Divisive Strategies for Predicting Non-Autonomous and Mixed Systems

K. Pawelzik
MPI für Strömungsforschung and SFB 185: Nonlinear Dynamics
Bunsenstr. 10, D-37073 Göttingen, Germany

K.-R. Müller and J. Kohlmorgen
GMD FIRST (German National Research Center for Computer Science)
Rudower Chaussee 5, D-12489 Berlin, Germany

Key words: time series, prediction, nonstationarity, divisive strategies, blind separation, competing experts

Abstract. We consider the problem of predicting time series originating from nonstationary and from mixed dynamical systems. It is shown that the complexity of finding representations for the dynamics of such systems can be drastically reduced if their composite nature is taken into account. Two paradigmatic cases are discussed and their solutions presented: jump processes and stationary mixtures. Examples demonstrate that divisive approaches can substantially improve predictions of time series compared to methods that model the dynamics globally.

1. Introduction

Time series from real systems rarely originate from unique autonomous dynamical systems. More common is the presence of additional noise or nonstationarities. Also the fact that data often are superpositions of different sources challenges attempts to model the systems by compact representations using, e.g., large neural nets as in [19].

In this contribution, we emphasize the importance of identifying the multiplicity of the underlying dynamical subsystems to build adequate models for such data. Two paradigmatic situations are discussed: jump processes and stationary mixtures.

Sudden changes of the dynamics constitute nonstationarities which occur in many complex systems. Examples include multistable dynamical systems that are switched by noise or control signals, non-autonomous systems that are externally switched, as e.g. technical systems, in which failures occur, and also ecological and economical systems [17, 14, 4].

Somewhat complementary to jump processes are mixtures in which the time series is a (time independent) weighted sum of several signals generated by independent systems called sources. This situation occurs most prominently in speech recognition where the relevant signal often is superimposed either by other voices or by disturbing non-speech

signals (e.g. music or noise). This is often referred to as cocktail party problem [9].

We show that for jumps and mixtures, estimations of the overall dynamics can be difficult if the composite structure of the system is not taken into account. In the worst case, the intrinsic dimensionality of the problem is given by the *sum of dimensions* of the component systems. On the other hand, if the signals are decomposed prior to modeling, the model complexity is determined by the complexity of the subsystem models, i.e. the dimensionality of the problem is bound by the largest dimension of the component systems.

In other words, compositions of the types discussed in this paper induce a strong curse of dimensionality and divisive approaches therefore are required for adequate modeling, especially if only a limited amount of data samples are available.

We will demonstrate this general effect with prediction problems of chaotic dynamical systems. Note, however, that our results are also relevant for more general stochastic dynamical systems.

The paper is arranged in the following manner. Section 2 briefly discusses the problem of modeling alternating dynamics and reviews an algorithm that performs an unsupervised segmentation of switching dynamics. We show that this approach not only analyzes and models the time series properly, it may also improve predictions significantly.

Section 3 considers the problem of mixtures. It turns out, that finding global models for the dynamics of simple mixtures may be practically impossible. However, the application of algorithms for blind separation, followed by modeling the sources independently, can make the problem tractable. We show that this divisive approach may lead to major improvements of predictions. We discuss our results in section 4.

2. Segmentation and Prediction of Switching Dynamics

Here we show, that systems with switching dynamics may be much easier to model, when the switch points are known. In particular, we will show that this corresponds to the fact, that an overall representation of the system may require an input space, which is D dimensional where $D = \sum d_i$ is the sum of the dimensions of the individual sub-dynamics.

To see this, consider input-output pairs $(x_t, y_t) = (x_t, f_l(x_t))$, where $t = 1, \ldots, T$, such that at each time step t there is a choice $l = l(t)$ of one of four maps. The maps are $f_1(x) = 4x(1-x), x \in [0,1]$ ("logistic map"), $f_2(x) = \{2x \text{ if } x \in [0,.5) \text{ and } 2(1-x), \text{ if } x \in [.5,1]\}$ ("tent map"), $f_3 = f_1 \circ f_1$ ("double logistic map") or $f_4 = f_2 \circ f_2$ ("double tent map"). $f \circ f$ denotes the iteration $f(f(x))$. Then, setting $x_0 = 0.529$

and $x_{t+1} = y_t$ provides a chaotic time series $\{x_t\}$ from the alternating dynamics $x_{t+1} = f_l(x_t)$, where $l = ((t/100) \bmod 4) + 1$, i.e. a new map is chosen after every 100 time steps.

While the individual dynamics are one-dimensional, the combined system can not be represented by considering only a one-dimensional input space, i.e. the most recent value (Fig.2(a)). Embedding, considering more than one past value, may resolve this ambiguity. For instance, embedding in two dimensions resolves f_1 and f_2 (Fig.2(b)). However, in this representation there are still ambiguities: there are values z_{t+2} which are not uniquely determined by z_{t+1} and z_t. These ambiguities are resolved by taking more past values into account (compare also [18]). Note, however, although in principle there is a solution, in practice it can be very difficult or even impossible to find it. The increased dimension of the input space from using large embedding dimensions, together with the increase in complexity of the representation in this space, may pose a severe problem for fitting global models. Particularly, when only a limited amount of data is given.

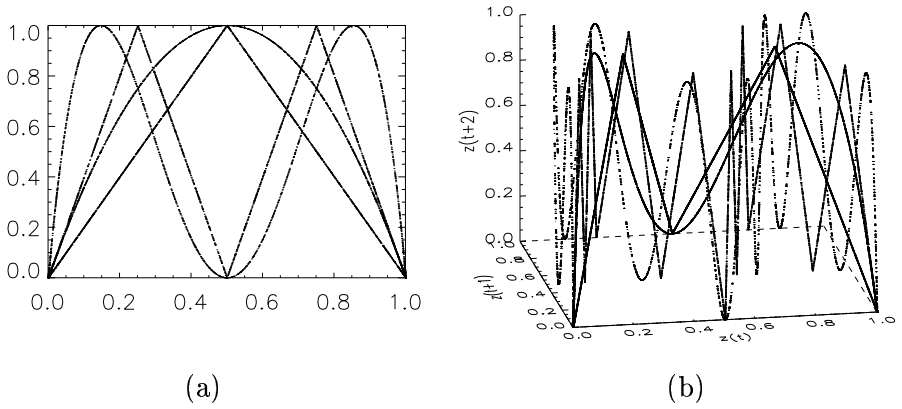


Figure 1. (a) The mapping to be learned by a single network with an input dimension of one. (b) The respective situation, when two inputs are considered.

To demonstrate this effect, we trained a Radial Basis Function (RBF) Network of the Moody-Darken type [11] to predict the time series. The first 1200 data points were used for training and the second 1200 for testing. Throughout this work, we used RBF networks, because they offer a robust and fast learning method. In this case, the RBF network had 120 gaussian basis functions (nodes), because this yielded the best prediction performance. The resulting root mean squared one-step prediction error (RMSE) on the test set was $e = 0.264, 0.149, 0.176, 0.246, 0.269$ for the respective embedding dimensions $m = 1, 2, 3, 4, 5$. As expected, the error increases for large dimensions.

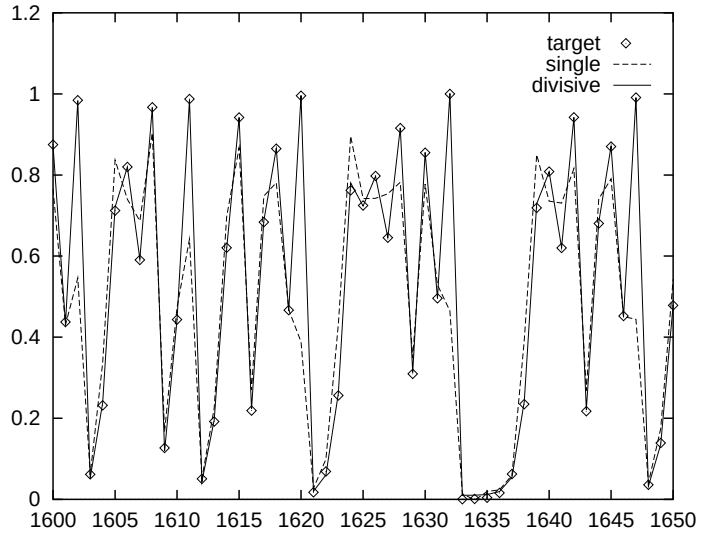


Figure 2. Comparison of the one-step-ahead predictions of the single RBF network (dashed line) and the competing experts approach (solid line). The expert network predicts the target dynamics (shown as points; sequence from the test set) perfectly, whereas the single network performs significantly worse.

If, in contrast, $l(t)$, i.e. the correct segmentation, is given, the optimal estimation of the system would consist in fitting four models $\tilde{f}_j$ for the respective segments of the data. In [6, 7, 13, 15], we described a method to obtain both the segmentation and the models for the respective underlying dynamics from a time series where no additional information about the dynamical modes is given, not even their number. The main idea is to train an ensemble of models on the data in such a way that they compete for the training data. The competition is based on the prediction performance of the individual models. Moreover, during training the competition is increased by decreasing a temperature parameter. Thereby, particular models become experts for certain segments, while superfluous models do not contribute any further. The result is an unsupervised segmentation together with an identification of the underlying systems.

To apply this divisive method to the time series above, 6 submodels are assumed. Each submodel is simulated by a Moody-Darken RBF network with 20 nodes. Within the divisive framework, a single input value to each network is sufficient, i.e. the embedding dimension is $m = 1$. During training, four RBF networks specialized each on a different map. The two other networks dropped out of the learning process and

do not appear in the final segmentation. The resulting prediction error of the ensemble is $e = 0.0395$. This is 3.77 times better than the best result achieved with the single global predictor. It clearly shows the superiority of the divisive strategy in this case (see Fig.2). Two further advantages from using the divisive approach are: (a) a lower embedding dimension, and (b) the underlying structure of the dynamics is readily available.

A more realistic data set is Data Set D of the Santa Fe Competition [19]. We applied the competing experts approach to this data set in [15]. The prediction performance on the test set was 10% better than the winning result for Data Set D, which was obtained by Zhang and Hutchinson ([19], pp. 219-241), who used a very large global network. The divisive method worked well, because this time series originated from a system which occasionally switched among distinct dynamical modes that were induced by several local minima in its potential.

As applications to real world data we considered the mode changes in physiological wake/sleep data [12] and in speech signals [13].

3. Unmixing and Predicting Mixed Signals

Now consider the case where the time series is generated by a mixture of sources [16]. Surprisingly, this situation is conceptionally very similar to the case of switching dynamics discussed in the last section.

As an example, consider the chaotic time series $x^1_{t+1} = f_1(x^1_t)$ and $x^2_{t+1} = f_2(x^2_t)$, where $f_1(x) = 4x(1-x)$ and $f_2(x) = 2x$ if $0 \leq x \leq 1/2$, $f_2(x) = 1 - 2x$ if $1/2 < x \leq 1$ are the logistic map and the tent map, respectively. Next, consider the simple mixture signal

$$z_{t+1} = f_1(x^1_t) + f_2(x^2_t) \ , \ t = 1, \ldots, T.$$

Obviously, there is no unique one-dimensional map representing this situation, i.e. there is no function $h : z_t \rightarrow z_{t+1}$. On the other hand, it has been proven that a system of dimension d can be reconstructed in principle from an observable using an embedding of dimension $m^* = 2d + 1$ [18]. In our case, the dimensionality of the system is the sum of the dimensions of the individual subsystems, $d = \sum d_i$, and we have $m^* = 5$. This means that there is a representation $z_{t+1} = h(z_t, z_{t-1}, \ldots, z_{t-m+1})$ as soon as $m \leq m^*$ is large enough.

Considering $m = 2$, the problem is far from solved for our example (Fig.3). We observe a 'cloud' instead of a function; z_{t+2} is not uniquely determined by z_{t+1} and z_t. Furthermore, the representation is strongly modulated. This comes as no surprise, since the complexity of representing the dynamics of one source (expressed by the order

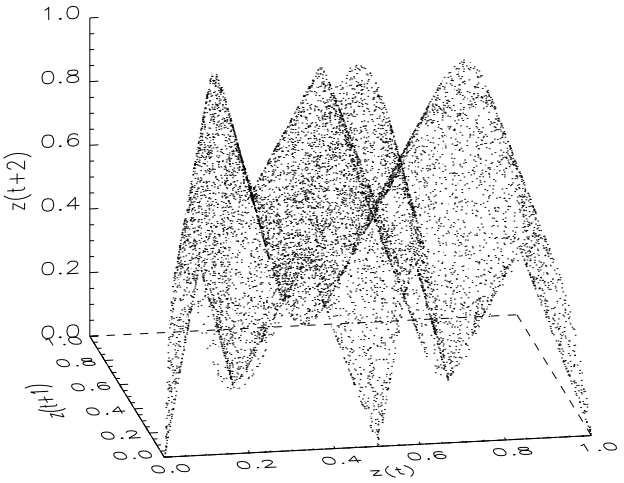


Figure 3. Three-dimensional embedding of an additive mixture $\{z_t\}$ of two independent chaotic time series. z_{t+2} is not uniquely determined by z_{t+1} and z_t.

of the polynomial) grows exponentially with iteration time. These two effects, the necessity of high-dimensional input spaces together with the increased complexity of the representation, make system identification and prediction highly tedious.

To demonstrate this, a simple RBF network with 100 nodes [11] is trained on the mixture signal z_t, using $T = 2000$ points. The prediction errors (normalized RMSE) are $e = 0.835,\ 0.593,\ 0.658,\ 0.767,\ 0.891$ for the embedding dimensions $m = 1, 2, 3, 4, 5$, respectively. Using the divisive method with two RBF networks of 50 nodes each and predicting the individual maps in $m = 1$, leads to errors that are $2-3$ orders of magnitude smaller, namely $e = 0.00141$ and $e = 0.00416$ (normalized RMSE) for the logistic map and the tent map, respectively. Obviously, mixing induces a severe dimensionality problem, which could be avoided if the underlying sources could be separated prior to modeling.

The above discussion emphasizes the importance of separating the sources underlying a signal. Unfortunately, this is very difficult for scalar signals. However, powerful methods for blind separation have been developed recently [1, 2, 3, 5, 10]. These apply to situations in which several observables, representing different mixtures of the underlying systems, are available. Such multivariate signals occur in many applications ranging from biology to economics.

To be explicit, let the time series be represented by a vector $\vec{z}_t = (z_t^1, \ldots, z_t^n)^T$ that is a linear mixture of the source vector $\vec{x}_t = (x_t^1, \ldots, x_t^n)^T$,

$$\vec{z}_t = M\vec{x}_t, \tag{1}$$

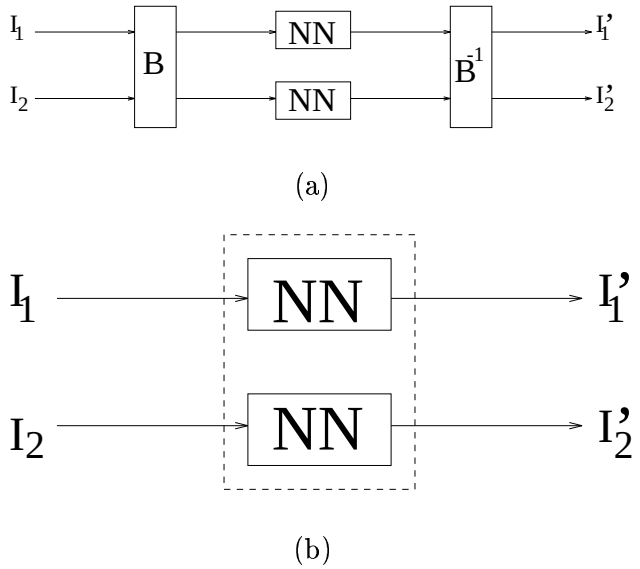


Figure 4. (a) The architecture of the unmix-predict-mix strategy. First, the input vector I is decomposed into independent components by a transformation B, which is obtained using a blind separation method. Then, these components are separately predicted by neural networks. Finally, the predicted signals are mixed together by the inverse transformation B^{-1}, which yields a prediction for future values of I. (b) The conventional approach: direct prediction of the signals, either separately or with one big network (dashed).

M is the mixing matrix. In this case, the blind separation methods mentioned above may reconstruct the original source signals, assuming that M is invertible and that the sources $\vec{x}_t$ are independent. The blind separation method presented in [10] applies to linearly independent sources and exploits linear autocorrelations. The other approaches [1, 2, 3, 5] rely on higher moments of mutual correlations and ignore the temporal coherence of the sources.

Instead of discussing the strengths and weaknesses of blind separation methods, we will demonstrate that their application can substantially improve time series prediction.

As a first example, consider a linear mixture of the maps f_1 and f_2, according to eq.(1). A time series of $T = 2000$ points is generated for each map and then mixed using

$$M = \begin{pmatrix} 1 & -0.53 \\ -0.87 & 1 \end{pmatrix}.$$

Next, using the blind separation method presented in [2], a separation matrix B is found as an estimate for M^{-1}. Applying the sepa-

ration matrix to the time series $\vec{z}_t$ yields good reconstructions for the source signals: $\tilde{x}_t^1$ and $\tilde{x}_t^2$. Then, two RBF networks, $\tilde{f}_1$ and $\tilde{f}_2$, with 20 nodes each, are trained on $\tilde{x}_t^1$ and $\tilde{x}_t^2$. Finally, the prediction errors for the mixed signals, z^1 and z^2, are computed from a test set, using the remixed estimate (cf. Fig.3)

$$\tilde{\vec{z}}_{t+1} = B^{-1}\tilde{F}[B\ \vec{z}_t], \quad \text{where} \quad \tilde{F}(\tilde{\vec{x}}_t) = (\tilde{f}_1(\tilde{x}_t^1), \tilde{f}_2(\tilde{x}_t^2)). \tag{2}$$

The errors are $e_1 = 0.0396$ and $e_2 = 0.0436$. It turned out that this approach improved predictions by a factor of 5.0 for z^1 and 5.9 for z^2, compared to the best results for the direct prediction of $\vec{z}_t$, using one RBF network with 100 nodes for each of the two components. The test set errors for the direct predictions are $e_1 = 0.314, 0.198, 0.231, 0.289, 0.304$, and $e_2 = 0.429, 0.259, 0.310, 0.398, 0.419$ for the embedding dimensions $m = 1, 2, 3, 4, 5$, respectively.

As a more realistic example, the well-known Mackey-Glass equation [8]

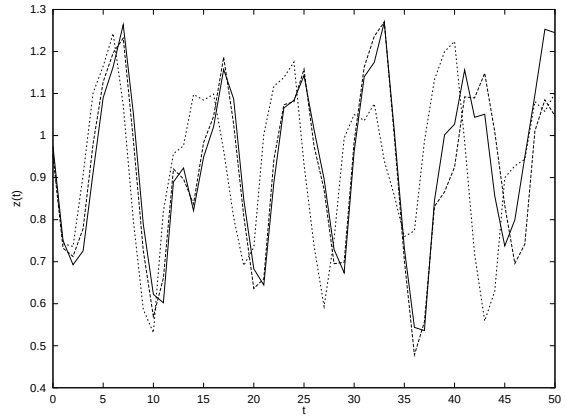
$$\frac{dx(t)}{dt} = -0.1x(t) + \frac{0.2x(t-\tau)}{1 + x(t-\tau)^{10}}, \tag{3}$$

originally introduced as a model for the irregular dynamics of blood cell production, is used. Two time series are generated using $\tau = 17$ and 23. The series are then subsampled with $\Delta t = 6$. The mixed signals of these two dynamical systems, using

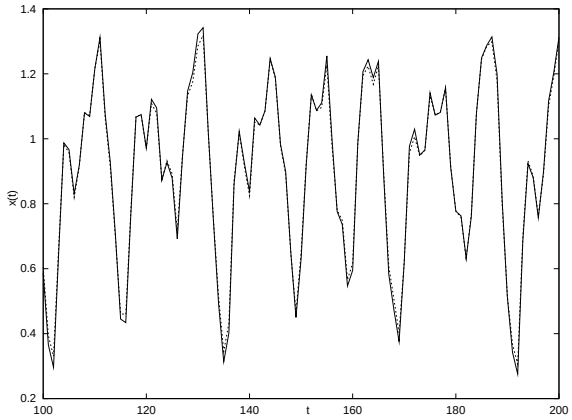
$$M = \begin{pmatrix} 0.3 & 0.7 \\ 0.8 & 0.2 \end{pmatrix},$$

provide the observables z_t^1 and z_t^2, $t = 1, \ldots, 1000$. First, these signals are modeled globally, using a single RBF network (m = 16 and 250 nodes). The RMSE errors computed from a test set are $e = 0.049$ and $e = 0.043$ for z^1 and z^2, respectively.

A subsequent blind separation of the mixtures, using the method presented in [10], proved to be very effective (dashed line in Fig.3(b)). As before, two RBF networks (m = 6 and 120 nodes each) are trained on the reconstructed source signals. Then, the test set error is computed for the remixed signals $\tilde{z}_t^1, \tilde{z}_t^2$, using the inverse of the estimated separation matrix, as in eq.(2). The errors are $e = 0.016$ and $e = 0.009$ for $\tilde{z}_t^1$ and $\tilde{z}_t^2$, respectively. The predictions improved by factors of 3.1 and 4.8. This is very important, because the prediction horizon also increases by the same factor, as seen from Fig.3(a).



(a)



(b)

Figure 5. Prediction of mixed Mackey-Glass sources. (a) Mixed signal z^1 (solid line), iterated prediction of a global radial basis function network (dots), and iterated prediction by the divisive approach, which involved blind separation (dashes). Clearly, the divisive strategy yields a better result than the global prediction. (b) Original source x^1 (solid line) in comparison to the estimate $\tilde{x}^1$ from blind separation (dashes), which is almost identical.

4. Discussion

By means of simple representative examples, we demonstrated that both alternating dynamical systems as well as mixing of dynamical sources can impose severe problems for system identification and prediction tasks. In particular, we showed that global representations of composed systems may suffer from a curse of dimensionality. This may

severely increase the complexity necessary for adequate modeling of the data. Therefore, we advocate divisive approaches.

In particular, we discussed a divisive method that segments time series according to distinct dynamical operating modes and simultaneously learns to predict the dynamics of the submodels. It was demonstrated that its application may lead to a significant improvement of predictions.

For signals which represent mixtures of independent dynamical systems, we showed that an *unmix-predict-mix* algorithm, using recently proposed methods for blind separation, can substantially improve predictions. The performance gain of this strategy for a prediction task in case of a mixture may be large and depends crucially on (a) the severity of the curse of dimensionality and (b) the accuracy of the estimation of the mixture matrix M. In an example of a high-dimensional time-delay differential system, our approach also obtains a more stable long-term prediction (Fig.3).

Future work will focus on further applications of divisive strategies to real data.[1]

Acknowledgements

We acknowledge support of the DFG (grant Ja379/51). Furthermore, we would like to thank A. Ziehe for fruitful discussions and acknowledge T. Bell, H. Yang and J.F. Cardoso for providing source code and valuable help.

References

1. Amari, S., Cichocki, A., Yang, H., A new learning algorithm for blind signal separation, *Advances in Neural Inf. Proc. Systems 8* (NIPS 95), D.S. Touretzky, M.C. Mozer and M.E. Hasselmo (eds.), MIT Press: Cambridge, MA (1996)
2. Bell, A.J., Sejnowski, T., An information-maximization approach to blind separation and blind deconvolution, *Neural Computation* 7, 1129-1159 (1995)
3. Cardoso, J.F., Laheld, B., Equivariant adaptive source separation, to appear in *IEEE Trans. on Signal Processing* (1996)
4. Hamilton, J. D. (1994). *Time Series Analysis*, Princeton University Press, New Jersey.
5. Jutten, C., Herault, J., Blind separation of sources, Part I: An adaptive algorithm based on neuromimetic architecture, *Signal Processing* 24, 1-10 (1991)
6. Kohlmorgen, J., Müller, K.-R., Pawelzik, K. (1994). Competing Predictors Segment and Identify Switching Dynamics. ICANN'94, Springer London, 1045–1048.

[1] Further information on related research can be found at:
`http://www.first.gmd.de/persons/Mueller.Klaus-Robert.html`.

7. Kohlmorgen, J., Müller, K.-R., Pawelzik, K. (1995). Improving short-term prediction with competing experts. ICANN'95, EC2 & Cie, Paris, 2:215-220.
8. Mackey, M., Glass, L., Oscillation and chaos in a physiological control system, Science **197**, 287 (1977).
9. von der Malsburg, C., The correlation theory of brain function, Int. Rep. 81-2, Max Planck Inst. f. Biophysical Chemistry, Göttingen (1981)
10. Molgedey, L., Schuster, H.G., Separation of a mixture of independent signals using time delayed correlations, *Phys. Rev. Lett.* 72, 23, 3634-3637 (1994)
11. Moody, J., C. Darken (1989). Fast Learning in Networks of Locally-Tuned Processing Units. Neural Computation **1**, 281-294, 1989.
12. Müller, K.-R., Kohlmorgen, J., Rittweger, J., Pawelzik, K. (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.
13. Müller, K.-R., Kohlmorgen, J., Pawelzik, K. (1995). Analysis of Switching Dynamics with Competing Neural Networks, *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, E78-A, No.10, 1306-1315.
14. Narendra, K. S., Mukhopadhyay, S. (1995). Intelligent Control using Neural Networks, *Intelligent Control Systems*, IEEE Press, 151–186.
15. 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.
16. Pawelzik, K., Müller, K.-R., Kohlmorgen, J. (1996). Prediction of Mixtures, ICANN '96: Proc. of the Int. Conf. on Artificial Neural Networks, LNCS 1112, Springer Berlin, 127-132.
17. Rabiner, L.R. (1988). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, Proc. IEEE, Vol **77**, 257–286.
18. Takens, F., Detecting strange attractors in turbulence, in: Rand, D., Young, L.-S., (Eds.), *Dynamical Systems and Turbulence*, Springer Lecture Notes in Mathematics, **898**, 366 (1981).
19. Weigend, A.S., Gershenfeld, N.A. (Eds.), *Time Series Prediction: Forecasting the Future and Understanding the Past*, Addison-Wesley (1994).