in: Proceedings of ICANN '96, Springer.

Prediction of Mixtures

K. Pawelzik[1], K.-R. Müller[2], J. Kohlmorgen[2]

[1] Inst. f. theo. Physik, Universität Frankfurt, 60054 Frankfurt/M., Germany
[2] GMD FIRST, Rudower Chaussee 5, 12489 Berlin, Germany

Abstract. The problem of predicting time series originating from mixtures of signals from independent dynamical systems is considered. We show that the problem of finding representations for the dynamics of such systems is hard if the mixing structure of the system is not taken into account. If, on the contrary, the sources can be unmixed in a preprocessing step the complexity of system identification may be drastically reduced. This is demonstrated using chaotic maps. It is shown that applications of methods for blind separation of sources can substantially improve both: prediction performance and prediction horizon.

1 Introduction

Time series from real systems originate from unique autonomous dynamical systems only under ideal circumstances. More common is the presence of additional noise, but also nonstationarities and the fact that data often are superpositions of different sources may challenge attempts to model the systems by compact representations e.g. using large neural nets (see e.g. [14, 16]).
For systems which switch their dynamics, methods for unsupervised segmentation have already been presented and it has been demonstrated that they can improve predictions [6, 7, 12, 13, 3, 17].
Here we are concerned with the other paradigmatic situation of compositional systems: mixtures of independent sources. This situation is present in many problems of time series analysis; most prominent maybe is the example of speech recognition where the relevant signal often is superimposed by other voices or non-speech signals.
When analysing the problem of system identification and prediction of mixtures, we found that its complexity may grow dramatically, if the mixed nature of the signal is not taken into account. If, in contrast, the sources can be separated, system identification can become feasible in cases which are otherwise intractable for the given data.

2 Curse of Dimensionality

Here we demonstrate this general effect with the simple example of mixtures of two chaotic time series. Consider the 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 $0 \leq x \leq 1/2$, and $f_2(x) = 2 - 2x$, if $1/2 < x \leq 1$, are the logistic map and the tent map respectively.

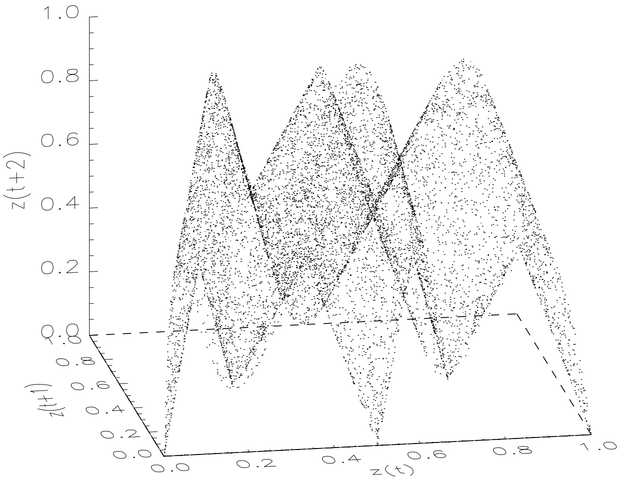


Fig. 1.

Next, consider the simple mixture signal $z_{t+1} = f_1(x_t^1) + f_2(x_t^2)$, $t = 1, \ldots, T$. 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 can in principle be reconstructed from an observable using an embedding of dimension $m^* = 2d + 1$ where d is the dimension of the underlying dynamics [15]. In our case the dimensionality of the system is the sum of the dimensions of the individual subsystems, i.e. $d = \sum d_i$ 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 is large enough.

Considering $m = 2$ we see that the problem is far from solved for our example (Fig. 1). We observe not only a 'cloud' instead of a function in this case, but also that the representation is strongly modulated. This comes as no surprise considering that already the complexity of representing the dynamics of one source (expressed e.g. by the order of the polynomial) grows exponentially with iteration time, the exponent being the positive Lyapunov exponent.

These two effects, the necessity of high dimensional input spaces together with the increasing complexity of the representation make system identification and prediction highly tedious even for this simple example. To show this, we use a simple radial basis function network $\tilde{h}$ with 100 centers [11] trained on the mixture signal z_t using T=2000 points. The root mean squared one-step prediction errors (RMSE) are $\epsilon = 0.314,\ 0.198,\ 0.231,\ 0.289,\ 0.304$ for the embeddings $m = 1, 2, 3, 4, 5$ respectively. Note that the nominally sufficient embedding of $m = 5$ did not entail an improvement of prediction. This is not surprising, considering the fact that the complexity of representing the dynamics in time delay coordinates grows *exponentially* with the embedding dimension, the exponent

being 2 in this case.
This result is particularly striking because predicting the individual maps in $m = 1$ using two radial basis function networks of 100 nodes each leads to errors which are $2 - 3$ orders of magnitude smaller, namely $e = 0.000222$ and $e = 0.000327$ for the logistic map and the tent map, respectively. Obviously, mixing induces a severe dimensionality problem, which can be avoided, if the underlying sources are separated prior to modeling.

3 Blind separation

Our previous example demonstrates the importance of separating the sources underlying a signal, if the dynamics of the time series has to be modeled. Fortunately, several powerful methods for blind separation [1, 2, 4, 5, 10] have been recently developed, which are very useful when several observables are available which represent different mixtures of the underlying systems. More explicitly, let the time series be represented by a vector $\mathbf{z}_t = (z_t^1, \ldots, z_t^n)^T$ with $\mathbf{z}_t = M\mathbf{x}_t$, i.e. a mixture of the source vector $\mathbf{x}_t = (x_t^1, \ldots, x_t^n)^T$. Then the above mentioned methods may reconstruct the original source signals without further assumptions except that M is invertible and that the sources are mutually independent. The method from [10] applies to linearly independent sources and exploits linear autocorrelations. The other approaches cited above rely on higher moments of mutual correlations and ignore the temporal coherence of the sources. Recent results indicate that modifications of these methods may be useful also in cases where the number of available signals is smaller than the number of underlying sources (T. Gramß, personal communication).

4 Predicting Mixtures

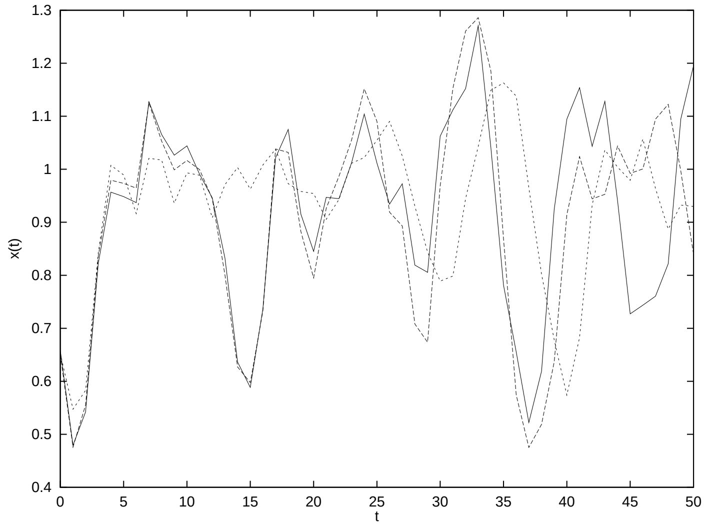
Instead of discussing the advantages and drawbacks of these methods we will here demonstrate that their application can substantially improve predictions. As a first example we used a mixture of the maps discussed in section 2. We generated series of $T = 2000$ points each and mixed them using

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

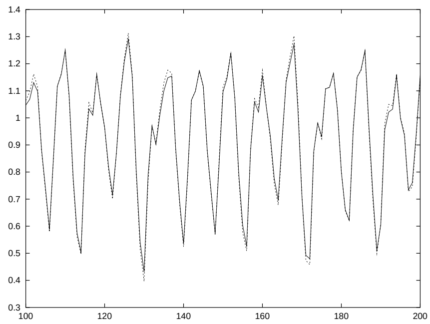
Applying a blind separation method [2, 10], we found a separation matrix B yielding good estimates for the sources $\tilde{x}_t^1$ and $\tilde{x}_t^2$. We then trained two radial basis function networks $\tilde{f}_1$ and $\tilde{f}_2$ with 20 nodes each on $\tilde{x}_t^1$ and $\tilde{x}_t^2$. Finally, we computed the prediction errors e_1 and e_2 for z^1 and z^2, respectively using

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

We found that this approach improved predictions by a factor of 4 compared to the best results for the direct prediction method. As a more realistic example



(a)



(b)

Fig. 2. *Prediction of mixed Mackey-Glass sources. (a) Mixed signal* z^1 *(solid line), iterative prediction of a global radial basis function network (dashed), and iterative prediction by the divisive approach which involved blind separation (long 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). Similar results were obtained for* z^2 *and* x^2 *(not shown here).*

we used the well-known Mackey-Glass equation [9]

$$\frac{dx(t)}{dt} = -0.1x(t) + \frac{0.2x(t-\tau)}{1+x(t-\tau)^{10}}, \tag{2}$$

which was originally introduced as a model for the irregular dynamics of blood cell production. We generated two time series of $T = 1000$ points each, sampled at time steps of length $\Delta t = 6$ using two parameter values $\tau = 17$ and $\tau = 23$ (Fig. 2b). The mixed sources of these two dynamical systems using

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

then provided the observables $\mathbf{z}_t$ (Fig. 2a). First, these signals were modeled

globally using a single radial basis function network (m= 6, 120 nodes each). The RMSE were computed from a test sample as $e = 0.068$ and $e = 0.055$ for z^1 and z^2, respectively.
A subsequent blind separation of the mixtures [2] proved to be very effective here (see dashed line in Fig. 2b). As before, we trained two radial basis function networks (m = 6, 120 nodes) on the reconstructed source signals and then computed the RMSE for the remixed signals $\tilde{z}_t^1, \tilde{z}_t^2$ using the inverse of the estimated unmixing matrix as in Eq. 1, which gave $e = 0.016$ and $e = 0.014$ respectively, i.e. the prediction improved by a factor of roughly 4 again. This is highly relevant, because it also substantially increases the prediction horizon as seen from Fig. 2a.

5 Summary and Discussion

It was demonstrated that mixing of dynamical sources can impose severe problems for system identification and prediction tasks. We showed that applications of blind separation methods can alleviate these problems, in particular the curse of dimensionality. Clearly, prior information, e.g. knowing that there is a mixing structure inherent in the signal, leads to a *divide-and-conquer* strategy, which in this case is in fact an *unmix-predict-mix* algorithm. The performance gain of this strategy for a prediction task is 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. It should be noted that also the prediction horizon is increased, i.e. we obtain a more stable long term prediction (cf. Fig. 2a).
Future research will focus on the application of our approach to real data.[3]

Acknowledgment: K.P. is supported by DFG (grant Pa 569/1-1). We thank T.Gramß and A.Ziehe for fruitful discussions. We 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 Information Processing 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. Bengio, Y., Frasconi, P. (1994). Credit assignment through time: Alternatives to backpropagation. NIPS 93, Morgan Kaufmann.
4. Cardoso, J.F., Laheld, B., Equivariant adaptive source separation, to appear in *IEEE Trans. on Signal Processing* (1996)

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

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.
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. Liebert, W., Pawelzik, K., Schuster, H.G., Optimal embeddings of chaotic attractors from topological considerations, Europhys. Lett. **14**, 521 (1991).
9. Mackey, M., Glass, L., Oscillation and chaos in a physiological control system, Science **197**, 287 (1977).
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., 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.
13. Pawelzik, K., Kohlmorgen, J., Müller, K.-R. (1996). Annealed Competition of Experts for a Segmentation and Classification of Switching Dynamics, *Neural Computation*, 8:2, 342-358.
14. Rumelhart, D.E., McClelland, J.L., *Parallel distributed processing*, MIT Press, Cambridge Massachusetts (1984).
15. 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).
16. Weigend, A.S., Gershenfeld, N.A. (Eds.), *Time Series Prediction: Forecasting the Future and Understanding the Past*, Addison-Wesley (1994)
17. Weigend, A.S., Mangeas, M., Nonlinear gated experts for time series: discovering regimes and avoiding overfitting, University of Colorado tech. Report. CU-CS-764-95 and submitted to *Machine Learning* (1995)

This article was processed using the LaTeX macro package with LLNCS style