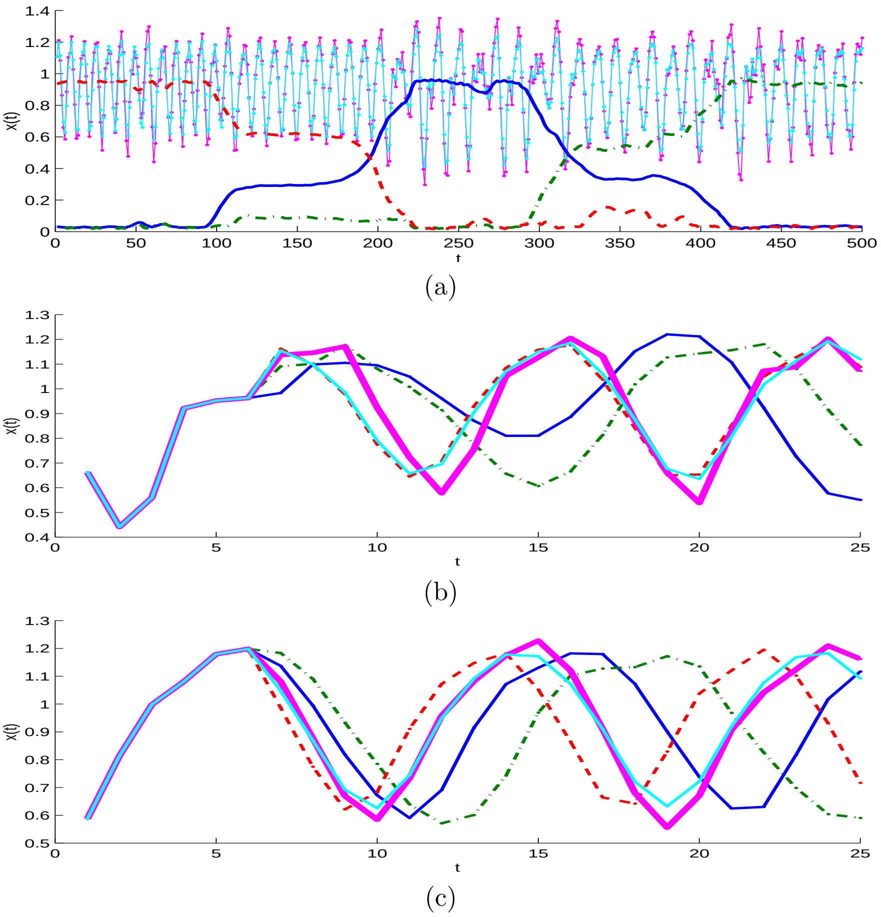
Next, we applied the barrier optimization method for the mixture model, using (7) and $N = 3$ predictors. The input to each predictor is a vector $\vec{x}_t$ of time-delay coordinates of the scalar time series $\{x_t\}$. The embedding dimension is $d = 6$ and the delay parameter is $\tau = 1$ on the subsampled data. The penalty parameter is annealed from $\beta = 0.5$ to 0.001. The other parameters are $\alpha = 2$, $C = 0.75$, $\sigma = 0.25$.

The result of the optimization is depicted in Fig. 1a. The three predictors have specialized on the prediction of the dynamics of modes A, B, and, C, respectively. The two intermediate mixture parts are represented as mixtures of the predictors and the found mixing proportions nicely agree with the real coefficients from $t = 100, \ldots, 200$. In the second mixture part, $t = 300, \ldots, 400$, the coincidence is similar, but not so perfect. However, considering the fact that only 500 data points of a rather complicated dynamical mixture system are given, the overall result is remarkably good. The long-term prediction performance by feeding back the single-step predictions into the predictors is shown in Fig. 1b,c. In Fig. 1b, the prediction starts at $t = 60$ (mode A). We iterated the predictors individually (thin black line, dash-dotted line, dashed line) and the whole ensemble, for which we used the mixing proportions found at $t = 60$ (grey line). Since the dashed predictor clearly dominates the mixture, its output is almost identical to that of the ensemble. Both generated continuations are similar to the target dynamics (thick line), whereas the continuations of the two other predictors are not. Fig. 1c shows the respective predictions for the mixed dynamics at $t = 150$. As expected, only the ensemble yields a good long-term prediction and the individual predictors do not. In fact, we found that the ensemble of predictors is able to reconstruct the dynamics of all five modes very well.

Next, we consider the case of continuously drifting dynamics. We used (10) with time-varying mixing factors $a(t) = 0.5 + 0.5 \sin(\pi t/100)$, $b(t) = 1 - a(t)$, and $c(t) = 0$. Thus, the generated time series is a continuous, sine-shaped drift between two modes A and B with the period 200. We used two experts and $T = 400$ data points. The other parameters were chosen as before. The respective result is shown in Fig. 2. The sine-shape was nicely found and the predictors even captured the dynamics of mode A and B, respectively, although there were no stationary parts of A or B in the data.

Wake/Sleep EEG

In [11], we analyzed EEG data recorded from the wake/sleep transition of humans. The objective was to provide an unsupervised method to detect the sleep onset and to give an approximation of the signal dynamics, ultimately to be used in diagnosis and treatment of sleep disorders. We applied the method proposed in this paper to the data in order to find out whether we

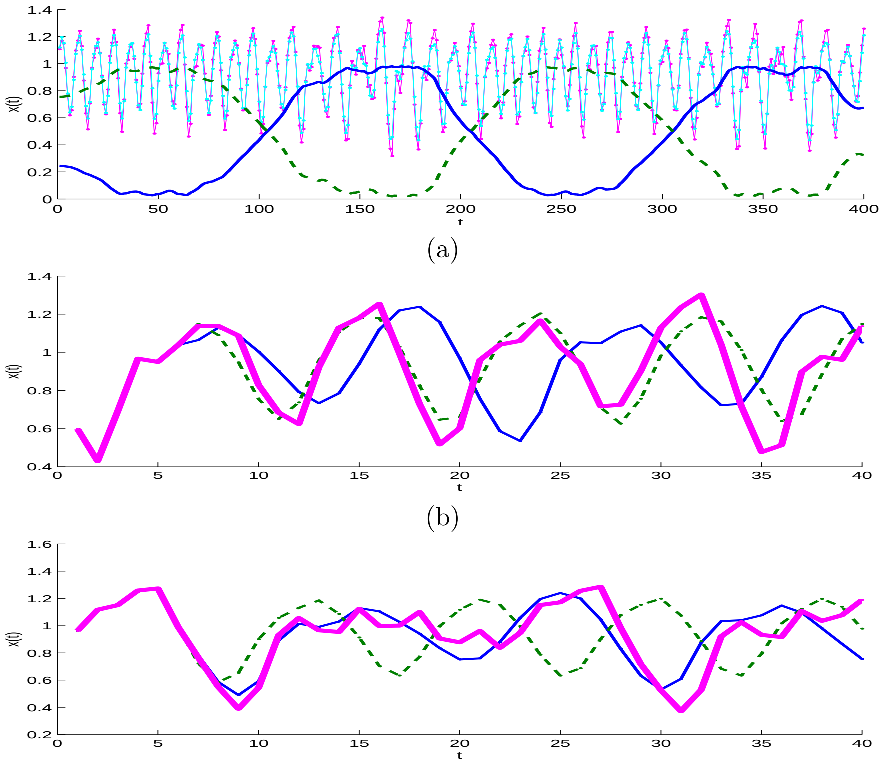


(a)

(b)

(c)

Figure 1: (a) Segmentation of a Mackey-Glass time series with two mixture states between $t = 100$ and 200, and $t = 300$ and 400. The prediction of the ensemble (thin grey line) is printed on top of the data (black dots). The obtained mixing proportions $p_{s,t}$ of the three expert predictors are plotted as dashed, dash-dotted, and solid line, respectively. They nicely correspond to the original proportions. (b) Iterated predictions of the individual experts (thin black line, dash-dotted line, dashed line) and of the whole ensemble (grey line) starting at $t = 60$ (mode A). The dashed predictor, and therewith the ensemble, fits the dynamics of mode A (thick line) very well. (c) Same as (b), but for $t = 150$ (mixture dynamics). Only the ensemble (grey line) properly predicts the long-term behavior of the system (thick line), whereas the individual predictors do not.



(a)

(b)

(c)

Figure 2: (a) Segmentation of a Mackey-Glass time series with a continuous sine-shaped drift between two operating modes. The prediction of the ensemble (thin line) is printed on top of the data (black dots). The obtained mixing proportions $p_{s,t}$ of the two experts are drawn as dashed and solid line. They largely agree with the sine-drift in the data. (b) Iterated predictions of the individual experts (dashed and thin solid line) for data that were not in the training set: a stationary time series from mode A (thick line). The dashed curve clearly resembles the dynamics of mode A. (c) Iterated predictions for a stationary time series from mode B. Here, the thin solid line is very similar to the dynamics of mode B (thick line).

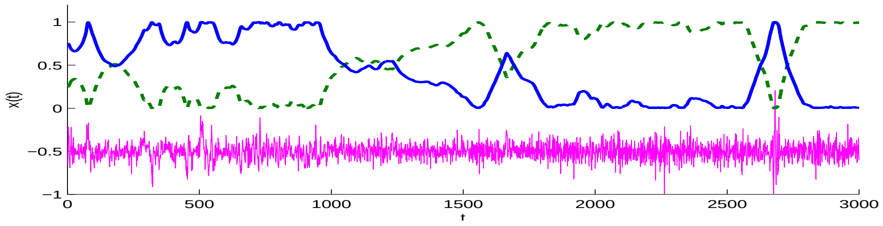


Figure 3: Segmentation of EEG data (thin line) from the wake/sleep transition. The obtained segmentation (above) is in good agreement with a manual segmentation by a medical expert and our previous analysis (see text).

could get similar results as in [11], which would support our previous findings.

The data was measured during an afternoon nap of a healthy human subject. As in [11], we analyzed data from a single-channel EEG recording from position O1. The embedding for the predictors was $\tau = 2$ and $d = 4$ on the raw 100 Hz data. In order to reduce the amount of data, we subsampled the obtained training data set by the factor 10 and chose a sequence of $T = 3000$ data points, such that the sleep onset is roughly at $t = 1000$. We applied the new method to this data set using two predictors. The penalty parameter was annealed from $\beta = 0.5$ to 0.001. The other parameters were $\sigma = 0.3$, $\alpha = 2$, and $C = 6$.

The resulting segmentation is depicted in Fig. 3. Roughly the first 1000 points are mainly assigned to one predictor (thick solid line). This corresponds to the wake phase. The next 2000 points are mainly assigned to the second predictor (dashed line), which corresponds to the sleeping phase. Moreover, there is a clear transition at the sleep onset at $t = 1000$, first to a mixture level of about 50%, then there is a decay of the first mixing proportion (solid line) to zero at $t \approx 1550$. This transition behavior nicely coincides with the results in [11]. However, the subsequent drift back to the wake-state predictor at $t \approx 1700$ is neither indicated in [11] nor in a manual segmentation by a medical expert, where $t = 1000, \ldots, 2000$ is assigned to sleep stage I. On the other hand, the more prominent transition to the wake-state predictor at $t = 2700$, is clearly indicated in the manual segmentation as an intermediate arousal. Note that the interval between $t = 50$ and 200, where the mixing proportions indicate a drift from the wake- towards the sleep-state predictor, is marked as artifact in the manual segmentation.

To summarize, except at $t \approx 1700$, the obtained segmentation of the EEG data is in good agreement with both the manual segmentation and the previous analysis in [11]. It demonstrates that our approach can find meaningful structure in complex real-world data.

SUMMARY AND DISCUSSION

A method for the unsupervised segmentation and identification of nonstationary drifting dynamics was presented. It applies to time series of dynamical systems that drift or switch among various operating modes. In contrast to previous approaches, a given time series does not necessarily need to contain stationary periods. Moreover, mixtures of more than two predictors are possible. On the other hand, if one uses more prediction experts than necessary, then the model has too many degrees of freedom and may fit the data in various ways. How to find the appropriate number of predictors efficiently is still an open question and so far requires the repeated application of the method with different numbers of predictors and then choosing the least complex ensemble among the solutions with the lowest error $E(\theta)$.

The application to wake/sleep EEG demonstrated that meaningful structure in real-world data can be found by this approach. We also expect useful applications of this method in other fields where complex, nonstationary dynamics plays an important role, like e.g. in climatology, in industrial applications, or in finance.

Acknowledgement We would like to thank J. Rittweger for providing the physiological data and acknowledge partial financial support of the Deutsche Forschungsgemeinschaft (grant Ja379/52).

REFERENCES

[1] Bengio, Y., Frasconi, P. (1995). An Input Output HMM Architecture. In: *NIPS'94: Advances in Neural Information Processing Systems* 7 (eds. G. Tesauro, D.S. Touretzky, T.K. Leen), Morgan Kaufmann, 427–434.

[2] Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*, Oxford University Press, NY.

[3] Cacciatore, T. W., Nowlan, S. J. (1994). Mixtures of Controllers for Jump Linear and Non-linear Plants. In *NIPS '93*, (eds. J.D. Cowan, G. Tesauro, J. Alspector), Morgan Kaufmann, 719–726.

[4] Cominetti, R., Dussault, J.-P. (1994). A stable exponential penalty algorithm with superlinear convergence. *J.O.T.A.*, 83:2.

[5] Fancourt, C., Principe, J. C. (1996). A Neighborhood Map of Competing One Step Predictors for Piecewise Segmentation and Identification of Time Series. In *ICNN '96: Proc. of the Int. Conf. on Neural Networks*, vol. 4, 1906–1911.

[6] Frisch, K.R. (1955). The logarithmic potential method of convex programming. Memorandum, University Institute of Economics, Oslo.

[7] Jacobs, R. A., Jordan, M. I., Nowlan, S. J., Hinton, G. E. (1991). Adaptive Mixtures of Local Experts. *Neural Computation* 3, 79–87.

[8] Kehagias, A., Petridis, V. (1997). Time Series Segmentation using Predictive Modular Neural Networks. *Neural Computation* 9, 1691–1710.

[9] Kohlmorgen, J., Müller, K.-R., Pawelzik, K. (1995). Improving short-term prediction with competing experts. In *ICANN '95: Proc. of the Int. Conf. on Artificial Neural Networks*, EC2 & Cie, Paris, 2:215–220.

[10] Kohlmorgen, J., Müller, K.-R., Pawelzik, K. (1997). Segmentation and Identification of Drifting Dynamical Systems. In *NNSP '97*, (eds. J. Principe, L. Giles, N. Morgan, E. Wilson), IEEE, 326–335.

[11] Kohlmorgen, J., Müller, K.-R., Rittweger, J., Pawelzik, K. (2000). Identification of Nonstationary Dynamics in Physiological Recordings, *Biological Cybernetics* 83(1), 73–84.

[12] Liehr, S., Pawelzik, K., Kohlmorgen, J., Müller, K.-R. (1999). Hidden Markov Mixtures of Experts with an Application to EEG Recordings from Sleep. *Theory in Biosciences* 118, 246–260.

[13] Mackey, M., Glass, L. (1977). Oscillation and Chaos in a Physiological Control System. *Science* 197, 287.

[14] Pawelzik, K., Kohlmorgen, J., Müller, K.-R. (1996). Annealed Competition of Experts for a Segmentation and Classification of Switching Dynamics. *Neural Computation* 8(2), 340–356.

[15] Shi, S., Weigend, A. S. (1997). Taking Time Seriously: Hidden Markov Experts Applied to Financial Engineering. In *CIFEr '97: Proc. of the Conf. on Computational Intelligence for Financial Engineering*, IEEE, NJ, 244–252.