Applications of nearest neighbours statistics

Jörg D. Wichard, Ulrich Parlitz and Werner Lauterborn

Drittes Physikalisches Institut, Georg-August-Universität Göttingen, D-37073 Göttingen, Germany
wichard@physik3.gwdg.de

Abstract— Based on an efficient method for finding nearest neighbours in the phase space of a dynamical system, several applications of nearest neighbour statistics are presented, including methods to detect nonstationarity in time series, to find optimal reconstruction parameters for delay-embeddings and to estimate the largest Lyapunov exponent of chaotic systems.

I. Introduction

Various methods of nonlinear time series analysis are based on exploring the neighbourhood of points in the (reconstructed) phase space of a dynamical system. However, looking for nearest neighbours in a straightforward way can be very time consuming for large data sets. A fast algorithm for finding nearest neighbours was proposed by Kantz and Schreiber [5]. It is based on a sorting-scheme, which arranges the data in a 2-dimensional grid, because nearest neighbours in two dimensions are good candidates for nearest neighbours in higher dimensional spaces.

II. Determining reconstruction parameters

Attractor reconstruction from scalar time series with delay coordinates is a frequently used method in the analysis of dynamical systems. It is based on the embedding theorem of Takens [15] and its extensions by Sauer et al. [13]. This allows us under several restrictions to reconstruct the attractor of a dynamical system up to diffeomorphic equivalence by a d-dimensional delay embedding. Let $\{x_n\}$ denote a time series from a dynamical system, then we build a d-dimensional delay vector with time lag τ as

$$\vec{x}_n = (x_n, x_{n+\tau}, \ldots, x_{n+(d-1)\tau}). \tag{1}$$

A. Determining proper delay times

The embedding theorems [13, 15] give some restrictions on choosing delay times, but they don't give a rule for a proper choice. It is well known, that the quality of attractor reconstruction depends very much on the delay time τ. If τ is too small the reconstructed states differ not much and the points are scattered around a straight line. If τ is too large chaotic attractors are folded and this may lead to self-intersections

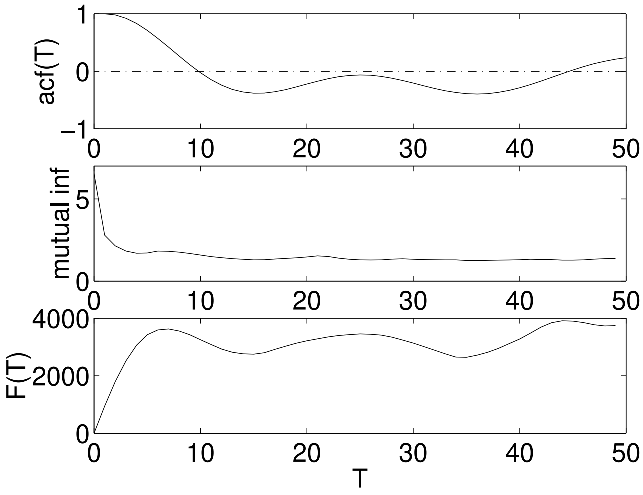

Figure 1: The autocorrelation (acf), mutual information and $F(T)$ vs. T for a time series from the Colpitts-Oscillator with 15000 samples. The first null of the acf lies at $T = 9$, the first minimum of the mutual information appears at $T = 4$, the first maximum of $F(T)$ is reached at $T = 7$.

of the reconstructed attractors and a loss of the one-to-one property of the reconstruction. Another aspect is the fact, that the window length $T = d\tau$ turned out to be a crucial quantity for characterizing the embedding. Many methods have been proposed for finding a good delay time, for an overview see [10] and the references therein. The most popular approach consists in a minimization of the redundancy of the coordinates of the reconstructed states using linear autocorrelation functions or information theoretic concepts like mutual information [3].
As an alternativ we present here an approach, where the redundancy of the components of the delay vector $\vec{x}_n$ is minimized by maximizing the mutual prediction error. Let $\{x_i\}_{i=1,\ldots,N}$ denote a time series from a dynamical system. We define the quantity $F(T)$ as

$$F(T) = \frac{1}{N-T} \sum_{i=1}^{N-T} \frac{|x_{i+T} - x_{nn+T}|}{|x_i - x_{nn}|}, \tag{2}$$

where x_{nn} is the nearest point to x_i, and all points with $|x_i - x_{nn}| = 0$ are excluded. The first maximum of $F(T)$ yields a candidate for a delay time. We applied this method to estimate the window length for an experimental time series from the Colpitts-Oscillator. This is an electronic circuit, that can operate in a chaotic regime. For a detailed description see ref. [6]. We compare our method with the autocorrelation and the mutual information of the times series, which are two well established criteria for choosing window lengths. The results are plotted in Fig.1 and we find, that the first minimum of $F(T)$ yields an optimal T that lies between the proposals of the two mentioned methods. Similar results have also been obtained with data from other chaotic systems.

B. Determining the embedding dimension

The method of false neighbours was proposed by several authors [2, 7, 9] to determine an optimal embedding dimension for scalar time series in the sense of Taken's theorem [15]. This method was refined by Cao [1] and it is based on the fact, that a wrong embedding dimension moves points close together in the embedding space, which are far apart in the original phase space. These points are called false neighbours and a good embedding should have no false neighbours.

Let $\vec{x}_n(d) = (x_n, x_{n+\tau}, \dots, x_{n+(d-1)\tau})$ be a d-dimensional reconstructed state in the sense of equation (1) and let $\vec{x}_n(d+1) = (x_n, x_{n+\tau}, \dots, x_{n+d\tau})$ be the state in embedding dimension $d+1$, then two states that are close together in d dimensions should be still close together in a $d+1$ dimensional space.

Now, let $\vec{x}_{NN(n,d)}(d+1)$ be the nearest neighbour of $\vec{x}_n(d)$ in d dimensions, but embedded in a $(d+1)$ dimensional space. We define the quantity

$$\Delta(n,d) = \frac{||\vec{x}_n(d+1) - \vec{x}_{NN(n,d)}(d+1)||}{||\vec{x}_n(d) - \vec{x}_{NN(n,d)}(d)||} \quad (3)$$

and compute the mean over all states

$$E(d) = \frac{1}{N}\sum_{n=1}^{N} \Delta(n,d). \quad (4)$$

We expect that $E(d)$ is minimized when the minimum embedding dimension is reached. This is shown in Fig. 2, where we plotted the quantity from equation (4) for different window lengths that we obtained from the three inspected methods in section II.A.

III. Estimating the largest Lyapunov exponent

Lyapunov exponents describe the exponential increase or decrease of small perturbations on the attractor of a dynamical system and they are invariant with respect to diffeomorphic transformations [11]. A nearest neighbour based method for estimating the largest

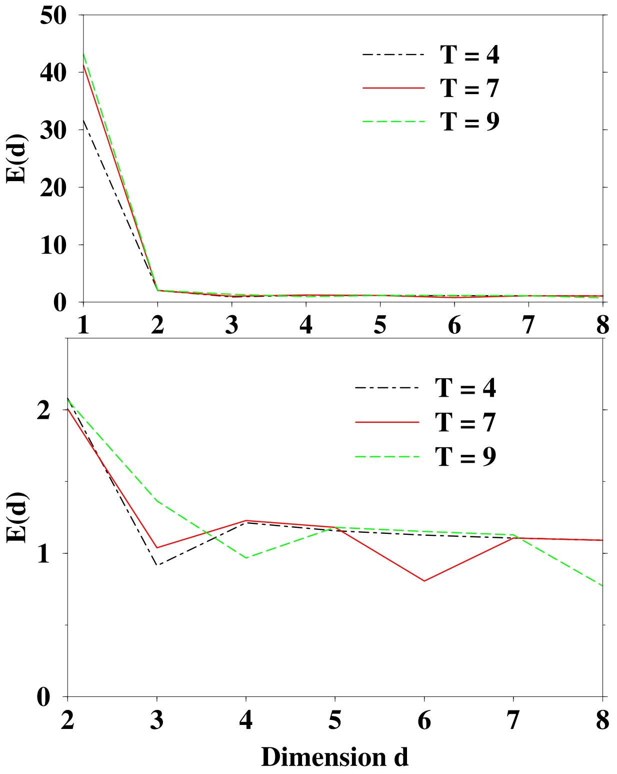

Figure 2: The quantity $E(d)$ from equation (4) vs. the dimension d for different window lengths T. We used 8000 samples of the experimental Colpitts data. The figure in the bottom shows the most important section of the upper figure. A clear minimum occurs for an embedding dimension $d = 3$ with a window length of $T = 7$.

Lyapunov exponent was proposed by Sato et al. [12], Kurths and Herzel [8] and Kantz [4], where the divergence of nearby trajectories on the attractor is studied. Let $\{x_n\}$ denote a time series from a dynamical system equidistantly sampled with sampling time t_s. After making a proper choice for the time lag τ and embedding dimension d, we can reconstruct a diffeomorphic image of the attractor in the original phase space by the use of d-dimensional delay coordinates $\vec{x}_n = (x_n, x_{n+\tau}, \dots, x_{n+(d-1)\tau})$. To each reconstructed state $\vec{x}_0$ we can find the nearest neighbour $\vec{y}_0$ on a segment of a nearby trajectory and calculate the difference vector $\vec{z}_0 = \vec{x}_0 - \vec{y}_0$. It is important to exclude points that are nearest neighbours close in time i.e. lying on the same segment of the trajectory. If the underlying dynamics shows chaotic behaviour, then the largest Lyapunov exponent λ_{max} will govern the time evolution of the difference vector $\vec{z}_0$. After a transient phase the vector $\vec{z}_k = \vec{x}_k - \vec{y}_k$ will point in the direction of λ_{max}. Then $||\vec{z}_0||$ will grow in the average with $\exp(t_s\lambda_{max}k)$ as long as the linear ap-

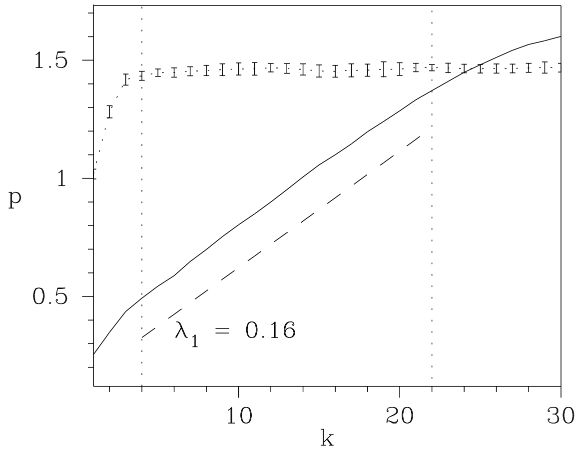

Figure 3: The prediction error $p(k)$ as defined in equation (5) vs. the number of steps k for an experimental time series from the Colpitts-Oscillator. The slope of the solid line in the linear region gives the largest Lyapunov exponent $\lambda_{max} = 0.16$.

proximation of the flow is sufficient. We choose N_{ref} reference points and inspect the quantity

$$p(k) = \frac{1}{t_s N_{ref}} \sum_{i=1}^{N_{ref}} \ln \frac{\|\vec{x}_{i+k} - \vec{y}_{i+k}\|}{\|\vec{x}_i - \vec{y}_i\|} \ . \tag{5}$$

The slope of the linear region in the $p(k)$ vs. k diagram should give us λ_{max}. In Fig.3 we give an example for the determination of the largest Lyapunov exponent λ_{max} of the Colpitts data by this method. The dotted curve is from phase randomized surrogate data [16] to detect nonlinearity. The error bars are for ± 5 standard deviations.

IV. Detecting Nonstationarity

A time series $\{X_t\}_{t\in\mathbb{Z}}$ is said to be weakly stationary, if

$$\begin{aligned} E(X_{t_1}) &= E(X_{t_1+k}) \\ cov(X_{t_1}, X_{t_2}) &= cov(X_{t_1+k}, X_{t_2+k}) \end{aligned} \tag{6}$$

for all $k, t_1, t_2 \in \mathbb{Z}$ [17]. A typical test for stationarity can be done by estimating a certain parameter for different sequences of the time series. If the variation of the parameter is larger than the expected statistical fluctuation, the time series can be viewed to be nonstationary. The following approach is based on a method of Schreiber [14], where the similarity between parts of the time series is calculated using cross predictions. But instead of exploring the cross prediction error we submit the mean distances of nearest neighbours to an examination. The time series $\{X_t\}$ is divided in N sequences $\{S_i\}_{i=1,...,N}$ of length L_i. These sequences are embedded in a d-dimensional space with time lag

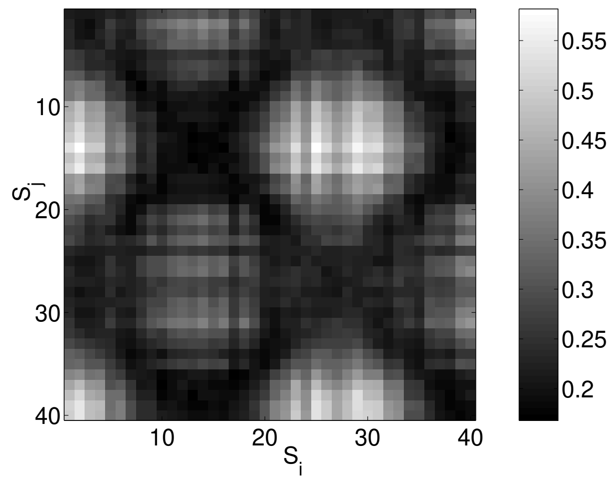

Figure 4: The average distance $\Delta(S_i, S_j)$ for the instationary Hénon-map encoded as gray-scales. The total length of the time series was $N_{total} = 40000$.

τ. Hence we obtain in the i-th sequence a set of embedded vectors $\vec{x}_n^i = (X_k, X_{k+\tau}, \ldots, X_{k+(d-1)\tau})$ with $n = 1, \ldots, L_i$. Now we look for the nearest neighbour $NN(\vec{x}_n^i,\ j)$ of the point $\vec{x}_n^i$ in the sequence S_j, with respect to the Euclidean norm. We define the average distance $\Delta(S_i, S_j)$ of S_i in relation to S_j as

$$\Delta(S_i, S_j) = \frac{1}{L_i} \sum_{n=1}^{L_i} \|\vec{x}_n^i - NN(\vec{x}_n^i,\ j)\| \ . \tag{7}$$

Note, that the average distance in general is not symmetric, i.e. $\Delta(S_i, S_j) \neq \Delta(S_j, S_i)$. As a numerical example we consider the Hénon-map with a slowly changing parameter

$$X_n = 1 - 1.4X_{n-1}^2 + (0.28 + 0.02\cos(\omega_n))X_{n-2}, \tag{8}$$

where $\omega_n = 3\pi n/N_{total}$ and N_{total} denotes the length of the time series. Furthermore, we removed the running mean and we normalize to unit running variance for a window lenth of 100 samples. The resulting time series $\{Y_n\}$ is given by

$$Y_n = \frac{X_n - \langle X \rangle_k}{\langle (X - \langle X \rangle_k)^2 \rangle_k} \ , \tag{9}$$

where $\langle \cdot \rangle_k$ denotes the avarage over the indices $\{$(n-k), ..., (n+k)$\}$ and $k = 50$. The results are plotted in Fig.4. The changing gray-scales correspond to the parameter drift in the time series, that can not be found by inspecting only the mean or the variance of the single time series sequences. This is shown in Fig.5, where the variation of the mean is plotted for each sequence. Let $\langle S \rangle$ be the mean of the whole time series and let $\langle S_i \rangle$ be the mean of the i-th sequence. The variation Δmean(i) is given by

$$\Delta\text{mean}(i) = \langle S_i \rangle - \langle S \rangle. \tag{10}$$

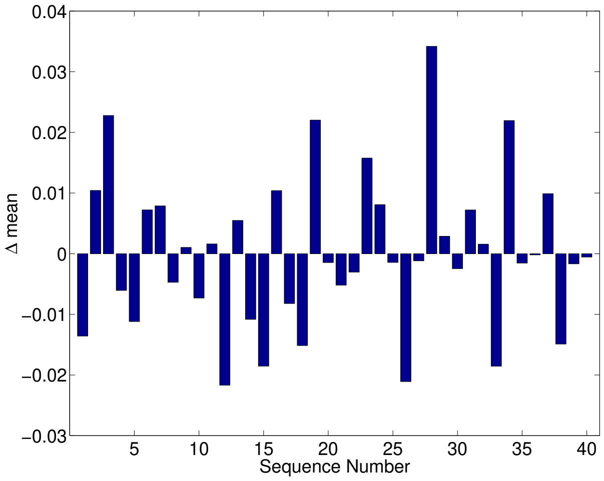

Figure 5: The variation of the mean for all segments, used for the computations that are shown in Fig.4.

Acknoledgement

The authors want to thank the members of the Nonlinear Dynamics group at the "Drittes Physikalisches Institut" for stimulating discussions. This work was supported by the *Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie* (grant 13 N 7038/9).

References

[1] L. Cao, "Practical method for determining the minimum embedding dimension of a scalar time series," *Physica D*, vol.110, pp.43–50, 1997.

[2] A. Cenys and K. Pyragas. "Estimation of the number of degrees of freedom from chaotic time series," *Phys. Lett. A*, vol.129, pp.227–230, 1988.

[3] A.M. Frazer, "Information and entropy in strange attractors" *IEEE Trans. Inf. Theory*, vol.35, pp.245–262, 1989.

[4] H. Kantz. "A robust method to estimate the maximal lyapunov exponent of a time series," *Phys. Lett. A*, vol.185, pp.77–87, 1994.

[5] H. Kantz and T. Schreiber, *Nonlinear time series analysis*, Cambridge University Press, Cambridge UK, 1997.

[6] M.P. Kennedy, "Chaos in the Colpitts Oscillator," *IEEE Trans. Circuits Syst.*, vol.41, pp.771–774, 1994.

[7] M.B. Kennel, Brown R. and H.D.I. Abarbanel, "Determining embedding dimension for phase-space reconstruction using a geometrical construction," *Phys. Rev. A*, vol.45, pp.3403–3411, 1992.

[8] J. Kurths and H. Herzel. "An attractor in solar time series," *Physica D*, vol.25, pp.165–172, 1987.

[9] W. Liebert and H.G. Schuster. "Proper choise of the time delay for the analysis of chaotic time series," *Phys. Lett. A*, vol.142, pp.107–111, 1989.

[10] M.T. Rosenstein, J.J. Collins and C.J. De Luca, "Reconstruction expansion as a geometry-based framework for choosing proper delay times," *Physica D*, vol.73, pp.82–98, 1994.

[11] D. Ruelle, *Chaotic evolution and strange attractors*, Cambridge University Press, Cambridge UK, 1989.

[12] S. Sato, M. Sano and Y. Sawada, "Practical methods of measuring the generalized dimension and largest Lyapunov exponent in high dimensional chaotic systems," *Prog. Theor. Phys.*, vol.77, pp.1–5, 1987.

[13] T. Sauer, J.A. Yorke and M. Casdagli, "Embedology," *J. Stat. Phys.*, vol.65, pp.579–611, 1991.

[14] T. Schreiber, "Detecting and analysing nonstationarity in a time series using nonlinear cross predictions," *Phys. Rev. Lett.*, vol.78, pp.843–846, 1997.

[15] F. Takens, "Detecting strange attractors in turbulence," In *Dynamical Systems and Turbulence*, Lect. Notes Math. Sringer-Verlag, Berlin, 1981.

[16] J. Theiler, S. Eubank, A. Longtin, B. Galdrikian, and J.D. Farmer. "Testing for nonlinearity in time series: the method of surrogate data," *Physica D*, vol.58, pp.77–94, 1992.

[17] H. Tong, *Non-linear Time Series*, Number 6 in Oxford Statistical Science Series, Oxford University Press, 1990.