TRACKING AND VISUALIZATION OF CHANGES IN HIGH-DIMENSIONAL NON-PARAMETRIC DISTRIBUTIONS

Jens Kohlmorgen
Fraunhofer FIRST.IDA
Kekuléstr. 7, D-12489 Berlin, Germany
E-mail: jek@first.fraunhofer.de

Abstract. Most real-world systems exhibit a non-stationary behavior, e.g., slow drifts due to wear or fast changes due to external influences. Extracting and quantifying these phenomena is often difficult due to the lack of a precise mathematical model of the underlying system. We here propose to model such high-level changes of a dynamical system solely on the basis of the observed measurements rather than by modeling the underlying system itself. In particular, we present a method to track and visualize changes in general data distributions. We approach the problem of how to represent continuous changes in high-dimensional non-parametric distributions by identifying anchor distributions and we model the transitions between those anchor distributions by defining a suitable similarity measure. Applications to a high-dimensional chaotic system and to a sleep-onset detection task in EEG demonstrate the efficiency of this approach.

INTRODUCTION

Non-stationary behavior is ubiquitous in many real-world systems like, for example, speech, climatological data, physiological recordings (EEG/MEG), industrial processes or financial markets. Methods for the analysis of time-varying dynamical systems are therefore important in many application areas. In [11], we introduced a method for time series from non-linear switching dynamics, where an ensemble of radial basis function predictors specializes on different dynamical regimes by increasing the competition among the predictors through a deterministic annealing scheme. Related approaches for switching dynamics were presented, e.g., in [1, 2, 3, 8, 12]. In [6], we extended the ability of the competing predictors method 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 (EEG and respiration) can be modeled more appropriately by a drifting dynamics

model [7]. Prediction of EEG signals, however, is a difficult task, which led us to the development of a segmentation method that does not depend on the predictability of the system but merely on time-dependent variations of the underlying *density* function of the data [5]. In [5], these changes are the criterion to find a segmentation in terms of a sequence of prototypical (static) density function estimates, each one optimally approximating the time-dependent density function in the corresponding segment.

Instead of providing merely a discrete segmentation, we here explore the possibility of extracting and visualizing the exact time course of the density changes. To this end, we propose a measure that quantifies the changes in the data distribution based on the obtained prototypical density estimates. The resulting "soft" segmentation is similar to the drift concept of the aforementioned prediction-based approach [6] and we illustrate the method by applications to artificially generated data and to an EEG sleep-onset detection task.

DENSITY TRACKING

Consider the problem of tracking a high-dimensional non-parametric density given just a sequence of data points that are drawn from an unknown time-varying density function. In order to quantify and visualize the evolution of such a high-dimensional system, it is desirable to obtain a *scalar* time-dependent quantity that represents its characteristic behavior. To this end, we propose to make use of the segmentation method in [5] and utilize the sequence of prototypical density estimates provided by that method as anchor points in the space of density functions. Density functions estimated between these prototypes can then be represented by a measure that indicates the relative distance to their preceding and succeeding anchor densities. In this way, the time course of the transition between the anchor densities can be modeled.

We first give a brief review of the segmentation approach in [5] and then introduce the distance measure that quantifies the drift between the anchor distributions. Let $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_T$, with $\vec{x}_t \in R^d$, be a sequence of data points that are drawn from an unknown time-varying density function. In order to track the changing density distribution of the data, we estimate a probability density function (pdf) in a sliding window of length W by using a Parzen window density estimator [10] with multivariate Gaussian kernels, centered on the data points[1] in the window $\{\vec{x}_{t-w}\}_{w=0}^{W-1}$,

$$p_t(\mathbf{x}) = \frac{1}{W} \sum_{w=0}^{W-1} \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left(-\frac{(\mathbf{x} - \vec{x}_{t-w})^2}{2\sigma^2}\right). \quad (1)$$

The kernel width σ acts as a smoothing parameter and its value is important

[1] In the following we use $\vec{x}$ to denote a specific vector-valued *point* and $\mathbf{x}$ to denote a vector-valued *variable*.

to obtain a good representation of the underlying distribution. We choose σ proportional to the mean distance of each $\vec{x}_t$ to its first k nearest neighbors, averaged over a sample set $\{\vec{x}_t\}$. We typically choose $k = d$, and – somewhat surprising – we didn't find the choice of k to be a critical issue in our simulations, where the k nearest neighbors averaging turned out to be a robust way to obtain a reasonable σ.

For the given data sequence, $\{\vec{x}_t\}_{t=1}^T$, we can obtain the corresponding sequence of pdfs, $\{p_t(\mathbf{x})\}_{t=W}^T$, according to Eq. (1). The unsupervised segmentation approach in [5] then yields a compact representation of this sequence of pdfs by extracting a small subset of pdfs, whose elements are called *prototypes*, and a sequence $\mathbf{s} = \{s(t)\}_{t=W}^T$ with $s(t) \in \{W, ..., T\}$, called *segmentation*, that assigns each pdf in the sequence to a prototype pdf, such that the cost function

$$o(\mathbf{s}) = \sum_{t=W}^{T} d(p_t, p_{s(t)}) \; + \; C \; n(\mathbf{s}) \tag{2}$$

is minimized. The function d quantifies the difference between a pdf p_t and its assigned prototype $p_{s(t)}$. In [5], the squared L_2-Norm, also called *integrated squared error* (ISE), $d(p_t, p_{\bar{t}}) = \int (p_t(\mathbf{x}) - p_{\bar{t}}(\mathbf{x}))^2 \, d\mathbf{x}$, is used, which has the advantage that it can be calculated analytically, if – as in our case – p_t and $p_{\bar{t}}$ are mixtures of Gaussians. By using Eq. (1), we obtain

$$\begin{aligned} d\left(p_t, p_{\bar{t}}\right) = \frac{1}{W^2 \left(4\pi\sigma^2\right)^{d/2}} \sum_{w,v=0}^{W-1} \Bigg[& \exp\left(-\frac{(\vec{x}_{\bar{t}-w} - \vec{x}_{\bar{t}-v})^2}{4\sigma^2}\right) \\ & -2\exp\left(-\frac{(\vec{x}_{\bar{t}-w} - \vec{x}_{t-v})^2}{4\sigma^2}\right) + \exp\left(-\frac{(\vec{x}_{t-w} - \vec{x}_{t-v})^2}{4\sigma^2}\right)\Bigg] \end{aligned} \tag{3}$$

The function n in Eq. (2) denotes the number of prototype changes in the sequence:[2] $n(\mathbf{s}) = \sum_{t=W}^{T-1}(1 - \delta_{s(t),s(t+1)})$. Thus, the cost function o penalizes the number of prototype changes in a sequence to an extent that is controlled by the regularization parameter $C > 0$. As a result, the cost function is biased towards large segments in the segmentation and in this way it favors compact representations. An appropriate value for the hyperparameter C, which thus controls the number of prototype pdfs, can be found by the method described in [4]. The optimal segmentation $\mathbf{s}^*$ with minimal costs $o(\mathbf{s}^*) = \min_{\mathbf{s}}\{o(\mathbf{s})\}$ can be computed efficiently in $O(T^2)$ time by *dynamic programming* [4]:

Optimal C–Segmentation Algorithm

1. Initialization ($t = W$):

$$\forall s \in \{W, ..., T\}: \;\; o_s(t) \; := \; d(p_t, p_s) \tag{4}$$

[2] In the following definition, $\delta_{i,j}$ denotes the Kronecker delta function, i.e. $\delta_{i,j} = 1$, if $i = j$, and $\delta_{i,j} = 0$ otherwise.

2. Recursion (for $t = W+1, ..., T$):

$$h \;:=\; \min_{s\in\{W,...,T\}} \Big\{\; o_s(t-1) \;\Big\} + C \tag{5}$$

$$\forall s \in \{W, ..., T\}\colon\;\; o_s(t) \;:=\; d(p_t, p_s) + \min\Big\{\; o_s(t-1),\; h \;\Big\} \tag{6}$$

3. Termination:

$$o^* \;:=\; \min_{s\in\{W,...,T\}} \Big\{\; o_s(T) \;\Big\} \tag{7}$$

The segmentation $\mathbf{s}^*$ that corresponds to the minimal costs o^* must then be obtained by backtracking through the sequence of states s that make up o^*.

We then use the resulting prototype pdfs as anchor distributions and quantify the evolution of the density function between two consecutive prototypes, p_a and p_b, in terms of the distance to the intermediate pdfs, p_t,

$$r_{[a,b]}(t) = \frac{d(p_t, p_a) - d(p_t, p_b)}{2\, d(p_a, p_b)} + \frac{1}{2} \tag{8}$$

We call $r_{[a,b]}(t)$ the *drift coefficient*. Its value is zero, if $p_t = p_a$, and it is one, if $p_t = p_b$. If an intermediate pdf has the same distance to both prototypes, p_a and p_b, then $r_{[a,b]}(t) = 0.5$. Note that the drift coefficient can also assume negative values, namely if the distance of p_t to p_b is so large that $d(p_t, p_b) > d(p_t, p_a) + d(p_a, p_b)$. This can be interpreted as p_t being "beyond" p_a rather than "between" p_a and p_b. Vice versa, the drift coefficient is larger than one, if p_t is "beyond" p_b, i.e. if $d(p_t, p_a) > d(p_t, p_b) + d(p_a, p_b)$.

In the following section, we apply the proposed drift measure to artificial and EEG data and illustrate its use.

APPLICATIONS

Drifting Uniform Distribution.

We start with a simple example to demonstrate the properties of the proposed drift measure. To this end, we generated a scalar time series drawn from a uniform distribution in the interval $[0, 1]$. After drawing 800 points, a linearly growing offset is added to the data, such that the sampling interval moves linearly towards $[3, 4]$ while another 800 points are drawn. Thereafter, 800 points are drawn from $[3, 4]$ (see Fig. 1). We then run the segmentation algorithm on the obtained time series, consisting of 2400 points, using a pdf window size $W = 100$. The resulting segmentation with two prototypes is shown in Fig. 1: the location of the two prototype pdf windows is depicted by thick black dots (indicating the center of the window) and the black line between the dots indicates the value of the drift coefficient for the intermediate pdf estimates. There is a gap in the line, which indicates that the pdf estimates in this region moved out of the neighborhood of both prototype pdfs. This condition is given by

$$d(p_t, p_a) > 0.95\; d(p_a, p_b) \quad \text{and} \quad d(p_t, p_b) > 0.95\; d(p_a, p_b), \tag{9}$$

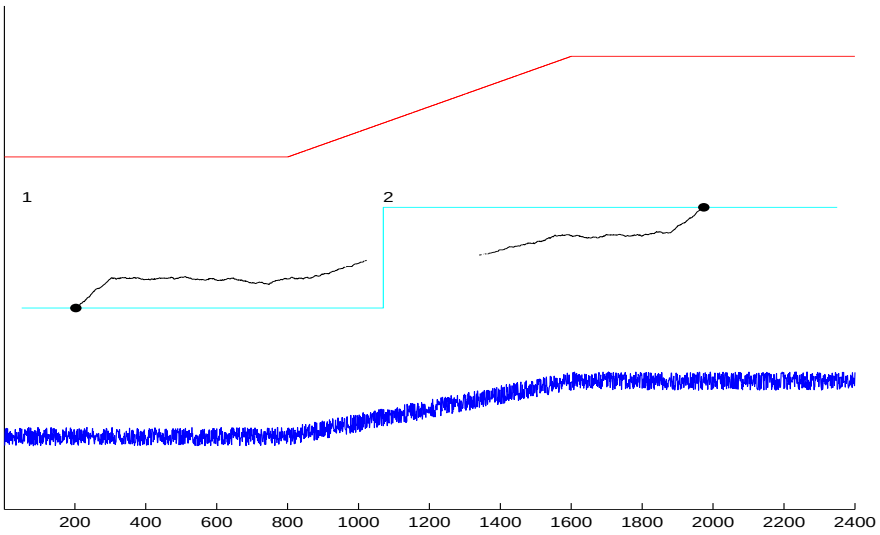

Figure 1: Drift segmentation of a drifting uniform distribution (lower curve): data drawn from a uniform distribution was shifted by a time-varying offset (top) and subsequently analyzed by the algorithm (middle curve). The drift coefficient (black line) indicates the deviation of intermediate pdf estimates in relation to the two obtained reference pdfs (black dots). The gap in the middle of the plot indicates that the pdf estimates moved out of the scope of both reference densities.

which we found to be a useful heuristic to indicate a drift away from *both* reference pdfs. In the example, all data points in the gap area are within the interval $[1, 3]$ and thus do not have any overlap with the data that make up the reference pdfs. Both prototype pdfs then have approximately the same (large) distance to the pdf estimates in the gap area and the drift coefficient therefore has a value close to 0.5. We found the blank-out rule in (9) a useful feature to distinguish this "out-of-range" situation from a real "in-between" density that also has a value of 0.5. This also highlights a property of the ISE: if the densities are non-overlapping, its value is independent of how far the probability masses actually are apart. Gaps occurring in the drift curve, due to the rule in (9), therefore suggests to use more anchor distributions in order to avoid that intermediate densities are totally disconnected from their corresponding anchor distributions. Fig. 2 shows the respective drift curve when using three anchor distributions for the given task. This solution was obtained by using a smaller regularization parameter C.

As can be seen in both Figures, 1 and 2, there is a systematic, almost linear increase of the drift coefficient in the vicinity of the prototype pdfs. This is due to the fact that the sequence of pdfs is estimated from overlapping windows of size W, which are shifted by just one data point at a time. Thus, only those pdfs that are at least W time steps away from each other are estimated from disjoint data sets. The drift coefficient for pdfs that are *more* than W time steps away from the prototypes, but still estimated from the same underlying distribution, indicates the sample variance – for pdfs

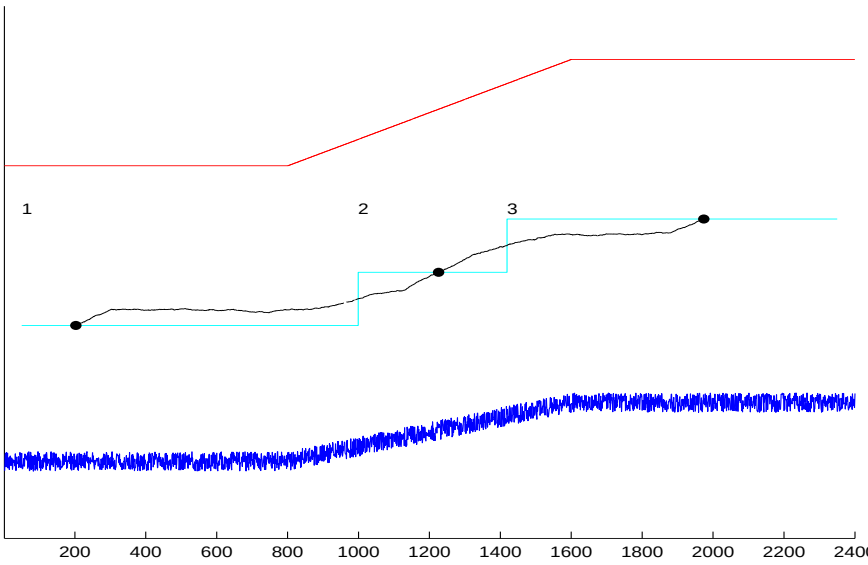

Figure 2: Drift segmentation as in Fig. 1, but with 3 reference pdfs: the central pdf estimates are now covered by the additional anchor pdf (indicated by the dot in the center) and the gap (Fig. 1) is closed.

that are closer, it indicates the window overlap. The actual drift of the data distribution is indicated between $t = 800$ and $t = 1600$. In Fig. 1, it is disrupted by the gap (due to the missing anchor pdf), and in Fig. 2 it is somewhat distorted by the window effect close to the central prototype (between $t = 1100$ and $t = 1300$).

Drifting Mackey-Glass Dynamics.

After having illustrated some properties and peculiarities of the new method, we now discuss a more realistic example that involves high-dimensional densities. It therefore cannot be analyzed by simply looking at the data as it might have been sufficient in the previous example. To this end, we generated a time series by using the Mackey-Glass delay differential equation,

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

It describes a high-dimensional chaotic system that is used as a model of blood cell regulation [9]. In our example, two stationary operating modes are established by using different delay parameters, $t_d = 17$ and 28. The transition between these two modes is performed by successively increasing (resp. decreasing) the delay parameter of the differential equation by $\Delta t_d = 1$ after every 80 time steps.[3] The drift from one stationary mode to the other thus takes 800 time steps in total ($t_d = 18, 19, ..., 27$, or vice versa). Each stationary mode then continues for another 800 time steps, after which the

[3]The time index of the data points refers to down-sampled data using a sampling step size $\Delta t = 6$.

dynamics drifts back to the first mode. A total of 7200 data points with five stationary segments and four drift segments was generated in this way (see Fig. 3). We then embedded the time series into a six-dimensional phase space by the method of time-delay coordinates [13]:

$$\vec{x}_t = (x(t), x(t-1), \ldots, x(t-5)). \tag{11}$$

The embedding aims to reconstruct the phase space of the dynamical system. Next, we applied the segmentation algorithm to the data using a pdf window of size $W = 200$. The resulting segmentation is shown as a grey line in Fig. 3. The location of the obtained prototype pdfs is again depicted by thick black dots. The time course of the drift coefficient is plotted as a black line between the dots and it reflects the evolution of the density of the Mackey-Glass attractor in the reconstructed phase space (see Fig. 4). The computed drift curve nicely resembles the time course of the time delay t_d in the Mackey-Glass equation, plotted in the upper graph of Fig. 3, which is remarkable given the fact that there exists only an indirect relationship between the two quantities.

Wake/Sleep Transition in EEG.

In [7], 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. The data was measured during an afternoon nap of a healthy human subject. As in [7], we analyzed data from a single EEG channel, recorded at position O1 of the 10-20 system. We embedded the raw 100 Hz signal into a 5-dimensional phase space,

$$\vec{x}_t = (x(t),\ x(t-2),\ x(t-4),\ x(t-6),\ x(t-8)). \tag{12}$$

To reduce the amount of data, we sub-sampled the embedded data by the factor 10 and then applied the segmentation algorithm to the region of the sleep onset using a pdf window of size $W = 50$.

Fig. 5 shows the resulting segmentation (top) and a manual segmentation by a medical expert (middle), which was worked out by using six physiological quantities (multi-channel EEG, EOG, ECG, heart rate, blood pressure, and respiration). The first segment obtained by the algorithm (labeled as '1') coincides with the phase where the subject was awake with the eyes open. The second segment (labeled as '2'), is assigned to the awake state with the eyes closed. Segment 3 corresponds to sleep stage I. The drift between the two prototype pdfs of segment 2 and 3 (black line between the dots) therefore corresponds to the transition from wake to sleep. The drift curve clearly resolves the changes in the EEG dynamics to much greater detail than the manual segmentation.[4] Note, however, that there is no ground-truth reference curve that can be compared with the obtained drift curve, which precludes a qualitative assessment.

[4]The segment labeled 'art.' in the manual segmentation denotes an artifact in one of the recorded channels.

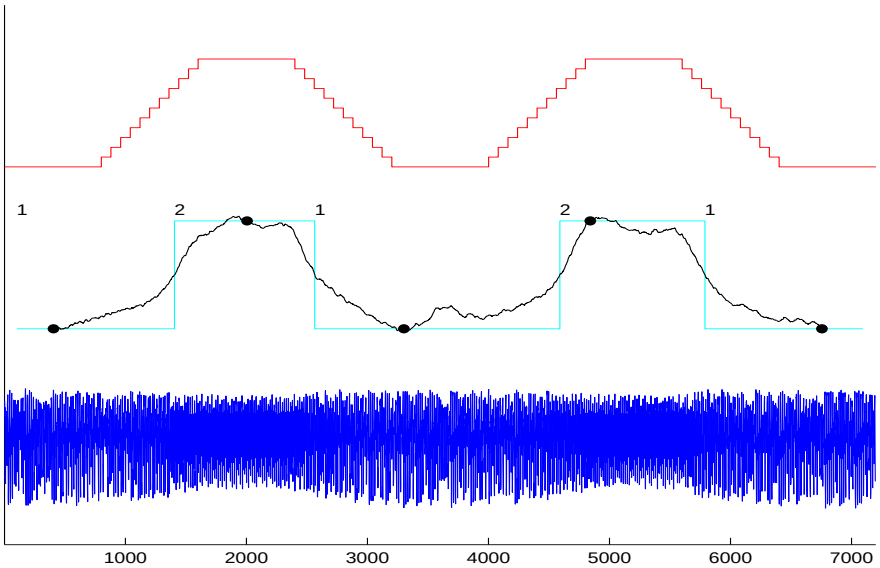

Figure 3: Analysis of a Mackey-Glass time series (bottom) that drifts between two different operating modes by varying the delay parameter t_d of the differential equation (top curve). The resulting drift curve (middle) nicely recovers the drift characteristics of the time series: the black line shows the time course of the drift coefficient between the obtained prototype pdfs (indicated by black dots).

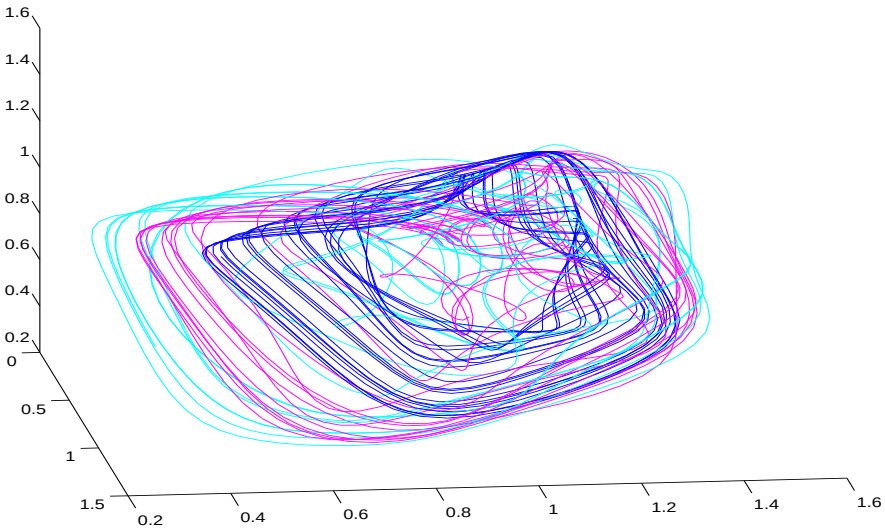

Figure 4: 3D-plot of three Mackey-Glass attractors, each one with a different delay parameter t_d: 17 (black line), 23 (grey line), and 28 (light grey). Each line covers a 200-points sequence used for the pdf estimates. The plot clearly indicates that the attractor is expanding when t_d is increasing. The pdf estimates stretch out accordingly.

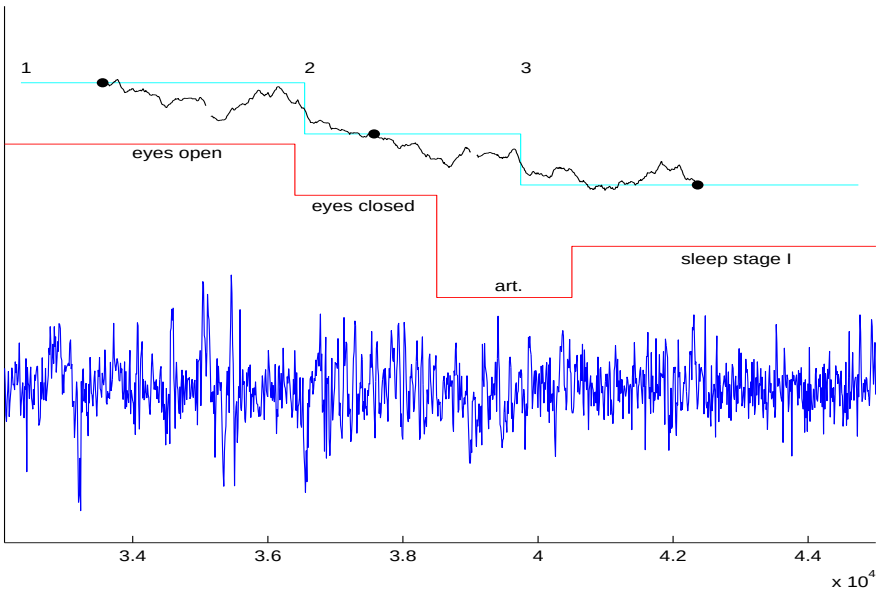

Figure 5: EEG analysis of an afternoon nap of a healthy human subject. The computer-generated drift segmentation (top) is performed on a single EEG channel (bottom: occipital-1), whereas the manual segmentation (middle) was worked out using six physiological quantities (cf. text).

DISCUSSION

We presented a method to track and visualize changes in high-dimensional non-parametric distributions. It was applied to artificially generated data and to EEG data from the wake/sleep transition. The method identifies anchor distributions and describes the transitions between those anchor distributions by means of a suitable distance measure. The use of other distance measures is certainly conceivable, but it might be computationally much more expensive. Here, we simply reuse the $d(p_t, p_{\bar{t}})$ values that are anyway computed for finding the anchor distributions. The method requires that the underlying data distribution changes sufficiently slowly, such that the relevant changes are not already averaged out within the pdf windows. Moreover, the window size should be large enough to sufficiently capture at least the *spatial extent* of each underlying anchor distribution. The density estimate doesn't have to be very precise though, unless the problem at hand requires it to resolve very fine-grained differences. In that case, a rather large window size would be necessary. In high dimensions a precise estimation of the density will easily become impractical though, since it requires an exceeding amount of data.

We expect useful applications of our method in fields where the analysis of complex non-stationary dynamics is highly relevant, like, e.g., in neurophysiology (EEG, MEG), climatology, or industrial applications.

Acknowledgments
I would like to thank S. Lemm for useful discussions and gratefully acknowledge J. Rittweger for providing the EEG data.

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] 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.

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

[4] Kohlmorgen, J. (2003). On Optimal Segmentation of Sequential Data. In: *Neural Networks for Signal Processing XIII*, IEEE, NJ, 449–458.

[5] Kohlmorgen, J., Lemm, S. (2001). An On-line Method for Segmentation and Identification of Non-stationary Time Series. In: *Neural Networks for Signal Processing XI*, IEEE, NJ, 113–122.

[6] Kohlmorgen, J., Müller, K.-R., Pawelzik, K. (1997). Segmentation and Identification of Drifting Dynamical Systems. In: *Neural Networks for Signal Processing VII*, IEEE, NJ, 326–335.

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

[8] 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.

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

[10] Parzen, E. (1962). On Estimation of a Probability Density Function and Mode. *Ann. Math. Statist.* 33(3), 1065–1076.

[11] 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.

[12] Ramamurti, V., Ghosh, J. (1999). Structurally Adaptive Modular Networks for Non-Stationary Environments. *IEEE Transactions on Neural Networks* 10(1), 152–160.

[13] Takens, F. (1981). Detecting Strange Attractors in Turbulence. In: Rand, D., Young, L.-S. (Eds.), *Dynamical Systems and Turbulence* 898, 366–381.