Hidden Markov Mixtures of Experts with an Application to EEG Recordings from Sleep

Stefan Liehr, Klaus Pawelzik

University of Bremen, Institute of Theoretical Neurophysics,
Kufsteiner Str., D–28334 Bremen, Germany

Jens Kohlmorgen, Klaus–Robert Müller

GMD FIRST,
Rudower Chaussee 5, D–12489 Berlin, Germany

Correspondence:

Stefan Liehr
Theoretische Physik, Abt. Neurophysik
Universität Bremen
Kufsteiner Str.
D–28334 Bremen, Germany
Tel.: +49-421-218-4460
Fax: +49-421-218-9104
email: sliehr@physik.uni-bremen.de

Key words: Time series, segmentation, nonstationarity, hidden Markov models, dynamical mode detection, EEG, sleep.

Summary: We present a framework for the analysis of time series from nonstationary dynamical systems that operate in multiple modes. The method detects mode changes and identifies the underlying subdynamics. It unifies the mixtures of experts approach and a generalized hidden Markov model with an input–dependent transition matrix. The adaptation of the individual experts and of the hidden Markov model is performed simultaneously. We illustrate the capabilities of our algorithm for chaotic time series and EEG recordings from human subjects during afternoon naps.

1 Introduction

It is a basic assumption of science that nature can be described by separating its complex structure into smaller parts which can be understood much more easily. On the other hand, it is well known that the interaction of even simple elements can produce very complex behavior. Often natural systems exhibit different kinds of nonstationary behavior generated by the interaction of coexisting subsystems. Therefore, prediction and classification of nonstationary dynamical systems may be performed better by identifying appropriate subdynamics and an early detection of the changes between these modes. Examples can be found in physics, biology, chemistry and climatology, but also in economic systems like financial markets.
Standard statistical techniques generally assume stationarity, i. e. they require that the underlying system is autonomous and does not change its parameters over time. If, however, the parameters of the system are varying in time, an analysis of the system can become very difficult. One approach to solve this problem was the application of efficient algorithms to short segments of the data, thereby monitoring possible changes in the characteristic quantities. These methods may suffer from the curse of dimensionality and other statistical problems that arise when estimating system parameters from small sample sizes.

A basic framework for dealing with nonstationarity is the mixtures of experts (ME) architecture, introduced in Jacobs et al. (1991). The mixtures of experts framework aims at separating the seemingly complex global behavior into a couple of lower dimensional subdynamics which can be modeled more easily.
To illustrate this approach consider the Lorenz system (Lorenz 1963). This rather simple nonlinear dynamical system exhibits switching between two different oscillatory modes where each each single oscillation can be described by an approximately linear dynamics near the corresponding fixed point. Two linear models would be a suitable choice in order to resolve the dynamical structure of the system. The nonlinearity of the system could be incorporated into a gating procedure that models the switching between the subdynamics. A central problem of using a set of experts is therefore the calculation of the activities of each expert — called the gating problem.
Many solutions have been proposed for dealing with the gating problem (Bengio and Frasconi 1995; Cacciatore and Nowlan 1994; Jacobs et al. 1991; Kehagias and Petridis 1997; Pawelzik et al. 1996; Shi and Weigend 1997; Weigend et al. 1995). In its original formulation, the mixtures of experts method can be applied to systems, where different regimes do not overlap in phase

space (i. e. the input space). The expert's activities are provided by a feed–forward gating network given the current location in phase space (Weigend et al. 1995). The use of a recurrent gating network (Cacciatore and Nowlan 1994) allows to distinguish also between overlapping regimes.

An alternative, non–recurrent approach to distinguish between overlapping regimes is the annealed competition of experts (ACE) method (Müller et al. 1995; Pawelzik et al. 1996). It has its roots in statistical mechanics and is a purely performance–driven concept, which considers a moving average prediction error for estimating the activities instead of using a gating network. An extension to the analysis of linear drifts between two dynamical modes was proposed later by using a dynamic programming approach based on a hidden Markov model (Kohlmorgen et al. 1997, 1998; Kohlmorgen 1998). This approach was successfully applied to the same EEG data sets that we investigate in this paper. A detailed comparison of the results, however, goes beyond the scope of this contribution.

All these approaches exhibit conceptual disadvantages. Some do not make use of all the available information for estimation of the expert activities (Jacobs et al. 1991; Weigend et al. 1995; Pawelzik et al. 1996; Kohlmorgen et al. 1997). Others use low-pass filters (Pawelzik et al. 1996), which can induce systematic delays when detecting switching events. Some approaches are inconsistent in training and application, because they include future information in order to calculate the expert's activities (Bengio and Frasconi 1995).

Here we present a novel framework that resolves the problems and inconstistencies of the previous methods mentioned above. It unifies the mixtures of experts approach and a generalized hidden Markov model with an input–dependent transition matrix: the Hidden Markov Mixtures of Experts (HMME). We apply a maximum likelihood learning method by using an Expectation–Maximization (EM) algorithm. The HMME approach is always trained in consistency with the later application: analysis or prediction. In the case of prediction, it can be used for an early detection of changes between dynamical regimes. These advantages are among the main differences to the IOHMM algorithm by Bengio and Frasconi (1995), who used a similar hybrid architecture. We illustrate our algorithm by using chaotic time series and EEG recordings from afternoon naps of healthy human subjects.

2 Hidden Markov Mixtures of Experts

2.1 The HMME architecture

We consider a modular dynamical system of K different models (experts), which are associated with discrete hidden dynamical states (modes). Figure 1 illustrates this concept. The overall prediction y_t^* of the modular system at time t is given by a linear superposition of the individual predictions y_t^k of each expert k depending on the input vector x_t:

$$y_t^*(x_t, \Theta) = \sum_{k=1}^{K} g_t^k(x_t, \Theta^G) y_t^k(x_t, \Theta^k) \quad \text{with} \quad \sum_{k=1}^{K} g_t^k(x_t, \Theta^G) = 1. \tag{1}$$

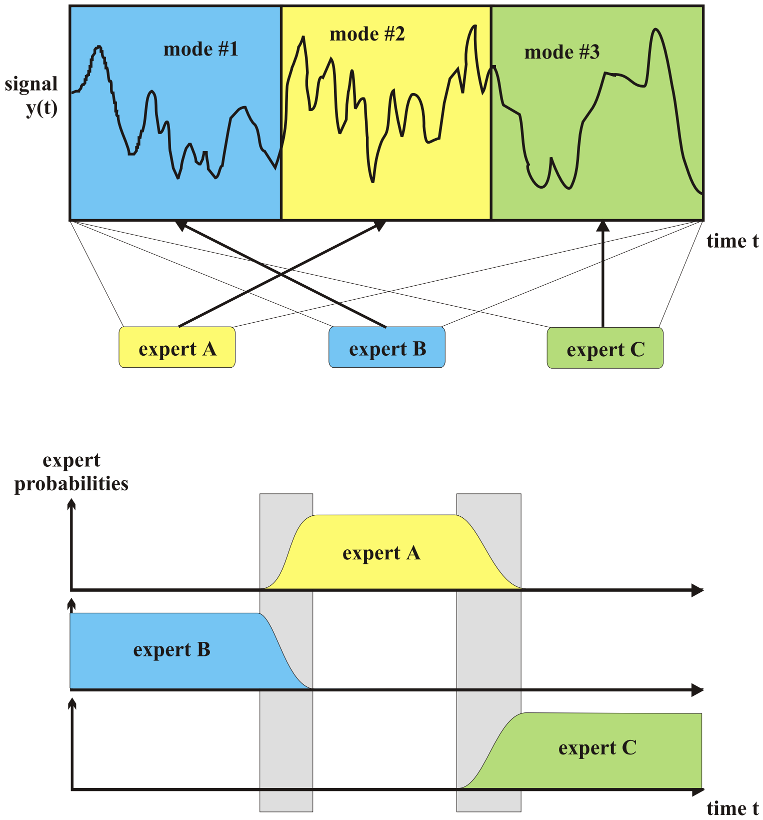


Figure 1: Illustration of the HMME architecture. The upper picture shows a caricature of a nonstationary time series consisting of three dynamical modes. The experts of a HMME system specialize on different dynamical modes. The picture below illustrates the corresponding probabilities g_t^k of the experts.

The dynamical state probabilities g_t^k are determined by the optimal trajectory through the hidden dynamical states. The calculation of this optimal path can be performed by using the theory of hidden Markov models (Rabiner 1988). In addition, we use an input–dependent transition matrix for dealing with phase space depending transition processes. The parameters of the HMME model are denoted by the overall parameter vector Θ which is composed of the individual parameter vectors Θ^k of the K experts and of the parameter vector Θ^G of the HMM.

2.2 The objective function

The objective function to be maximized is the likelihood L for observing the given input and output sequences $X = (x_1, \ldots, x_T)$ and $Y = (y_1, \ldots, y_T)$, the sequence of hidden dynamical modes $Q = (q_1, \ldots, q_T)$ and their mixed approximations $Y^* = (y_1^*, \ldots, y_T^*)$ depending on the model parameters Θ, i. e. the expected negative log–likelihood

$$\begin{aligned} R(\Theta) &= -\left\langle \log L(\Theta) \right\rangle_X & (2) \\ &= -\left\langle \log P\left(Y, Y^*, Q | X, \Theta\right) \right\rangle_X & (3) \end{aligned}$$

has to be minimized. The expectation $\langle . \rangle_X$ must be taken over the distribution of all possible input sequences. Using the Markov assumptions with one–step memory, the distribution of outputs given the inputs can be factorized into sums of products of two types of factors, output probabilities and transition probabilities:

$$P\left(Y, Y^*, Q | X, \Theta\right) = P\left(q_0 | \Theta\right) \prod_{t=1}^{T} P\left(y_t, y_t^* | q_t, x_t, \Theta\right) P\left(q_t | q_{t-1}, x_t, \Theta\right) \quad (4)$$

For consideration of the distribution of possible sequences of hidden dynamical modes it is common to introduce indicator variables z_t^k which are given by the sequence Q with $z_t^k = 1$ if $q_t = k$, and $z_t^k = 0$ if $q_t \neq k$, and q_0 being the starting state. This yields

$$\begin{aligned} R(\Theta) &= -\left\langle \log \prod_{k=1}^{K} P(q_0 = k|\Theta)^{z_1^k} + \log \prod_{t=1}^{T} \prod_{k=1}^{K} \left(P(y_t, y_t^* | q_t = k, x_t, \Theta)^{z_t^k} \cdot \right.\right. \\ &\qquad \left.\left. \prod_{k'=1}^{K} P(q_t = k | q_{t-1} = k', x_t, \Theta)^{z_t^k z_{t-1}^{k'}} \right) \right\rangle_X & (5) \\ &= -\sum_{k=1}^{K} g_1^k \log P(q_0 = k|\Theta) - \frac{1}{T} \sum_{t=1}^{T} \sum_{k=1}^{K} g_t^k \log P(y_t, y_t^* | q_t = k, x_t, \Theta) \\ &\quad - \frac{1}{T} \sum_{t=1}^{T} \sum_{k,k'=1}^{K} h_t^{k|k'} g_{t-1}^{k'} \log P(q_t = k | q_{t-1} = k', x_t, \Theta) & (6) \end{aligned}$$

where $g_t^k = \langle z_t^k \rangle_x$ and $h_t^{k|k'} g_{t-1}^{k'} = \langle z_t^k z_{t-1}^{k'} \rangle_x$ are the expected values of the indicator variables and their product. We assume Gaussian error distributions with the variance σ^2 of the estimated noise level of the data:

$$\begin{aligned} P(y_t, y_t^* | q_t{=}k, x_t, \Theta) &= P(y_t^* | y_t, x_t, \Theta) P(y_t | q_t{=}k, x_t, \Theta) && (7) \\ &= \frac{1}{2\pi\sigma^2} \exp\left(-\frac{(y_t - y_t^*)^2 + (y_t - y_t^k)^2}{2\sigma^2} \right) && (8) \end{aligned}$$

Further we assume equally distributed initial probabilities $g_0^k = P(q_0 = k|\Theta) = \frac{1}{K}$, $1 \le k \le K$. Using the notation $a_t^{k|k'} = a_t^{k|k'}(x_t, \Theta) := P(q_t = k | q_{t-1}{=}k', x_t, \Theta)$ from Rabiner (1988), the objective function can be written as

$$R(\Theta) = \frac{1}{2\sigma^2} \sum_{k=1}^{K} p^k \left(E^{*2} + E^{k^2} \right) + \sum_{k,k'=1}^{K} C^{k,k'} + \log\left(2\pi\sigma^2 K \right) \qquad (9)$$

with the mean squared errors of the mixture system E^{*2} and the experts E^{k^2}

$$E^{*2} = \frac{1}{T} \sum_{t=1}^{T} (y_t - y_t^*)^2 \quad \text{and} \quad E^{k^2} = \frac{1}{Tp^k} \sum_{t=1}^{T} g_t^k (y_t - y_t^*)^2 \qquad (10)$$

$$p^k = \frac{1}{T} \sum_{t=1}^{T} g_t^k \qquad (11)$$

and the entropy $C^{k,k'}$ between expected and modeled conditional activations of mode k and mode k'

$$C^{k,k'} = -\frac{1}{T} \sum_{t=1}^{T} h_t^{k|k'} g_{t-1}^{k'} \log a_t^{k|k'}. \qquad (12)$$

2.3 Training procedure

Training is performed by a generalized expectation–maximization (GEM) algorithm (Dempster et al. 1977), because in general there is no way to analytically maximize the objective function $R(\Theta)$. The E–step consists of estimating the probabilities, the M–step adapts the models by minimizing the objective function using gradient descent. Since in the M–step the probabilities are considered to be constant, the derivatives of the objective function can be simplified drastically. Because the output sequence Y plays the role of a target value, the adaptation of the experts is a supervised learning problem. On the other hand there is no direct desired value given for adaptation of the

HMM model. Therefore, the detection of dynamical states and their switching behavior is an unsupervised learning problem. The partial derivatives with respect to the constituents of the parameter vector Θ are

$$\begin{aligned} \frac{\partial R(\Theta)}{\partial \Theta^k} &= -\frac{1}{T}\frac{1}{\sigma^2}\sum_{t=1}^{T} g_t^k \left((y_t - y_t^*) + (y_t - y_t^k)\right)\frac{\partial y_t^k}{\partial \Theta^k} && (13) \\ \frac{\partial R(\Theta)}{\partial \Theta^G} &= -\frac{1}{T}\sum_{t=1}^{T}\sum_{k,k'=1}^{K} g_{t-1}^{k'}\left(\frac{h_t^{k|k'}}{a_t^{k|k'}} - 1\right)\frac{\partial a_t^{k|k'}}{\partial \Theta^G}. && (14) \end{aligned}$$

Equation (14) is calculated using the method of Lagrange multipliers for incorporating the normalization conditions $\sum_{k=1}^{K} a_t^{k|k'} = 1$ and $\sum_{k,k'=1}^{K} h_t^{k|k'} = 1$. Furthermore, this equation is equivalent to the derivative of the Kullback–Leibler distance between the HMM transition matrix and the estimated conditional probabilities $h_t^{k|k'}$. Thus, the adaptation of the transition matrix can be interpreted as minimizing this measure.

During training, the HMM learns to predict the expected conditional probabilities $h_t^{k|k'}$, which are calculated according to the theory of hidden Markov models using the forward and backward probabilities $\alpha_t^k := P(y_1^t, q_t = k|x_1^t, \Theta^k)$ and $\beta_t^k := P(y_{t+1}^T, q_t = k|x_t^T, \Theta^k)$:

$$h_t^{k|k'} = \frac{\beta_t^k r_t^k h_t^{k|k'} \alpha_{t-1}^{k'}}{\sum_l \beta_t^l r_t^l h_t^{l|l'} \alpha_{t-1}^{l'}} \qquad (15)$$

using the observation probabilities

$$r_t^k = \frac{\exp\left(-\frac{1}{2\sigma^2}(y_t - y_t^k)^2\right)}{\sum_l \exp\left(-\frac{1}{2\sigma^2}(y_t - y_t^l)^2\right)} \qquad (16)$$

For classification tasks, the calculation of expected state probabilities g_t^k can be performed using the normalized HMM state probability that contains past and future information:

$$g_t^k \overset{acausal}{\longrightarrow} \hat{\gamma}_t^k = \frac{\beta_t^k \alpha_t^k}{\sum_l \beta_t^l \alpha_t^l} \qquad (17)$$

In the case of prediction, however, no future information is available. For consistency between training and application of the algorithm in that case, the expected state probabilities have to be calculated by an iterated procedure of a–priori and a–posteriori probabilities:

$$\begin{aligned} g_t^k \overset{causal}{\longrightarrow} p_t^{k,prior} &= \sum_{k'=1}^{K} a_t^{k|k'} \cdot p_{t-1}^{k,post} && (18) \\ p_t^{k,post} &= \frac{r_t^k \cdot p_t^{k,prior}}{\sum_l r_t^l \cdot p_t^{l,prior}} && (19) \end{aligned}$$

The a–posteriori probability can be identified with the normalized forward probabilities $\hat{\alpha}_t^k = \alpha_t^k / \sum_l \alpha_t^l$.

2.4 Simulated annealing

The method of simulated annealing (Kirkpatrick et al. 1983; Kirkpatrick 1984) is suitable for optimization problems, where a global optimum is hidden among many local extrema. The standard scheme for finding minima (maxima) is going downhill (uphill) as far as possible. This often leads to a local but not necessarily global extremum.

Transfered to the adaptation of the HMME, without annealing, the algorithm most probably gets stuck in a non–optimal segmentation of the data. Simulated annealing introduces a probabilistic component in the training process by using the temperature–like parameter σ^2 of equation (8) with a small amount of noise on the HMME parameters Θ. In our context, the annealing parameter can be interpreted as a competition factor. First, at high temperature (without competition), all data points are uniformly distributed among the experts. Finally, at low temperature (hard competition), each data point is associated with only one expert exclusively. The essence of the process is a slow decrease of the temperature, allowing ample time for redistribution of the experts. The annealing process ensures a slow evolution of the shape of the objective function in parameter space and allows the system to get out of a local extremum. With this method the experts successively specialize in a hierarchical manner via a series of phase transitions (Pawelzik et al. 1996), an effect which has also been analyzed in the context of clustering (Rose et al. 1990).

3 Detection and prediction of subdynamics

In order to demonstrate the performance of the algorithm, we first applied it to the Lorenz system (Lorenz 1963) which is given by a set of three coupled differential equations

$$\begin{aligned} \dot{X} &= -\sigma X + \sigma Y \\ \dot{Y} &= -XZ + rX - Y \\ \dot{Z} &= XY - bZ. \end{aligned} \tag{20}$$

With the chosen parameters, $\sigma = 16$, $b = 4$, and $r = 45.92$, the dynamics exhibits a switching behavior between oscillations around two fix–points. The system is globally nonlinear, with the strongest nonlinearity near the switching area from one oscillatory wing to the other, while each single oscillation can be assumed to be approximately linear near the corresponding fix–point. Therefore, we choose two linear experts and a nonlinear radial basis functions network of Moody Darken type (Moody and Darken 1989) for modeling the HMME transition matrix:

$$y_t^k = \Theta_0^k + \sum_{n=1}^{d} \Theta_n^k x_{t,n} \tag{21}$$

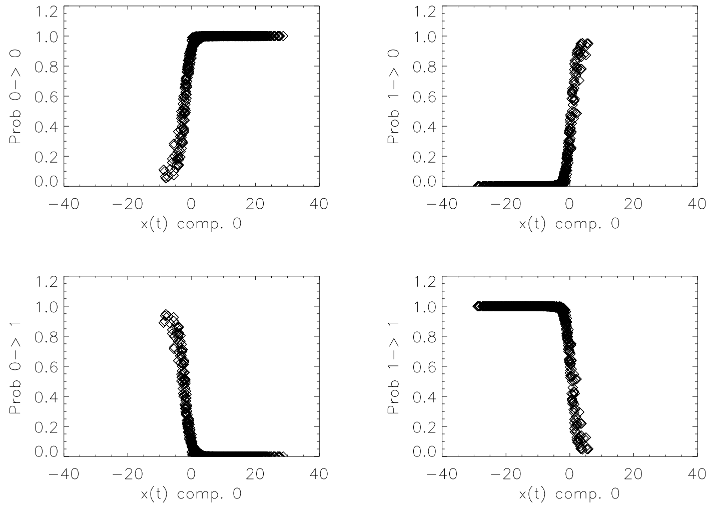


Figure 2: Input–dependent transition matrix of the HMME model learned from the Lorenz dynamics. Each of the four pictures shows the transition probability Prob($k' \to k$) from expert $\#k'$ to expert $\#k$ depending on the X component of the state vector.

$$
\begin{aligned}
a_t^{k|k'} \quad &= \quad \sum_{m=1}^{M} w_m^{k,k'} \frac{\gamma_m(x_t)}{\sum_{m'} \gamma_{m'}(x_t)} \qquad (22)\\
&\text{with } \gamma_m(x_t) = e^{-\frac{(x_t - z_m)^2}{2\sigma_m{}^2}}\\
&\text{and } \Theta^G = \{w_m^{k,k'}, z_m, \sigma_m\}
\end{aligned}
$$

The RBF–network consists of $M = 20$ centers. The nonlinearity is thus only incorporated into the gating procedure. The input and output of the experts are given by the state vector (X, Y, Z) of the Lorenz system.

Figure 2 shows the final estimation of the input–dependent transition matrix of the HMME. The functional dependency of the four elements of the transition matrix is projected onto the plane (X/Prob) between the X–coordinate of the state vector and the transition probability. Obviously, each expert specializes on one oscillation and the transition matrix forces the probability evolution to follow that segmentation of the dynamics. The projection reflects also the switching behavior of the Lorenz dynamics near $X = 0$. The algorithm can follow even short–term mode changes which is shown in Figure 3 (see also Liehr et al. 1999).

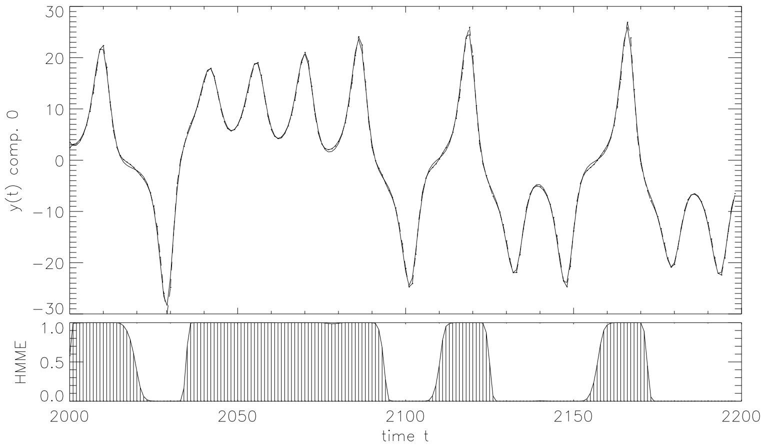


Figure 3: The upper picture shows a part of the Lorenz dynamics, the small lower panel shows the probability evolution of one of the two experts of the HMME and demonstrates the ability to detect changes between the different dynamical regimes of that process very quickly.

4 Analysis of EEG data

To illustrate the performance of our algorithm on experimental data we analyzed EEG recorded during afternoon naps of a healthy human. The objective was to give a detailed description of the signal dynamics with a high time resolution, ultimately to detect the sleep onset in an unsupervised manner.

In general practice, the analysis of sleep and the segmentation into different modes depends strongly on the specific experience and intuition of the medical expert who conveys it. Moreover, manual analysis is rather time consuming. Furthermore, disagreements of classification between different medical experts are of the order 10-15%.

Our HMME algorithm might be a tool in order to avoid these problems. First, it provides an objective method with an exactly defined performance function for segmentation of the data and for modeling the individual dynamical modes. Second, after training, the analysis of new data is very fast and might be performed online while recording the signal. The EEG is a signal with high time resolution of typically 100 Hz up to 1000 Hz. Therefore, it potentially allows to determine the sleep onset more accurately than other physiological signals like the EOG. The EEG data we used here, was first analyzed in (Kohlmorgen et al. 1997). For a more detailed analysis, see (Kohlmorgen 1998; J. Kohlmorgen and Pawelzik 1999).

For this study, we analyzed two single–channel EEG recordings from one subject, sampled with 100 Hz, both shown in the top of Figures 4 and 5. In order to reconstruct the dynamical space, we embedded the EEG time series $\{s_t\}$ with an embedding dimension $d = 50$ and a delay of $\tau = 2$ (20 ms). Thus, the input vector is $x_t = (s_t, s_{t-\tau}, \ldots, s_{t-(d-1)\tau})$ and the output or target value is $y_t = s_{t+\tau}$. The method of reconstruction is based on fundamental theoretical work and common in nonlinear dynamics (Packard et al. 1980; Takens 1981). Our choice of embedding parameters was motivated to capture the most important frequency domains in the power spectrum of the EEG signal.

We used a set of four linear expert models and the transition matrix was modeled by an RBF network, as given in equations (21) and (22). The RBF network consists of $M = 5$ centers. For smoothing out probability fluctuations of g_t^k and emphasizing the dynamical structure, a short low-pass filter of 1 s was used for illustration purposes.

The HMME was trained on the first data set NP-11, then we used it for segmentation of the same data and an additional test data set NP-13, which was not presented during training.

The segmentation of NP-11 using the HMME that was trained on this data set is shown in Figure 4. For comparison, the manual segmentation of the

EEG data is given in the second plot of the same figure. The filled region shows the 1 s low-pass filtered segmentation of the HMME. A remarkable result is the dominance of only two experts, expert #4 during the wake state and expert #1 during the sleep state. They receive the largest amount of data and are sufficient for explaining the main dynamical parts. The sleep onset at 6.5 min and the final arousal at 16 min are indicated correctly in the segmentation. The second of the two intermediate arousals at 12 min is likewise indicated correctly. The first intermediate arousal at 9 min, however, is not as prominent as the second one.

Comparing the HMME segmentation with the segmentation of the medical expert we see that the probability of expert #3 shows some coincidences with regions of the EEG recording which were classified as artifacts (2.5 min, 5 min, 6.5 min) but it shows also small peaks of high probability which can not be interpreted within the given manual segmentation. Expert #2 shows a higher occurance of activity in regions of changes in the EEG like the intermediate or final arousal, but it is also hard to interpret its behavior. In order to assess the generalization ability of the approach with respect to its segmentation capabilities, we applied the HMME for the segmentation of the data set NP-13. As shown in Figure 5, the overall structure of the obtained drift segmentation is again in good agreement with the hand labeling. Likewise, the responsibilities of the experts are the same as mentioned above. In particular, two short intermediate arousals during the nap are nicely represented in the segmentation by the probabilities of experts #1 and #4. Thus, one advantage of reusing a previously trained HMME is that one only has to label the experts once, after training.

Dataset NP-11

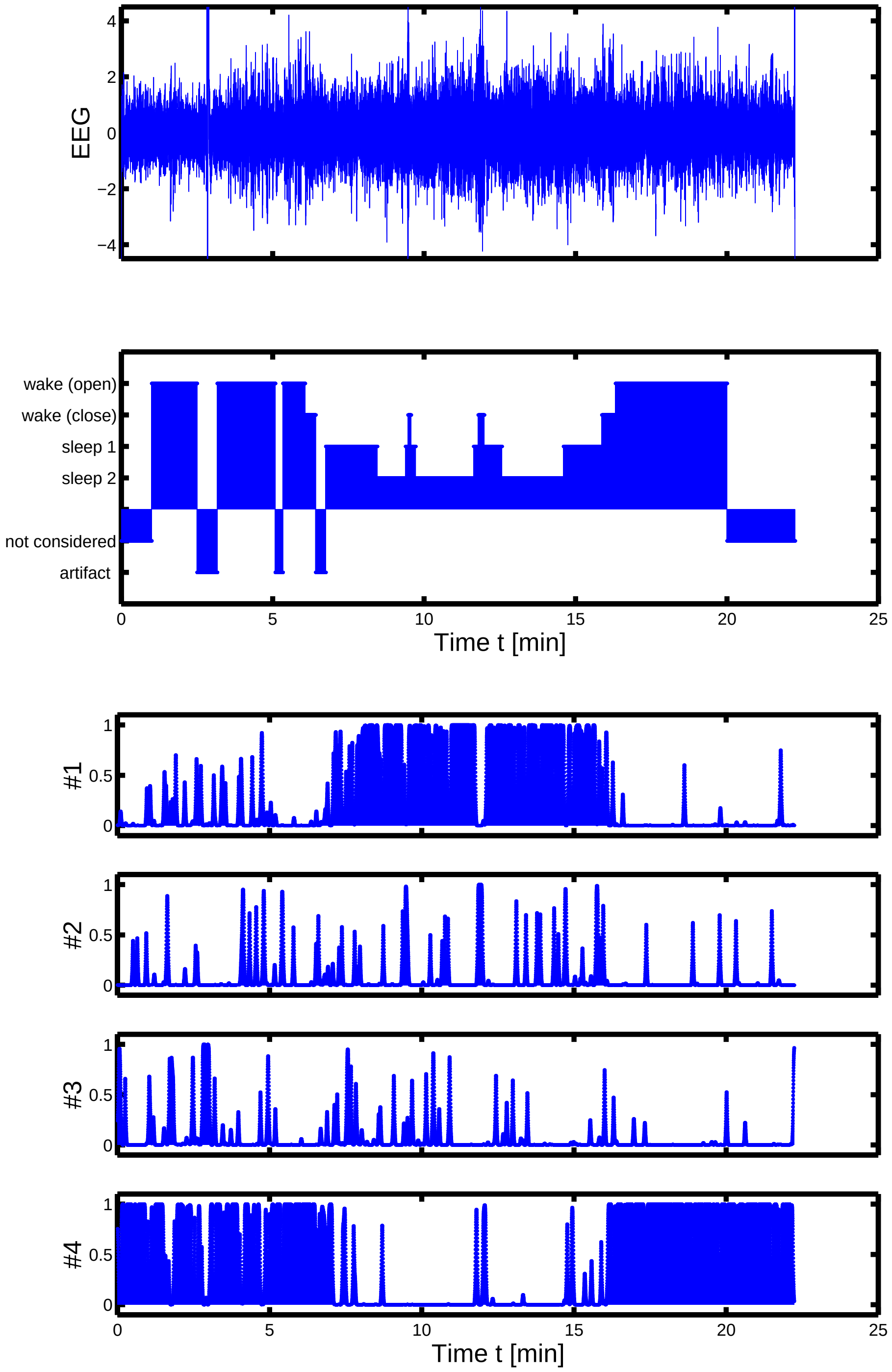
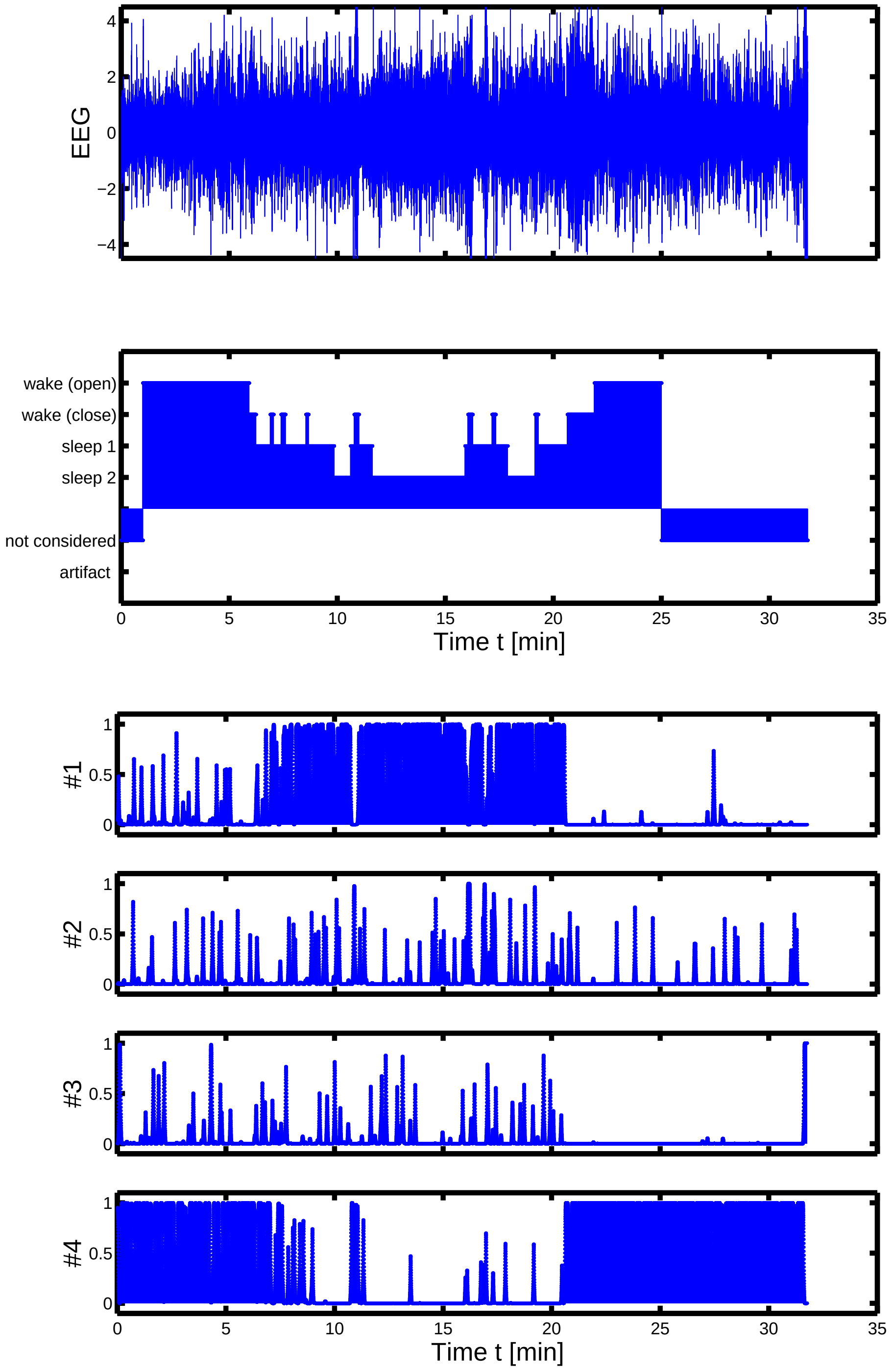


Figure 4: Dataset NP-11. *Top:* A single channel EEG-O1 (occipital-1) recording of an afternoon nap of about 20 min is used for training the HMME. The sampling time was 10 ms. *Second plot:* Result of a manual segmentation performed by a medical expert. It is based on the following eight physiological signals (each EEG channel corresponds to a recording at the indicated electrode position): EEG-O1 (occipital-1), EEG-O2 (occipital-2), EEG-F3 (frontal-3), EOG (electrooculogram), ECG (electrocardiogram), heart rate, blood pressure and respiration. Two wake states with eyes opened and eyes closed, and sleep stages 1 and 2 are classified. Some time regions have not been considered or are classified as artifacts. *Four panels below:* HMME segmentation of the NP-11 EEG recording. The evolution of the four individual expert probabilities g_t^k are marked by #1 to #4. Probabilities are smoothed with a Gaussian filter of 1 s standard deviation. Transitions from wake state to sleep onset and back are detected clearly by transitions between expert #4 and expert #1. Also the short intermediate arousals can be resolved. Experts #2 and #3 specialize on artifacts and very localized dynamical structures which can not be interpreted on the considered time scale.

Figure 5: Dataset NP-13. *Top:* Another single channel EEG-O1 (occipital-1) of an afternoon nap of about 30 min is used for testing the generalization performance. The sampling time of the EEG recording was again 10 ms. *Second plot:* Manual segmentation based on the same physiological signals as explained in Figure 4. *Four panels below:* The four panels show the segmentation of the NP-13 EEG recording in order to test the generalization performance of the HMME. Probabilities g_t^k are smoothed with a Gaussian filter of 1 s standard deviation. The segmentation by the HMME shows again a good agreement with the manual segmentation. Both short arousals at 11 min and 16–18 min are clearly indicated.

Dataset NP-13



5 Conclusion

A combined supervised and unsupervised method for identification and segmentation of nonstationary dynamics was presented. It applies to time series of dynamical systems that alternate among different operating modes. An application to EEG data demonstrated that dynamical structure can be resolved by our approach in an unsupervised manner.

We would like to emphasize that the method neither needs prior information whether the time series contains multiple modes, nor what the dynamics of the operating modes looks like. Instead of using a single but complex predictor, we apply a divide and conquer strategies which forces a set of competing predictors to specialize on sub–sequences of the data. Thereby, a segmentation of the data and an identification of the individual dynamics is developed simultaneously.

The experiments on chaotic time series showed the proof of concept for our HMME algorithm. In case of physiological wake/sleep data, the results are in so far encouraging as our mathematical model worked well on real-world data and was capable to identify wakefulness and transitions from wake to sleep in EEG recordings.

We consider our method as a first step towards new algorithms suitable not only for better EEG analysis but also for the analysis of complex nonstationary time series in general.

Acknowledgement: We acknowledge support of the Deutsche Forschungsgemeinschaft (grants Pa569/2–1 and Ja379/52). Jörn Rittweger is gratefully acknowledged for many helpful discussions and for providing the physiological data.

References

Bengio, Y. and P. Frasconi (1995). An input/output HMM architecture. In G. Tesauro, D. Touretzky, and T. Leen (Eds.), *NIPS'94: Advances in Neural Information Processing Systems 7*, pp. 427–434. Cambridge, MA: Morgan Kaufmann, MIT Press.

Cacciatore, T. W. and S. J. Nowlan (1994). Mixtures of controllers for jump linear and non–linear plants. In *NIPS'93: Advances in Neural Information Processing Systems*, pp. 719–726. Morgan Kaufmann, MIT Press.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum–likelihood from incomplete data via the EM algorithm. *Journal of Royal Statistical Society B 39*, 1–38.

J. Kohlmorgen, K.-R. Müeller, J. R. and K. Pawelzik (1999). Identification of non-stationary dynamics in physiological recordings. *Biological Cybernetics 83*, 73–84.

Jacobs, R. A., M. I. Jordan, S. J. Nowlan, and G. E. Hinton (1991). Adaptive mixtures of local experts. *Neural Computation 3*, 79–87.

Kehagias, A. and V. Petridis (1997). Time series segmentation using predictive modular neural networks. *Neural Computation 9*, 1691–1710.

Kirkpatrick, S. (1984). Optimization by simulated annealing – quantitative studies. *J. Stat. Phys. 34*, 975–986.

Kirkpatrick, S., C. D. Gelatt, and M. P. Vecchi (1983). Optimization by simulated annealing. *Science 220*, 671–680.

Kohlmorgen, J. (1998, July). *Analyse schaltender und driftender Dynamik mit neuronalen Netzen.* Ph.D. thesis (in german), GMD Berlin, GMD Research Series 22/1998, ISBN 3-88457-346-2, Sankt Augustin.

Kohlmorgen, J., K.-R. Müller, and K. Pawelzik (1997). Segmentation and identification of drifting dynamical systems. In J. Principe, L. Giles, N. Morgan, and E. Wilson (Eds.), *NNSP '97: IEEE Workshp on Neural Networks for Signal Processing*, pp. 326–335. IEEE.

Kohlmorgen, J., K.-R. Müller, and K. Pawelzik (1998). Analysis of drifting dynamics with neural network hidden markov models. In *NIPS '97: Advances in Neural Information Processing Systems 10*, pp. 735–741. MIT Press.

Liehr, S., K. Pawelzik, J. Kohlmorgen, S. Lemm, and K. R. Müller (1999). Hidden Markov mixtures of experts for prediction of non–stationary dynamics. In *NNSP'99 Workshop on Neural Networks for Signal Processing*, pp. 195–204. IEEE, NY.

Lorenz, E. N. (1963). Deterministic nonperiodic flow. *J. Atmos. Sci. 20*, 130–141.

Moody, J. and C. J. Darken (1989). Fast learning in networks of locally–tuned processing units. *Neural Computation 1*, 281–294.

Müller, K.-R., J. Kohlmorgen, J. Rittweger, and K. Pawelzik (1995). Analysing physiological data from the wake–sleep state transition with competing predictors. In *NOLTA'95: Las Vegas Symposium on Nonlinear Theory and its Applications*, pp. 223–226.

Packard, N. H., J. P. Crutchfield, and J. D. Farmer (1980). Geometry from a time series. *Physical Review Letters 45*, 712–716.

Pawelzik, K., J. Kohlmorgen, and K.-R. Müller (1996). Annealed competition of experts for a segmentation and classification of switching dynamics. *Neural Computation 8*, 342–358.

Rabiner, L. R. (1988). A tutorial on hidden Markov models and selected applications in speech recognition. In A. Waibel and K. Lee (Eds.), *Readings in Speech Recognition*, pp. 267–296. Morgan Kaufmann.

Rose, K., E. Gurewitz, and G. Fox (1990). Statistical mechanics and phase transitions in clustering. *Physical Review Letters 65*, 945–948.

Shi, S. and A. S. Weigend (1997). Taking time seriously: Hidden Markov experts applied to financial engineering. In *Proceedings of the IEEE/IAFE Conference on Computational Intelligence for Financial Engineering*, New York, pp. 244–252.

Takens, F. (1981). Detecting strange attractors in turbulence. *Lecture Notes in Math. 898*, 366–381.

Weigend, A. S., M. Mangeas, and A. N. Srivastava (1995). Non–linear gated experts for time series: discovering regimes and avoid overfitting. *International Journal of Neural Systems 6*, 373–399.