

# Identification of nonstationary dynamics in physiological recordings

J. Kohlmorgen<sup>1</sup>, K.-R. Müller<sup>1</sup>, J. Rittweger<sup>2</sup>, K. Pawelzik<sup>3</sup>

<sup>1</sup> GMD FIRST, Rudower Chaussee 5, 12489 Berlin, Germany

<sup>2</sup> Institute for Physiology, FU Berlin, 14195 Berlin, Germany

<sup>3</sup> Institute for Theoretical Neurophysics, University of Bremen, Kufsteinerstrasse, 28334 Bremen, Germany

Received: 5 May 1999 / Accepted in revised form: 28 December 1999

**Abstract.** We present a novel framework for the analysis of time series from dynamical systems that alternate between different operating modes. The method simultaneously segments and identifies the dynamical modes by using predictive models. In extension to previous approaches, it allows an identification of smooth transition between successive modes. The method can be used for analysis, diagnosis, prediction, and control. In an application to EEG and respiratory data recorded from humans during afternoon naps, the obtained segmentations of the data agree with the sleep stage segmentation of a medical expert to a large extent. However, in contrast to the manual segmentation, our method does not require a priori knowledge about physiology. Moreover, it has a high temporal resolution and reveals previously unclassified details of the transitions. In particular, a parameter is found that is potentially helpful for vigilance monitoring. We expect that the method will generally be useful for the analysis of nonstationary dynamical systems, which are abundant in medicine, chemistry, biology and engineering.

## 1 Introduction

The analysis of time series from nonlinear dynamical systems made considerable progress during the last decades (Mayer-Kress 1986; Kantz and Schreiber 1997). In particular, the notion of deterministic chaos, concepts like the fractal dimensionality of attractors, and measures of predictability and complexity (Badii and Politi 1997) allowed a deeper understanding of complex phenomena. A major breakthrough came with the discovery that a measured time series carries the information necessary to estimate the above-mentioned quantities (Takens 1981). In particular, the method of *embedding* using time-delay coordinates, first introduced by Packard et al. (1980),

provided a general tool for the identification and analysis of complex systems in terms of low-dimensional chaotic systems. Besides the fact that the statistical techniques in this regard need a reasonable amount of data, they also generally assume stationarity: they require that the underlying system is autonomous and does not change its parameters over time. If, however, the parameters of the system are drifting or externally switched from time to time, then an analysis of the system can become very difficult. One approach to solve this problem is the application of algorithms to short segments of the data, thereby monitoring possible changes in the characteristic quantities. However, such methods may suffer from the curse of dimensionality and other statistical problems that arise when estimating from few data points.

Here we suggest a different approach. We assume that the dynamics observed from a given system can be described by a discrete set of different dynamics, which we call modes, and by transitions between these modes. This is the case when parameters drift or switch among these modes, or, for example, in high-dimensional systems that exhibit moderately complex dynamics most of the time, intermittently disrupted by irregular bursts (Kaneko 1989), after which a different but again relatively simple dynamics is exhibited. We believe that similar nonstationarities are present in many complex systems.

As an intuitive example consider speech. The speech signal is given by time-varying air pressure originating from a dynamical system, the articulatory organ, which is 'externally' driven through a sequence of configurations. Given the signal, it would be very useful to identify those segments that are well characterized by a particular mode and by transitions between them (Müller et al. 1995). Another example is provided by physiological data from humans in different stages of wakefulness. In particular, the electroencephalogram (EEG) exhibits a complex dynamics that contains different characteristic waveforms during different stages of wakefulness and sleep.

Nonstationarities of this kind constitute the problem this work aims to solve: to find segments in the data that

are well described by a quasi-stationary dynamics and to characterize the transitions between these modes. Once the segmentation is found, the estimated modes can be interpreted, for example, as phonemes in the above example of speech, or as sleep stages in the case of physiological data from a sleep laboratory.

In previous work we developed an approach for analyzing change points in *switching* dynamics from speech (Müller et al. 1994, 1995a), medical (Müller et al. 1995b), and chaotic signals (Kohlmorgen et al. 1994, 1995; Kohlmorgen 1998). Analysis of switching dynamics has also been considered by several other authors (Cacciatore and Nowlan 1994; Hamilton 1994; Bengio and Frasconi 1995; Weigend and Mangeas 1995; Kehagias and Petridis 1997; Shi and Weigend 1997). In our present study, we propose a novel unsupervised technique with a high time resolution for the analysis of time series that include *drifts*: with this method it is possible to describe a mode change not simply as an abrupt switching, but, if appropriate, also as a smooth transition from one dynamics to another. In Sect. 2, we address the unsupervised analysis of drifting time series within the framework of competing prediction experts. In Sect. 3, we apply the algorithm to synthetic data and show that the method reveals the dynamical structure that is hidden in the data and yields the desired representation of the dynamics. The analysis of physiological data from the sleep onset in Sect. 4 demonstrates that our method reveals drifts also in natural data. Moreover, the unsupervised extraction of dynamical structure in physiological recordings may contribute to a better understanding of the transient behavior that emerges when falling asleep. Finally, we discuss the results and give an outlook.

## 2 A method for the detection of drifting dynamics

The analysis of drifting dynamics is performed in two steps. First, an unsupervised hard segmentation method, the Annealed Competition of Experts algorithm (ACE) (Pawelzik et al. 1996; Kohlmorgen 1998), is applied. In this approach, an ensemble of competing prediction experts  $f_i, i = 1, \dots, N$ , is trained on a given time series. The optimal choice of function approximators  $f_i$  depends on the specific application. In this work, we use radial basis function (RBF) networks of the Moody-Darken type (Moody and Darken 1989) as predictors, because they offer a fast and robust learning method. The second step consists in the application of a drift segmentation algorithm that is based on a hidden Markov model.

### 2.1 Training the experts

Consider a time series that consists of pairs of input and target data,  $\{(\vec{x}_t, y_t)\}$ . In particular, the target data might be a future value of a scalar time series  $\{x_t\}$ , for example,  $y_t = x_{t+\tau}$ , and the input data might be a vector of past values  $\vec{x}_t = (x_t, x_{t-\tau}, \dots, x_{t-(d-1)\tau})$ . This is the usual

formulation of a time series prediction problem. The parameter  $d$  is called the embedding dimension and  $\tau$  is called the delay parameter. Note that the extension to multivariate time series is straightforward. For readability, however, we restrict ourselves to the scalar notation in the following.

At each time step  $t, 1 \leq t \leq T$ , each expert  $i$  provides a prediction  $\hat{y}_t^i = f_i(\vec{x}_t)$  for the target  $y_t$ . In our applications, we use normalized RBF networks (Moody and Darken 1989) as prediction experts,

$$f_i(\vec{x}) = \frac{\sum_{h=1}^H w_{ih} g_h(\vec{x})}{\sum_{h=1}^H g_h(\vec{x})}, \quad (1)$$

where each  $g_h(\vec{x})$  is a Gaussian basis function,

$$g_h(\vec{x}) = \exp\left(-\frac{\|\vec{x} - \vec{\mu}_h\|^2}{2\sigma_h^2}\right). \quad (2)$$

The number of basis functions,  $H$ , determines the complexity of the model and has to be set in advance. As suggested in Moody and Darken (1989), the centers  $\vec{\mu}_h$  are determined by a clustering algorithm on the input data,<sup>1</sup> and the widths  $\sigma_h$  are then computed for each  $\vec{\mu}_h$  as the distance to the nearest neighboring cluster center. The  $w_{ih}$  are adapted during the training process, which will be derived in the following.

Under a Gaussian assumption, the probability that a particular predictor  $i$  might have produced the observed data  $y_t$  is given by

$$p(y_t|i) = \sqrt{\frac{\beta}{\pi}} \exp(-\beta(y_t - f_i(\vec{x}_t))^2). \quad (3)$$

If we assume that the experts are mutually exclusive and exhaustive, we can write  $p(y_t) = \sum_{i=1}^N p(y_t|i)p(i)$ . We further assume that the experts are – a priori – equally probable,

$$p(i) = 1/N. \quad (4)$$

To train the experts, we want to maximize the likelihood that the ensemble would have generated the time series. This can be done by a gradient method. For the derivative of the log-likelihood  $\log L = \log(p(y_t))$  with respect to the output of an expert, we get

$$\frac{\partial \log L}{\partial f_i} \propto \left[ \frac{\exp(-\beta(y_t - f_i(\vec{x}_t))^2)}{\sum_{j=1}^N \exp(-\beta(y_t - f_j(\vec{x}_t))^2)} \right] (y_t - f_i(\vec{x}_t)). \quad (5)$$

According to Bayes' rule the term in brackets is the posterior probability that expert  $i$  is the correct choice for the given data,  $y_t$ , that is,  $p(i|y_t)$ ,

$$p(i|y_t) = \frac{p(y_t|i)p(i)}{\sum_{j=1}^N p(y_t|j)p(j)}, \quad (6)$$

where  $p(y_t|i)$  and  $p(i)$  are given by (3) and (4). Therefore, we can simply write (5) as

<sup>1</sup> We use an on-line variant of the K-means algorithm (Duda and Hart (1973)).

$$\frac{\partial \log L}{\partial f_i} \propto p(i|y_t)(y_t - f_i(\vec{x}_t)) . \quad (7)$$

The experts are trained by expectation-maximization (EM) (Dempster et al. 1977). The E-step consists in estimating the probabilities  $p(i|y_t)$ . The M-step then adapts the parameters of the experts by a gradient ascent on the log-likelihood function to maximize the likelihood of the model. Since in the M-step the  $p(i|y_t)$  are considered to be constant, the computation of the derivative of the log-likelihood with respect to the expert's parameters is much simpler. The learning rule for the  $w_{ih}$  in the normalized RBF case, for example, is given by

$$\Delta w_{ih} \propto \frac{\partial \log L}{\partial f_i} \frac{\partial f_i}{\partial w_{ih}} \quad (8)$$

$$\Delta w_{ih} = \eta p(i|y_t)(y_t - f_i(\vec{x}_t)) \frac{g_h(\vec{x})}{\sum_{k=1}^H g_k(\vec{x})} . \quad (9)$$

In each M-step, the learning rule is applied successively for each data point in the training set. The rule can be interpreted as a weighting of the learning rate  $\eta$  by the expert's relative prediction performance  $p(i|y_t)$ : the expert with the best prediction is allowed to make the largest training step.

The learning rule in (9), however, can often be insufficient for obtaining the correct segmentation and therewith a low prediction error. Without explicitly incorporating an assumption about the switching frequency of the dynamical modes, a variety of switching dynamical systems are conceivable as the origin of a given time series. For example, one can imagine a system that is switching to a new mode at each time step. In the current framework, the choice of models is already limited by the number of predictors, and the predictors typically only allow for relatively simple and smooth mappings. Yet, it is often still possible to fit the data in various ways and the training process is likely to select a wrong model and to get stuck in local minima of the error function (Pawelzik et al. 1996; Kohlmorgen 1998). Constraining the training process to find only those models with a relatively low switching rate solves the problem in cases where the dynamics does indeed switch at low rates. We do this by incorporating some inertia into the predictor weighting scheme in the learning rule in (7), thereby introducing the concept of *memory* into the framework. We simply replace the probability that a particular predictor has generated a *single* data point,  $p(i|y_t)$ , by the probability that the predictor has generated *all the data points in a temporal neighborhood* of the current data point,  $p(i|y_t^\Delta)$ , with  $y_t^\Delta = (y_{t-\Delta}, \dots, y_{t+\Delta})$ . Using Bayes' rule, we get

$$p(i|y_t^\Delta) = \frac{p(y_t^\Delta|i)p(i)}{p(y_t^\Delta)} , \quad (10)$$

where

$$p(y_t^\Delta) = \sum_{j=1}^N p(y_t^\Delta|j)p(j) . \quad (11)$$

For simplicity, we assume  $p(y_t^\Delta|i) = \prod_{t'=\Delta}^{t+\Delta} p(y_{t'}|i)$ . Using (3, 4), we get

$$p(i|y_t^\Delta) = \frac{\exp(-\beta \sum_{t'=\Delta}^{t+\Delta} (y_{t'} - f_i(\vec{x}_{t'}))^2)}{\sum_{j=1}^N \exp(-\beta \sum_{t'=\Delta}^{t+\Delta} (y_{t'} - f_j(\vec{x}_{t'}))^2)} . \quad (12)$$

Thus, instead of using individual prediction errors  $e_i^t = (y_t - f_i(\vec{x}_t))^2$  [in (5)], the incorporation of a low switching rate assumption leads to the use of low-pass filtered errors in the learning rule

$$E_i^t = \sum_{t'=\Delta}^{t+\Delta} e_i^{t'} . \quad (13)$$

In (11) we actually assume at each time step that the whole sequence  $y_t^\Delta$  was generated by exactly one predictor. This simplification, that is, assuming probability 0 for sequences of length  $\Delta$  that contain a switching event, leads to the box-type filter, which might be replaced by a weighted low-pass filter

$$E_i^t = \sum_{t'=\Delta}^{t+\Delta} G(t' - t) e_i^{t'} \quad (14)$$

to model the switching probabilities more realistically. Yet, without any knowledge about the switching characteristics of the time series, (13) seems to be the simplest and at the same time computationally least expensive way to include memory. Heuristically, (13) is in analogy to evolutionary inertia, since once a predictor has performed better than its competitors, it also has an advantage for temporally adjacent data points. Note that to use this framework for forecasting, the proposed a-causal filter needs to be replaced by a causal filter that only uses error information from the past. An alternative approach was presented, for example, in Liehr et al. (1999), where memory was included into a system of competing predictors by means of a Markov process.

For the purpose of segmentation, it might seem to be most desirable to choose  $\beta$  large. Indeed, one could consider  $\beta = \infty$ , which corresponds to a hard competition (winner-takes-all) and guarantees an unambiguous segmentation (Kohlmorgen et al. 1994; Müller et al. 1994, 1995a). We found, however, that the use of hard competition right from the beginning of the training process does not always lead to a sufficient diversification of the predictors. The final result might strongly depend on the choice of initial parameters, which may lead to poor local optima in the likelihood  $L$  (Pawelzik 1996; Kohlmorgen 1998).

We solve this initialization problem by *adiabatically* increasing the degree of competition. For  $\beta = 0$ , the predictors equally share all the data for training. Increasing  $\beta$  enforces the competition, thereby driving the predictors to a specialization on different subsets of the data. Diversification occurs at particular 'temperatures'  $\theta = 1/\beta$  and the network parameters separate abruptly, resolving the underlying structure to more detail (Pawelzik 1996; Kohlmorgen 1998). These phase transitions are indicated by a drop of the weighted root mean squared error (RMSE),

$$E = \sqrt{\sum_{t=1}^T \sum_{i=1}^N p(i|y_t^\Delta) e_i^2} \quad (15)$$

and have been described within a statistical mechanics formalism in the context of hierarchical clustering (Rose et al. 1990). Note that a careful increase of  $\beta$  is crucial when fine differences of underlying functions need to be resolved.

As a prerequisite of this method, mode changes should occur infrequently, that is, between two mode changes the dynamics should operate stationarily in one mode for a certain number of time steps. The application of this method then yields a (hard) segmentation of a time series into different operating modes together with prediction experts for each mode. In case of a drift between two modes, the respective segment tends to be subdivided into several parts, because a single predictor is not able to handle the nonstationarity (cf. Fig. 1 in Sect. 3.1). In the next section, we present a method that is able to deal more appropriately with such drifts.

## 2.2 A hidden Markov model for drift segmentation

The second stage of the algorithm takes the drift into account. We propose a segmentation algorithm that allows us to model drifts between two stationary modes by combining the two respective predictors,  $f_i$  and  $f_j$ . The drift is modeled by a weighted superposition

$$f(\vec{x}_t) = a(t)f_i(\vec{x}_t) + (1 - a(t))f_j(\vec{x}_t), \quad 0 \leq a(t) \leq 1, \quad (16)$$

where  $a(t)$  is a mixing coefficient. We will now define a hidden Markov model (HMM) that allows us to use the Viterbi algorithm (cf. Rabiner 1990) for the detection of this kind of drifting dynamics. For this purpose, however, a discretization of the mixing proportions is necessary.

An HMM consists of (1) a set  $S$  of states, (2) a matrix  $\mathbf{A} = \{p_{\tilde{s},s}\}$  of state transition probabilities, (3) an observation probability distribution  $p(y|s)$  for each state  $s$ , which is a continuous density in our case, and (4) the initial state distribution  $\pi = \{\pi_s\}$ . For a thorough introduction to HMMs, see Rabiner (1990) and the references therein.

The construction of  $S$ , the set of states, is the crucial point of this approach. The set  $S$  consists of ‘pure’ states, which represent single predictors, and mixture states, which represent particular mixtures of two predictors. We first consider the subset  $P$  of pure states. Each state  $s \in P$ ,  $|P| = N$ , represents one of the predictors  $f_{k(s)}$  trained in the first step. The function  $k(s)$  simply returns the index of a predictor assigned to a given state. The predictions in  $P$ -states are thus performed by single predictors. Next, consider the subset  $M$  of mixture states. Each state  $s \in M$  represents a linear mixture of two predictors  $f_{i(s)}$  and  $f_{j(s)}$ .<sup>2</sup> Given a state  $s \in S$ ,  $S = P \cup M$ , the prediction of the overall system is then performed by

$$g_s(\vec{x}_t) = \begin{cases} f_{k(s)}(\vec{x}_t); & \text{if } s \in P \\ a(s)f_{i(s)}(\vec{x}_t) + b(s)f_{j(s)}(\vec{x}_t); & \text{if } s \in M. \end{cases} \quad (17)$$

For each mixture state  $s \in M$ , the coefficients  $a(s)$  and  $b(s)$  need to be determined together with the respective network indices  $i(s)$  and  $j(s)$ . For computational feasibility, the number of mixture states has to be restricted. Our intention is to allow for drifts between any two network outputs of the previously trained ensemble. We choose  $a(s)$  and  $b(s)$  such that  $0 < a(s) < 1$  and  $b(s) = 1 - a(s)$ . Next, a discrete set of  $a(s)$  values needs to be chosen. For simplicity, we use equally distant steps,

$$a_r = \frac{r}{R+1}, \quad r = 1, \dots, R, \quad (18)$$

where  $R$  is the number of intermediate mixture levels. A given resolution  $R$  between any two out of  $N$  nets yields a total number of mixed states  $|M| = RN(N-1)/2$ . If, for example, the resolution  $R = 32$  is used and we assume  $N = 8$ , then there are  $|M| = 896$  mixture states, plus  $|P| = N = 8$  pure states.

As a second step, the transition matrix  $\mathbf{A} = \{p_{\tilde{s},s}\}$  has to be chosen. It determines the transition probability for each pair of states. In principle, this matrix can be found using a training procedure, as, for example, the Baum-Welch method (Rabiner 1990). However, this is hardly feasible in this case, because of the immense size of the matrix. In the above example, the matrix  $\mathbf{A}$  has  $(896 + 8)^2 = 817,216$  elements that would have to be estimated. Such an exceeding number of free parameters is practically intractable for adaptive methods given only a limited amount of data. Therefore, we use a fixed matrix. In this way, prior knowledge about the dynamical system can be incorporated. In our applications, we either allow for switches or smooth drifts between two predictors, such that a (monotonous) drift from one net to another is a priori as likely as a switch. All the other transitions are disabled by setting  $p_{\tilde{s},s} = 0$ . The definitions for  $p(y|s)$  and  $\pi$  are straightforward. Following (3) and (4), we again assume Gaussian noise

$$p(y|s) = \sqrt{\frac{\beta}{\pi}} \exp(-\beta(y - g_s)^2) \quad (19)$$

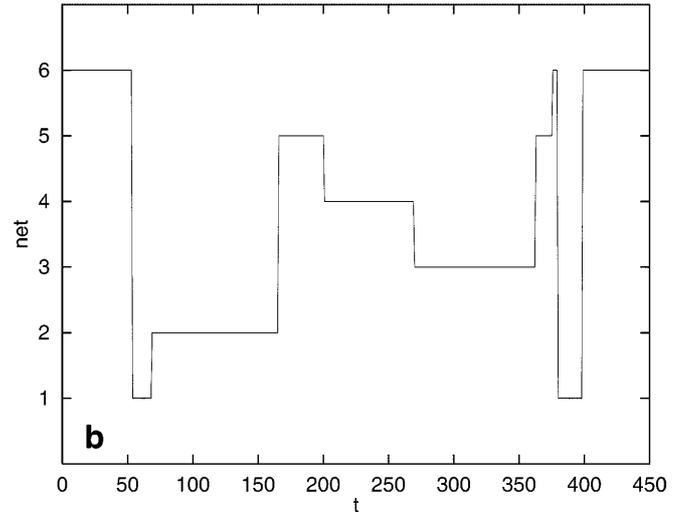
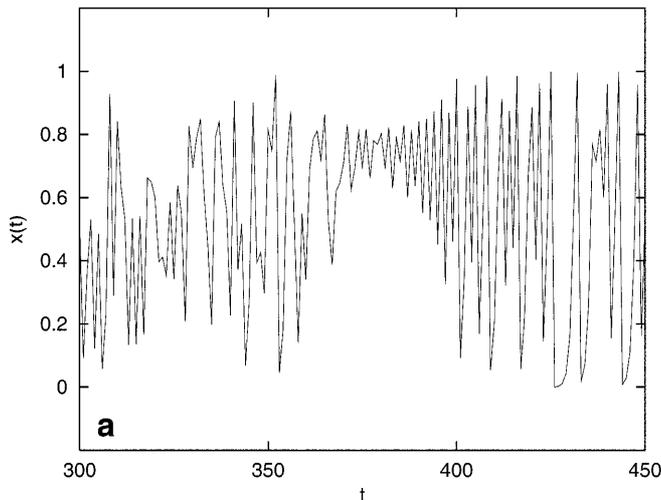
and equally probable initial states,  $\pi_s = 1/|S|$ .

The Viterbi algorithm (Rabiner 1990) can now be applied to the above HMM without any further training of the HMM parameters. It directly yields the drift segmentation of a given time series, that is, the most likely state sequence (the sequence of predictors or linear mixtures of two predictors) that might have generated the time series – in our case, using the assumption that mode changes occur either as (smooth) drifts or as infrequent switches.

## 2.3 The drift segmentation algorithm

In the following, we present a dynamic programming technique, which is equivalent to the Viterbi algorithm,

<sup>2</sup> Note that, in principle, each state might also represent a *non-linear* mixture of *arbitrarily* many predictors.



**Fig. 1.** **a** A part of the training data, generated by the chaotic return maps  $f_1$  and  $f_4$ . First,  $f_4$  is iterated from  $t = 300$  to  $t = 350$ . Then, there is a drift to  $f_1$  between  $t = 350$  and  $t = 400$ . After  $t = 400$ ,  $f_1$  is iterated. **b** The final segmentation into training subsets, obtained by the competitive training procedure. Shown are the first 450 data

and which computes the sequence of nets and linear mixtures of nets much more efficiently. Instead of computing the most likely HMM state sequence in terms of probabilities  $p$ , we compute the state sequence in terms of costs  $C = -\log(p)$ . In this way, we can replace products by sums and avoid numerical problems (Rabiner 1990). To this end, we compute the cost function  $C^*$ , which is the sum of squared prediction errors plus transition costs for the best matching state sequence. This sequence can be obtained between two points in time,  $t_0$  and  $t_{\max}$ , by recursively computing, for all  $s \in S$  and all  $t = t_0, \dots, t_{\max}$ , the cost  $C_s(t)$  of the most likely state sequence that might have produced the part of the time series considered so far,  $\{x_{t_0}, \dots, x_t\}$ , and whose state at time  $t$  is  $s$ . By using the squared prediction error of the pure or mixed network output,  $\varepsilon_s(t) = (y_t - g_s(\vec{x}_t))^2$ , the recursion can be formulated as follows:

$$C_s(t_0) = \varepsilon_s(t_0) , \quad (20)$$

$$C_s(t) = \varepsilon_s(t) + \min_{\hat{s} \in S} \{C_{\hat{s}}(t-1) + T(\hat{s}, s)\}, \quad t = t_0 + 1, \dots, t_{\max} \quad (21)$$

$$C^* = \min_{s \in S} \{C_s(t_{\max})\} . \quad (22)$$

Here,  $T(\hat{s}, s)$  is the transition cost for jumping from state  $\hat{s}$  to state  $s$ . Note that the transition costs are in analogy to the transition probabilities in the HMM, that is, the choice of the transition matrix  $T$  determines the transition probabilities between the states. The resulting segmentation sequence is obtained by backtracking through the sequence of states that make up  $C^*$  (cf. Rabiner 1990).

### 3 Detecting drifts in synthetic data

To illustrate the basic idea of our learning algorithm, we first discuss a simple example of drifting chaotic

points. This segmentation cannot represent the drift. The stationary parts,  $f_1$  in  $[0, 50]$  and  $[400, 450]$ ,  $f_2$  in  $[100, 150]$ ,  $f_3$  in  $[200, 250]$ ,  $f_4$  in  $[300, 350]$ , are predicted by nets 6, 2, 4, and 3, respectively. The nonstationary drift parts in between are shared among all predictors, including nets 1 and 5

dynamics. It is followed by an application to a drifting system of the Mackey-Glass model of blood cell regulation.

#### 3.1 Drifting chaos

Consider a chaotic time series  $\{x_t\}$ , where  $x_{t+1} = f(x_t)$ , Fig. 1a. Four operating modes are established by using four different chaotic maps:

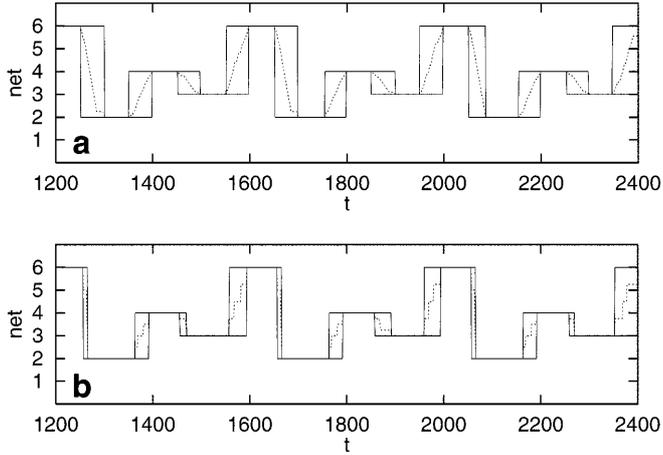
$$\begin{aligned} f_1(x) &= 4x(1-x), \quad x \in [0, 1] && \text{(logistic map)} \\ f_2(x) &= f_1(f_1(x)) && \text{(double logistic map)} \\ f_3(x) &= 2x, \quad \text{if } x \in [0, 5) \\ &\quad \text{and } 2(1-x), \quad \text{if } x \in [.5, 1] && \text{(tent map)} \\ f_4(x) &= f_3(f_3(x)) && \text{(double tent map)} \end{aligned}$$

For the first 50 time steps,  $f_1$  is applied recursively, starting with  $x_0 = 0.5289$ . After  $t = 50$  time steps, the dynamics is drifting from  $f_1$  to  $f_2$  using

$$f(x_t) = (1 - a(t))f_1(x_t) + a(t)f_2(x_t), \quad a(t) = \frac{t - t_a}{t_b - t_a} , \quad (23)$$

with  $t_a = 50$  and  $t_b = 100$ . The drift is linear in time and takes another 50 time steps. Then, the system runs stationarily in mode  $f_2$  for the following 50 time steps, whereupon it is drifting to  $f_3$  in the same fashion as before, and so on. At  $t = 350$ , the system starts to drift back from  $f_4$  to  $f_1$  and the cycle starts again at  $t = 400$ .

The first step of analysis consists in applying the ACE approach, described in Sect. 2.1, to the first 1,200 data points of the generated time series. An ensemble of six RBF predictors competes for the data during the training phase. Each predictor contains 20 Gaussian basis functions. After training, four predictors have specialized each on a different chaotic map, and the other two



**Fig. 2.** **a** The segmentation obtained by the drift algorithm on the test data, using the resolution  $R = 32$ . Shown is the sequence of nets as a function of time. The *dotted line* indicates the evolution of the mixing coefficient  $a(t)$  of the respective nets. For example, between  $t = 1350$  and  $1400$  it denotes a drift from net 2 to net 4, which in this case turns out to be a linear drift, as expected. The segmentation almost perfectly reproduces the behavior of the dynamical system. **b** A segmentation with a low resolution,  $R = 3$ , can only traverse the drift parts in three steps

predictors appear to have specialized on the drift parts. This can be observed in the final segmentation of the competition procedure, shown in Fig. 1b. Clearly, the true dynamical structure of the system cannot be represented by the switching model.

Next, the drift segmentation algorithm is applied to the next 1,200 data points (test data set) using all six previously trained networks. It perfectly reproduces the

behavior of the dynamics, as seen in Fig. 2a for the resolution  $R = 32$ : a linear drift between four stationary operating modes is extracted from the data in an unsupervised manner. Figure 2b is included to demonstrate the effect of a lower resolution.

### 3.2 A drifting Mackey-Glass system

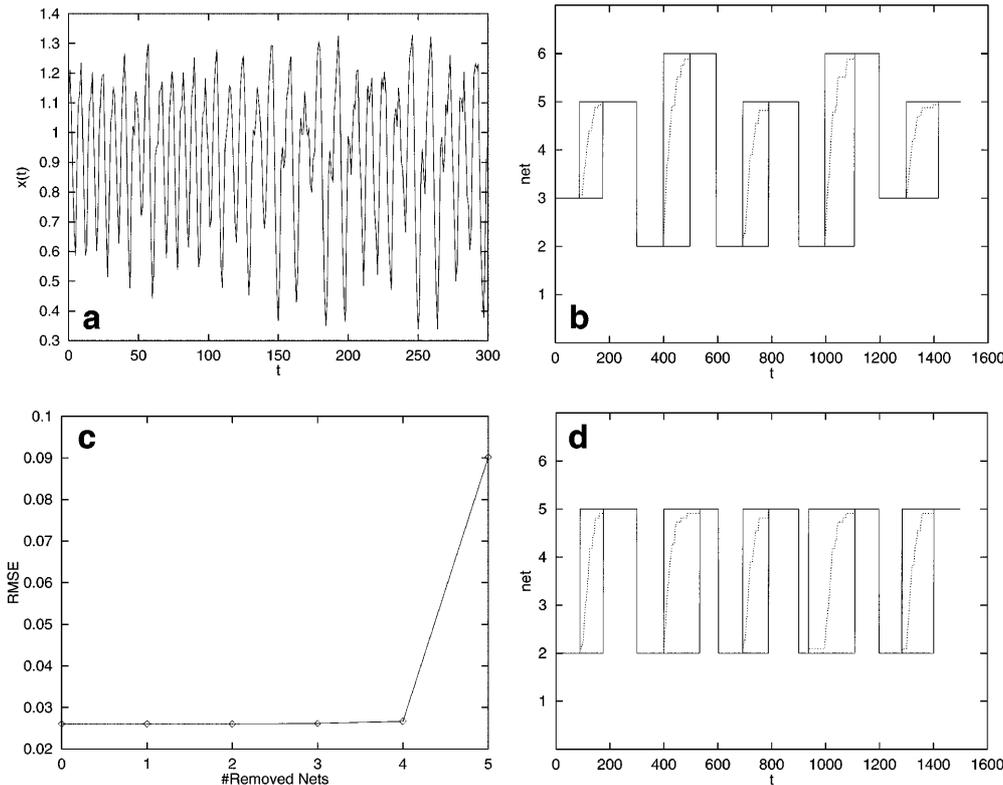
Consider a high-dimensional chaotic system generated by the Mackey-Glass equation, which originally was developed as a model for respiratory regulation and haematopoiesis (Mackey and Glass 1977)

$$\frac{dx(t)}{dt} = -0.1x(t) + \frac{0.2x(t-t_d)}{1+x(t-t_d)^{10}}. \quad (24)$$

Two stationary operating modes, A and B, are established by using different delays,  $t_d = 17$  and  $23$ , respectively. After operating 100 time steps in mode A (with respect to a subsampling step size  $\tau = 6$ ), the dynamics drifts to mode B. The drift takes another 100 time steps. It is performed by mixing the equation for  $t_d = 17$  and  $23$  during the integration of (24). The mixture is generated according to (16), using an exponential drift

$$a(t) = \exp\left(\frac{-4t}{100}\right), \quad t = 1, \dots, 100. \quad (25)$$

After the drift, the system runs stationary in mode B for the next 100 time steps, whereupon it *switches* back to mode A at  $t = 300$ , and the loop starts again (Fig. 3a). The competing experts algorithm is applied to the first 1,500 data points of the generated time series,



**Fig. 3.** **a** The drifting Mackey-Glass time series. The dynamical system operates in mode A for the first 100 time steps. Then, the dynamics drifts to mode B during the next 100 steps and remains stationary in B. After  $t = 300$ , the system switches back to mode A and the cycle starts again. **b** The resulting drift segmentation invokes four nets. This is because two nets became experts for mode A, and two others for mode B. **c** Increase of the prediction error when predictors are successively removed. Although no further training has been performed, up to four predictors can be removed without a significant increase of the prediction error. **d** The two remaining predictors model the dynamics of the time series properly

using an ensemble of six 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 RBF predictors consist of 40 basis functions each.

After training, nets 2 and 3 have specialized on mode A, nets 5 and 6 on mode B. This can be seen in the drift segmentation in Fig. 3b. Moreover, the removal of four nets does not increase the RMSE of the prediction significantly (Fig. 3c), which correctly indicates that two predictors completely describe the dynamical system. The sequence of nets to be removed is obtained by repeatedly computing the RMSE of all  $n$  subsets with  $n - 1$  nets each and selecting the subset with the lowest RMSE of the respective drift segmentation. The segmentation of the remaining nets, 2 and 5, nicely reproduces the evolution of the dynamics, as shown in Fig. 3d.

#### 4 Analysis of physiological data

In this section, we present results on the analysis of physiological recordings from afternoon naps of healthy humans. We analyzed the EEG and the respiratory signal using the method presented in Sect. 2. We used single-channel recordings for the computer-based analysis and did not incorporate any medical expert knowledge into the algorithm. We compare the results with a manual segmentation performed by a medical expert, which is based on eight physiological signals (EEG: O1, O2, F3, EOG, ECG, heart rate, blood pressure, respiration).

##### 4.1 Physiological background

In the past decades, the impact of sleep on physiological and pathophysiological functions of humans has been established. Relevant sleep disorders have been identified, in which the transitions between wakefulness and different sleep stages are affected. From the neurophysiological point of view, a re-organization of the neuronal network in the reticular formation of the brain stem prevails (Steriade and McCarley 1990). Here, the crucial structures are found that regulate and integrate cardiovascular, respiratory, and somatomotor systems and vigilance (Langhorst et al. 1983). Recent studies have shown that the discharge behavior of reticular neurons, that is, their mode of neuronal processing, depends strongly on their level of activity (Lambertz and Langhorst 1995). To keep in line with the arguments of the paper, the reorganization during sleep onset thus could be thought of as a ‘switching to a different mode’ of neuronal processing.

The typical findings during falling asleep comprise – besides blurring of consciousness – a slowing of cortical activity and heart rate, a decrease of the arterial blood pressure, reduction and instability (Trinder et al. 1992) of respiration, a decrease in metabolic rate, reduction of

muscle tone, and the occurrence of slow eye movements (SEM) mainly in the horizontal plane. Consequently, modern sleep polysomnography comprises EEG, electrocardiogram (ECG), a respiratory trace, electroculogram (EOG), and, if possible, the measurement of arterial blood pressure.

Much of the analysis of sleep recordings is based on a good segmentation of the recordings and classification to different sleep stages (Rechtschaffen and Kales 1968). Methods for segmentation and classification of EEG have been proposed, for example, in (Praetorius et al. 1977; Creutzfeldt et al. 1985; Kemp et al. 1987). For the physiological understanding of the process of sleep onset, but possibly also for the diagnosis of sleep disorders, it would be highly desirable to access not only the occurrence, but also the *time course* of the transitions. Several techniques using time-varying linear models were suggested for tracking the signal characteristics in EEG (Isaksson and Wennberg 1976; Bohlin 1977; Kaipio and Karjalainen 1997; Hiltunen et al. 1999). Here we present a different approach: we first determine dynamical basis modes in an unsupervised manner, using nonlinear models, and, in a second step, estimate the drift between these modes. We apply our method to EEG recordings and respiration data and obtain both a segmentation into dynamical modes and the time course of the drift between them.

##### 4.2 Analysis of respiratory data

We analyzed recordings of five healthy persons, all of them 20–40 years old and nonsmokers. For each of them the thoracic excursions from three successive afternoon naps were recorded with an extrathoracic strain belt. The training method described in Sect. 2 was applied to each of the 15 time series using the embedding dimension  $d = 4$  and the delay parameter  $\tau = 7$  on the given 10-Hz data. We took eight RBF networks with 20 Gaussian basis functions as prediction experts. After training, all 15 trained ensembles of predictors were used for the segmentation of all 15 data sets, which thus can be subdivided into three different classes: (1) segmentation of the recording that was used for training, (2) segmentation of the 2 other recordings from the same subject (sleeper dependent), and (3) segmentation of the 12 recordings from the other subjects (sleeper independent). Moreover, the agreement of the hand labeling with the machine segmentation is measured on different levels of resolution. The coarsest resolution is a segmentation into wakefulness (W) and sleep (S), finer resolutions also distinguish between two different wake states, eyes open (W1) and eyes closed (W2), and sleep stage I (S1) and sleep stage II (S2). Other sleep stages do not occur in the recorded afternoon naps.

The results are depicted in Table 1. Considering the segmentation of the training set data, 69.7% of the data points are segmented in agreement with the manual segmentation; if we consider only the segmentation into wake and sleep phase the agreement is 86.33%. This is remarkable, considering the fact that the two segmen-

**Table 1.** Average percentage of agreement between manual and machine segmentation of respiratory data. ‘Training set’ denotes the segmentation of the training set, ‘sleeper dependent’ denotes the segmentation of test data of the same sleeper. ‘Sleeper independent’ corresponds to the segmentation of recordings from other sleepers. The percentage of agreement with respect to four categories – eyes open (W1), eyes closed (W2), sleep stage I (S1), and sleep stage II (S2) – is quoted in the *left column*. The agreement is significantly larger if no distinction between S1 and S2 is required (*middle column*). The results for merging also W1 and W2 into one class, that is distinguishing only between wakefulness and sleep, are reported in the *right column*

Respiration	W1-W2-S1-S2	W1-W2-S	W-S
Training set	69.70	79.08	86.33
Sleeper dependent	51.80	63.92	75.24
Sleeper independent	34.02	45.72	59.53

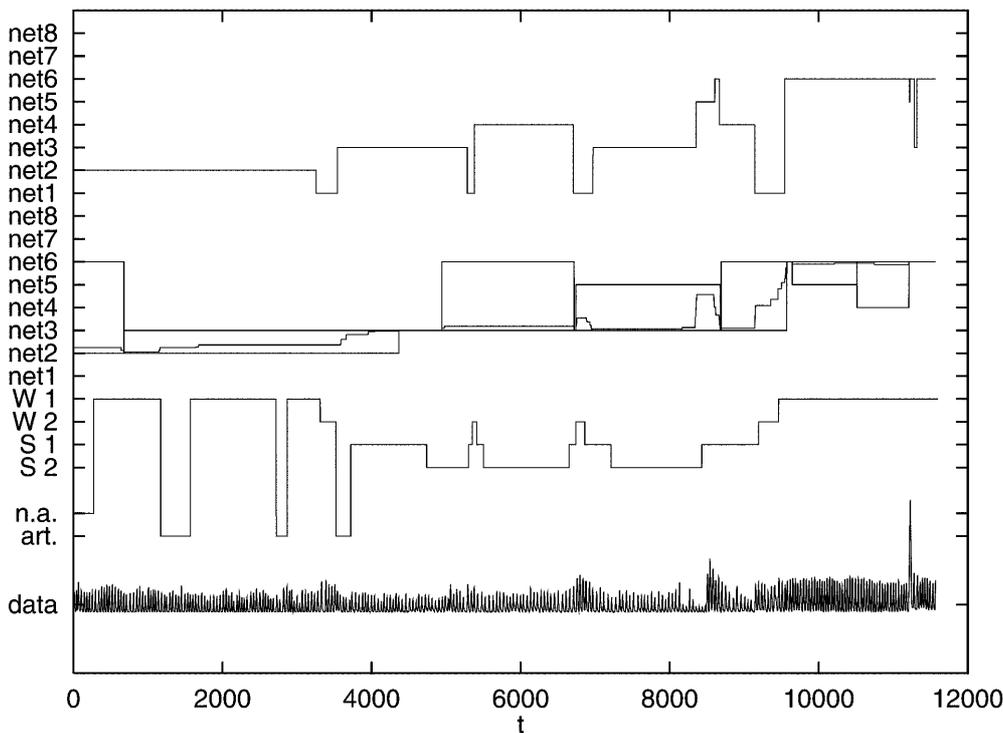
tations were obtained in rather different ways: the manual segmentation is based on standard criteria defined in Rechtschaffen and Kales (1968) using all available physiological signals, mainly EEG and EOG, whereas the machine segmentation is purely data driven using only the respiration signal and no additional expert knowledge. Systematic deviations may also arise from this fact, since the typical signs of sleep onset in cortical activity and in other organ systems may not be fully simultaneous (Koella 1982).

It also must be remembered that standard segmentation criteria were not formulated on the basis of physiological findings that clearly separate the stages from each other. Therefore, manual segmentation of sleep recordings depends on the persons who convey it; the inter-personal agreement is in the same order as between machine and human expert in our study (Kubicki et al. 1982).

The distinction between wake and sleep, for example, the detection of the sleep onset, can clearly be made by our algorithm. As expected, it is not possible for our method to distinguish between sleep stages I and II from respiratory data. A separation of stages I and II on the basis of the respiratory signal is also not defined in the standard criteria, and, in fact, for a human expert the distinction between sleep stage I and II cannot be made given only respiratory data.

The results in Table 1 show that the dynamic models obtained from a certain recording cannot be used for the segmentation of new recordings without a significant loss of performance, especially if applied to recordings from other subjects. Obviously, the respiration during sleep changes for the same individual on different days, but even more striking are the differences between several individuals. Thus, it is preferable to adapt the prediction experts to each recording individually. In practice, this should not be a problem, since training the experts on 12,000 data points can be performed in about 20 seconds on a SUN Ultrasparc-I.

A sample segmentation is depicted in Fig. 4. The upper line shows the segmentation that is obtained when only *switching* between experts (nets) is allowed. As the result of the segmentation algorithm, net 2 is responsible for the W1 state (eyes open) at the beginning of the recording. Net 1 corresponds to W2 (eyes closed), nets 3 and 4 are active during sleep stages I and II (S1 and S2), and net 6 is responsible for the W1 dynamics after the arousal. The transition from S2 to S1 after  $t = 8000$  comes along with a broad but temporary increase of respiration activity (shown in the data), which is captured by net 5. It is, however, not represented in the manual segmentation. Apart from that we can conclude that the switch segmentation is in good agreement with



**Fig. 4.** Comparison of switch segmentation (*top*), drift segmentation (*middle*), and a manual segmentation by a medical expert (*bottom*) of respiration data from a single experiment (time scale: 100 ms). W1 and W2 indicate two wake states in the manual analysis; S1 and S2 indicate sleep stages I and II, respectively (*n.a.* no assessment, *art.* artifacts)

the manual reference segmentation, even though, as expected, no clear-cut distinction can be made between S1 and S2 in this example.

The drift segmentation shown in Fig. 4 (middle) gives more detailed information about the evolution of the dynamics. In the beginning, it shows a long-term drift from the wake state predictor, net 2, towards the sleep predictor, net 3: the dotted line indicates the evolution of the mixing coefficient  $a(t)$  between the two predictors denoted by solid horizontal lines.

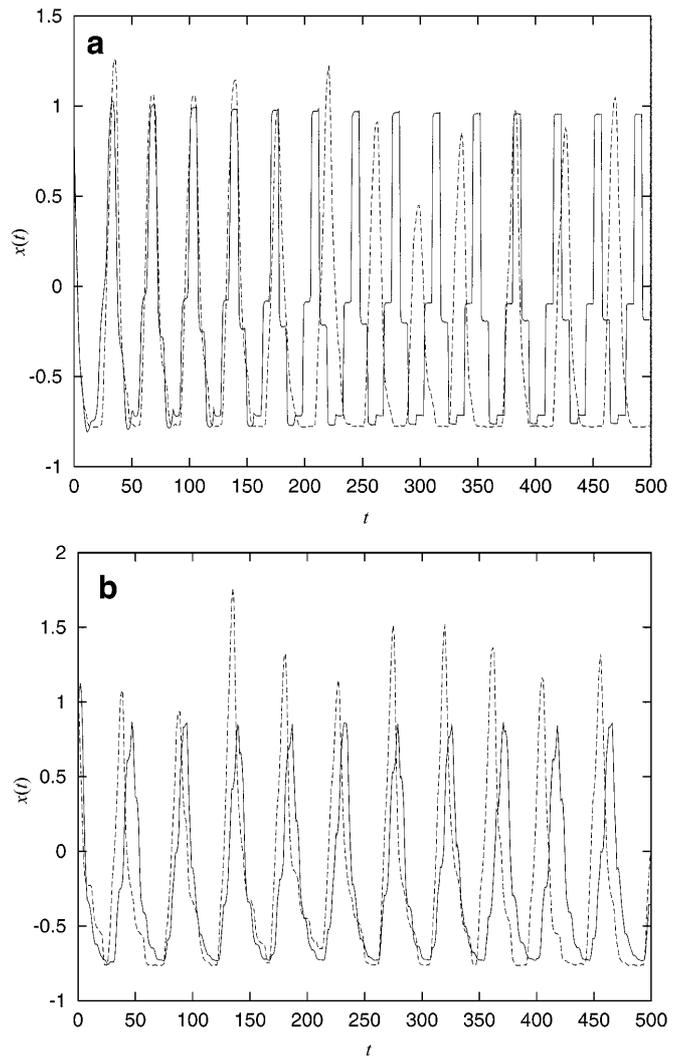
At the sleep onset before  $t = 4000$ , there is a fast change in the mixing proportion in favor of net 3. At  $t = 4500$ , net 3 represents the dynamics solely. The final arousal before  $t = 10,000$  is represented as a drift from sleep state predictor 3 to wake state predictor 6, with a fast change of the mixing coefficient at the arousal point. Both transitions nicely comply with the manual segmentation. On the other hand, only the second intermediate arousal after  $t = 6000$  is indicated (there is a slight drift towards net 5) whereas the first one, before  $t = 6000$ , is not.

After having obtained very reasonable segmentations into wake and sleep states, we now consider the dynamics that the competing predictors have learned. In other words, how well has the nonlinear dynamics of the signal been captured by the predictors, that is, was it possible to identify the inherent dynamics of the physiological states? To illustrate the prediction performance of the method for the respiratory signal, the predictors for the wake and sleep phase are iterated *autonomously* from a point in the wake or sleep state, respectively. The iteration yields the long-term prediction for 500 time steps (50 seconds) shown in Fig. 5a and b. Although the networks are only trained to predict the next data point, they have captured the underlying dynamics to such an extent that reliable predictions are possible for 180 time steps (18 seconds) into the future for the wake state, and for 350 time steps (35 seconds) for the sleep state.

### 4.3 EEG analysis

The EEG allows us to determine the sleep onset more accurately and easily than the respiratory signal. This is because EEG is a very informative signal measured with a high time resolution of typically 100 Hz up to 1000 Hz. In this study we analyzed EEG signals from O1 (occipital-1), recorded from a single subject. We used a set of eight RBF predictors with six radial basis functions and the embedding  $d = 4$  and  $\tau = 2$  on the raw 100-Hz data. Such an ensemble was trained for each data set (eeg11, eeg12, eeg13) and it was then used for the segmentation of all three recordings. As expected, the deviation from the manual segmentation is less than in the case of respiration data (Table 2).

For the switch segmentation of the EEG we found that the predictor ensemble did not distinguish well between sleep stage I and II; in most cases only a single predictor was responsible for both stages. This result is reflected in Table 2. The leftmost column, which asks for the distinction of the two sleep stages, shows significantly



**Fig. 5.** Iterating the predictors responsible **a** for the wake state and **b** for the sleep state (*solid lines*) results in a good accordance with the true continuation (*dashed lines*) of the respiratory signal for **a** 180 and **b** 350 time steps (time scale: 100 ms)

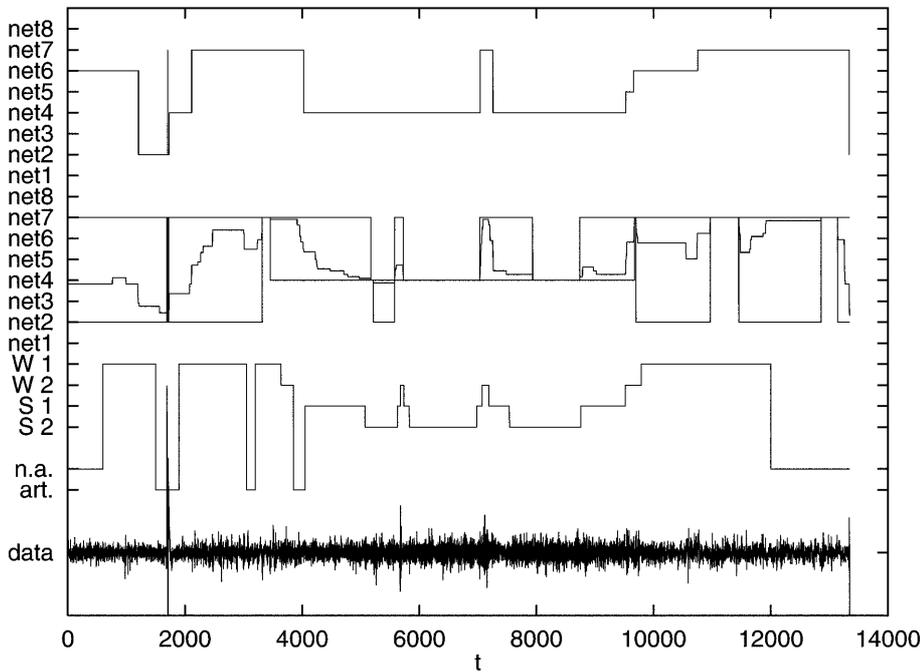
**Table 2.** Percentage of agreement between manual and machine segmentation for three EEG recordings from a single subject. A set of RBF prediction experts was trained for each of the recordings (subtables eeg11, eeg12, eeg13). Each set was then used for the segmentation of all three recordings (11, 12, 13). The agreement of the hand labeling with our machine segmentation is again measured on different levels of resolution (cf. Table 1)

EEG	W1-W2-S1-S2	W1-W2-S	W-S
eeg11			
11	70.87	92.12	96.44
12	67.31	77.82	91.09
13	60.93	82.40	90.67
eeg12			
11	67.04	94.55	98.99
12	83.63	86.19	89.21
13	57.65	84.04	91.58
eeg13			
11	72.77	94.22	98.51
12	69.82	77.97	80.18
13	73.92	86.78	95.10

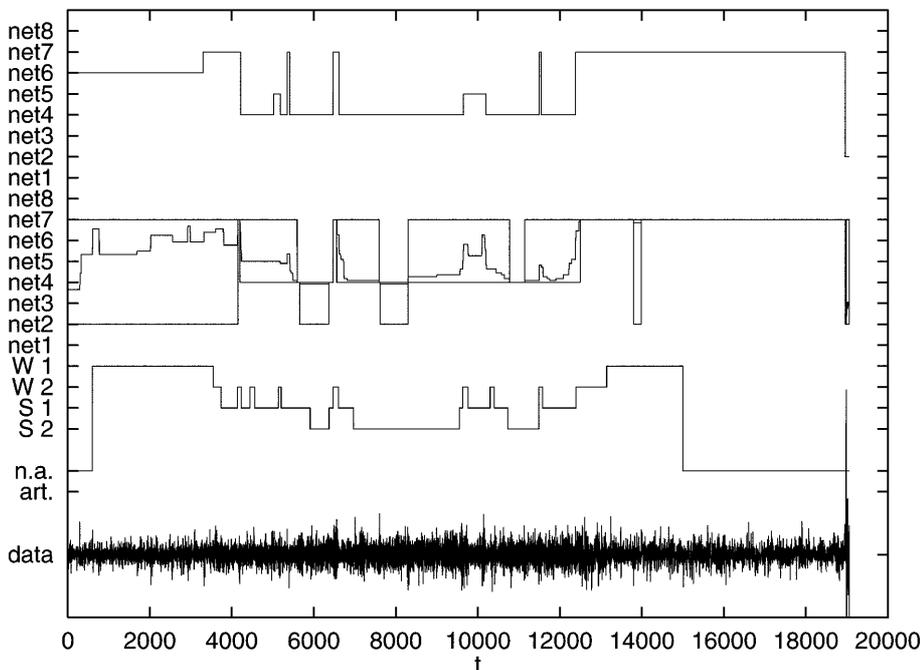
less agreement with the hand labeling than the column in the middle, which does not make this distinction.

The drift segmentations, however, provide a different view (Figs. 6, 7). In contrast to the switch segmentations (upper lines), the drift segmentations (middle) distinguish between sleep stage I and II as follows: in both figures the sleep stage II dynamics is represented by net 4, whereas sleep stage I is not represented by an individual net but simply by a *linear mixture* of net 4 and a wake state predictor, net 7. Note, however, that net 4 has a much larger contribution to the mixture, which explains why it is selected for both stages in the switch segmentation.

In detail, Fig. 6 shows switch and drift segmentation of eeg11 using experts that were trained on this data set. The sleep onset at  $t = 4000$  and the final arousal before  $t = 10,000$  are indicated correctly in both segmentations. The drift segmentation, however, reveals a much more detailed dynamical structure of the EEG data. The sleep onset is represented as a kind of exponential drift from a wake state predictor, net 7, to the sleep state predictor, net 4. The arousal is introduced at  $t = 9000$  by a slight drift back towards net 7. This mixing proportion holds until the arousal point at  $t = 9500$  is reached. There, a jump-like change in the mixing proportion gives significantly more weight to wake state net 7. Finally, at



**Fig. 6.** Switch segmentation (*top*) and drift segmentation (*middle*) for a single-channel EEG recording of an afternoon nap (eeg11, time scale: 100 ms). The manual segmentation by a medical expert is shown at the bottom



**Fig. 7.** Segmentation results for EEG recording eeg13 using predictors trained on recording eeg11 (same sleeper)

$t = 10,000$ , a mixture of two wake state nets, 2 and 7, represents the dynamics. The two intermediate arousals are likewise indicated by drifts towards net 7. Clearly, the segmentation is in good agreement with the human expert, and, due to the ability to represent mixture states, interesting structure in the dynamics of transitions is found.

The prediction performance of the experts, however, is poor. Therefore, in contrast to the respiratory signal, long-term predictions do not yield reasonable results. We found that the predictors trained on the EEG can only be used to *discriminate* between different modes; they did not capture the individual dynamics properly. This result, however, is not surprising, since the EEG is a very complex signal.

To assess the generalization ability of the set of predictors with respect to its segmentation capability, we used the ensemble that was trained on eeg11 for the segmentation of eeg13. As shown in Fig. 7, the overall structure of the obtained switch and drift segmentation is again in good agreement with the hand labeling. In particular, two short intermediate arousals before and after  $t = 10,000$  are nicely represented in the drift segmentation. Moreover, nets 2 and 7, previously identified as wake state predictors in the drift segmentation, are again selected for modeling the wake states. Likewise, net 4 is again responsible for sleep stage II. Thus, one advantage of reusing previously trained experts is that one only has to label the experts once, after training, instead of labeling the experts each time a new data set needs to be segmented. On the other hand, if the experts are trained on each data set, the segmentation performance is clearly better (cf. Table 2).

## 5 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. The method needs neither prior information on *whether* the time series contains multiple modes that switch or drift in time, nor on *what* the dynamics of the operating modes look like. Instead of using a single but complex predictor, we apply a divide-and-conquer strategy that forces a set of competing predictors to specialize on subsequences of the data. Thereby, a segmentation of the data and an identification of the individual dynamics are developed simultaneously.

An application to physiological wake/sleep data demonstrated that drifting dynamics can be found in natural systems. We therefore believe that it is important to consider this aspect of data description. In the case of wake/sleep data, the results are so far encouraging, as a mathematical model was capable of identifying wakefulness, sleep stage I, and with certain reservations also stage II, from the EEG after the standard criteria by Rechtschaffen and Kales (1968) without using any medical expert knowledge. The errors of the model were in the same order of magnitude as they would occur

when the data are subjected to analysis by a human expert. Most interestingly, a fairly good segmentation was possible by analysis of the respiratory recordings alone.

The obtained drift segmentations reveal many more details of the dynamical structure than hard segmentations. In particular, the criteria of Rechtschaffen and Kales do not consider transitions between stages at all. The extracted drift curve may be interpreted as a high-resolution vigilance curve. To verify this and to find out how individual the vigilance curve is or whether a vigilance curve can be used as a diagnostic tool to distinguish between sleep disorders is subject to further investigations and experiments. Potential applications to vigilance monitoring are numerous. For example, the extracted drift might be used to give an alarm below a certain vigilance level of, for example, human controllers performing their tasks.

Sleep concerns all organ systems, but not all organ systems change their state in the same manner (Rittweger 1999). In the future it may therefore be another fruitful field for investigations to analyze differential transitions in different organ systems. Combined with parameters extracted from the model (e.g. drift velocity), clinically useful information may be obtained.

We would like to emphasize that our method is not restricted to sleep data; it can be applied as well to other physiological data like EEG recordings during epileptic seizures, ECG data in various diagnostic contexts, hormone levels, and so forth, where interesting dynamical transitions take place in relation to functional state changes. We also expect useful applications of our algorithm in other fields where complex, nonstationary dynamics plays an important role, for example, in climatology, in industrial applications, or in finance.

*Acknowledgement.* We acknowledge partial financial support of the Deutsche Forschungsgemeinschaft (grants Pa569/2-1 and Ja379/52).

## References

- Badii R, Politi A (1997) Complexity. Cambridge University Press, Cambridge
- Bengio Y, Frasconi P (1995) An input output HMM architecture. In: Tesauro G, Touretzky DS, Leen TK (eds) NIPS'94: advances in neural information processing systems 7. MIT Press, Cambridge, pp 427–434
- Bohlin T (1977) Analysis of EEG signals with changing spectra using a short-word Kalman estimator. Math Biosci 35: 221–259
- Cacciatore TW, Nowlan SJ (1994) Mixtures of controllers for jump linear and non-linear plants. In: Cowan JD, Tesauro G, Alspector J (eds) NIPS'93: advances in neural information processing systems 6. Morgan Kaufmann, San Mateo, pp 719–726
- Creutzfeldt O-D, Bodenstein G, Barlow JS (1985) Computerized EEG pattern classification by adaptive segmentation and probability density function classification. Clinical evaluation. Electroencephalogr Clin Neurophysiol 60: 373–393
- Dempster A, Laird N, Rubin D (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc B 39: 1–38
- Duda RO, Hart PE (1973) Pattern classification and scene analysis. Wiley, New York

- Hamilton JD (1994) Time series analysis. Princeton University Press, Princeton, NJ
- Hiltunen J, Karjalainen PA, Partanen J, Kaipio JP (1999) Estimation of the dynamics of event-related synchronization changes in EEG. *Med Biol Eng Comput* 37: 307–315
- Isaksson A, Wennberg A (1976) Spectral properties of nonstationary EEG signals, evaluated by means of Kalman filtering: application examples from a vigilance test. In: Kellaway P, Petersen I (eds) *Quantitative analytic studies in epilepsy*. Raven Press, New York, 389–402
- Kaipio JP, Karjalainen PA (1997) Estimation of event-related synchronization changes by a new TVAR method. *IEEE Trans Biomed Eng* 44(8): 649–656
- Kaneko K (1989) Chaotic but regular positive-negative switch among coded attractors by cluster-size variation. *Phys Rev Lett* 63: 219
- Kantz H, Schreiber T (1997) *Nonlinear time series analysis*. (Cambridge Nonlinear Science Series 7) Cambridge University Press, Cambridge
- Kehagias A, Petridis V (1997) Time series segmentation using predictive modular neural networks. *Neural Comput* 9: 1691–1710
- Kemp B, Grönwald EW, Janssen AJMW, Franzen JM (1987) A model based monitor of human sleep stages. *Biol Cybern* 57: 365–378
- Koella WP (1982) A modern neurobiological concept of vigilance. *Experientia* 38: 1426–1437
- Kohlmorgen J (1998) Analyse schaltender und driftender Dynamik mit neuronalen Netzen. (GMD Research Series 22/1998) GMD, Sankt Augustin
- Kohlmorgen J, Müller K-R, Pawelzik K (1994) Competing predictors segment and identify switching dynamics. In: Marinaro M, Morasso PG (eds) *ICANN'94: Proceedings of the International Conference on Artificial Neural Networks*. Springer, London Berlin Heidelberg, pp 1045–1048
- Kohlmorgen J, Müller K-R, Pawelzik K (1995) Improving short-term prediction with competing experts. In: Fogelman-Soulie F, Gallinari P (eds) *ICANN'95: Proceedings of the International Conference on Artificial Neural Networks*. EC2 & Cie, Paris, 2: 215–220
- Kubicki S, Herrmann WM, Höller L, Scheuler W (1982) Kritische Bemerkungen zu den Regeln von Rechtschaffen und Kales ueber die visuelle Auswertung von EEG-Schlafableitungen. *Z EEG-EMG* 13: 51–60
- Lambertz M, Langhorst P (1995) Cardiac rhythmic patterns in neuronal activity are related to the firing rate of neurons. I. Brainstem reticular neurons of dogs. *J Auton Nerv Syst* 51: 153–163
- Langhorst P, Schulz B, Schulz G, Lambertz M (1983) Reticular formation of the lower brainstem. A common system for cardiorespiratory and somatomotor functions: discharge patterns of neighbouring neurons influenced by cardiovascular and respiratory afferents. *J Auton Nerv Syst* 9: 411–432
- Liehr S, Pawelzik K, Kohlmorgen J, Müller K-R (1999) Hidden Markov mixtures of experts with an application to EEG recordings from sleep. *Theo Biosci* 118: 246–260
- Mackey M, Glass L (1977) Oscillation and chaos in a physiological control system. *Science* 197: 287
- Mayer-Kress G (ed) (1986) *Dimensions and entropies in chaotic systems*. Springer, Berlin Heidelberg New York
- Moody J, Darken C (1989) Fast learning in networks of locally-tuned processing units. *Neural Comput* 1: 281–294
- Müller K-R, Kohlmorgen J, Pawelzik K (1994) Segmentation and identification of switching dynamics with competing neural networks. In: Kim M-W, Lee S-Y (eds) *ICONIP'94: Proceedings of the International Conference on Neural Information Processing*, KAIST, Taejon, Korea, pp 213–218
- Müller K-R, Kohlmorgen J, Pawelzik K (1995a) Analysis of switching dynamics with competing neural networks. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, E78-A, No. 10, pp 1306–1315
- Müller K-R, Kohlmorgen J, Rittweger J, Pawelzik K (1995b) Analysing physiological data from the wake-sleep state transition with competing predictors. In: *NOLTA'95: International Symposium on Nonlinear Theory and its Applications*, IEICE, Tokyo, pp 223–226
- Packard NH, Crutchfield JP, Farmer JD, Shaw RS (1980) Geometry from a time series. *Phys Rev Lett* 45: 712–716
- Pawelzik K, Kohlmorgen J, Müller K-R (1996) Annealed competition of experts for a segmentation and classification of switching dynamics. *Neural Comput* 8: 340–356
- Praetorius HM, Bodenstein G, Creutzfeldt O-D (1977) Adaptive segmentation of EEG records: a new approach to automatic EEG analysis. *Electroencephalogr Clin Neurophysiol* 42: 84–94
- Rabiner LR (1990) A tutorial on hidden Markov models and selected applications in speech recognition. In: Waibel A, Lee K (eds) *Readings in speech recognition*. Morgan Kaufmann, San Mateo, Calif., pp 267–296
- Rechtschaffen A, Kales A (1968) *A manual of standardized terminology: techniques and scoring systems for sleep stages of human subjects*. US Government Printing Office, Washington DC
- Rittweger J (1999) Common slow modulation of respiration, arterial blood pressure and cortical activity during sleep onset while napping. *Clin Physiol* 19: 3
- Rose K, Gurewitz E, Fox G (1990) Statistical Mechanics and phase transitions in clustering. *Phys Rev Lett* 65: 945–948
- Shi S, Weigend AS (1997) Taking time seriously: hidden Markov experts applied to financial engineering. In: *Proceedings of the IEEE/IAFE Conference on Computational Intelligence for Financial Engineering*, IEEE, NJ, pp 244–252
- Steriade M, McCarley RW (1990) *Brainstem control of wakefulness and sleep*. Plenum Press, New York
- Takens F (1981) Detecting strange attractors in turbulence. In: Rand D, Young L-S (eds) *Dynamical systems and turbulence*. (Springer Lecture Notes in Mathematics, vol 898) Springer, Berlin Heidelberg New York, pp 366–381
- Trinder J, Whitworth F, Kay A, Wilkin P (1992) Respiratory instability during sleep onset. *J Appl Physiol* 73: 2462–2469
- Weigend AS, Mangeas M (1995) Nonlinear gated experts for time series: discovering regimes and avoiding overfitting. *Int J Neural Syst* 6: 373–399