Distances of Time Series Components by Means of Symbolic Dynamics

Karsten Keller[1] and Katharina Wittfeld[2]

[1] Mathematical Institute, Wallstraße 40, 23560 Lübeck

[2] Department of Mathematics and Computer Sciences, Jahnstraße 15a, 17487 Greifswald

Abstract

In this note we describe a simple method for visualizing time-dependent similarities and dissimilarities between the components of a high-dimensional time series. On the base of symbolic dynamics, the time series is turned into a series of matrices whose rows quantify pattern types in the components of the original series. For different scales we introduce distances between the components via the obtained pattern type distributions and approximate them in a one-dimensional manner. The method is illustrated for 19-channel EEG data.

1 Introduction

During the last years the method of *symbolic dynamics* has been established in the qualitative analysis of time series (e.g., see Ebeling & Nicolis [1992], Schwarz et al. [1993], Finney et al. [1998], Daw et al. [1998], beim Graben et al. [2000], beim Graben et al. [this issue]). The idea behind this method is rather simple: Instead of considering the exact state of a system at some time, one is interested in a coarse-grained description. The state space is decomposed into a small number of pieces, and states being contained in the same piece are identified.

We use symbolic dynamics in the context of high-dimensional time series. In particular, we apply the ideas of symbolic dynamics for analyzing and visualizing EEG data. An EEG (electroencephalogram) records the electrical activity of the brain. In epilepsy it is used to detect and to localize abnormal brain behavior. EEG data are mostly high-dimensional and very large since an EEG can be recorded from many electrodes and over a long time. Usually, the medical expert visually inspects the EEG recordings. He is looking for special wave forms and has to distinguish between brain-relevant patterns and artefacts. This is generally a very time-consuming procedure. So one is interested in an automatic processing, at least for a first rough view on the data.

In the following we propose a method based on introducing time-dependent distances between the time series components and approximating them in a one-dimensional manner. First we explain the ideas taken from symbolic dynamics. In Sec. 2 we introduce the distances and describe their one-dimensional approximation. Here we generalize a method known as Correspondence Analysis. Further, we illustrate our approach for EEG data being related to epileptic activity. Sec. 3 is devoted to the mathematical details. We want to refer to the approach by Steuer et al. [this issue], which in some points is similar to our method.

Symbolic dynamics: the one-dimensional case. Let us explain the idea of symbolic dynamics in more detail, in the often considered context of a *delay embedding* (compare Takens [1981]). Given a one-dimensional theoretical time series $(x_t)_{t\in T=\{\ldots,-2,-1,0,1,2,\ldots\}}$, a *delay* $\tau = 1, 2, 3, \ldots$, and some *dimension* d, one considers the delay embedding map

$$t \in T \mapsto (x_{t-(d-1)\tau}, x_{t-(d-2)\tau}, \ldots, x_{t-\tau}, x_t)$$

which assigns to each time t a d-dimensional vector containing the value at t and some history in the form of the values at $t-\tau, t-2\tau, \ldots, t-(d-1)\tau$. The d-dimensional space is decomposed into n pieces with names $1, 2, \ldots, i, \ldots, n-1, n$ - the *symbols* -, and for each $t \in T$ the symbol σ_t of the piece containing $(x_{t-(d-1)\tau}, x_{t-(d-2)\tau}, \ldots, x_{t-\tau}, x_t)$ is determined. In this way the original time series $(x_t)_{t\in T}$ is turned into the symbolic time series $(\sigma_t)_{t\in T}$. The main task is the analysis of the symbol frequencies.

Example 1. Consider a time series $\ldots, x_0 = 5, x_1 = 7, x_2 = 3, x_3 = 1, x_5 = 9, \ldots$ and let $\tau = 1$ and $d = 2$. The first bisecting line divides the 'delay embedding' plane into two parts, where that above the line is assigned the symbol 1 and the other one the symbol 2. So $\sigma_t = 1$ if $x_{t-1} < x_t$ and $\sigma_t = 2$ if $x_{t-1} > x_t$. (For simplicity, we assume that $x_{t-1} = x_t$ is impossible.) Thus the obtained symbol sequence is $\ldots, \sigma_1 = 1, \sigma_2 = 2, \sigma_3 = 2, \sigma_4 = 1, \ldots$.

Note that the type of symbolic dynamics described by Example 1 is used in Steuer et al. [this issue].

Ordinal patterns. Clearly, there are infinitely many ways for defining a decomposition of the state space in order to symbolize time series. For practical reasons, it is important that the obtained symbol sequences contain as much information on the original time series as possible, and particularly if the given data sets are large, the determination of the symbols should be easy and fast from the computational viewpoint.

Following an idea of Bandt & Pompe [2002], we propose a symbolization on the base of *ordinal patterns*, which describe the up and down in the time series in a more subtle way than in Example 1. However, the subsequent general considerations do not need a special decomposition of the state space. Let us explain the concept of an ordinal pattern by an example.

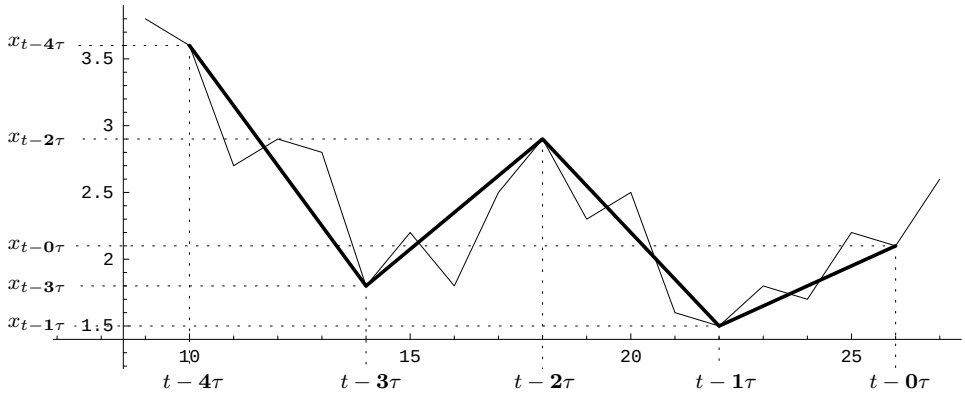

Figure 1: Ordinal pattern

Example 2. Fig. 1 shows a part of a fictive time series. (The values given for times $9, 10, \ldots, 26, 27$ are connected by thin line segments.) We want to demonstrate what is meant by the ordinal pattern at time $t = 26$ given for delay $\tau = 4$ and dimension $d = 5$. The times providing the 'delay vector' satisfy

$$\begin{array}{ccccccccc} t-\mathbf{0}\tau & > & t-\mathbf{1}\tau & > & t-\mathbf{2}\tau & > & t-\mathbf{3}\tau & > & t-\mathbf{4}\tau \\ \| & & \| & & \| & & \| & & \| \\ 26 & & 22 & & 18 & & 14 & & 10 \end{array} .$$

On the other hand, the values considered at these times are in the order

$$x_{t-\mathbf{4}\tau} > x_{t-\mathbf{2}\tau} > x_{t-\mathbf{0}\tau} > x_{t-\mathbf{3}\tau} > x_{t-\mathbf{1}\tau},$$

what can be coded in the form $(\mathbf{4}(=d-1), \mathbf{2}, \mathbf{0}, \mathbf{3}, \mathbf{1})$. The latter is no more than a 'spatial' permutation of the 'time steps' $0, 1, 2, \ldots, 4$.

Generally, in a one-dimensional time series $(x_t)_{t\in T}$ the *ordinal pattern* of order d at time t is the permutation $(r_0, r_1, \ldots, r_{d-1})$ of $(0, 1, \ldots, d-1)$ if the values at t, $t-\tau$, $\ldots$, $t-(d-2)\tau$, $t-(d-1)\tau$ are ordered as follows:

$$x_{t-r_0\tau} \geq x_{t-r_1\tau} \geq \ldots \geq x_{t-r_{d-2}\tau} \geq x_{t-r_{d-1}\tau}.$$

In order to get a unique result, we set $r_{l-1} > r_l$ in the case that $x_{t-r_{l-1}\tau} = x_{t-r_l\tau}$. For example, if the value x_{14} of the data illustrated by Fig. 1 would be changed to the given value at time 22, one would have two candidates of ordinal patterns: $(4, 2, 0, 3, 1)$ and $(4, 2, 0, 1, 3)$. The above setting fixes the first one.

Clearly, each ordinal pattern defines a connected piece of the d-dimensional space. Since there exist $d!$ permutations, we have $d!$ ordinal patterns. We identify each of them with exactly one of the 'symbols' $j = 1, 2, \ldots, n = d!$. The way of assignment does not play a role because we are not interested in the patterns themselves but in their distribution. Note that the computation of ordinal patterns only needs a few number of comparisons.

The multidimensional case. Now suppose we are given an m-dimensional time-series

$$(\mathbf{x}_t)_{t\in T} = ((x_t^i)_{i=1}^m)_{t\in T}$$

with components $(x_t^i)_{t\in T}$, which we want to call the i-th *channels*. (This is useful in view to the EEG analysis.) For some delay τ and some dimension d, each of the channels is transformed into a symbolic time series as described above. In the time window $[t-\delta+1, t]$ given for fixed *window length* δ, count the symbols $j = 1, 2, \ldots, n$ for each channel $i = 1, 2, \ldots, m$, and divide the results by $m\delta$. The so obtained relative frequencies p_{ij} form matrices

$$P = P_t = P_t(d, \tau, \delta) = \begin{pmatrix} p_{11} & \ldots & p_{1j} & \ldots & p_{1n} \\ \vdots & & \vdots & & \vdots \\ p_{i1} & \ldots & p_{ij} & \ldots & p_{in} \\ \vdots & & \vdots & & \vdots \\ p_{m1} & \ldots & p_{mj} & \ldots & p_{mn} \end{pmatrix},$$

which are the base for our further considerations. Clearly, $\sum_{i,j=1}^{m,n} p_{ij} = 1$.

The distances mentioned at the beginning are distances between the rows of these matrices. They depend on a scaling parameter α and measure similarities and dissimilarities between the symbol distributions in the different channels. The intention is that these distances are robust - in particular, they neglect small phase transitions -, and allow a good one-dimensional approximation and visualization of changes in the originally given time series.

2 Distances on different scales

Let us first motivate the special choice of the distance measures given below.

Contingency as mean squared profile length. The *contingency*, also called φ^2-measure, is often used for quantifying distributional inhomogeneities between the rows (or columns) of a relative frequency matrix. Let us have a closer look to the contingency in the situation described above. For this let t be fixed and consider the total relative frequencies $p_{\cdot j} = \sum_{i=1}^m p_{ij}$ of the symbols $j = 1, 2, \ldots, n$ related to the matrix $P = P_t$. Here we assume that $p_{\cdot j} \neq 0$ for all j. Clearly, since the number of measurements is the same for all channels, we have $p_{i\cdot} = \sum_{j=1}^n p_{ij} = \frac{1}{m}$ for the analogue 'relative channel frequencies'.

Assuming that the $p_{\cdot j}$ are not vanishing, the contingency of the matrix $P = P_t$ is given by

$$\begin{aligned} \varphi^2 = \varphi_t^2 = \sum_{i,j} \frac{(p_{ij} - p_{i\cdot}p_{\cdot j})^2}{p_{i\cdot}p_{\cdot j}} &= \frac{1}{m}\sum_{i=1}^m \left(\sum_{j=1}^n \frac{(mp_{ij} - p_{\cdot j})^2}{p_{\cdot j}} \right) \\ &= \frac{1}{m}\sum_{i=1}^m \left(\sum_{j=1}^n p_{\cdot j} \left(\frac{mp_{ij}}{p_{\cdot j}} - 1 \right)^2 \right). \end{aligned} \tag{1}$$

So the last term in (1) shows that φ^2 can be interpreted in a nice way: Considering *profiles*

$$\mathbf{a}_i = (a_{ij})_{i=1}^n = \left(\frac{mp_{ij}}{p_{\cdot j}} - 1 \right)_{i=1}^n \tag{2}$$

and assigning n-dimensional vectors $\mathbf{a} = (a_j)_{j=1}^n$ the weighted Euclidean length $\sqrt{\sum_{j=1}^n p_{\cdot j} a_j^2}$, this term is no more than the mean of the squared profile lengths.

Beyond this, the distance defined by $\sqrt{\sum_{j=1}^n p_{\cdot j}(a_j - \tilde{a_j})^2}$ for two n-dimensional vectors $\mathbf{a} = (a_j)_{j=1}^n$ and $\tilde{\mathbf{a}} = (\tilde{a_j})_{j=1}^n$ allows particularly to measure differences between the profiles. This is the base of *Correspondence Analysis*, which was developed primarily in the 70's by Benzécri and his students (see Greenacre [1984]) for visualizing categorial variables given a contingency matrix.

In Correspondence Analysis profile components are weighted according to the total frequencies of the symbols j. In order to be able to emphasize the role of larger or smaller frequencies we more generally consider a scalar product

$$\langle \mathbf{a}, \tilde{\mathbf{a}} \rangle^\alpha := \sum_{j=1}^n p_{\cdot j}^\alpha a_j \tilde{a_j} \tag{3}$$

for n-dimensional vectors $\mathbf{a} = (a_j)_{j=1}^n$ and $\tilde{\mathbf{a}} = (\tilde{a_j})_{j=1}^n$ providing length

$$\|\mathbf{a}\|^\alpha = \sqrt{\sum_{j=1}^n p_{\cdot j}^\alpha a_j^2}$$

of a vector $\mathbf{a} = (a_j)_{j=1}^n$ and the distance

$$\|\mathbf{a} - \tilde{\mathbf{a}}\|^\alpha = \sqrt{\sum_{j=1}^n p_{\cdot j}^\alpha (a_j - \tilde{a_j})^2}$$

of two vectors $\mathbf{a} = (a_j)_{j=1}^n$ and $\tilde{\mathbf{a}} = (\tilde{a_j})_{j=1}^n$. Here α is a positive parameter. The larger α is, the more the components related to large total symbol frequencies are emphasized, and the more those related to small frequencies are suppressed. On the other hand, decreasing α increases the influence of components related to small frequencies.

Remark 1. The profile lengths $\|\mathbf{a}_i\|^\alpha$ can be interpreted as distances of the symbol distributions from the total symbol distributions (taken over all channels). The origin is indeed the centroid of the $\mathbf{a}_i$, i.e. it coincides with $\frac{1}{m}\sum_{i=1}^m \mathbf{a}_i$. Note, that for $\alpha = 1$ there is a completely equivalent description of the considered distances in the simplex of all distributions on an n-point set (see Greenacre [1984], Lauffer & Keller [2002]).

Remark 2. In the case that the symbol distributions for the channels are not too different in a certain sense, it holds

$$\frac{\varphi_t^2}{2} \approx H_t - \frac{1}{m}\sum_{i=1}^{m} H_t^i, \tag{4}$$

where $H_t = -\sum_{j=1}^{n} p_{\cdot j} \ln p_{\cdot j}$ and $H_t^i = -\sum_{j=1}^{n} m p_{ij} \ln(m p_{ij})$ are the Shannon entropies related to the total symbol distribution and the i-th channel symbol distributions, respectively (see Lauffer & Keller [2002]). This can be interpreted in the way that at time t the amount of inhomogeneity between the channel symbol distributions is proportional to the difference of the total complexity and the mean channel complexity.

It is important to note that for all EEG experiments we did on the base of ordinal patterns, formula (4) has turned out to be valid. We also want to mention that the Shannon entropy on the base of ordinal patterns, called *permutation entropy* by Bandt and Pompe, is interesting both from the practical and the theoretical viewpoint (see Bandt & Pompe [2002], Bandt et al. [2002]).

One-dimensional profile approximations. Usually the system of profiles spans a high-dimensional space. In order to visualize profile distances in dependence on time t, we use one-dimensional approximations of the profile system for all times t of interest. This is done on the base of Singular Value Decomposition (SVD). We do not want to go into details at this place. For more mathematical background we refer to Sec. 3.

Given $\alpha > 0$, the mean squared profile length $\frac{1}{m}\sum_{i=1}^{m}(\|\mathbf{a}_i\|^\alpha)^2$ is a good measure of overall inhomogeneity between the channel symbol distributions. Here recall that according to the considerations above

$$\varphi^2 = \varphi_t^2 = \frac{1}{m}\sum_{i=1}^{m}(\|\mathbf{a}_i\|^1)^2. \tag{5}$$

So, for getting a good one-dimensional profile representation, it is natural to ask for a direction explaining a maximal possible amount of mean squared profile length. More precisely, given $\alpha > 0$ and a line L through the origin, one has the following decomposition of the mean squared profile length:

$$\frac{1}{m}\sum_{i=1}^{m}(\|\mathbf{a}_i\|^\alpha)^2 = \frac{1}{m}\sum_{i=1}^{m}(\|\mathbf{a}_i^L\|^\alpha)^2 + \frac{1}{m}\sum_{i=1}^{m}(\|\mathbf{a}_i - \mathbf{a}_i^L\|^\alpha)^2.$$

Here $\mathbf{a}_i^L$ denotes the projection of $\mathbf{a}_i$ onto L. The first term on the right of the equality can be interpreted as the part of mean squared profile length *explained* by L and the second term as the part of mean squared profile length *not explained* by L. An optimal line L_{opt} - explaining as much mean squared profile length as possible - is defined by

$$\frac{1}{m}\sum_{i=1}^{m}(\|\mathbf{a}_i^{L_{opt}}\|^\alpha)^2 \geq \frac{1}{m}\sum_{i=1}^{m}(\|\mathbf{a}_i^L\|^\alpha)^2 \tag{6}$$

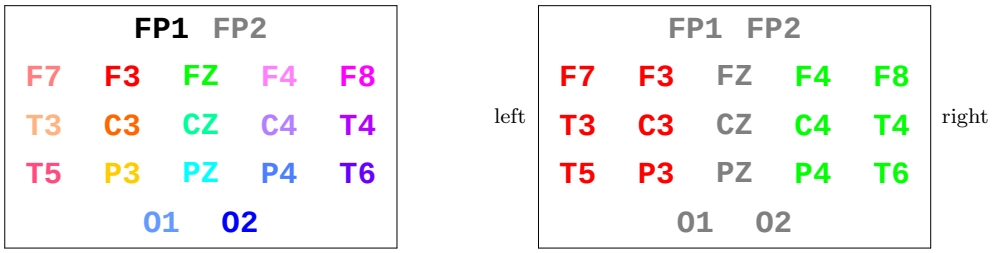

Figure 2: 10-20-system of electrode placement

for all lines L through the origin.

L_{opt} can be identified with the real line such that the origin corresponds to 0 and the distance between two points on L_{opt} is the absolute difference of the corresponding real numbers. Then the projection $\mathbf{a}_i^{L_{opt}}$ of each profile $\mathbf{a}_i$ corresponds to a number w_i. The time-dependent *representation vectors* $\mathbf{w}_t^\alpha = (w_i)_{i=1}^m$ are the base for visualizing the channels (via the related profiles) in a coordinate system (see Fig. 4):

Horizontally a t-axis is drawn, and for the times t the components of $\mathbf{w}_t^\alpha$ are given in vertical direction. So the i-th channels are represented by a curve showing w_i in dependence on time. If two curves are near to another (far from each other) at some time, this indicates that the symbol distributions obtained from the corresponding channels are similar (dissimilar) at this time.

Application to EEG data. We have applied the method for exploring 19-channel scalp EEG data from children with epileptic disorders. The experiments have shown that it can be used for visualizing qualitative temporal changes and spatial differences in an EEG, supplementarily to the methods in Lauffer & Keller [2002]. In particular, the variation of the parameter α can be helpful.

Let us illustrate this by showing the numerical results for an EEG data set obtained from an eight years old boy with epilepsy. The corresponding data were collected according to the international 10-20-system of electrode placement. The diagrams in Fig. 2 show the positions of the electrodes, where FP1 and FP2 are the two most frontal ones. We use two colorings, one distinguishing all channels, and one showing the electrodes on the right hemisphere in green and those on the left hemisphere in red. The sampling rate was 256 Hz, and the data were filtered by

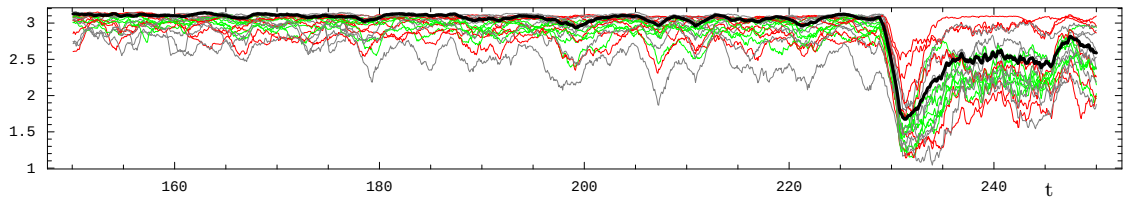

Figure 3: Permutation entropies

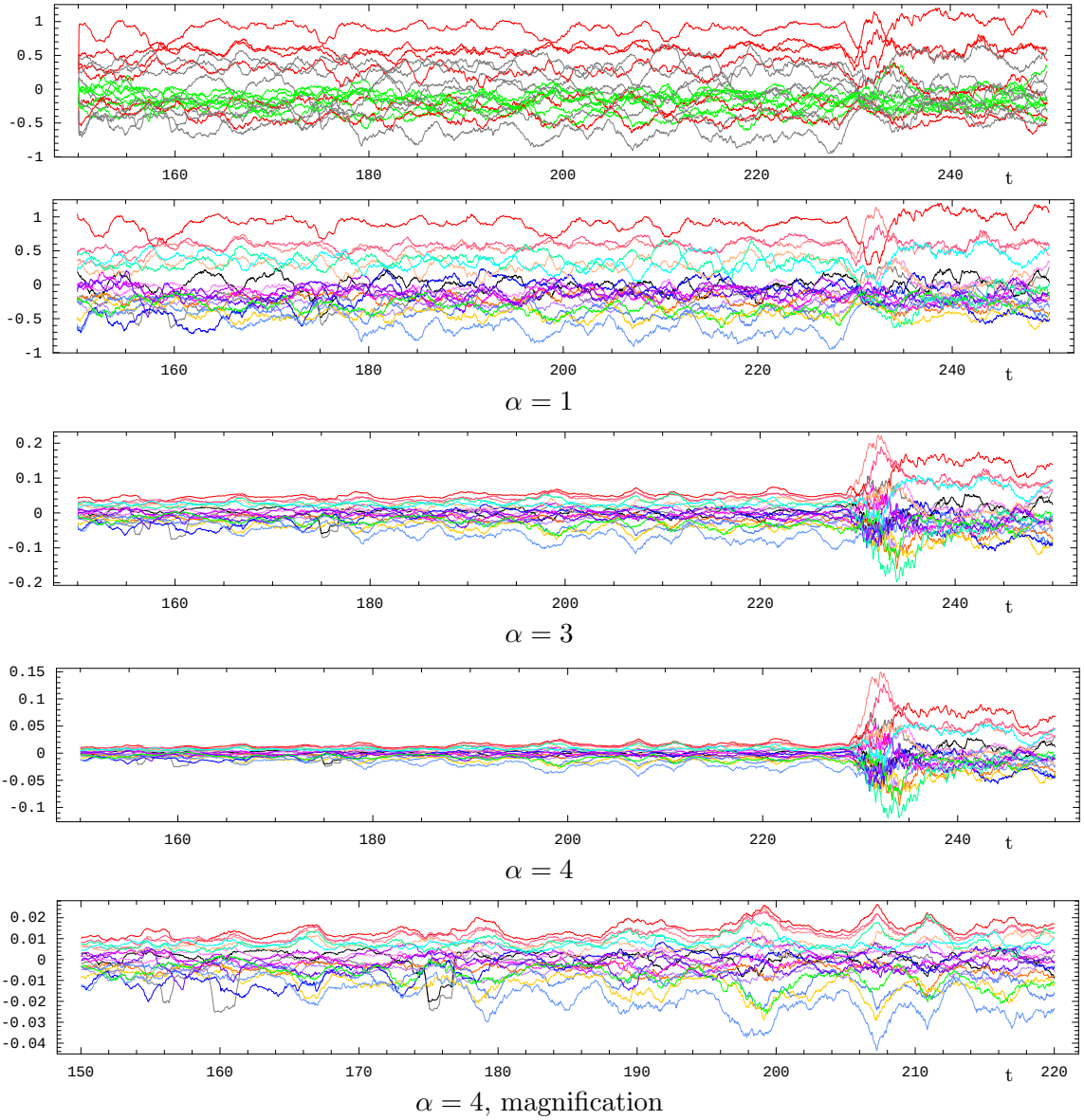

Figure 4: Channel representation

a bandpass $(0.3 - 70$ Hz). In the following graphics the time is given in seconds (1 second corresponds to 256 data points). We have used ordinal patterns of order $d = 4$ for a window length $\delta = 2s (= 512$ points) and a delay $\tau = \frac{1}{256} (= 1$ point).

The boy the EEG was taken from has lesions predominately in the left temporal lobe of the brain, resulting from a connatal toxoplasmosis. (Toxoplasmosis is an infection that comes from parasites found in animal feces or undercooked meat and that when contracted by a pregnant woman can pose serious risks to her unborn baby.) The EEG of the boy contains much epileptic activity. Here we consider an interval of 100 s.

From the original data it is known that there was a generalized epileptic seizure, i.e. a seizure that takes place throughout the entire brain, starting at 228 s. This is

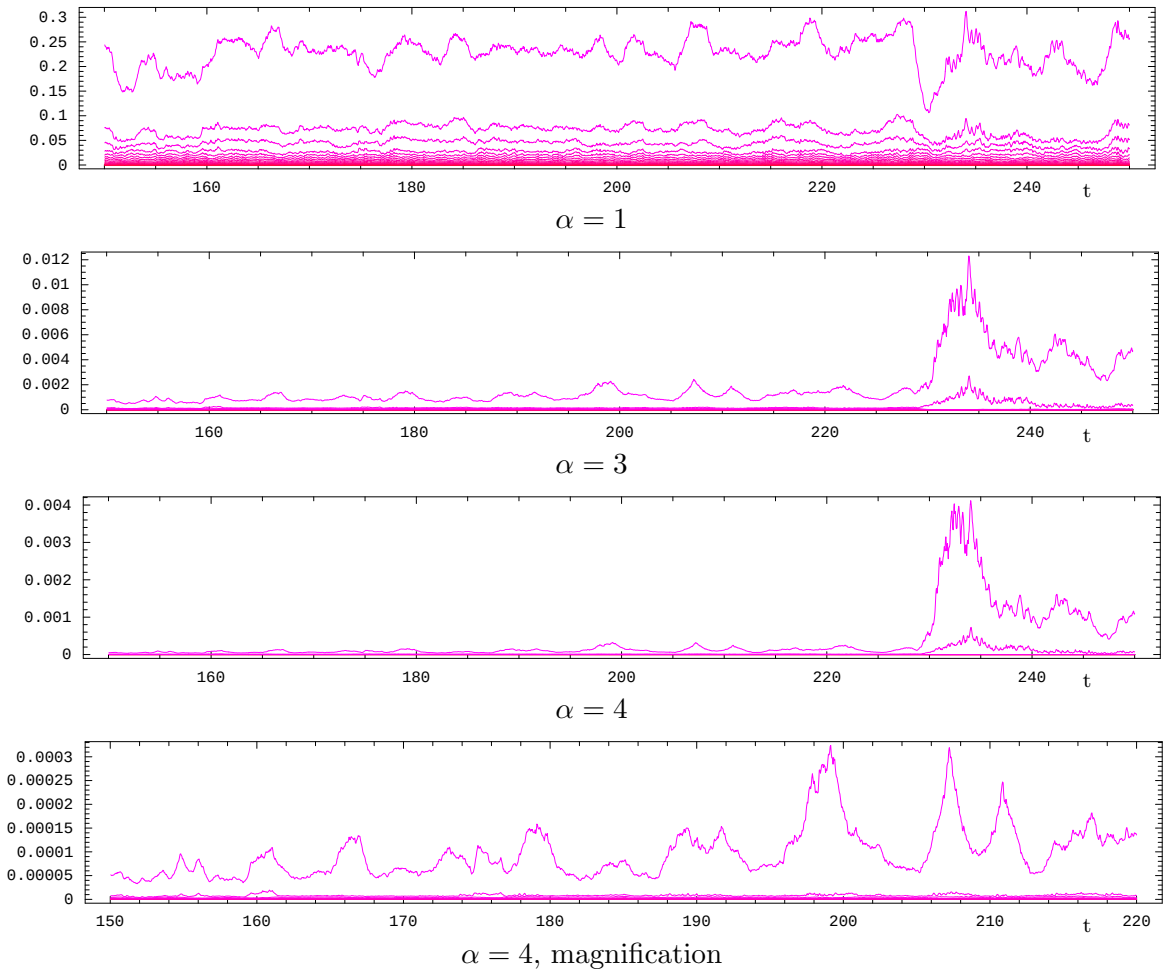

Figure 5: Mean squared profile lengths and non-explained parts

reflected by Fig. 3, which illustrates the complexity of ordinal pattern distributions of the channels related to the 19 electrodes. Here the coloring is given according to the right diagram in Fig. 2. A fat black curve is added in order to illustrate the complexity of the total ordinal pattern distribution. As described by Remark 1, the complexity measure used is the Shannon entropy. One sees a loss of complexity related to the epileptic seizure in all channels and in the whole. Phenomena of that type were generally discussed by different authors, for example, see Lehnertz et al. [2000]. For a discussion based on the ordinal pattern setting we refer to Lauffer & Keller [2002].

Here we want to concentrate on the similarities and dissimilarities between the channels. Fig. 4 provides one-dimensional channel representations for $\alpha = 1, 3, 4$. The two first drawings on the top giving the representations for $\alpha = 1$ coincide up to the coloring. The use of green and red shows more similarity of the channels on the right side than on the left (bad) side, it is however remarkable that the whole plots do not change much in the period of epileptic activity (compare the completely different situation in Fig. 3). We also refer to the relationship between the plots for $\alpha = 1$ and in Fig. 3 being given according to Remark 2.

Whereas for $\alpha = 1$ there is only a weak indication of the epileptic seizure starting at 228 s, one sees a dramatic change of the representation vector for $\alpha = 3$ and in a stronger way for $\alpha = 4$. Here the drawings indicate higher profile distances and more changing mutual positions of the 'channels' in the representation during the period of epileptic activity. Recall that for large α the distance measurement is strongly concentrated on patterns with large total frequencies. We have magnified the left part of the representation for $\alpha = 4$. Note the slowly increasing 'channel distances' prior the epileptic seizure for the given data set.

For the goodness of the one-dimensional approximations see Fig. 5 and the text below Remark 3. At this place only note that in the drawings in Fig. 5 the first curves from the top show the mean squared profile lengths, being the contingency for $\alpha = 1$ (see (5)). Clearly, there is a strong relationship between the representation vectors at a given time and the corresponding mean squared profile lengths.

3 Singular Value Decomposition (SVD)

We now provide the mathematical details. Here, beyond one-dimensional profile approximations, we consider approximations for general dimension. This gives a bit more insight into the methodology and could be useful for further investigations.

Consider the matrix

$$A = A_t = \left(\frac{p_{ij}}{p_{i\cdot} p_{\cdot j}} - 1 \right)_{i,j=1}^{m,n}$$

whose rows are the profiles $\mathbf{a}_i$ (see (2)). Further, let $C^\alpha = C_t^\alpha$ and R be the $n \times n$ and $m \times m$ diagonal matrices with diagonal $p_{\cdot 1}^\alpha, p_{\cdot 2}^\alpha, \ldots, p_{\cdot n}^\alpha$ and $\frac{1}{m}, \frac{1}{m}, \ldots, \frac{1}{m}$, respectively. Then the scalar product on $\mathbb{R}^n$ given by (3) can be written as $\langle \mathbf{a}, \tilde{\mathbf{a}} \rangle^\alpha = \mathbf{a}^T C^\alpha \tilde{\mathbf{a}}$. We also want to consider the scalar product $\langle \cdot, \cdot \rangle$ on $\mathbb{R}^m$ defined by

$$\langle \mathbf{b}, \tilde{\mathbf{b}} \rangle = \mathbf{b}^T R \tilde{\mathbf{b}} = \frac{1}{m} \sum_{i=1}^{m} b_i \tilde{b_i}$$

for $\mathbf{b} = (b_i)_{i=1}^m$ and $\tilde{\mathbf{b}} = (\tilde{b_j})_{i=1}^m$. The corresponding norm, which we denote by $\| \cdot \|$, only scales the usual Euclidean norm by $\sqrt{\frac{1}{m}}$. However, for some further calculations it will be convenient to use the matrix R. The approximation described above is based on the *Singular Value Decomposition (SVD)* of a matrix.

Simple SVD. Each real $m \times n$ matrix M of rank k can be written in the form $M = UDV^T$, where D is a $k \times k$ diagonal matrix having diagonal $\lambda_1, \lambda_2, \ldots, \lambda_k$ with $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_k > 0$, and U and V are $m \times k$ and $n \times k$ matrices, such that $U^T U$ and $V^T V$ form identity matrices (compare Greenacre [1984]). This decomposition has the following geometric consequence, which was found by Eckart & Young [1936] and is central for the subsequent considerations:

Let $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k \in \mathbb{R}^m$ and $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k \in \mathbb{R}^n$ be the columns of U and V, respectively. Then for each $l \leq k$ the vectors $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_l$ (resp. $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_l$) span a linear subspace S of $\mathbb{R}^m$ (resp. $\mathbb{R}^n$) which is *optimal* in the sense that the sum of quadratic Euclidean distances between the row vectors (resp. column vectors) of M and their projections onto S is minimal (or, equivalently, the sum of quadratic projection lengths is maximal) for all subspaces of dimension not exceeding l. So S approximates distances between the row vectors (resp. column vectors) of M in an optimal l-dimensional way.

Generalized SVD. We need the SVD of the matrix A in a form which respects the weights on $\mathbb{R}^m$ and $\mathbb{R}^n$ described by R and C^α, respectively: a matrix U^α consisting of rows $\mathbf{u}_1^\alpha, \mathbf{u}_2^\alpha, \ldots, \mathbf{u}_k^\alpha$ orthonormal in $(\mathbb{R}^m, \langle\cdot,\cdot\rangle)$, a matrix V^α consisting of columns $\mathbf{v}_1^\alpha, \mathbf{v}_2^\alpha, \ldots, \mathbf{v}_k^\alpha$ orthonormal in $(\mathbb{R}^n, \langle\cdot,\cdot\rangle^\alpha)$ and a diagonal matrix D^α as D above, such that

$$A = U^\alpha D^\alpha (V^\alpha)^T, \tag{7}$$

are looked for. Note that the case $\alpha = 1$ provides one of the known approaches to Correspondence Analysis. The decomposition (7), in a more general context called generalized SVD, can easily be deduced from the SVD. Indeed, the matrices are given by $U^\alpha := R^{-\frac{1}{2}}U, V^\alpha := (C^\alpha)^{-\frac{1}{2}}V$ and $D^\alpha := D$, where UDV^T is the SVD of $M := R^{\frac{1}{2}}A(C^\alpha)^{\frac{1}{2}}$. Moreover, it holds

$$R^{\frac{1}{2}}AC^\alpha A^T R^{\frac{1}{2}} = MM^T = U(D^\alpha)^2 U^T$$

and

$$(C^\alpha)^{\frac{1}{2}}A^T RA(C^\alpha)^{\frac{1}{2}} = M^T M = V(D^\alpha)^2 V^T.$$

Since $(U^\alpha)^T RU^\alpha$ and $(V^\alpha)^T C^\alpha V^\alpha$ form identity matrices, this implies

$$AC^\alpha A^T RU^\alpha = U^\alpha (D^\alpha)^2 \tag{8}$$

and

$$A^T RAC^\alpha V^\alpha = V^\alpha (D^\alpha)^2. \tag{9}$$

Note that according to (7) U^α and V^α are linked by $V^\alpha = A^T RU^\alpha (D^\alpha)^{-1}$ and $U^\alpha = AC^\alpha V^\alpha (D^\alpha)^{-1}$. Formulae (8) and (9) say no more than that the columns of the matrices U^α and V^α are eigenvectors of $AC^\alpha A^T R = \frac{1}{m}AC^\alpha A^T$ and $A^T RAC^\alpha = \frac{1}{m}A^T AC^\alpha$, respectively. So, using the statement of Eckart and Young, the following is not hard to show, when $\mathbf{b}_i; i = 1, 2, \ldots, m$ are the columns of A and $\mathbf{x}^S$ denotes the projection of a vector $\mathbf{x}$ onto a subspace S. Note that the statements (I') and (II') are not used and not discussed subsequently, but we give them for completeness.

Theorem. If A has rank k, for each $\alpha > 0$ there exist orthonormal vectors $\mathbf{v}_1^\alpha, \mathbf{v}_2^\alpha, \ldots, \mathbf{v}_k^\alpha \in (\mathbb{R}^n, \langle\cdot,\cdot\rangle^\alpha)$ with the same span as $\{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_m\}$, orthonormal vectors $\mathbf{u}_1^\alpha, \mathbf{u}_2^\alpha, \ldots, \mathbf{u}_k^\alpha \in (\mathbb{R}^m, \langle\cdot,\cdot\rangle)$ with the same span as $\{\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_n\}$, and numbers $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_k > 0$, such that for all $l = 1, 2, \ldots, k$ the following is satisfied:

(I) $\sum_{r=1}^{l}\lambda_r^2=\frac{1}{m}\sum_{i=1}^{m}\|\mathbf{a}_i^{span(\mathbf{v}_1^\alpha,\mathbf{v}_2^\alpha,\ldots,\mathbf{v}_l^\alpha)}\|^\alpha\geq\frac{1}{m}\sum_{i=1}^{m}\|\mathbf{a}_i^S\|^\alpha$ for all linear subspaces S of $\mathbb{R}^n$ of dimension l, i.e. the span of $\{\mathbf{v}_1^\alpha,\mathbf{v}_2^\alpha,\ldots,\mathbf{v}_l^\alpha\}$ is optimal for dimension l,

(I') $\sum_{r=1}^{l}\lambda_r^2=\sum_{j=1}^{n}p_{\cdot j}^\alpha\|\mathbf{b}_j^{span(\mathbf{u}_1^\alpha,\mathbf{u}_2^\alpha,\ldots,\mathbf{u}_l^\alpha)}\|\geq\sum_{j=1}^{m}p_{\cdot j}^\alpha\|\mathbf{b}_j^S\|$ for all linear subspaces S of $\mathbb{R}^m$ of dimension l,

(II) $\lambda_l\mathbf{u}_l^\alpha=(\langle\mathbf{a}_i,\mathbf{v}_l^\alpha\rangle^\alpha)_{i=1}^m$,

(II') $\lambda_l\mathbf{v}_l^\alpha=(\langle\mathbf{b}_j,\mathbf{u}_l^\alpha\rangle)_{j=1}^n$,

(III) $\mathbf{u}_l^\alpha$ and $\mathbf{v}_l^\alpha$ are eigenvectors for the eigenvalue λ_l^2 of the matrices $\frac{1}{m}AC^\alpha A^T$ and $\frac{1}{m}A^TAC^\alpha$, respectively. ■

Remark 3. The special form of A does not play any role, in other words, the theorem remains true for each real $m\times n$ matrix A of rank k with rows $\mathbf{a}_i$ and columns $\mathbf{b}_j$.

The statements given need some explanation. First of all, by (I) the mean squared profile length is $\sum_{r=1}^{k}\lambda_r^2$, where - when we have empirical data - A has usually rank $k=m$ under the assumption $m<n$. So $\sum_{r=1}^{l}\lambda_r^2$ quantifies the part of the mean squared profile length *explained by* (an optimal subspace of) *dimension* l.

The drawings in Fig. 5 being related to those in Fig. 4 show the mean squared profile lengths and their parts not explained by dimensions $1,2,\ldots,m$ from above to below. According to formulae (1) and (2), in the case $\alpha=1$ the first curve from the top shows the contingency. Generally, the difference between the values of the first and the second curve provides the part of mean squared profile length explained by dimension one, i.e. the amount of inhomogeneity visualized by the drawings in Fig. 4. The more this difference approaches the whole mean squared profile distance the better is the one-dimensional approximation of profile distances.

We add two graphics showing $\lambda_1^2,\lambda_2^2,\lambda_3^2,\ldots$ related to the mean squared profile length for some fixed time and for varying α (see Fig. 6, left), and their logarithms (see Fig. 6, right). (The data set considered is the same as at the end of Sec. 2, and the time taken is $t=210s$, there was however no special reason for that choice.) The graphics illustrate that for α large enough, λ_1^2 - being the leading eigenvalue of $\frac{1}{m}AC^\alpha A^T$ and $\frac{1}{m}A^TAC^\alpha$ -, dominates $\lambda_2^2,\lambda_3^2,\ldots$ more and more when α increases. This is not surprising since the matrix $\frac{1}{m}A^TAC^\alpha$ decomposes into $\frac{1}{m}A^TA$ and C^α, and so for increasing α the role of the j-th column in $\frac{1}{m}A^TAC^\alpha$ increases if $p_{\cdot j}$ is a maximal total symbol frequency.

The theorem above says that $\{\mathbf{v}_1^\alpha,\mathbf{v}_2^\alpha,\ldots,\mathbf{v}_k^\alpha\}$ forms a base of the span of $\{\mathbf{a}_1,\mathbf{a}_2,\ldots,\mathbf{a}_m\}$, and by (II) the coordinate of the profile $\mathbf{a}_i$ in the $\mathbf{v}_l^\alpha$-direction is equal to the i-th component of $\lambda_l\mathbf{u}_l^\alpha$. Thus the first l coordinates of an optimal approximation of $\mathbf{a}_i$ in a subspace of dimension l are given by the i-th components of $\lambda_1\mathbf{u}_1^\alpha,\lambda_2\mathbf{u}_2^\alpha,\ldots,\lambda_l\mathbf{u}_l^\alpha$. Note that by (III) the directions defined by the $\mathbf{v}_l^\alpha$ and $\mathbf{u}_l^\alpha$ are unique if all λ_l are different. One can assume that this is true for 'real' data sets.

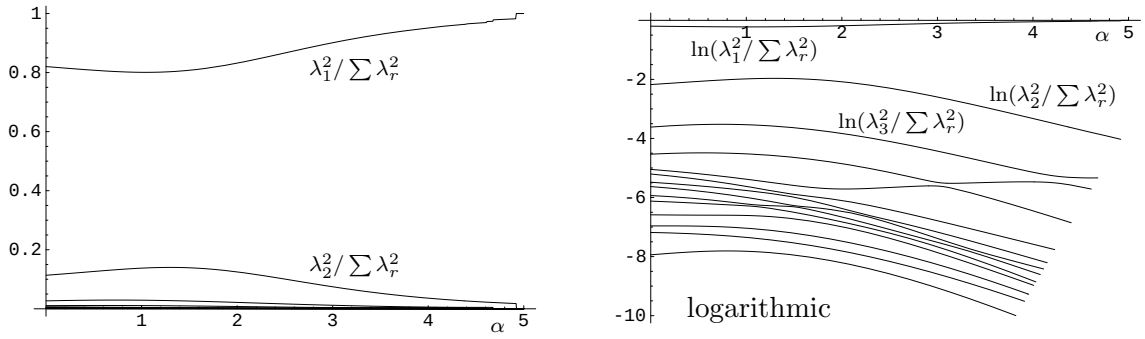

Figure 6: Eigenvalue ratios

One-dimensional profile approximations. In particular, since in the real data case the directions corresponding to the largest eigenvalue can be considered to be unique, a representation vector $\mathbf{w}^\alpha$ as described below (6) must be equal to $\lambda_1\mathbf{u}_1^\alpha$ or $-\lambda_1\mathbf{u}_1^\alpha$. In other words, $\mathbf{w}^\alpha$ is unique up to the signum. $\mathbf{w}^\alpha = \mathbf{w}_t^\alpha$ can numerically be obtained in a very simple way:

Consider the matrix $O_t = \frac{1}{m}A_tC_t^\alpha A_t^T$. The iterates of a suitable (random) vector $\mathbf{w_0}$ with respect to the operator

$$\mathbf{w} \mapsto \frac{O_t(\mathbf{w})}{\|O_t(\mathbf{w})\|}\sqrt{\frac{\|O_t(\mathbf{w})\|}{\|\mathbf{w}\|}} = \frac{O_t(\mathbf{w})}{\sqrt{\|\mathbf{w}\|\,\|O_t(\mathbf{w})\|}}$$

converge to (one of the) $\mathbf{w}_t^\alpha$ (compare II). The reason for this is that λ_1^2 is the leading eigenvalue of the matrix O_t, hence multiplication with this matrix emphasizes the eigendirection of λ_1^2. Clearly, $\frac{O_t(\mathbf{w})}{\|O_t(\mathbf{w})\|}$ has length 1, and in course of the iteration $\frac{\|O_t(\mathbf{w})\|}{\|\mathbf{w}\|}$ approaches to λ_1^2. Note that the initial vector for the iteration must not be orthogonal to this initially unknown eigendirection, which is usually satisfied for a random vector.

Having a good approximation of a first representation vector $\mathbf{w}_t^\alpha$, it is usually a rather good approximation of (one) $\mathbf{w}_{t+s}^\alpha$ for small s. So it can be taken as the initial value for getting a better approximation of $\mathbf{w}_{t+s}^\alpha$. Now clearly one iterates with respect to the operator related to O_{t+s}. So representation vectors can successively be obtained from the first one by increasing t step by step, with the advantage that the distance of successive representation vectors is small: 'the signum does not jump'. (In the EEG situation described below, for $s = 1$ one iteration in each step is usually enough.)

4 Conclusions

We have discussed a simple multivariate method for visualizing qualitative temporal changes in a high-dimensional time-series and differences between its components.

The method, which combines symbolic dynamics and a generalized version of Correspondence Analysis, is based on counting pattern type frequencies. For each time of interest, it quantifies how inhomogeneous the system of time series components is and provides a one-dimensional representation of this system. A scaling parameter allows to differentiate between the components with respect to a specific weighting of the pattern frequencies.

When the underlying symbolic dynamics only refers to the ordinal structure of the given time series as proposed by Bandt & Pompe [2002], the method is fast and robust. Then it can be applied to very long time series of large dimension, in particular to EEG data sets. The application to 19-channel scalp EEG data from children with epileptic disorders has shown a potential of the method to visualize long-term qualitative changes and local differences in brain-electrical activity. Here the scaling parameter introduced can play a substantial role. In order to go beyond the explorative level and to get reliable results, a general modelling approach would be important.

Acknowledgement. I would like to thank Heinz Lauffer from the Department of Pediatric Medicine of the University Greifswald for many fruitful discussions and for providing the EEG data.

References

Bandt, C., Keller, G., & Pompe, B. [2002] "Entropy of interval maps via permutations" *Nonlinearity* **15**, 1595–1602.

Bandt, C. & Pompe, B. [2002] "Permutation entropy: A natural complexity measure for time series" *Phys. Rev. Lett.* **88**, 174102.

beim Graben, P., Jurish, B., Saddy, D., & Frisch, S. [this issue] "Language processing by dynamical systems" *Int. J. Bif. Chaos*.

beim Graben, P., Saddy, J. D., Schlesewsky, M., & Kurths, J. [2000] "Symbolic dynamics of event-related brain potentials" *Phys. Rev. E* **62**, 5518–5541.

Daw, C. S., Finney, C. E. A., Nguyen, K., & Halow, J. S. [1998] "Symbol statistics: a new tool for understanding multiphase flow phenomena" in: *International Mechanical Engineering Congress & Exposition (Anaheim, California)* pp. 221–229.

Ebeling, W. & Nicolis, G. [1992] "Word frequency and entropy of symbolic sequences: a dynamical perspective" *Chaos Solitons Fractals* **2**(6), 635–650.

Eckart, C. & Young, G. [1936] "The approximation of one matrix by another of lower rank" *Psychometrica* **1**, 211–218.

Finney, C. E. A., Green, J. B., & Daw, C. S. [1998] "Symbolic time-series analysis of engine combustion measurements" SAE Paper No. 980624.

Greenacre, M. J. [1984] Theory and applications of correspondence analysis. (Academic Press Inc. London).

Lauffer, H. & Keller, K. [2002] "Symbolic analysis of high-dimensional time-series" Preprint (Greifswald).

Lehnertz, K., Arnhold, J., Grassberger, P., & Elger (eds.), C. [2000] Chaos in brain? (World Scientific).

Schwarz, U., Benz, A., Kurths, J., & Witt, A. [1993] "Analysis of solar spike events by means of symbolic dynamics methods" *Astronomy and Astrophysics* **277**, 215–224.

Steuer, R., Ebeling, W., Bengner, T., Dehnicke, C., Hättig, H., & Meencke, H.-J. [this issue] "Entropy and complexity analysis of intracranially recorded EEG" *Int. J. Bif. Chaos*.

Takens, F. [1981] "Detecting strange attractors in turbulence" in: *Dynamical systems and turbulence, Warwick 1980 (Coventry, 1979/1980)* pp. 366–381 (Springer Berlin).