Publication Series of the John von Neumann Institute for Computing (NIC)

NIC Series Volume 24

John von Neumann Institute for Computing (NIC)

Alexander Kraskov

Synchronization and Interdependence Measures and their Applications to the Electroencephalogram of Epilepsy Patients and Clustering of Data

Dissertation (PhD thesis)

NIC Series Volume 24 ISBN 3-00-013619-3

Research Group Complex Systems

Die Deutsche Bibliothek – CIP-Cataloguing-in-Publication-Data
A catalogue record for this publication is available from Die Deutsche Bibliothek

Publisher: NIC-Directors
Distributor: NIC-Secretariat
Research Centre Jülich
52425 Jülich
Germany
Internet: www.fz-juelich.de/nic
Printer: Graphische Betriebe, Forschungszentrum Jülich

NIC Series Volume 24 ISBN 3-00-013619-3

Department of Physics
University of Wuppertal

Synchronization and Interdependence Measures and their Applications to the Electroencephalogram of Epilepsy Patients and Clustering of Data

Dissertation (PhD thesis)

by

Alexander Kraskov

February 2004

(WUB-DIS 2004-1)

Zusammenfassung

Das Hauptziel der vorliegenden Doktorarbeit ist die vergleichende Untersuchung, Weiterentwicklung und Anwendung verschiedener Ansätze zur Messung von Synchronisation und Interdependenz zwischen Zeitreihen. Die behandelten Ansätze beinhalten die lineare Kreuzkorrelation und zwei aus der Informationstheorie abgeleitete Maße, Mutual Information und Transfer-Entropie. Des weiteren werden zwei Indices für Phasensynchronisation sowie vier Versionen nichtlinearer Interdependenzmae untersucht.

Im ersten Teil dieser Arbeit werden zwei neue Schätzer für Mutual Information und Transfer-Entropie vorgestellt. Diese weisen im Gegensatz zu den bisher gebräuchlichen Schätzern minimale systematische und statistische Fehler auf.

Anschließend werden im zweiten Teil zwei verschiedene Methoden der Phasenextraktion, zum einen basierend auf der Hilbert-Transformation und zum anderen auf der Wavelet-Transformation, theoretisch miteinander verglichen.

Im dritten Teil werden die verschiedenen Maße zur Analyse elektroenzephalographischer Aufzeichnungen von Epilepsie-Patienten herangezogen. Zunächst wird in einer umfassenden Studie die Eignung der verschiedenen Maße zur Lokalisierung des epileptischen Fokus untersucht. Anschließend werden die dabei gewonnenen Resultate mit Hilfe einer speziellen bivariaten Surrogatdatenmethode getestet.

Im letzten Teil dieser Arbeit wird eine neue auf der Mutual Information basierende Methode der hierarchischen Klassifizierung präsentiert. Anhand einer Anwendung auf zwei Beispiele biologischer Daten (dem EKG einer schwangeren Frau sowie DNS verschiedener Spezies) wird illustriert, dass diese Methode zu guten Ergebnissen für Daten verschiedenster Herkunft führen kann.

First referee:	Prof. Dr. Peter Grassberger, Universität Wuppertal
Second referee:	Prof. Dr. Andreas Klümper, Universität Wuppertal
Day of submission:	30. January 2004
Day of oral examination:	27. February 2004

Contents

Chapter 1

Introduction

The history of the very interesting phenomenon of *synchronization* started in the seventeenth century when the famous Dutch researcher Christiaan Huygens observed perfect agreement between the oscillating motions of two clocks hanging from a common support [41, 89]. The word *"synchronization"* came to many contemporary languages from ancient Greece. The etymology of this word is very simple, it consists of two parts $\sigma\upsilon\nu$ (syn = common) and $\chi\rho o\nu o\varsigma$ (chronos = time). In a direct translation the verb "synchronize" means "to happen at the same time" or "to agree in time". This translation can be taken as a first approximation to the definition of synchronization because it contains one of the main features, namely coincidence in time, i.e., synchronous motion. One can argue, that this condition is too weak, especially, if coincidence happens very rarely. In our opinion the repeated coincidence over long time, i.e., lasting synchronous motion, is usually a consequence of synchronization. Therefore, in this thesis we will use synchronization and synchronous motion as synonymous.

Later, in the beginning of the twentieth century, synchronization phenomena were studied by W.H. Eccles and J.H. Vincent in the context of electrical and radio engineering development. In their experiments the adjustment of the frequencies of two coupled triode generators with initially different frequencies was demonstrated. A few years later E. Appleton [11] and B. van der Pol extended the experiments of Eccles and Vincent and also undertook theoretical investigations of synchronization phenomena. Van der Pol derived his famous equation, the first example of a non-linear self-oscillating system [124]. Moreover, van der Pol together with van der Mark proposed an electrical model of the human heart consisting of three coupled relaxation oscillators [125].

A new stage of synchronization studies started some decades after the discovery of deterministic chaos [71]. In the early 1980s, the notion of synchronization was extended to the case of interacting chaotic oscillators [31, 86, 4, 84]. Deterministic chaos is characterized by sensitivity to initial conditions, i.e., trajectories starting from very close points diverge exponentially. Therefore, synchronization between chaotic oscillators was not expected.

However, extensive experimental investigations have proven its existence. Prominent examples include electronics [87, 39, 83], laser dynamics [28, 105, 119], solid state physics [85], plasma physics [97], communication [20, 45] and chaos control [91, 107].

The simplest form of synchronization occurs if the states of systems exactly coincide in time. This type of synchronization is usually referred to as identical synchronization. It can be observed if the coupling strength between identical systems is high enough [31, 86]. In this case coincidence means that the two states are identical. One can easily extend the notion of coincidence to a more complicated functional relation. This leads us to the concept of generalized synchronization, which was introduced for unidirectionally coupled systems in Refs. [4, 106]. Quite often the most reliable information about interacting systems is contained in the phases of each system. An entrainment of these phases is fundamental for phase synchronization. For chaotic oscillators it was first described in Refs. [101, 88, 80].

The variety of synchronization concepts spurred the development of many different approaches aiming at a quantification of the degree of synchronization between two systems or rather between two time series measured from the respective systems. Mutual information is one of them [37, 21]. It is zero if and only if two random variables are strictly independent. This distinctive feature singles out mutual information among other measures. Different estimators for mutual information were proposed in the literature but all of them have significant systematical errors. This problem has motivated us to develop two new families of estimators with a minimal bias. These estimators will be introduced in the first original part of this thesis.

Topological approaches to quantify generalized synchronization include the method of mutual false nearest neighbors [106] and the index based on non-linear mutual predictions [108] as well as more recent measures like the non-linear interdependencies [13] and synchronization likelihood [116]. Different ways to quantify phase synchronization have been proposed in [120, 72]. In this context, the notion of a phase is very important and different approaches for its extraction from time series have been developed. Two of the most important techniques use the Hilbert transform [101] and the wavelet transform [57]. In the literature a theoretical comparison of these two methods was still missing. This comparison along with an extended discussion about the ambiguity of phase definition constitutes the second original part of this thesis.

A challenging application for measures of synchronization is the study of neuronal dynamics, since synchronization phenomena have been increasingly recognized as a key feature for establishing the communication between different regions of the brain [126, 29, 127]. On the other hand synchronization plays an important role for pathological processes such as Parkinson's disease or epilepsy. A unique window to neuronal dynamics is given by electroencephalographic (EEG) recordings from epilepsy patients undergoing pre-surgical diagnostics [69]. To yield sufficient information for an unequivocal localization of the seizure-generating structure (epileptic focus) in the brain, sometimes multichannel recordings using intracranial monitoring techniques are acquired. In this case the EEG is recorded directly from the surface of the brain and from specific structures within the brain [26]. The investigation of these recordings by means of linear and non-linear time series analysis techniques can help to further understand the spatio-temporal dynamics of the epileptic brain (see, e.g., [62]). In particular, synchronization and de-synchronization phenomena play an important role in the epileptic process. This motivated us to carry out a comprehensive comparison of the different measures of synchronization with respect to their capability to detect the side of the epileptic focus. To test the degree to which the obtained results are specifically related to synchronization phenomena we applied a bivariate surrogate data analysis. This study constitutes the third original part of this thesis.

All measures of phase synchronization and nonlinear interdependence test for similarities between two systems using only one dimensional times series, whereas mutual information can be applied to objects of any dimension. This feature puts mutual information between solely bivariate approaches and multivariate approaches, like e.g., independent component analysis (ICA) [42]. In application to the analysis of multichannel EEG recordings multivariate methods can be used to derive different kinds of spatial and temporal information. For example, grouping the different channels for a more precise localization of the epileptic focus or classification of the intervals preceding an epileptic seizure and the intervals far away from any seizure activity can be of great value in epilepsy research. An attempt to retrieve this type of information can be undertaken with the help of clustering methods [27, 43]. In the last original part of this thesis we propose a new method for a hierarchical clustering of data based on the grouping property of mutual information. We show two examples of its application to data from genetics (mitochondrial DNA sequences of mammals) and cardiology (electrical activity of the heart of a pregnant woman).

This thesis is organized as follows: First, in Chapter 2 an introduction to synchronization and its different notions is given. In Chapter 3 different approaches to quantify synchronization phenomena are presented. Along with traditional methods such as linear cross-correlation and coherence functions (Sec. 3.1), nonlinear approaches with information theoretical background, namely mutual information (Sec. 3.2) and transfer entropy (Sec. 3.3) and methods developed in the framework of nonlinear time series analysis (Secs. 3.4 and 3.5) are introduced. In Sec. 3.2.2 new estimators for mutual information and transfer entropy are presented. Sections 3.4.3 and 3.4.4 contain a comparative study of different phase extraction methods. In Chapter 4 all measures of synchronization and interdependence introduced in Chapter 3 are applied to electroencephalographic time series measured from the brain of epilepsy patients (Section 4.1). In particular, the localization of the epileptic focus is addressed in Section 4.2. The bivariate surrogate data techniques are described and applied in Sec. 4.2.2. In Chapter 5 a new algorithm for hierarchical clustering based on mutual information is presented. In Section 5.1 the algorithm is formulated and in Section 5.3 two its applications are discussed. Finally, the conclusions of this thesis are drawn in Chapter 6.

Chapter 2

What is synchronization?

Following Pikovsky *et al.* we will understand synchronization as an *adjustment of rhythms of oscillating objects due to their weak interaction* [89]. Let us explain in more detail what exactly we understand under the terms used in this definition. Interaction can be realized for instance through a coupling. The coupling can be either unidirectional or bidirectional. In the latter case one could expect *mutual synchronization*, both systems adjust their rhythms to each other. In the former case one usually speaks about synchronization by an external force, where the forcing system is also called a *driver* and the driven system is called a *response*. The rhythm of the response is adjusted to the rhythms of the driver.

One of the main properties which distinguish a synchronization phenomenon for instance from a resonance is the existence of own rhythms for each oscillating object, even when not driven at all. If the rhythm of a response is only induced by a driver (as it is the case with the resonance) then it is not possible to treat it as synchronization. Moreover, own rhythms should exist also for a noninteractive case, i.e., a system under consideration can in principle be separated into different subsystems all of which have their own rhythms. A prominent example is the hare-lynx cycle, a well-known ecological phenomenon in which one cannot speak about synchronization between the two populations because the hare-lynx ecological system cannot be separated into independent oscillating subsystems (either the lynxes will die without food or the hare population will explode). Nevertheless, in such a large system as a human brain, which is known to contain approximately 10^{11} neurons with a total of 10^{14} to 10^{15} synaptic connections, one can still investigate the synchronization phenomenon between different brain regions. Considering the strength of an interaction one can say that as soon as it gets strong (very large values of coupling) one cannot speak of two interacting systems but rather of one combined system. That is why the word “weak” appears in the definition of synchronization.

In the literature three main types of synchronization are usually distinguished, namely *identical synchronization*, *phase synchronization*, and *generalized synchronization*.

2.1 Identical synchronization

The simplest case of synchronization is identical or complete synchronization. It is observed if the states of coupled systems coincide in the limit $t \to \infty$ [31, 86]. For the two systems X and Y with state vectors $\mathbf{x}(t)$ and $\mathbf{y}(t)$ it means

$$\lim_{t\to\infty}[\mathbf{x}(t) - \mathbf{y}(t)] = 0. \tag{2.1}$$

Identical synchronization is a special case of many other types of synchronization which is obtained for sufficiently strong coupling, however, only in the case when the systems have identical parameters. Otherwise, if the parameters of the coupled systems slightly mismatch, the states can come close to each other but still remain different.

2.2 Phase synchronization

Phase synchronization is based on the notion of the phase of oscillation, originating from the phase of a harmonic motion. In physics the term “phase” carries many different meanings. For instance the phase in “phase transition” has nothing to do with the phase in “phase space”. In the following we will always use the term “phase” in the sense of a phase of an oscillation.

The phase is a very specific variable of a motion. If we consider the behavior of an oscillating object in a coordinate system rotating with an angular velocity of oscillations then the motion will be represented in state space by a single point. Moreover, any perturbation of the phase along the trajectory is equivalent to a change in time. Since autonomous systems are time invariant, the phase perturbation neither grows nor decays. The phase of an oscillator can be considered as the variable that corresponds to the zero Lyapunov exponent. That is why the phase can be very easily adjusted by an external action, i.e., synchronization can occur.

Phase synchronization is a natural concept for the description of two coupled linear (harmonic) or non-linear oscillators or any other system for which the definition and determination of a phase is straightforward. Only recently has this concept been also applied to chaotic oscillators [101, 88, 80] and even further extended to the analysis of almost arbitrary dynamics (cf. [102]).

Mathematically, phase synchronization can be defined as the entrainment of the phases:

$$|n\varphi_x(t) - m\varphi_y(t)| \leq const, \tag{2.2}$$

where $\varphi_x(t)$ and $\varphi_y(t)$ are the phases extracted from systems X and Y, n and m being integers. Different methods of phase extraction will be considered in detail in the Sec. 3.4.

The phases do not have to be defined in this Section on the circle from 0 to 2π but should rather be unfolded.

Phase synchronization has fewer constraints than identical synchronization, i.e., only the phases have to be locked whereas the amplitudes can remain chaotic and may even be uncorrelated.

2.3 Generalized synchronization

As already stated in Sec. 2.1, identical synchronization which is characterized by complete coincidence of states of the systems can only occur for identical systems. If the systems are not identical it is still possible to speak about synchronization if one system (response) is following the other (driver), although in a weaker sense. This phenomenon is usually called generalized synchronization [106, 4, 84].

Let us consider two systems which are unidirectionally coupled:

$$\begin{aligned} \mathbf{x}(t+1) &= \mathbf{f}(\mathbf{x}(t)), \\ \mathbf{y}(t+1) &= \mathbf{g}(\mathbf{x}(t), \mathbf{y}(t)). \end{aligned} \tag{2.3}$$

Here **x** is a driver, and **y** is a response. One speaks about generalized synchronization if the state of the response is completely defined by the state of the driver, i.e., there exists a function $\mathcal{G}$ such that

$$\mathbf{y} = \mathcal{G}(\mathbf{x}). \tag{2.4}$$

The function $\mathcal{G}$ does not need to be smooth. In fact, Pyragas [92] defined the cases of smooth and non-smooth transformations as strong and weak synchronization, respectively (see also Ref. [40]).

The existence of generalized synchronization means that dynamics of **y** is completely defined by the dynamics of **x**. It becomes possible only if the dynamics of **y** is stable, i.e., the maximal Lyapunov exponent[1] corresponding to **y** is negative. This condition is necessary and sufficient if one excludes the case of multistability.

The relation between phase synchronization and generalized synchronization can not be defined in general. First, in Ref. [83] it was claimed that generalized synchronization implies phase synchronization i.e., the phase synchronization appears first with increasing of the interaction strength. Later, for several examples the reverse order was found [130].

It is possible to apply the techniques designed to detect generalized synchronization not to the systems directly but rather to their phases. This type of synchronization was described in Ref. [60] and was called *generalized phase synchronization*.

[1] This Lyapunov exponent is also sometimes called conditional Lyapunov exponent [84].

2.4 Once again about identical synchronization

Let us consider the identical synchronization from a slightly different point of view. The usual identical synchronization introduced in Sec. 2.1 can schematically be represented as

$$X \Longleftrightarrow Y. \tag{2.5}$$

Here systems X and Y *speak* to each other.

The next possible case when two identical systems might synchronize is if they are driven by a common driver (unidirectional coupling). Two responses which only *listen* to the same driver but are not coupled directly might still be identically synchronized with each other. The driver Z can be a system completely different from X and Y (it can even be a random noise), and clearly X and Z are not synchronized identically. The schematic representation is the following

$$X \longleftarrow Z \longrightarrow Y. \tag{2.6}$$

In this case one usually speaks about generalized synchronization between the driver Z and the response X (or Y) as described in the previous Section. The identical synchronization between two responses can be used as a criterion for generalized synchronization (auxiliary [3] or replica [84] system approach). The existence of identical synchronization between non interacting, but driven systems is already non trivial. The situation will become more difficult if we let the systems X and Y influence the driver through a back coupling. This can be schematically represented as

$$X \Longleftrightarrow Z \Longleftrightarrow Y. \tag{2.7}$$

Here the responses X and Y do not only *listen* to the driver Z but also *talk* to it but not with each other. Strictly speaking, the subsystems are neither responses nor drivers, but for convenience we will still refer to X and Y as responses and to Z as a driver. Also in this case the identical synchronization between X and Y is possible, although the responses X and Y are not identically synchronized with the driver Z.

This type of identical synchronization can be also used as a criterion for the *partial generalized synchronization* between the driver Z and the response X (or Y) [131].

Chapter 3

Measures of interdependence and synchronization

In this Chapter we will review different methods for the detection of interdependencies and synchronization. We will also introduce new estimators for mutual information and transfer entropy, which serve as measures of dependence and predominant direction of interaction, respectively. All these methods will be applied later in this thesis to experimental time series mainly of biological nature, e.g., electrocardiograms or electroencephalograms.

In their discussion about detection of synchronization from experimental data Pikovsky *et al.* [89] distinguish two types of experiments, namely “active” and “passive” ones. In the next chapters, where we will apply methods for detection of interdependencies and synchronization to biological time series, we will obviously deal with experiments of the “passive” type. In such experiments a “tuning knob” is not available. It is not always possible to control in detail the parameters of the systems and/or the interaction strength.

One naturally arising question is whether one can detect synchronization by analyzing bivariate data from passive experiments. A very important property of synchronization is adjustment and it is a process and not a state. That is why the answer to the question is that in general such detection is not possible. Nevertheless, a synchronization analysis may provide useful information on the interrelation or interdependence of systems.

This Chapter has the following structure. First, the linear cross-correlation will be introduced in Sec. 3.1. In Secs. 3.2 and 3.3 the two measures with information theoretical background, namely mutual information and transfer entropy, will be discussed. Here new estimators for mutual information will be presented. Measures of phase synchronization as well as the relation between some of them will be discussed in Sec. 3.4. Finally, in Sec. 3.5 measures of generalized synchronization based on the relation between attractors reconstructed in time-delay state space will be described.

3.1 Cross-correlation

A physical process X can be described either in the *time domain*, by the value of some quantity $x(t)$ as a function of time, or alternatively in the *frequency domain*, where the process X is specified by giving a complex function $\hat{x}(\omega)$ of frequency ω. One goes backward and forward between these two representations by means of the *Fourier transform* (FT),

$$\hat{x}(\omega) = (\mathcal{F}x)(\omega) \quad = \quad \int_{-\infty}^{\infty} x(t)e^{i\omega t}dt, \tag{3.1}$$

$$x(t) = (\mathcal{F}^{-1}\hat{x})(t) \quad = \quad \frac{1}{2\pi}\int_{-\infty}^{\infty} \hat{x}(\omega)e^{-i\omega t}d\omega, \tag{3.2}$$

where $(\mathcal{F}x)$ denotes the FT and $(\mathcal{F}^{-1}\hat{x})$ denotes its inverse [18].

The cross-correlation function of two functions $x(t)$ and $y(t)$ is defined as a function of a time *lag* τ:

$$c_{xy}(\tau) = \int_{-\infty}^{\infty} x(t+\tau)y(t)dt. \tag{3.3}$$

In the frequency domain one can define the cross-spectrum function which is a complex function of frequency,

$$\tilde{c}_{xy}(\omega) = (\mathcal{F}x)(\omega)\cdot(\mathcal{F}y)^{*}(\omega), \tag{3.4}$$

where the asterisk denotes complex conjugation. One can prove the correlation theorem:

$$c_{xy} = (\mathcal{F}\tilde{c}_{xy}) \quad \text{and} \quad \tilde{c}_{xy} = (\mathcal{F}^{-1}c_{xy}). \tag{3.5}$$

This result shows that multiplying the FT of one function by the complex conjugate of the FT of the other gives the FT of their cross-correlation. The cross-correlation of a function with itself is called its *autocorrelation* and the cross-spectrum of a function with itself is called its *autospectrum*. In this case the correlation theorem (Eq.(3.5)) becomes the well known Wiener-Kninchin theorem. In addition, one can also introduce the normalized cross-spectrum, which is usually called *coherence* function. It is the cross-spectrum normalized by the autospectra of each function:

$$\Gamma_{xy}(\omega) = \frac{|\langle\tilde{c}_{xy}(\omega)\rangle|}{\sqrt{\langle\tilde{c}_{xx}(\omega)\rangle}\sqrt{\langle\tilde{c}_{yy}(\omega)\rangle}}, \tag{3.6}$$

where $\langle\cdot\rangle$ denotes the ensemble average.

In applications one usually deals with a finite amount of measurements offered by experimentalists. Assume that $x_0,\ldots,x_{N-1}$ and $y_0,\ldots,y_{N-1}$ are two simultaneously measured stationary time series, which have zero mean and unit variance. The estimate of the cross-correlation is then defined[1] as a function of the time lag $\tau = -(N-1), \ldots, 0, \ldots, N-1$:

$$c_{xy}(\tau) = \begin{cases} \frac{1}{N-\tau} \sum_{i=1}^{N-\tau} x_{i+\tau} y_i, & \tau \geq 0 \\ \\ c_{yx}(-\tau), & \tau < 0. \end{cases} \tag{3.7}$$

The cross-correlation is normalized to the range from minus one (complete anti-synchronization) to one (complete synchronization). The value of cross-correlation near zero indicates linear independence of systems. The estimator for the cross-correlation could give non-zero values for two completely linearly independent system. That is why a significance threshold for the estimated cross-correlation should be taken into account (e.g., the Bartlett estimator [15, 18]). Nevertheless, the cross-correlation is one of the simplest and mostly used measures of synchronization between two systems, although it is not sensitive to nonlinear dependencies.

The estimation of the coherence function (Eq.(3.6)) is not easy. The ensemble averaging is not possible and we have to replace it by the time average assuming the stationarity of underlying dynamics. Usually one divides time series into segments and estimates the cross-spectrum and the autospectra for each segment, and then takes the average over these segments. The choice of the segment length, a window function (e.g., Bartlett, Welch) is not always obvious and depends on the problem in hand.

While the coherence function is a function of the frequency ω, it is a very useful measure when one is interested in the synchronization related to certain frequency ranges only, e.g., in the classical EEG frequency bands (cf. [69]).

In applications (see Chapter 4) we will use two measures of linear synchronization, namely

$$C_0 = c_{xy}(0), \tag{3.8}$$

i.e., the cross-correlation at zero time lag and the maximum cross-correlation:

$$C_{max} = \max_{\tau}\{|c_{xy}(\tau)|\}. \tag{3.9}$$

Both of these are symmetric measures.

3.2 Mutual information

Information theoretical measures [21, 37] like Shannon and Kolmogorov entropies are widely used to analyze nonlinear systems. In particular, they are used to characterize the

[1] Here and in the following we will use the same notations for mathematical quantity and their estimates.

degree of randomness of time sequences, and to quantify the difference between two probability distributions. Statistical dependence between signals is often estimated by their mutual information (MI).

Among the measures of interdependence between random variables, MI is singled out by its close ties to Shannon entropy and the theoretical advantages derived from this. In contrast to the linear cross-correlation, it is sensitive also to dependencies which do not manifest themselves in the cross-correlation. Indeed, MI is zero if and only if the two random variables are strictly independent. The latter is also true for quantities based on Renyi entropies [95], and these are often easier to estimate (in particular if their order is 2).

First, the Shannon entropy is defined for *discrete* random variables. Assume that one has a discrete random variable X, with $p_X(x) = \mathrm{prob}(X = x)$, where x is one of the possible states of X. Then the Shannon entropy is defined as

$$H(X) = -\sum_x p_X(x) \log p_X(x). \tag{3.10}$$

The Shannon entropy is the *average information* about X. The base of the logarithm determines the units in which information is measured. In particular, taking the base two leads to information measured in bits. In the following we always will use natural logarithms.

To define an entropy for a continuous variable X with density[2] μ_X one first introduces some binning ('coarse-graining'), artificially defining thereby a discrete random variable. If x is a vector with dimension m and each bin has Lebesgue measure Δ, then $p_X(x) \approx \mu_X(\tilde{x})\Delta^m$ with $\tilde{x}$ chosen suitably in bin x. According to Eq.(3.10) the entropy for the binned variable will be[3]

$$H_{\mathrm{bin}}(X) \approx \tilde{H}(X) - m \log \Delta, \tag{3.11}$$

where the *differential entropy* $\tilde{H}(X)$ is given by

$$\tilde{H}(X) = -\int dx \; \mu_X(x) \log \mu_X(x). \tag{3.12}$$

Notice that $H_{\mathrm{bin}}(X)$ is a true (average) information and is thus non-negative, but $\tilde{H}(X)$ is not an information and can be negative. Also, $\tilde{H}(X)$ is not invariant under homeomorphisms (smooth and uniquely invertible maps) $x \to \phi(x)$.

The MI between two discrete random variables X and Y with marginal probabilities $p_X(x) = \mathrm{prob}(X = x)$ and $p_Y(y) = \mathrm{prob}(Y = y)$, and with joint probability $p(x, y) = \mathrm{prob}(X = x, Y = y)$ is defined as

$$I(X,Y) = \sum_{x,y} p(x,y) \;\log \frac{p(x,y)}{p_X(x)p_Y(y)}. \tag{3.13}$$

[2] Here, it is assumed that the density of X exists as a "smooth function".

[3] If X lives on a fractal set m, then for $\Delta \to 0$ one has $H_{\mathrm{bin}}(X) \sim -D_I \log \Delta$, where D_I is its information dimension.

For continuous variables one can use a similar binning to find that $I(X,Y)$ remains finite and is independent of Δ, giving

$$I(X,Y) = \lim_{\Delta\to 0} I_{\text{bin}} = \iint dxdy\ \mu(x,y)\ \log \frac{\mu(x,y)}{\mu_X(x)\mu_Y(y)}\ , \tag{3.14}$$

where $\mu(x,y)$ is the joint density and $\mu_X(x) = \int dy\mu(x,y)$ and $\mu_Y(y) = \int dx\mu(x,y)$ are the marginal densities of X and Y. It is assumed that the integrals written above exist in some mathematical sense. In particular, it is always assumed that $0\log(0) = 0$, and therefore one does not have to assume that densities are strictly positive.

Despite of being the sum of entropies

$$I(X,Y) = H(X) + H(Y) - H(X,Y) \tag{3.15}$$

the MI is invariant under homeomorphisms $x \to \phi(x)$ and $y \to \psi(y)$.

PROOF: If $J_X = ||\partial X/\partial X'||$ and $J_Y = ||\partial Y/\partial Y'||$ are the Jacobi determinants then

$$\mu'(x',y') = J_X(x')J_Y(y')\mu(x,y) \tag{3.16}$$

and similarly for the marginal densities, which gives

$$\begin{aligned} I(X',Y') &= \iint dx'dy'\mu'(x',y') \log \frac{\mu'(x',y')}{\mu'_x(x')\mu'_y(y')} \\ &= \iint dxdy\ \mu(x,y)\ \log \frac{\mu(x,y)}{\mu_x(x)\mu_y(y)} \\ &= I(X,Y)\ . \end{aligned} \tag{3.17}$$

For n random variables $X_1, X_2 \dots X_n$, the MI is defined as

$$I(X_1,\dots,X_n) = \sum_{k=1}^{n} H(X_k) - H(X_1,\dots,X_n). \tag{3.18}$$

This quantity is often referred to as (generalized) redundancy, in order to distinguish it from different "mutual informations" which are constructed analogously to higher order cumulants, but we shall not follow this usage.

3.2.1 Estimation of Mutual Information

In applications, one usually has the data available in form of a statistical sample. To estimate $I(X,Y)$ one starts from N bivariate measurements $z_i = (x_i, y_i),\ i = 1,\dots,N$ which are assumed to be iid (independent identically distributed) realizations. The aim is to estimate $I(X,Y)$ from the set $\{z_i\}$ alone, without knowing explicitly the densities μ, μ_x, and μ_y.

The most straightforward and widespread approach for estimating MI consists in partitioning the supports of X and Y into bins of finite size, and approximating Eq.(3.14) by the finite sum

$$I(X,Y) \approx I_{\text{bin}}(X,Y) \equiv \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p_X(x)p_Y(y)}. \tag{3.19}$$

An estimator of $I_{\text{bin}}(X,Y)$ is obtained by simply counting the numbers of points falling into the various bins. If n_x (n_y) is the number of points falling into the x-th bin of X (y-th bin of Y), and n_{xy} is the number of points in their intersection, then we approximate $p_X(x) \approx n_x/N$, $p_Y(y) \approx n_y/N$, and $p(x,y) \approx n_{xy}/N$. It has already been mentioned that I_{bin} converges to $I(X,Y)$ if we first let $N \to \infty$ and then let all bin sizes tend to zero, if all densities exist as proper (not necessarily smooth) functions.

The bin sizes used in Eq.(3.19) do not need to be the same for all bins. Optimized estimators [30, 23] use adaptive bin sizes which are essentially geared at having equal numbers n_{xy} for all pairs (x,y) with non-zero measure. While such estimators are much better than estimators using fixed bin sizes, they still have systematic errors which result from approximating $I(X,Y)$ by $I_{\text{bin}}(X,Y)$, and from approximating (logarithms of) probabilities by (logarithms of) frequency ratios. The latter bias could presumably be minimized by using corrections for finite n_x resp. n_{xy} [35, 33].

Kernel techniques is an attractive alternative to binning a distribution which is discussed thoroughly in the literature (cf. [114]). The main assumption is that the probability density is smooth enough such that structure below a certain kernel band width may be ignored. The simplest possibility is to estimate the density at a point $\tilde{x}$ by the number of points in a box centered at $\tilde{x}$ of size ϵ divided by its volume. Rather than simply counting the points, one can give them distance-dependent weights using some kernel function. In order to reconstruct the density of X at an arbitrary point $\tilde{x}$, the general form of a kernel estimator is given by:

$$\hat{\mu}(\tilde{x}) = \frac{1}{N\epsilon} \sum_{x}^{N} K(\frac{\tilde{x}-x}{\epsilon}) \tag{3.20}$$

The non-negative function $K(x)$ determines the distance-dependent weight of each point. One of the most common kernels is the Gaussian kernel $K(u) = \frac{1}{\sqrt{2\pi}} e^{-(1/2)\,u^2}$. In case of the rectangular kernel $K(u) = \frac{1}{2}$, $(|u| < 1)$ this estimator is equivalent to the method mentioned in the previous paragraph[4]. The parameter ϵ is called the band width and determines the scale below which structure is ignored. If K satisfies

$$\int K(u)du = 1, \quad \int uK(u)du = 0, \quad \int u^2 K(u)du < \infty \tag{3.21}$$

and μ is twice differentiable, then at each point x the convergence $\hat{\mu}(\tilde{x}) \to \mu(\tilde{x})$ for $\epsilon \to 0$ holds on average [114]. With a finite number of points, the convergence of course may not be seen since for small ϵ, statistical fluctuations become important.

[4] In binning, the points $\tilde{x}$ would form a regular lattice with step ϵ which they usually do not do for kernel methods.

If some parametric families are assumed for the distributions, it is sufficient to estimate these parameters and then calculate MI analytically or numerically from the known functional form of the density.

3.2.2 New Estimators

In this Section new estimators [52] for MI will be presented. We first review the derivation of a related differential Shannon entropy estimate [34, 46, 129] since the estimators for MI are obtained by very similar arguments.

Let X be a continuous random variable with values in some metric space, i.e., there is a distance function $||x - x'||$ defined between any two realizations of X, and let the density $\mu(x)$ exist as a proper function. Differential Shannon entropy is defined as

$$H(X) = -\int dx \mu(x) \log \mu(x) \ . \tag{3.22}$$

Our aim is to estimate $H(X)$ from a random sample $(x_1 \dots x_N)$ of N realizations of X.

The first step is to realize that Eq.(3.22) can be understood (up to the minus sign) as an average of $\log \mu(x)$. If we had unbiased estimators $\widehat{\log \mu(x)}$ of the latter, we would have an unbiased estimator

$$\hat{H}(X) = -N^{-1} \sum_{i=1}^{N} \widehat{\log \mu(x_i)} \ . \tag{3.23}$$

In order to obtain the estimate $\widehat{\log \mu(x_i)}$, we consider the probability distribution $P_k(\epsilon)$ for the distance between x_i and its k-th nearest neighbour. The probability $P_k(\epsilon)d\epsilon$ is equal to the probability that there is one point within the distance $r \in [\epsilon/2, \epsilon/2 + d\epsilon/2]$ from x_i, that there are $k-1$ other points at smaller distances, and that $N-k-1$ points have larger distances from x_k. Let us denote by p_i the mass of the ϵ-ball centered at x_i, $p_i(\epsilon) = \int_{||\xi - x_i|| < \epsilon/2} d\xi \mu(\xi)$. Using the trinomial formula we obtain

$$P_k(\epsilon)d\epsilon \ = \ \frac{(N-1)!}{1!(k-1)!(N-k-1)!} \times \frac{dp_i(\epsilon)}{d\epsilon} d\epsilon \ \times \ p_i^{k-1} \times (1-p_i)^{N-k-1} \tag{3.24}$$

or

$$P_k(\epsilon) = k \binom{N-1}{k} \frac{dp_i(\epsilon)}{d\epsilon} \ p_i^{k-1} (1-p_i)^{N-k-1} \ . \tag{3.25}$$

One easily checks that this is correctly normalized, $\int d\epsilon P_k(\epsilon) = 1$. Similarly, one can obtain the expectation value of $\log p_i(\epsilon)$

$$\begin{aligned} \mathsf{E}(\log p_i) = \int_0^\infty d\epsilon \ P_k(\epsilon) \log p_i(\epsilon) \ &= \ k \binom{N-1}{k} \int_0^1 dp \ p^{k-1} (1-p)^{N-k-1} \log p \\ &= \ \psi(k) - \psi(N) \ . \end{aligned} \tag{3.26}$$

Here, $\psi(x)$ is the digamma function, $\psi(x) = \Gamma(x)^{-1}d\Gamma(x)/dx$. It satisfies the recursion $\psi(x+1) = \psi(x)+1/x$ and $\psi(1) = -C$ where $C = 0.5772156\ldots$ is the Euler-Mascheroni constant. For large x, $\psi(x) \approx \log x - 1/2x$. The expectation value is taken here over the positions of all other $N-1$ points, with x_i kept fixed. For any ϵ, an estimator for $\log \mu(x)$ is obtained by assuming that $\mu(x)$ is constant within the entire ϵ-ball. The latter gives

$$p_i(\epsilon) \approx c_d \epsilon^d \mu(x_i) \ . \tag{3.27}$$

where d is the dimension of x, and c_d is the volume of the d-dimensional unit ball. For the maximum norm one simply has $c_d = 1$, while for the Euclidean norm $c_d = \pi^{d/2}/\Gamma(1 + d/2)/2^d$.

Using Eqs.(3.26) and (3.27) one obtains

$$\log \mu(x_i) \approx \psi(k) - \psi(N) - d \ \mathsf{E}(\log \epsilon(i)) - \log c_d \ , \tag{3.28}$$

which finally leads to Kozachenko-Leonenko estimator for differential Shannon entropy [46]

$$\hat{H}(X) = -\psi(k) + \psi(N) + \log c_d + \frac{d}{N}\sum_{i=1}^{N} \log \epsilon(i) \tag{3.29}$$

where $\epsilon(i)$ is twice the distance from x_i to its k-th neighbour.

From this derivation it is obvious that Eq.(3.29) would be unbiased, if the density $\mu(x)$ were strictly constant. The only approximation is in Eq.(3.27). For points on a torus (e.g., when x is a phase) with a strictly positive density one can easily estimate the leading corrections to Eq.(3.27) for large N. One finds that they are $O(1/N^2)$ and that they scale, for large k and N, as $\sim (k/N)^2$. In most other cases (including, e.g., Gaussians and uniform densities in bounded domains with a sharp cut-off) numerical simulations suggest that the error is $\sim k/N$ or $\sim k/N \log(N/k)$.

Mutual information could be obtained by estimating in this way $H(X)$, $H(Y)$ and $H(X, Y)$ separately and using [21]

$$I(X, Y) = H(X) + H(Y) - H(X, Y). \tag{3.30}$$

But this would mean that the errors made in the individual estimates would presumably not cancel, and therefore we proceed differently.

Indeed we will present two slightly different algorithms, both based on the above idea. Both use for the space $Z = (X, Y)$ the maximum norm,

$$||z - z'|| = \max\{||x - x'||, ||y - y'||\}, \tag{3.31}$$

while any norms can be used for $||x-x'||$ and $||y-y'||$ (they need not be the same, as these spaces can be completely different). Let us denote by $\epsilon(i)/2$ the distance from z_i to its k-th

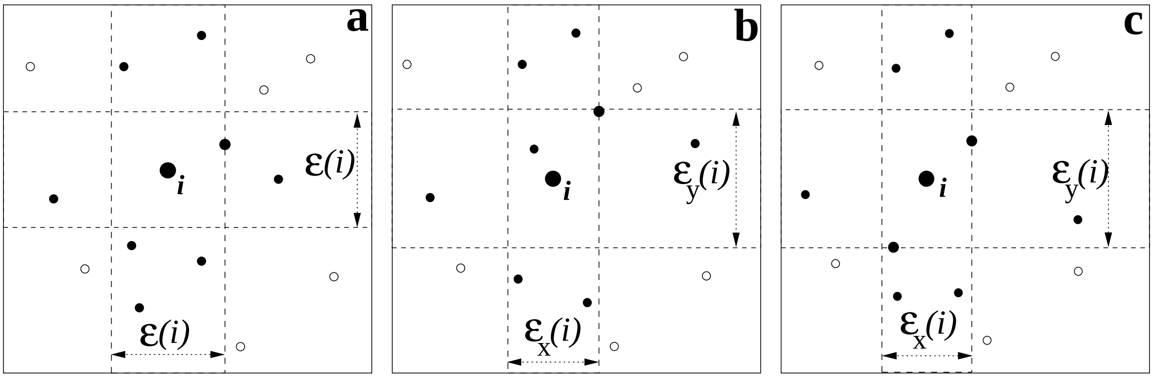

Figure 3.1: Panel (a): Determination of $\epsilon(i)$, $n_x(i)$ and $n_y(i)$ in the first algorithm, for $k = 1$ and some fixed i. In this example, $n_x(i) = 5$ and $n_y(i) = 3$.
Panels (b),(c): Determination of $\epsilon_x(i)$, $\epsilon_y(i)$, $n_x(i)$ and $n_y(i)$ in the second algorithm, for $k = 2$. Panel (b) shows a case where $\epsilon_x(i)$ and $\epsilon_y(i)$ are determined by the same point, while panel (c) shows a case where they are determined by different points. In this examples, $n_x(i) = 6$ and $n_y(i) = 4$.

neighbor, and by $\epsilon_x(i)/2$ and $\epsilon_y(i)/2$ the distances between the same points projected into the X and Y subspaces. Obviously, $\epsilon(i) = \max\{\epsilon_x(i), \epsilon_y(i)\}$.

In the first algorithm, we count the number $n_x(i)$ of points x_j whose distance from x_i is strictly less than $\epsilon(i)/2$, and similarly for y instead of x. This is illustrated in Fig. 1a. Notice that $\epsilon(i)$ is a random (fluctuating) variable, and therefore also $n_x(i)$ and $n_y(i)$ fluctuate.

Alternatively, in the second algorithm, we replace $n_x(i)$ and $n_y(i)$ by the number of points with $||x_i - x_j|| \leq \epsilon_x(i)/2$ and $||y_i - y_j|| \leq \epsilon_y(i)/2$ (see Figs. 3.1b and 3.1c).

For both algorithms, we will use the estimator Eq.(3.29) for $H(X,Y)$. Replacing d by $d_X + d_Y$ and c_d by $c_{d_X} c_{d_Y}$ we obtain

$$\hat{H}(X,Y) = \psi(k) - \psi(N) - \log(c_{d_X} c_{d_Y}) - \frac{d_X + d_Y}{N} \sum_{i=1}^{N} \log \epsilon(i) \ . \tag{3.32}$$

In order to obtain $I(X,Y)$ we have to subtract this from the estimates for $H(X)$ and $H(Y)$. For the latter we could use Eq.(3.29) directly with the same k. But as we said above, this would mean that we would effectively use different distance scales in the joint and marginal spaces. For any fixed k, the distance to the k-th neighbour in the joint space will be larger than the distances to the neighbours in the marginal spaces. Since the bias in Eq.(3.29) from the non-uniformity of the density depends of course on these distances, the biases in $\hat{H}(X)$, $\hat{H}(Y)$, and in $\hat{H}(X,Y)$ would not cancel.

To avoid this, we notice that Eq.(3.29) holds for *any* value of k, and that we do not have to choose a fixed k when estimating the marginal entropies. Assume, as in Fig. 3.1a, that the k-th neighbour of x_i is on the one of the vertical sides of the square of size $\epsilon(i)$. In this case, if there are altogether $n_x(i)$ points within the vertical lines $x = x_i \pm \epsilon(i)/2$, then $\epsilon(i)/2$ is the distance to the $(n_x(i)+1)-$st neighbour of x_i, and

$$\hat{H}(X) = \frac{1}{N}\sum_{i=1}^{N}\psi(n_x(i)+1) - \psi(N) - \log c_{d_X} - \frac{d_X}{N}\sum_{i=1}^{N}\log\epsilon(i)\ . \tag{3.33}$$

For the other direction (the y direction in Fig. 3.1a) this is not exactly true, i.e., $\epsilon(i)$ is not exactly equal to twice the distance to the $(n_y(i)+1)-$st neighbour, if $n_y(i)$ is analogously defined as the number of points with $||y_j - y_i|| < \epsilon(i)/2$. Nevertheless we can consider Eq.(3.33) also as a good approximation for $H(Y)$, if we replace everywhere X by Y in its right hand side (this approximation becomes exact when $n_y(i) \to \infty$, and thus also when $N \to \infty$). If we do this, subtracting $\hat{H}(X,Y)$ from $\hat{H}(X) + \hat{H}(Y)$ leads directly to

$$I^{(1)}(X,Y) = \psi(k) - \langle\psi(n_x+1) + \psi(n_y+1)\rangle + \psi(N). \tag{3.34}$$

These arguments can easily be extended to m random variables and lead to

$$\begin{aligned} I^{(1)}(X_1, X_2, \ldots, X_m) &= \psi(k) + (m-1)\psi(N) \\ &- \langle\psi(n_{x_1}) + \psi(n_{x_2}) + \ldots + \psi(n_{x_m})\rangle. \end{aligned} \tag{3.35}$$

The main drawback of $I^{(1)}$ is that the Kozachenko-Leonenko estimator is used correctly in only one marginal direction. This seems unavoidable if one wants to stick to isotopic "balls", i.e., to (hyper)cubes in the joint space. In order to avoid it we have to switch to (hyper)rectangles. Let us first discuss the case of two marginal variables X and Y, and generalize later to m variables $X_1, \ldots, X_m$. As illustrated in Figs. 3.1b and 3.1c, there are two cases to be distinguished (all other cases, where more points fall onto the boundaries $x_i \pm \epsilon_x(i)/2$ and $y_i \pm \epsilon_y(i)/2$, have zero probability; see however the third paragraph of Sec. 3.2.3): Either the two sides $\epsilon_x(i)$ and $\epsilon_y(i)$ are determined by the same point (Fig. 3.1b), or by different points (Fig. 3.1c). In either case we have to replace $P_k(\epsilon)$ by a 2-dimensional density,

$$P_k(\epsilon_x, \epsilon_y) = P_k^{(b)}(\epsilon_x, \epsilon_y) + P_k^{(c)}(\epsilon_x, \epsilon_y) \tag{3.36}$$

with

$$P_k^{(b)}(\epsilon_x, \epsilon_y) = \binom{N-1}{k}\frac{d^2[q_i^k]}{d\epsilon_x d\epsilon_y}(1-p_i)^{N-k-1} \tag{3.37}$$

and

$$P_k^{(c)}(\epsilon_x, \epsilon_y) = (k-1)\binom{N-1}{k}\frac{d^2[q_i^k]}{d\epsilon_x d\epsilon_y}(1-p_i)^{N-k-1}\ . \tag{3.38}$$

Here, $q_i \equiv q_i(\epsilon_x, \epsilon_y)$ is the mass of the rectangle of size $\epsilon_x \times \epsilon_y$ centered at (x_i, y_i), and p_i is as before the mass of the square of size $\epsilon = \max\{\epsilon_x, \epsilon_y\}$. The latter is needed since by using the maximum norm we guarantee that there are no points in this square which are not inside the rectangle.

Again we verify straightforwardly that P_k is normalized, while we get instead of Eq.(3.26)

$$\begin{aligned} \mathsf{E}(\log q_i) &= \iint_0^\infty d\epsilon_x d\epsilon_y P_k(\epsilon_x, \epsilon_y) \log q_i(\epsilon_x, \epsilon_y) \\ &= \psi(k) - 1/k - \psi(N) \ . \end{aligned} \tag{3.39}$$

Denoting by $n_x(i)$ and $n_y(i)$ the number of points with distance less *or equal* to $\epsilon_x(i)/2$ resp. $\epsilon_y(i)/2$, we get

$$I^{(2)}(X,Y) = \psi(k) - 1/k - \langle \psi(n_x) + \psi(n_y) \rangle + \psi(N). \tag{3.40}$$

For the generalization to m variables we have to consider m-dimensional densities $P_k(\epsilon_{x_1}, \dots, \epsilon_{x_m})$. The number of distinct cases (analogous to the two cases shown in Figs. 3.1b and 3.1c) proliferates as m grows, but fortunately we do not have to consider all these cases explicitly. One sees easily that each of them contributes to P_k a term

$$\propto \frac{d^m[q_i^k]}{d\epsilon_{x_1} \dots d\epsilon_{x_m}} \ (1 - p_i)^{N-k-1} \tag{3.41}$$

The direct calculation of the proportionality factors would be extremely tedious (we did it for $m = 3$), but it can be avoided by simply demanding that the sum is correctly normalized. This gives

$$P_k(\epsilon_{x_1}, \dots, \epsilon_{x_m}) = k^{m-1} \binom{N-1}{k} \frac{d^m[q_i^k]}{d\epsilon_{x_1} \dots d\epsilon_{x_m}} \times \ (1 - p_i)^{N-k-1} \ . \tag{3.42}$$

Calculating again $\mathsf{E}(\log q_i) = \psi(k) - (m-1)/k - \psi(N)$ analytically and approximating the density by a constant inside the hyper-rectangle, we obtain finally

$$\begin{aligned} I^{(2)}(X_1, X_2, \dots, X_m) &= \psi(k) - (m-1)/k + (m-1)\psi(N) \\ &- \langle \psi(n_{x_1}) + \psi(n_{x_2}) + \dots + \psi(n_{x_m}) \rangle \ . \end{aligned} \tag{3.43}$$

Before leaving this Section, we should mention that we slightly cheated in deriving $I^{(2)}(X,Y)$ (and its generalization to $m > 2$). Assume that in a particular realization we have $\epsilon_x(i) < \epsilon_y(i)$, as in Fig. 3.1b,3.1c. In that case we know that there cannot be any point in the two rectangles $[x_i - \epsilon_y(i)/2, x_i - \epsilon_x(i)/2] \times [y_i - \epsilon_y(i)/2, y_i + \epsilon_y(i)/2]$ and $[x_i + \epsilon_x(i)/2, x_i + \epsilon_y(i)/2] \times [y_i - \epsilon_y(i)/2, y_i + \epsilon_y(i)/2]$ (see Fig. 3.2). While we have taken this correctly into account when estimating $H(X,Y)$ (where it was crucial), we have neglected it in $H(X)$ and $H(Y)$. There, the corrections are of order $O(1/n_x)$ and $O(1/n_y)$ and should vanish for $N \to \infty$. It could be that their net effect vanishes because they

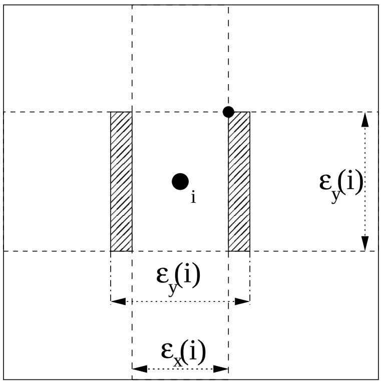

Figure 3.2: There cannot be any points inside the shaded rectangles. For the second method, this means that the estimates of the marginal entropy $H(X)$ ($H(Y)$) should be modified, since part of the area outside (inside) the stripe with ϵ_x (ϵ_y) is forbidden. This is neglected in Eq.(3.40).

contribute with opposite signs to $H(X)$ and $H(Y)$. But we have no proof for it. Anyhow, due to the approximation of constant density within each rectangle we cannot expect our estimates to be exact for finite N, and any justification ultimately relies on numerics.

In general, both formulas give very similar results. For the same k, Eq.(3.34) gives slightly smaller statistical errors (because $n_x(i)$ and $n_y(i)$ tend to be larger and have smaller relative fluctuations), but have larger systematic errors. The latter is only severe if we are interested in very high dimensions where $\epsilon(i)$ typically tends to be much larger than the marginal $\epsilon_{x_j}(i)$. In that case the second algorithm seems preferable. Otherwise, both can be used equally well.

3.2.3 Implementations Details

- Mutual information is invariant under reparametrization of the marginal variables. If $X' = F(X)$ and $Y' = G(Y)$ are homeomorphisms, then $I(X,Y) = I(X',Y')$ (see (3.17)). This can be used to rescale both variables first to unit variance. In addition, if the distributions are very skewed and/or rough, it might be a good idea to transform them such as to become more uniform (or at least single-humped and more or less symmetric). Although this is not required, strictly spoken, it will in general reduce errors.
- When implemented straightforwardly, the algorithm spends most of the CPU time for searching neighbours.
  - In the most naive version, we need two nested loops through all points which gives a CPU time of order $O(N^2)$. While this is acceptable for very small data sets (say $N \leq 300$), fast neighbour search algorithms are needed when dealing with larger sets.

  - Let us assume that X and Y are scalars. An algorithm with complexity $O(N\sqrt{k\ N})$ is then obtained by first ranking the x_i by magnitude (this can be done by any sorting algorithm such as quicksort), and co-ranking the y_i with them [90]. Nearest neighbours of (x_i, y_i) can then be obtained by searching x-neighbours on both sides of x_i and verifying that their distance in y direction is not too large. Neighbours in the marginal subspaces are found even easier by ranking both x_i and y_i. Most results in this Chapter were obtained using this method which is suitable for N up to a few thousands.

  - The fastest (but also the most complex) algorithm is obtained by using grids ('boxes') [36]. Indeed, we use three grids: A 2-dimensional one with box size $O(\sqrt{k/N})$ and two 1-dimensional ones with box sizes $O(1/N)$. First the k neighbours in 2-d space are searched using the 2-d grid, then the boxes at distances $\pm\epsilon$ from the central point are searched in the 1-d grids to find n_x and n_y. If the distributions are smooth, this leads to complexity $O(\sqrt{k}N)$. This last algorithm is comparable in speed to the algorithm of [23]. For all three versions of our algorithm it costs only little additional CPU time if one evaluates, along with $I(X,Y)$ for some $k > 1$, also the estimators for smaller numbers of neighbors.

- Empirical data usually are obtained with a resolution of a few (e.g., 12 or 16) binary digits, which means that many points in a large set may have identical coordinates. In that case the numbers $n_x(i)$ and $n_y(i)$ need no longer to be unique (the assumption of continuously distributed points is violated). If no precautions are taken, any code based on nearest neighbour counting is then bound to give wrong results. The simplest way out of this dilemma is by adding very low amplitude noise to the data ($\approx 10^{-10}$, say, when working with double precision) to break this degeneracy. We found this to give satisfactory results in all cases.

- Often, MI is estimated after *rank ordering* the data, i.e., after replacing the coordinate x_i by the rank of the i-th point when sorted by magnitude. This is equivalent to applying a monotonic transformation $x \to x',\ y \to y'$ to each coordinate which leads to a strictly uniform empirical density, $\mu'_x(x') = \mu'_y(x') = (1/N)\sum_{i=1}^{N} \delta(x' - i)$. For $N \to \infty$ and $k \gg 1$ this clearly leaves the MI estimate invariant. But it is not obvious that it leaves invariant also the estimates for finite k, since the transformation is not smooth at the smallest length scale. We found numerically that rank ordering gives correct estimates also for small k, if the distance degeneracies implied by it are broken by adding low amplitude noise as discussed above. In particular, both estimators still gave zero MI for independent pairs (see below). Although rank ordering can reduce statistical errors, we did not apply it in the following tests, and we did not study the properties of the resulting estimators in detail.

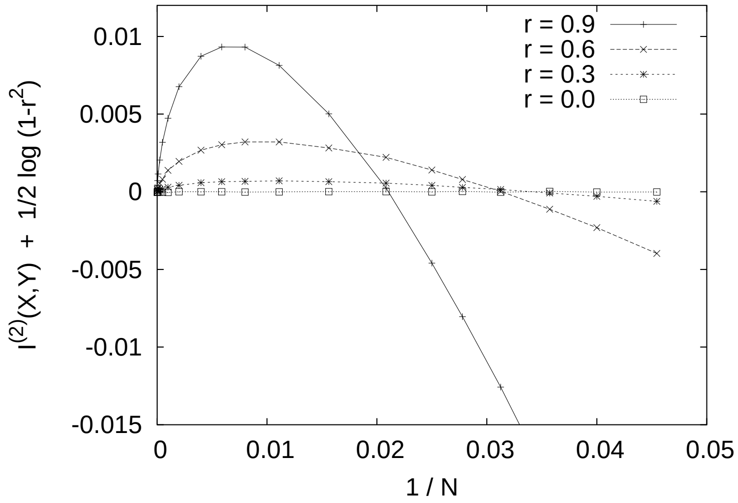

Figure 3.3: Estimates of $\langle I^{(2)}(X,Y)\rangle - I_{\text{exact}}(X,Y)$ for Gaussians with unit variance and covariances r, plotted against $1/N$. In all cases $k = 1$. The number of realizations is $> 2\times 10^6$ for $N <= 1000$, and decreases to $\approx 10^5$ for $N = 40000$. Error bars are smaller than the size of the symbols.

3.2.4 Results

3.2.4.1 Two-dimensional distributions

We shall first discuss applications of our estimators to correlated Gaussians, mainly because we can in this way most easily compare our results with analytic results and with previous numerical analyses. In all cases we shall deal with Gaussians of unit variance and zero mean. For m such Gaussians with covariance matrix σ_{ik} $i, k = 1 \dots m$, one has

$$I(X_1, \dots .X_m) = -\frac{1}{2}\log(\det(\sigma)) \; . \qquad (3.44)$$

For $m = 2$ and using the notation $r = \sigma_{XY}$, this gives

$$I_{\text{exact}}(X,Y) = -\frac{1}{2}\log(1-r^2) \; . \qquad (3.45)$$

In Fig. 3.3 we show the systematic errors $\langle I^{(2)}(X,Y)\rangle - I_{\text{exact}}(X,Y)$ for various values of r, obtained from a large number of realizations (typically $10^5 - 10^7$). We show only results for $k = 1$, plotted against $1/N$. Results for $k > 1$ are similar. To a first approximation $I^{(1)}(X,Y)$ and $I^{(2)}(X,Y)$ depend only on the ratio k/N.

The most conspicuous feature seen in Fig. 3.3, apart from the fact that indeed $\langle I^{(2)}(X,Y)\rangle - I_{\text{exact}}(X,Y) \to 0$ for $N \to \infty$, is that the systematic error is compatible with zero for $r = 0$, i.e., when the two Gaussians are uncorrelated. We checked this with high statistics runs for many different values of k and N (*a priori* one should expect that systematic errors become large for very small N), and for many more distributions (exponential, uniform, etc.). In all cases we found that both $I^{(1)}(X,Y)$ and $I^{(2)}(X,Y)$ become exact for independent variables. Moreover, the same seems to be true for higher order MI. We thus have the following conjecture.

Conjecture: Eqs.(3.34) and (3.40) are exact for independent X and Y, i.e., $I^{(1)}(X,Y) = I^{(2)}(X,Y) = 0$ if and only if $I(X,Y) = 0$.

We have no proof for this very surprising result. We have numerical indications that, moreover,

$$\frac{|I^{(1,2)}(X,Y) - I(X,Y)|}{I(X,Y)} \leq \mathrm{const} \tag{3.46}$$

as X and Y become more and more independent, but this is much less clean and therefore much less sure.

In Fig. 3.4 we compare values of $\langle I^{(1)}(X,Y)\rangle$ (left panel) with those for $\langle I^{(2)}(X,Y)\rangle$ (right panel) for different values of N and for $r = 0.9$. The horizontal axes show k/N (left) and $(k - 1/2)/N$ (right). Except for very small values of k and N, we observe scaling of the form

$$I^{(1)}(X,Y) \approx \Phi(\frac{k}{N}) \ , \quad I^{(2)}(X,Y) \approx \Phi(\frac{k-1/2}{N}) \ . \tag{3.47}$$

This is a general result and is found also for other distributions. The scaling of $I^{(1)}(X,Y)$ with k/N results simply from the fact that the number of neighbors within a fixed distance would scale $\propto N$, if there were no statistical fluctuations. For large k these fluctuations should become irrelevant, and thus the MI estimate should depend only on the ratio k/N. For $I^{(2)}(X,Y)$ this argument has to be slightly modified, since the smaller one of ϵ_x and ϵ_y is determined (for large k where the situation illustrated in Fig. 3.1c dominates over that in Fig. 3.1b) by $k - 1$ instead of k neighbors.

The fact that $I^{(2)}(X,Y)$ for a given value of k is between $I^{(1)}(X,Y)$ for $k - 1$ and $I^{(1)}(X,Y)$ for k is also seen from the variances of the estimates. In Fig. 3.5 we show the standard deviations, again for covariance $r = 0.9$. These statistical errors depend only weakly on r. For $r = 0$ they are approximately 10% smaller. As seen from Fig. 3.5, the errors of $I^{(2)}(X,Y;k)$ are roughly half-way between those of $I^{(1)}(X,Y;k-1)$ and $I^{(1)}(X,Y;k)$. They scale roughly as $\sim \sqrt{N}$, except for very large k/N. Their dependence on k does not follow a simple scaling law. The fact that statistical errors increase when k decreases is intuitively obvious, since then the width of the distribution of ϵ increases too. Qualitatively the same dependence of the errors was observed also for different distributions. For practical applications, it means that one should use $k > 1$ in order to reduce statistical errors, but too large values of k should be avoided since then the increase of systematic errors outweighs the decrease of statistical ones. We propose to use typically $k = 2$ to 4, except when testing for independence. In the latter case we do not have to worry about systematic errors, and statistical errors are minimized by taking k to be very large (up to $k \approx N/2$, say).

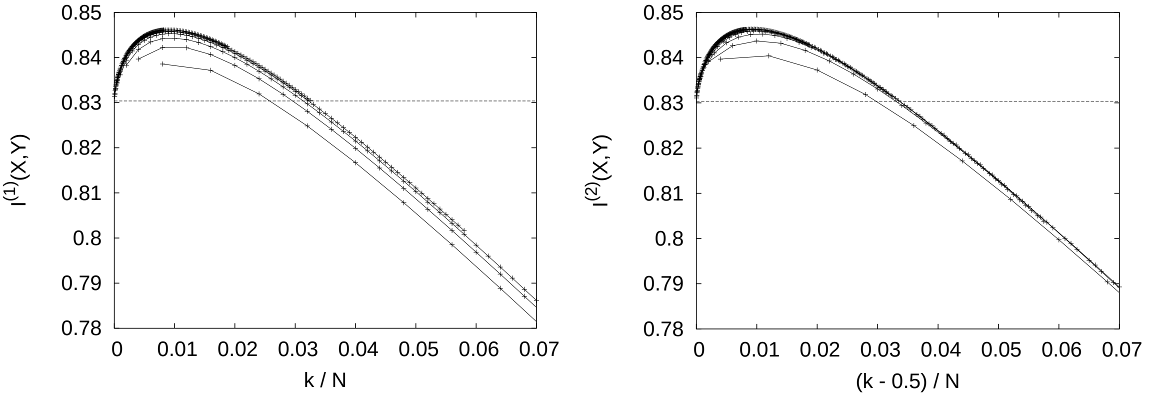

Figure 3.4: Mutual information estimates $I^{(1)}(X, Y)$ (left panel) and $I^{(2)}(X, Y)$ (right panel) for Gaussian deviates with unit variance and covariance $r = 0.9$, plotted against k/N (left panel) resp. $(k - 1/2)/N$ (right panel). Each curve corresponds to a fixed value of N, with $N = 125, 250, 500, 1000, 2000, 4000, 10000$ and 20000, from bottom to top. Error bars are smaller than the size of the symbols. The dashed line indicates the exact value $I(X, Y) = 0.830366$.

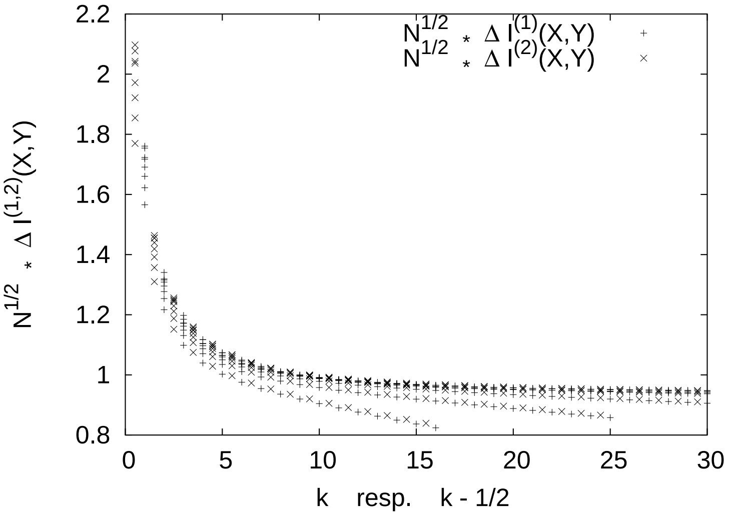

Figure 3.5: Standard deviations of the estimates $I^{(1)}(X, Y)$ (+) and $I^{(2)}(X, Y)$ (×) for Gaussian deviates with unit variance and covariance $r = 0.9$, multiplied by $\sqrt{N}$ and plotted against k ($I^{(1)}(X, Y)$) resp. $k - 1/2$ ($I^{(2)}(X, Y)$). Each curve corresponds to a fixed value of N, with $N = 125, 250, 500, 1000, 2000, 4000, 10000$ and 20000, from bottom to top.

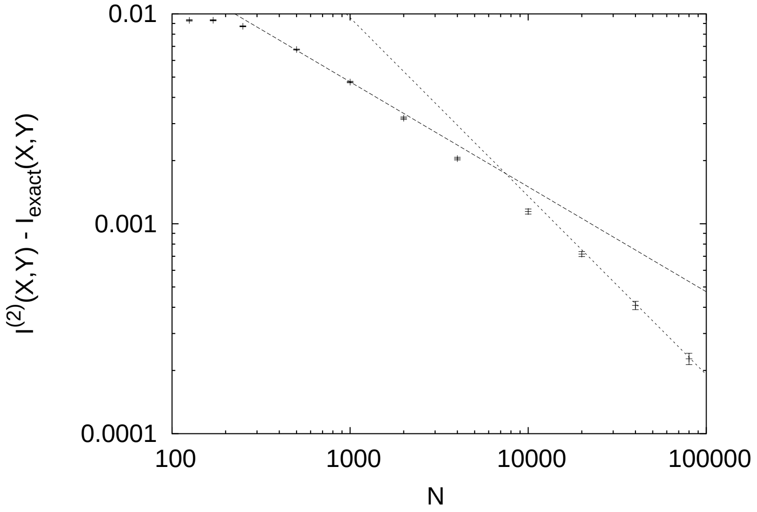

Figure 3.6: Systematic error $\langle I^{(2)}(X,Y)\rangle - I_{\mathrm{exact}}(X,Y)$ for $k = 3$ plotted against N on a log-log scale, for $r = 0.9$. The dashed lines are $\propto N^{-0.5}$ and $\propto N^{-0.85}$.

The above shows that $I^{(1)}(X,Y)$ and $I^{(2)}(X,Y)$ behave very similarly. Also CPU time needed to estimate them is nearly the same. In the following, we shall only show data for one of them, understanding that everything holds also for the other, unless the opposite is said explicitly.

For $N \to \infty$, the systematic errors tend to zero, as they should. From Figs. 3.3 and 3.4 one might conjecture that $\langle I^{(1,2)}(X,Y)\rangle - I_{\mathrm{exact}}(X,Y) \sim N^{-1/2}$, but this is not true. Plotting this difference on a double logarithmic scale (Fig. 3.6), we see a scaling $\sim N^{-1/2}$ for $N \approx 10^3$, but faster convergence for larger N. It can be fitted by a scaling $\sim 1/N^{0.85}$ for the largest values of N reached by our simulations, but the true asymptotic behaviour is presumably just $\sim 1/N$.

In Fig. 3.7 we show how the *relative* systematic errors behave for Gaussians when $r \to 0$. More precisely, we show $I^{(1,2)}(X,Y)/I^{(1,2)}_{\mathrm{exact}}(X,Y)$ for $k = 1$, plotted against N for four different values of r. Obviously these data converge, when $r \to 0$, to a finite function of N. We have observed the same also for other distributions, which leads to a conjecture stronger than the conjecture Eq.(3.46). Assume that we have a one-parameter family of 2-d distributions with densities $\mu(x,y;r)$, with r being a real-valued parameter. Assume also that μ factorizes for $r = r_0$, and that it depends smoothly on r in the vicinity of r_0, with $\partial\mu(x,y;r)/\partial r$ finite. Then we propose that for many distributions (although not for all!)

$$I^{(1,2)}(X,Y)/I_{\mathrm{exact}}(X,Y) \to F(k,N) \qquad (3.48)$$

for $r \to r_0$, with some function $F(k,N)$ which is close to 1 for all k and all $N \gg 1$, and which converges to 1 for $N \to \infty$. We have not found a general criterion for which families of distributions we should expect Eq.(3.48).

Examples of several other two dimensional distributions can be found in Ref.[52].

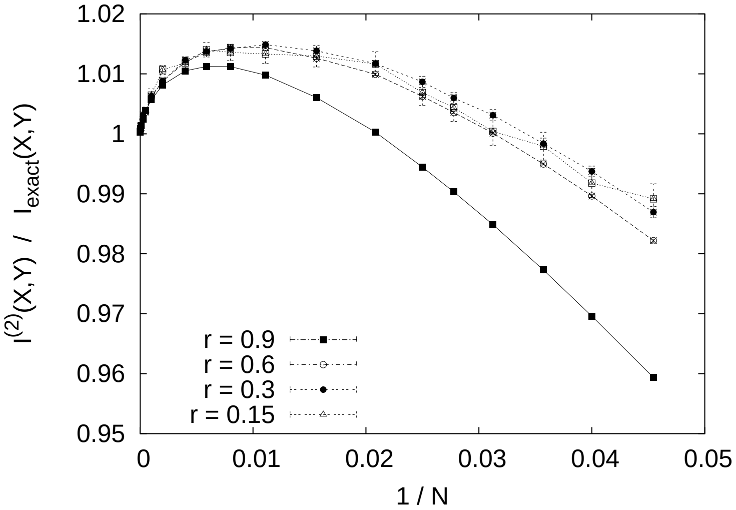

Figure 3.7: Ratios $I^{(2)}(X,Y)/I_{\text{exact}}(X,Y)$ for $k = 1$ plotted against $1/N$, for Gaussians with unit variance and covariance r.

3.2.4.2 High dimensional distributions

In higher dimensions we shall only discuss applications of our estimators to *m* correlated Gaussians, because as in the case of two dimensions this is easily compared to analytically derived values (Eq.(3.44)) and to previous numerical results [22]. As already shown above for 2-d distributions (Fig. 3.7) our estimates seem to be exact for independent random variables. We choose the same one-parameter family of 3-d Gaussian distributions with all the correlation coefficients equal to r as in Ref. [22]. In Fig. 3.8 we show the behavior of the *relative* systematic errors of both proposed estimators. One can easily see that the data converge for $r \to 0$, i.e., when all three Gaussians become independent. This supports the conjecture made in the previous subsection. In addition, in Fig. 3.8 one can see the difference between the estimators $I^{(1)}$ and $I^{(2)}$. For intermediate numbers of the points, $N \sim 100 - 200$, the $I^{(1)}$ estimator has lower systematic error. Apart from that, $I^{(2)}$ evaluated for N is roughly equal to $I^{(1)}$ evaluated for $2N$, reflecting the fact that $I^{(2)}$ effectively uses smaller length scales as discussed already for $d = 2$.

To compare our results in high dimension with the ones presented in Ref. [22] we shall calculate not the MI of all variables $I(X_1, X_2, ..., X_m)$ but the MI between two variables, namely an $(m-1)$ dimensional vector X^{m-1} and a scalar X_m, i.e, $I(X^{m-1}, X_m)$. For estimation of this MI we can use the same formulas as for the 2-d case (Eq.(3.34) and Eq.(3.40), respectively) where n_x is defined as the number of points in the $(m-1)$ dimensional stripe of (hyper)cubic cross section.

In Fig. 3.9 we show the average values of $I^{(1,2)}$. They are in very good agreement with the theoretical ones for all three values of the correlation coefficient r and all dimensions tested here (in contrast, in Ref. [22] the estimators of MI significantly deviate from the theoretical

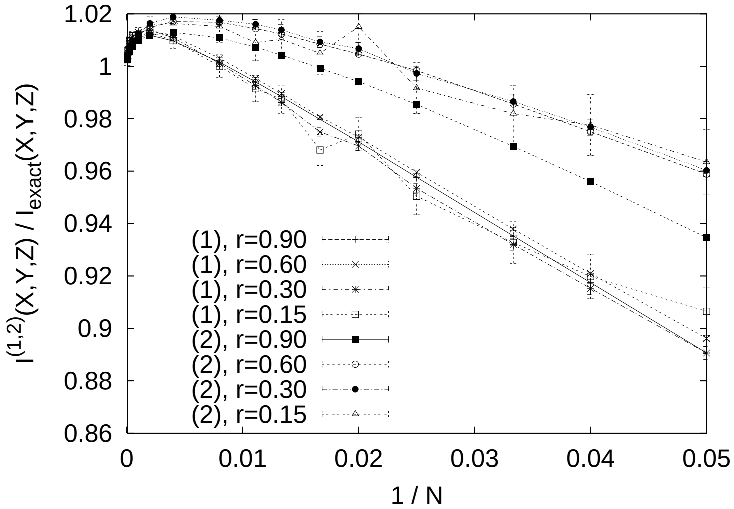

Figure 3.8: Ratios $I^{(1,2)}(X,Y,Z)/I_{\text{exact}}(X,Y,Z)$ for $k = 1$ plotted against $1/N$. All Gaussians have unit variance and all non-diagonal elements in the correlation matrix $\sigma_{i,k}, i \neq k$ (correlation coefficients) take the value r.

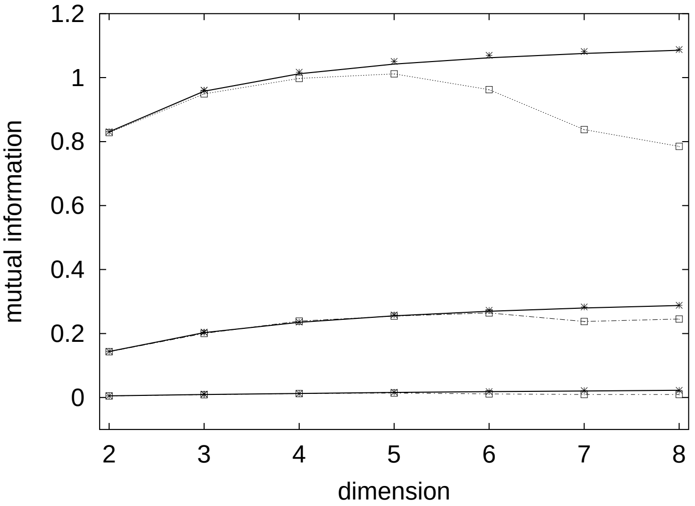

Figure 3.9: Averages of $I^{(1,2)}(X^{m-1}, X_m))$ for $k = 1$ plotted against d, for 3 different values of $r = 0.1, 0.5, 0.9$. The sample size is 50000, averaging is done over 100 realizations (same parameters as in Ref. [22], Fig. 1). Full lines indicate theoretical values, pluses $(+)$ are for $I^{(1)}$, crosses $(\times)$ for $I^{(2)}$. Squares and dotted lines are read off from Fig. 1 of Ref. [22].

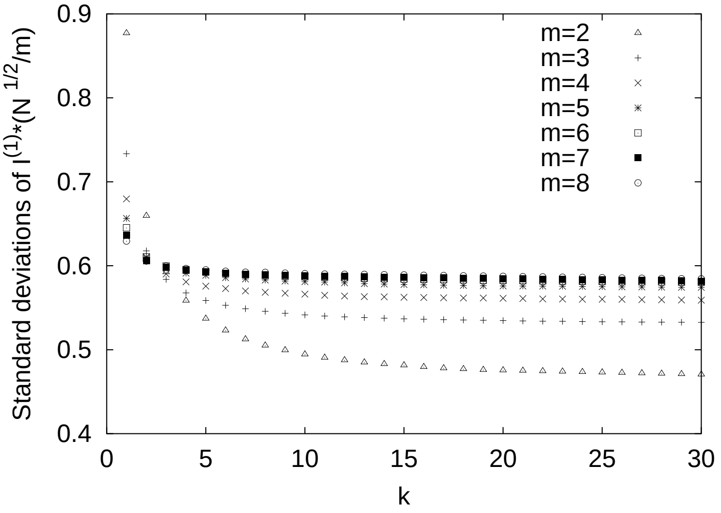

Figure 3.10: Standard deviations of the estimate $I^{(1)}$ for Gaussian deviates with unit variance and covariance $r = 0.9$, multiplied by $\sqrt{N}/m$ and plotted against k. Each curve corresponds to a fixed value of dimension m. Number of samples is $N = 10000$.

values for dimensions ≥ 6). It is impossible to distinguish (on this scale) between estimates $I^{(1)}$ and $I^{(2)}$.

In Fig. 3.10, statistical errors of our estimate are presented as a function of the number of neighbours k. More precisely, we plotted the standard deviation of $I^{(1)}$ multiplied by $\sqrt{N}/m$ against k for the case where all correlation coefficients are $r = 0.9$. Each curve corresponds to a different dimension m. The data scale roughly as $\sim m/\sqrt{N}$ for large dimensions. Moreover, these statistical errors seem to converge to finite values for $k \to \infty$. This convergence becomes faster for increasing dimensions. The same behavior is observed for $I^{(2)}$.

3.3 Transfer Entropy

Once the interdependence between two systems has been established, one usually is faced with the next question, namely what is the predominant direction of interaction, or in other words who is a driver and who is a response.

Mutual information cannot directly be applied for this purpose because it is symmetric. In order to analyze driver-response relationships, time-delayed mutual information was proposed [128]. Later it was shown that quantities based on transition probabilities allow better detection of information transfer in a system [44, 111]. An appropriate relative entropy was introduced in Ref. [111], called transfer entropy (TE). In this section TE will be described.

Assume that X_i and Y_j are random iid variables[5]. Let us define random variables consisting of words of length k, as

$$X_i^{(k)} = (X_i, \dots, X_{i-k+1}), \qquad Y_j^{(k)} = (Y_j, \dots, Y_{j-k+1}). \tag{3.49}$$

If X_i and Y_j are discrete random variables one can define along with the simple probabilities $p_i(x) = \mathrm{prob}(X_i = x)$ and $p_j(y) = \mathrm{prob}(Y_j = y)$ the transition probabilities

$$\begin{aligned} p_{i+1}(x_{i+1}|x_i^{(k)}, y_j^{(l)}) &= \mathrm{prob}(X_{i+1}|X_i^{(k)} = x_i^{(k)}, Y_j^{(l)} = y_j^{(l)}), && (3.50) \\ p_{j+1}(y_{j+1}|y_j^{(l)}, x_i^{(k)}) &= \mathrm{prob}(Y_{j+1}|Y_j^{(l)} = y_j^{(l)}, X_i^{(k)} = x_i^{(k)}), && (3.51) \end{aligned}$$

where $x_i^{(k)} = (x_i, \dots, x_{i-k+1})$ is the state of $X_i^{(k)}$ and $y_i^{(j)} = (y_j, \dots, y_{j-l+1})$ is the state of $Y_j^{(l)}$. The transition probability denotes the probability of finding X_{i+1} in state x_{i+1} when $X_i^{(k)}$ is in state $x_i^{(k)}$ and $Y_j^{(l)}$ is in $y_j^{(l)}$, and similarly for Y_{j+1}.

TE is closely related to conditional entropy. Suppose the future state x_{i+1} of X_i depends on the k past states $x_i^{(k)}$, but not on the l past states $y_j^{(l)}$, then the generalized Markov property $p(x_{i+1}|x_i^{(k)}, y_j^{(l)}) = p(x_{i+1}|x_i^{(k)})$ holds. If there is any such a dependency, it can be quantified by the TE, which is obtained by comparing two conditional entropies

$$\begin{aligned} T(X_{i+1}|X_i^{(k)}, Y_j^{(l)}) &= H(X_{i+1}|X_i^{(k)}) - H(X_{i+1}|X_i^{(k)}, Y_j^{(l)}) \\ &= \sum_{x,y} p(x_{i+1}, x_i^{(k)}, y_j^{(l)}) \log \frac{p(x_{i+1}|x_i^{(k)}, y_j^{(l)})}{p(x_{i+1}|x_i^{(k)})}. && (3.52) \end{aligned}$$

In words, the information flow T from system Y to X is the information about the future of X_i retrieved from both $X_i^{(k)}$ and $Y_j^{(j)}$ minus the information about it retrieved only from $X_i^{(k)}$. From the construction of the above formula one can easily see that TE is asymmetric under exchange of X and Y. Information flow from system X to system Y will be defined as

$$T(Y_{j+1}|Y_j^{(l)}, X_i^{(k)}) = \sum_{x,y} p(y_{j+1}, y_j^{(l)}, x_i^{(k)}) \log \frac{p(y_{j+1}|x_j^{(l)}, y_i^{(k)})}{p(y_{j+1}|y_j^{(l)})}. \tag{3.53}$$

The transfer entropy can be expressed as a sum of mutual informations

$$\begin{aligned} T(X_{i+1}|X_i^{(k)}, Y_j^{(l)}) &= I(\left(X_{i+1}, X_i^{(k)}\right), Y_j^{(l)}) - I(X_i^{(k)}, Y_j^{(l)}), && (3.54) \\ &\text{or} \\ T(X_{i+1}|X_i^{(k)}, Y_j^{(l)}) &= I(\left(X_i^{(k)}, Y_j^{(l)}\right), X_{i+1}) - I(X_i^{(k)}, X_{i+1}), && (3.55) \end{aligned}$$

where the notation $I((X, Y), Z)$ means MI between two variables (X, Y) and Z.

[5] In this Section the indices i and j are time indices, i.e., $i + 1$ is the next moment with respect to the i-th moment

Rewriting T as a sum of Shannon entropies one obtains alternatively

$$T(X_{i+1}|X_i^{(k)},Y_j^{(l)}) = H(X_i^{(k)},Y_j^{(l)}) - H(X_{i+1}^{(k+1)},Y_j^{(l)}) + H(X_{i+1}^{(k+1)}) - H(X_i^{(k)}). \quad (3.56)$$

Consider now continuous random variables. Because TE is the difference of MIs, the binned TE converges under refinement (bin size $\Delta \to 0$) to

$$T(X_{i+1}|X_i^{(k)},Y_j^{(l)}) = \iiint \mu(x_{i+1},x_i^{(k)},y_j^{(l)}) \log \frac{\mu(x_{i+1}|x_i^{(k)},y_j^{(l)})}{\mu(x_{i+1}|x_i^{(k)})} dx_{i+1} dx_i^{(k)} dy_i^{(l)}, \quad (3.57)$$

where μ are the densities in corresponding spaces (the same holds for $T(Y_{j+1}|Y_j^{(l)})$).

It shares with MI the property of invariancy under homeomorphisms $x \to \phi(x)$ and $y \to \psi(y)$.

Theoretically all of the formulas (Eqs.(3.54), (3.55), (3.56)) can be used equally, but they will give different errors if one uses them for numerical estimation. To check this we choose a system of equations considered in Ref. [44]. The system consists of two correlated Gaussians processes, namely

$$\begin{cases} X_{i+1} &= \alpha X_i + \eta_i^X, \\ Y_{i+1} &= \beta Y_i + \gamma X_i + \eta_i^Y, \end{cases} \quad (3.58)$$

where η^X and η^Y are uncorrelated normal random numbers. For this system transfer entropies (with $k = l = 1$) can be calculated exactly. For any parameters α, β, γ there is no information transfer from system Y to system X, i.e., $T(X_{i+1}|X_i, Y_i) \equiv 0$, while $T(Y_{(i+1)}|Y_i, X_i) > 0$. We choose one set of parameters $\alpha = 0.5, \beta = 0.6, \gamma = 0.4$ and calculate $T(Y_{(i+1)}|Y_i, X_i)$. The theoretical value is ≈ 0.092280.

In Fig. 3.11a relative systematic errors for the transfer entropy from system X to system Y are shown. More precisely, we plot $\langle T(Y_{(i+1)}|Y_i, X_i)\rangle / \, T_{\text{exact}}$. For the calculation of the average transfer entropy both formulas Eq.(3.54) and Eq.(3.55) (with $j = i$ and exchanging X and Y) were used[6]. The mutual informations appearing in these equations were calculated using our new estimator Eq.(3.40) for the first term and Eq.(3.43) for the second one (k was equal to one in both cases). The slight difference between the results of different equations is more pronounced for long realizations (more then 500 points). It occurs because the exact values of mutual informations appearing in Eq.(3.55) ($I((Y_i, X_i), Y_{i+1}) \approx 0.3900$ and $I(Y_i, Y_{i+1}) \approx 0.2977$) are higher than the ones in Eq.(3.54) ($I((Y_{i+1}, Y_i), X_i) \approx 0.1179$ and $I(Y_i, X_i) \approx 0.0256$) and therefore, also the statistical errors for the same length of the realization in Eq.(3.55) are higher.

In Fig. 3.11b the absolute value of the transfer entropy from system Y to X is shown. By construction this value should be equal to zero for Eq.(3.58). The estimated values are very

[6] The Eq.(3.56) is not shown in the Figure because of huge systematical errors

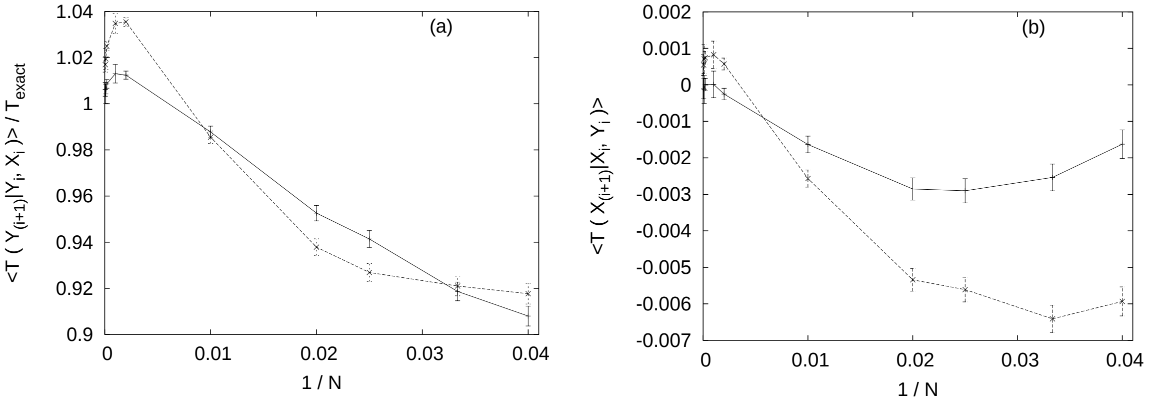

Figure 3.11: Solid lines correspond to Eq.(3.54), dashed lines correspond to Eq.(3.55) (a) relative systematic errors for the average transfer entropy from X to Y. (b) Average transfer entropy from Y to X. In both cases averaging was done over 2×10^6 for $N < 500$, over 10^5 for $N < 10000$, and over 10^4 for other values of N.

close to zero, but not as close as in the case of mutual information, where we did not have any visible bias for the completely independent processes. Unfortunately, this property is not extended to the transfer entropy. All the mutual informations entering the equations for $T(X_{(i+1)}|X_i, Y_i)$ are non-zero, so they just cancel each other resulting in zero transfer entropy. Again the values of the transfer entropy obtained using Eq.(3.54) are more precise. It can be explained with the same arguments as before.

These results have preliminary character and can be extended, e.g., the comparison with other methods for the estimation of the transfer entropy (cf. Ref. [44]) can be done.

3.4 Phase synchronization

In this Section measures of phase synchronization will be discussed. As the name suggest phases play the main role in this type of synchronization phenomenon and the extraction of the phases is a crucial point for the detection of phase synchronization.

3.4.1 Extracting phase via the Hilbert transform

Suppose one has a continuous real signal $u(t)$. Assume it is decomposed into sinusoidal oscillations and a slowly varying amplitude

$$u(t) = a(t) \cos \varphi(t). \tag{3.59}$$

However, it is not straightforward how to separate the known function $u(t)$ into the factors $a(t)$ and $\cos \varphi(t)$. It becomes easier if we assume that $u(t)$ is the real part of an *analytic* *function* of the complex variable $z = t + i\tau$,

$$w(t) = u(t) + iv(t) = a(t)\, e^{i\,\varphi(t)}, \tag{3.60}$$

$w(t)$ is called *analytic signal* [7] [32, 82, 123].

In the analytical signal concept one chooses as the imaginary part $v(t)$ the *Hilbert transform* (HT) of $u(t)$:

$$v(t) \equiv (\mathcal{H}u)(t) = \frac{1}{\pi} p.v. \int_{-\infty}^{+\infty} \frac{u(t')}{t - t'}\, dt' \tag{3.63}$$

(here $p.v.$ denotes the Cauchy principal value[8]). This choice is unambiguous. Let us denote by $\varphi_x^H(t)$ the *instantaneous phase* of $x(t)$ defined via the HT. The phase is given by

$$\varphi_x^H(t) = \arctan \frac{(\mathcal{H}x)(t)}{x(t)}. \tag{3.65}$$

Already on the fact that the HT of a sine is a minus cosine, and the HT of a cosine is a sine, one can see that HT performs a phase shift of $-\pi/2$ $(\sin(x - \frac{\pi}{2}) = -\cos(x)$, $\cos(x - \frac{\pi}{2}) = \sin(x))$. Thus, the HT can be considered as an ideal filter with unitary amplitude response and phase response of $-\pi/2$. For non-harmonic signals it performs a $-\pi/2$ phase shift for every spectral component of $u(t)$.

Noting that HT is a convolution of the signal with $1/\pi t$ and using the convolution theorem one can write

$$\mathcal{H}x = \mathcal{F}^{-1}\left[\frac{1}{i}\, sgn(\omega) \cdot \mathcal{F}x\right], \tag{3.66}$$

where $\mathcal{F}x$ denotes the Fourier transform of x, $\mathcal{F}^{-1}$ its inverse, the function $sgn(\omega) = \frac{\omega}{|\omega|}$.

The easiest way to compute the HT is to perform fast FT (FFT) of the original time series, shift the phase of every frequency component by $-\pi/2$ and apply the inverse FFT.

[7] Consider a complex function $\psi(z)$ of a complex variable z. This function may be written as a complex function of two real variables:

$$\psi(z) = \psi(t,\tau) = u(t,\tau) + i\, v(t,\tau), \qquad (t,\tau) \in \mathbf{R}^2,\ z = t + i\tau. \tag{3.61}$$

The function $\psi(z) = u(t,\tau) + i\; v(t,\tau)$ is called the *analytic function* if

$$\frac{\partial u}{\partial t} = \frac{\partial v}{\partial \tau}; \quad \frac{\partial u}{\partial \tau} = -\frac{\partial v}{\partial t}. \tag{3.62}$$

These equations are called Cauchy-Riemann equations.

An *analytic signal* is defined as a complex function of the real variable t in the form $\psi(t) = u(t,0) + i\; v(t,0)$, therefore it represents the values of analytic function $\psi(z)$ taken along the real axis.

[8]

$$v(t) = \frac{1}{\pi} \lim_{\epsilon \to 0, A \to \infty} \left(\int_{-A}^{-\epsilon} + \int_{\epsilon}^{A} \frac{u(t')}{t - t'}\, dt' \right) \tag{3.64}$$

3.4.2 Extracting phase via the wavelet transform

Another method used to extract the phases from time series is based on the *wavelet transform* (WT) and has recently been introduced by Lachaux et al. [57, 56]. In their approach the phase was determined by the convolution of the signal with the complex Morlet wavelet:

$$\Psi^{(\mathrm{o})}(t) = e^{i\omega_0 t} \cdot e^{-t^2/2\sigma^2}. \tag{3.67}$$

However, this commonly used “Morlet wavelet” is not a wavelet at all, because it does not fulfill the admissibility condition[9].

In our work we compared the method of Lachaux with the one based on the HT [48, 49, 94]. However, for the former we used a slightly different wavelet,

$$\Psi(t) = (e^{i\omega_0 t} - e^{-\omega_0^2\sigma^2/2}) \cdot e^{-t^2/2\sigma^2}. \tag{3.68}$$

Here ω_0 is the frequency of the wavelet, σ denotes the width of the peak in the spectrum of wavelet, and the term $e^{-\omega_0^2\sigma^2/2}$ is a correction term which was introduced in Ref. [38]. The FT of the corrected wavelet is $(\mathcal{F}\Psi)(\omega) = \sigma\sqrt{2\pi} \cdot \left(e^{-\sigma^2(\omega-\omega_0)^2/2} - e^{-\sigma^2(\omega^2+\omega_0^2)/2}\right)$, giving $(\mathcal{F}\Psi)(0) = 0$.

Instead of the parameter σ we used the number of significant oscillations n_c. In Fig. 3.12 the real part of the wavelet is plotted for two different values of n_c. An approximate oscillation is rated as “significant” if its amplitude is larger then 1% of the value at $t = 0$. The parameter σ can then be written as $\sigma = n_c/6\omega_0$. The smaller the parameter σ, the narrower is the peak in the spectrum of the wavelet, and the width of the spectrum for fixed σ does not dependent on the frequency ω_0. In our calculations we fix n_c and vary ω_0 (and hereby also σ), which allows us to have a wider spectrum for larger frequencies. This seems to be more reliable in real applications.

The convolution of $x(t)$ with $\Psi(t)$ yields a complex time series of wavelet coefficients $W_x(t)$

$$(\mathcal{W}_\Psi x)(t) = W_x(t) = (\Psi \circ x)(t) = \int \Psi(t')\, x(t-t')\, dt' = A_x^W(t) \cdot e^{i\varphi_x^W(t)}, \tag{3.69}$$

from which the phase can be defined as

$$\varphi_x^W(t) = \arctan \frac{\mathrm{Im}[W_x(t)]}{\mathrm{Re}[W_x(t)]}. \tag{3.70}$$

3.4.3 Comparison of phase extraction methods

In this Section we will discuss the differences and similarities between the two methods for phase extraction, based on HT and on WT, respectively.

[9] Indeed, as seen from its FT $(\mathcal{F}\Psi^{(\mathrm{o})})(\omega) = \sigma\sqrt{2\pi}\, e^{-\sigma^2(\omega-\omega_0)^2/2}$ it has a non-zero component at $\omega = 0$.

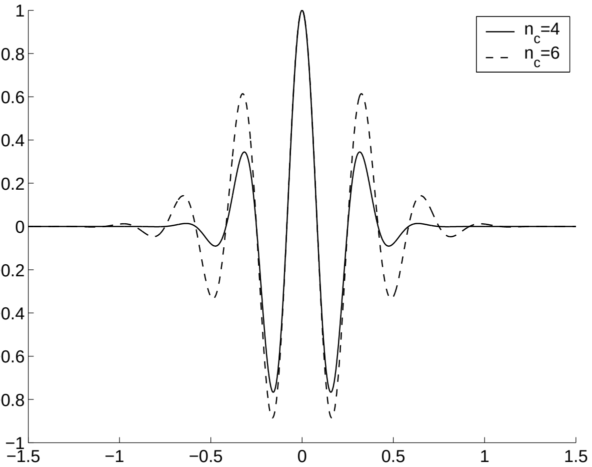

Figure 3.12: Real part of the corrected complex Morlet wavelet.

The phase defined using the corrected wavelet (Eq.(3.68)) is independent of the mean of the signal. This is not the case if one uses methods based on the HT or on the simple complex Morlet wavelet $\Psi^{(\mathrm{o})}$ without this correction (Eq.(3.67)). The phase defined using HT transform is very sensitive to the mean value of the signal. For stationary signals without slow components this problem can be solved by subtracting the mean value from the signal. The case of non-stationary signals or signals with slow components is, however, more problematic. One can try to filter out "uninteresting" parts of the signal but the choice of the appropriate filter is not always obvious. Therefore, for such signals as electroencephalograms which are intrinsically non-stationary, the use of corrected complex Morlet wavelet is more appropriate.

In principle, one can regard the WT as a filtering procedure. Let us first apply WT and then define the phase of the real part of the wavelet coefficients using the HT

$$\varphi^H_{\mathrm{Re}[W_x]}(t) = \arctan\frac{(\mathcal{H}\mathrm{Re}[W_x])(t)}{\mathrm{Re}[W_x(t)]}. \tag{3.71}$$

Numerically we found that this phase is very similar to the phase defined via WT, i.e., $\varphi^H_{\mathrm{Re}[W_x]}(t) \approx \varphi^W_x(t)$. This can be explained by approximate analyticity of the complex Morlet wavelet. A wavelet is called "analytic", if it is (i) the real part of an analytic function $\psi(z)$ of $z = t + i\tau$ taken along the real axis t and (ii) $\mathcal{H}\mathrm{Re}[\Psi] \equiv \mathrm{Im}[\Psi]$. For such an analytical wavelet, the phase of the signal via WT is given by

$$\varphi^W_x(t) = \arctan\frac{\mathrm{Im}[W_x(t)]}{\mathrm{Re}[W_x(t)]} = \arctan\frac{(\mathcal{W}_{\mathrm{Im}[\Psi]}x)(t)}{(\mathcal{W}_{\mathrm{Re}[\Psi]}x)(t)} = \arctan\frac{(\mathcal{W}_{(\mathcal{H}\mathrm{Re}[\Psi])}x)(t)}{(\mathcal{W}_{\mathrm{Re}[\Psi]}x)(t)}. \tag{3.72}$$

Changing the order of the wavelet and the Hilbert transform in the numerator, and noting that the real (imaginary) part of complex wavelet transform is equivalent to the wavelet transform with the real (imaginary) part of the wavelet, one gets Eq.(3.71). It means that

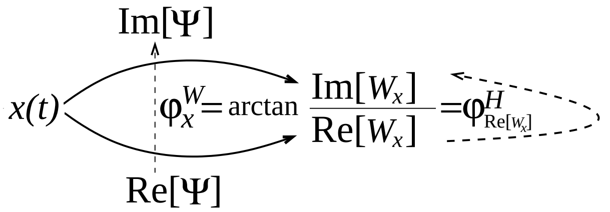

Figure 3.13: Solid lines denote the wavelet transform, dashed lines denote the Hilbert transform.

for the complex *analytic* wavelet $\varphi^H_{\mathrm{Re}[W_x]}(t) \equiv \varphi^W_x(t)$. All these formulas are illustrated in Fig. 3.13. Therefore, the method of defining the phase via WT is a combination of the filtering realized by the convolution with the real part of the wavelet and HT.

$$\text{Wavelet Transform} \equiv \text{Filtering} + \text{Hilbert Transform}$$

The fact that the corrected Morlet wavelet is very close to being analytic is shown in Fig. 3.14.

It is important to remark that the previous result is not limited to complex Morlet wavelet and can be extended to other wavelet functions. In particular, from a real wavelet function $\Psi(t)$ we can construct an analytic signal by using the Hilbert transform, i.e., $\Psi'(t) \equiv \Psi(t) + i\,(\mathcal{H}\Psi)(t)$. Then, from $W_x(t) = (\Psi' \circ x)(t)$ we can define a phase using Eq.(3.70). The important advantage is that we have the freedom of defining the phase from a particular wavelet function, chosen from available wavelets according to the signal to be studied. This can be interesting in cases where defining a phase via HT is troublesome or if conventional filters are not well suited.

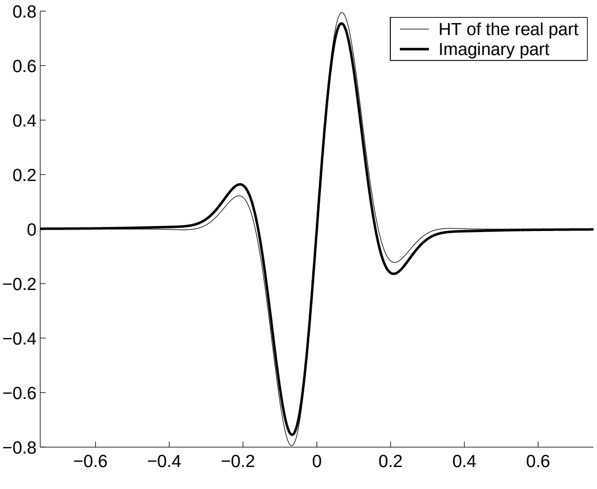

Figure 3.14: Imaginary part of a complex Morlet wavelet and Hilbert transform of the real part.

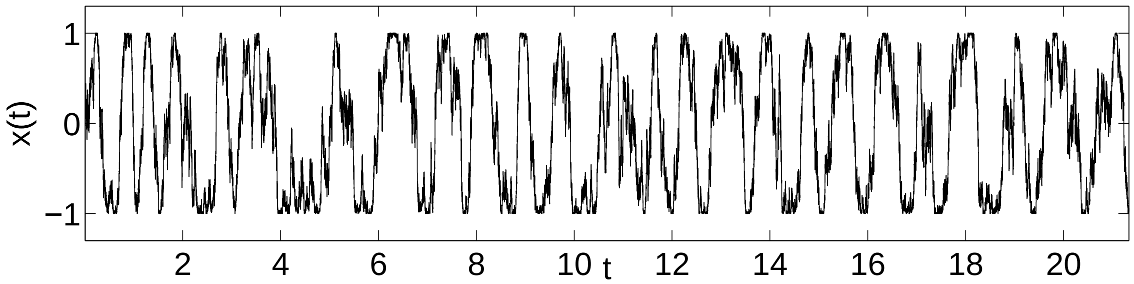

Figure 3.15: Part of the signal analyzed in Fig. 3.16.

3.4.4 Discussion. Phase extraction

In general, all the methods for defining a phase aim to find suitable variables in which the signal can be represented as a rotation around the origin. The phase can be identified with the angle of the vector drawn from the origin to the corresponding point on the trajectory. In the case of the HT method these variables are the signal itself and its Hilbert transform, for the method based on the WT they are the real and imaginary parts of the wavelet coefficients. If in an experiment more than one variable is available, one can try to use them directly or in some combination, e.g., the phase portrait of the Rössler system [104] plotted in x and y coordinates corresponds to rotations around the origin. For the Lorenz system [71] one can similarly plot $\sqrt{x^2+y^2}$ against z. This is not a complete list of methods available in the literature, look for example Refs. [98, 103].

One naturally arising question is whether all these approaches provide in general the same phase [47]. Unfortunately, they often do not. This is illustrated by the following example.

In Fig. 3.15 we show part of a signal. A phase portrait obtained by plotting $x(t)$ against $x(t+\tau)$ is shown in Fig. 3.16a, a similar phase portrait using the Hilbert transform in Fig. 3.16b, and the more smooth phase portrait using the wavelet transform in Fig. 3.16c. In neither case one clearly sees a point around which the orbit circles, thus neither allows a clear and robust definition of the phase. The spectrum, obtained with a Welch window, is shown in Fig. 3.17. A prominent peak is seen, but this peak is not sufficiently sharp and thus a unique angular frequency seems not obtainable. One can try several other methods popular in signal analysis, but we argue that none of them will lead to a robust determination of a phase.

And yet – there is a simple and clear-cut phase that enters in this example. The signal shown in Fig. 3.15 is generated by a random process defined as

$$x(t) = \cos(\phi(t)) \tag{3.73}$$

with the phase performing a biased random walk (cf. [103]),

$$d\phi(t)/dt = \omega + \eta(t) \tag{3.74}$$

where $\eta(t)$ is δ-correlated white noise,

$$\langle \eta(t) \rangle = 0, \qquad \langle \eta(t)\eta(t') \rangle = D\delta(t-t'). \tag{3.75}$$

The parameter values used in Figs. 3.15 and 3.16 are $\omega = 1$ Hz and $D = 5$, and the integration was made with the step $\delta t = 0.0001$. The delay used in Fig. 3.16a was $\tau = 0.015$.

If the noise variance D were much smaller, we would not have problems. The problems arise since we chose a rather large D such that the phase is not monotonically increasing. Instead there are long intervals during which the phase decreases, leading to "fake" loops in Fig. 3.16abc. The phase portrait obtained using wavelet transform looks more promising,

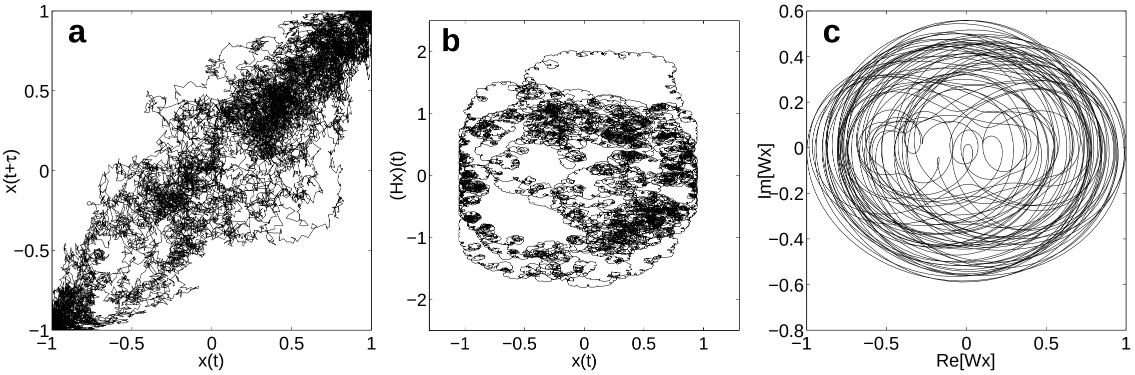

Figure 3.16: Phase portraits of the signal shown in Fig. 3.15, obtained by plotting (a) $x(t)$ against $x(t+\tau)$ with $\tau = 0.015$, (b) $x(t)$ against its Hilbert transform and (c) real part against imaginary part of wavelet coefficients obtained using corrected complex Morlet wavelet introduced in Sec. 3.4.2 ($\omega_0 = 1Hz, n_c = 1$).

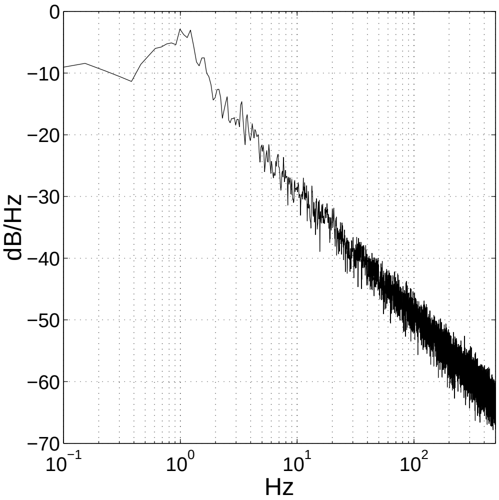

Figure 3.17: The spectrum of the signal shown in Fig. 3.15.

because it has less “fake” loops and is smoother. Moreover, if one takes the parameter n_c for example equal to 20 (this will correspond to a very narrow filter) one will avoid “fake” loops completely. But the phase of the signal defined with such transform has nothing to do with the original phase dynamics, because any details are washed out by the filter.

Our point is *not* that presently popular methods for phase extraction cannot distinguish between such phase reversals (or even just sudden slow-downs of the instantaneous phase velocity) and “true” amplitude variations[10]. Rather, we want to stress that there is no way *in principle* to distinguish between them. Thus, attempts to improve phase extraction methods in similarly ambiguous situations are likely to lead to ambiguous results, even if this ambiguity might be hidden.

Ways to avoid these ambiguities can be found only by restricting ourselves to what we accept as a sensible phase definition. One could argue, e.g., that a basic intuitive feature of a phase is its *continuous temporal progression*, i.e., positivity of the instantaneous phase velocity. Demanding this would mean that there is no possibility at all to define a phase for the above model, and the same would be true for a large class of signals.

Does this mean that such a requirement is too restrictive to be useful? We believe not. One traditional way out of the dilemma when phases should be defined for arbitrary signals is Fourier analysis. One decomposes the signal into harmonic components, and can then define phases for each component (or, when the signal is decomposed into frequency bands, for each band). What we propose is to decompose signals more generally into components with *positive* but not necessarily constant (as in a Fourier decomposition) phase velocities. This added freedom might allow much more physically relevant decompositions. Indeed, we do not have to invent any new example for this, since the best example demonstrating the power of such an approach is known since nearly four hundred years: progress in understanding planetary motion was only possible when Kepler replaced the decomposition into the harmonic epicycles of Ptolemaeus and Copernicus by a decomposition into elliptic motions, which are just of the type advocated here. A general ansatz with monotonically increasing phases in the spirit of the above discussion would be $x(t) = f(\phi_1(t), \ldots, \phi_n(t))$ with $\dot{\phi}_i(t) > 0$ for $i = 1, \ldots, n$. Details of such a decomposition will of course depend on the problem at hand, and we cannot give any general algorithm. But the possibility and the eventual usefulness of such an approach should be kept in mind.

3.4.5 Indices of phase synchronization

Let us come back now to the standard notion of phase. To quantify phase synchronization, several indices were introduced in the literature (see, e.g., [120, 102, 77]). All of them

[10] A hint at the actual structure of the present time sequence is obtained from the fact that it has many degenerate extrema: while there are many local maxima and minima with fluctuating amplitudes, the absolute extrema are all at $x(t) = \pm 1$. One might use such kind of information in special cases like the present one, but this will hardly lead to a universal and robust algorithm.

quantify in a slightly different way the deviation between the observed phase difference distribution and the one expected for independent systems. In our analysis we mostly used two of them, the first is an index introduced in Ref. [120] and based on the Shannon entropy, the second one is an index based on circular variance [77].

Suppose the phases $\varphi_x(t)$ and $\varphi_y(t)$ of the signals X and Y are extracted with any method (the methods need not necessarily be the same). Then it is possible to define the phase difference as

$$\phi_{xy}(t) = \varphi_x(t) - \varphi_y(t). \tag{3.76}$$

These are the common steps in defining both indices. The phase synchronization *index based on circular variance* is then given by

$$\gamma = \left| \left\langle e^{i\phi_{xy}(t)} \right\rangle_t \right|, \tag{3.77}$$

where $\langle \cdot \rangle_t$ denotes averaging over time.

In the literature this index has been introduced by Mormann and colleagues using the term *mean phase coherence* [77], and independently by Rosenblum and colleagues [102]. They refer to it as the *intensity of the first Fourier mode* of the phase distribution.

The phase synchronization *index based on Shannon entropy* is defined as

$$\sigma = \frac{S_{\max} - S}{S_{\max}}, \tag{3.78}$$

where S denotes the Shannon entropy of the distribution of the phase differences ϕ_{xy}, and S_{max} is the maximum entropy which corresponds to a uniform distribution of the phase difference[11]. Uniformity is assumed to correspond to the distribution of the phase difference for independent systems, although it is not quite correct. Instead of the uniform distribution, S_{max} should be estimated from independent pairs of phases. Since in general the phase distribution is not uniform, the distribution of phase differences is not either. For example, for relaxation oscillators the phase variable is changing not uniformly, thus the phase difference between two oscillators from which at least one is a relaxation oscillator will be non uniform also in case of independent oscillations. One of the way to avoid this problem is to switch to the rank ordered phases. Later in this thesis (see Chapter 4) we will check this idea on the example of EEG data.

In interpreting experimental results for σ and γ, one should always keep in mind that, if the phase velocities depend on the phase then the phase synchronization indices are not zero also for independent signals. Thus one should always compare results for σ and γ with estimates obtained from independent *surrogates* [122].

Both indices of phase synchronization introduced in this Section are symmetric by definition and therefore are not suited to exploit the directionality of interaction. Just recently

[11] The Shannon entropy is usually estimated using a binning technique.

two asymmetric extensions of the concept of phase synchronization have been proposed, the first one by Rosenblum et al. [100, 99] and the second one, an information-theoretic approach, by Palus and Stefanovska [81]. The former approach has been adapted for the application to short and noisy time series in Ref. [115]. The property of asymmetry is shared by the measures of non-linear interdependence introduced in the following Section.

3.5 Non-linear interdependencies

In this Section measures of generalized synchronization will be described. We will concentrate on the measures of non-linear interdependencies introduced by Arnhold and colleagues in Ref. [13]. This measures are related to earlier attempts to detect generalized synchronization like the method of mutual false nearest neighbors [106] and the mutual cross predictability introduced in Ref. [108], but they are optimized for robustness against noise and imperfections in the data. In Ref. [13], a number of other variants were also discussed. Some of these variants were found to be inferior, and systematically tested later in Ref. [93].

From the time series measured in two systems X and Y, let us reconstruct delay vectors [118] $\mathbf{x}_n = (x_n, \ldots, x_{n-(m-1)\tau})$ and $\mathbf{y}_n = (y_n, \ldots, y_{n-(m-1)\tau})$, where $n = 1, \ldots N$, m is the embedding dimension and τ denotes the time lag. Let $r_{n,j}$ and $s_{n,j}$, $j = 1, \ldots, k$, denote the time indices of the k nearest neighbors of $\mathbf{x}_n$ and $\mathbf{y}_n$, respectively.

For each $\mathbf{x}_n$, the mean squared Euclidean distance to its k neighbors is defined as

$$R_n^{(k)}(X) = \frac{1}{k} \sum_{j=1}^{k} \left(\mathbf{x}_n - \mathbf{x}_{r_{n,j}}\right)^2 \tag{3.79}$$

and the *Y-conditioned* mean squared Euclidean distance is defined by replacing the nearest neighbors by the equal time partners of the closest neighbors of $\mathbf{y}_n$ (see Fig. 3.18),

$$R_n^{(k)}(X|Y) = \frac{1}{k} \sum_{j=1}^{k} \left(\mathbf{x}_n - \mathbf{x}_{s_{n,j}}\right)^2. \tag{3.80}$$

Suppose the point cloud $\{\mathbf{x}_n\}$ has an average squared radius $R(X) = \frac{1}{N} \sum_{n=1}^{N} R_n^{(N-1)}(X)$, then $R_n^{(k)}(X|Y) \approx R_n^{(k)}(X) \ll R(X)$ if the systems are strongly correlated, while $R_n^{(k)}(X|Y) \approx R(X) \gg R^{(k)}(X)$ if they are independent. Accordingly, we can define an interdependence measure $S^{(k)}(X|Y)$ [13] as

$$S^{(k)}(X|Y) = \frac{1}{N} \sum_{n=1}^{N} \frac{R_n^{(k)}(X)}{R_n^{(k)}(X|Y)}. \tag{3.81}$$

Since $R_n^{(k)}(X|Y) \geq R_n^{(k)}(X)$ by construction, we have

$$0 < S^{(k)}(X|Y) \leq 1. \tag{3.82}$$

Low values of $S^{(k)}(X|Y)$ indicate independence between X and Y, while high values indicate synchronization.

Following Ref. [13, 93] one can define another non-linear interdependence measure $H^{(k)}(X|Y)$ as

$$H^{(k)}(X|Y) = \frac{1}{N} \sum_{n=1}^{N} \log \frac{R_n(X)}{R_n^{(k)}(X|Y)} \tag{3.83}$$

It is zero if X and Y are completely independent, while it is positive if closeness in Y implies closeness in X for equal time partners. It would be negative if close pairs in Y corresponded mainly to distant pairs in X. This is very unlikely but not impossible. Therefore, $H^{(k)}(X|Y) = 0$ suggests that X and Y are independent, but does not prove it. This is one main difference between $H^{(k)}(X|Y)$ and the mutual information, which is defined in Sec. 3.2.

In a study on coupled chaotic systems [93], H was more robust against noise and easier to interpret than S, but with the drawback that it is not normalized. Therefore, in Ref. [94] a new measure which uses a different way of averaging was proposed:

$$N^{(k)}(X|Y) = \frac{1}{N} \sum_{n=1}^{N} \frac{R_n(X) - R_n^{(k)}(X|Y)}{R_n(X)}. \tag{3.84}$$

It is normalized (but as in the case of H, it can be slightly negative) and, in principle, more robust than S. The measure N reaches its maximum value of 1, only if $R_n^{(k)}(X|Y) \equiv 0$. This will not happen even if X and Y are identically synchronized (except in the case of periodic signals). This small drawback was corrected in Ref. [6], where one more way of normalizing ratios of distances was proposed,

$$M^{(k)}(X|Y) = \frac{1}{N} \sum_{n=1}^{N} \frac{R_n(X) - R_n^{(k)}(X|Y)}{R_n(X) - R_n^{(k)}(X|Y)}. \tag{3.85}$$

The opposite interdependencies $S(Y|X)$, $H(Y|X)$, $N(Y|X)$ and $M(Y|X)$ are defined in complete analogy and they are in general not equal to $S(X|Y)$, $H(X|Y)$, $N(Y|X)$ and $M(Y|X)$, respectively. The asymmetry of S, H, N and M is the main advantage over other non-linear measures such as the mutual information and the phase synchronization. This asymmetry can give information about driver-response relationships [13, 93, 109], but can also reflect the difference of properties of each dynamics [13, 93].

Figure 3.18 illustrates how the non-linear interdependence measures work. Let us consider a Lorenz and a Rössler system that are independent (upper case, no coupling) and a second

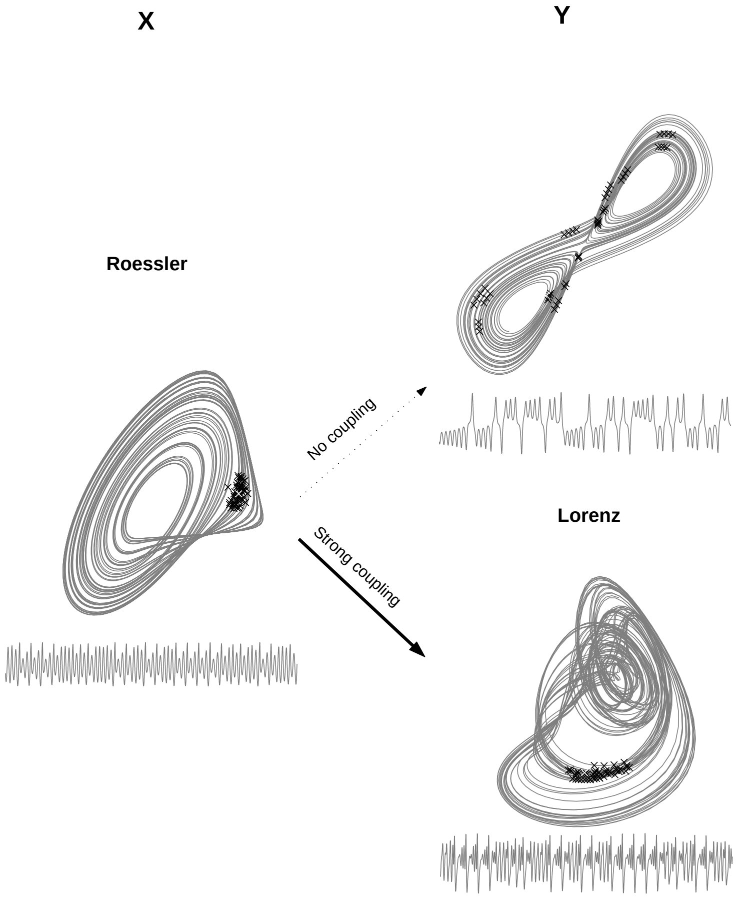

Figure 3.18: Basic idea of the non-linear interdependence measures. The size of the neighborhood in one of the systems, say X, is compared with the size of its mapping in the other system. The example shows a Lorenz system driven by a Rössler system with zero coupling (upper case) and with strong coupling (lower case). Below each attractor, the corresponding time series is shown. The black crosses on the left panel indicate 50 nearest neighbors of the point x_n shown by the white cross. Assume that the nearest neighbors of x_n are at discretely sampled times $r_{n,1}, r_{n,2}, \ldots, r_{n,50}$. On the right panels, the crosses indicate the states at the same times n (white crosses) and $r_{n,1}, r_{n,2}, \ldots, r_{n,50}$ (black crosses). For upper case $S(Y|X) = 0.001$ and $H(Y|X) = 0.056$, for lower case $S(Y|X) = 0.275$ and $H(Y|X) = 3.694$. The $(X|Y)$ interdependencies are calculated in the same way, starting with a neighborhood in Y.

case with the Rössler driving the Lorenz via a strong coupling (lower plot). For a detailed study of synchronization between these systems refer to [93]. Given a neighborhood in one of the attractors, we see how this neighborhood maps to the other. If the systems are synchronized, the point cloud is still a small neighborhood (lower plot). On the other hand, if the points are spread over the attractor (upper plot), the systems are independent. The four measures S, H, N and M are just different ways of normalizing ratios of distances.

Chapter 4

Application to the EEG of epilepsy patients

In this Chapter the interdependence and synchronization measures introduced in Chapter 3 are applied to electroencephalographic (EEG) recordings of epilepsy patients. Due to the physiological and pathophysiological variations in the brain these electroencephalographic time series represent a prominent example of biological data showing a rich and diverse appearance. Since synchronization plays an important role for pathological processes such as epilepsy, the EEG recorded from epilepsy patients represents a great challenge for the application of methods derived from the theory of dynamical systems and, in particular, from the theory of synchronization.

Applications of several bivariate synchronization measures to the problem of epileptic focus localization have been reported in the literature [59, 77, 12]. For instance, in Ref. [77] one of the indices of phase synchronization was applied. In Refs. [59, 12] an application of different measures of interdependence was described. Making use of only one synchronization measure is the common feature of all these studies, and a comparative study of different measures was still missing. In such studies a rigorous statistical validation of the obtained results is very important. Even more important is to test to which degree the results obtained using synchronization analysis are specifically related to synchronization phenomena. This can be done with the powerful method of bivariate surrogate data analysis.

This Chapter is organized as follows. First, in Sec. 4.1 a short introduction to the disease epilepsy and to the method of the electroencephalography is given. In Section 4.2 a detailed analysis of the spatial variability of neuronal synchronization is presented, and the potential of synchronization analysis for the localization of epileptic foci is investigated. The bivariate surrogate data techniques are described and applied in Sec. 4.2.2. Finally, conclusions are drawn in Sec. 4.3.

4.1 Epilepsy and the electroencephalogram

The word *'epilepsy'* is derived from the Greek verb $\varepsilon\pi\iota\lambda\alpha\mu\beta\alpha\nu\varepsilon\iota\nu$ (*epilamvanein* = 'to be seized' or 'to be attacked'). In ancient medicine epilepsy was considered as a "sacred disease" rather than as a mental disorder. But already the famous ancient Greek physician Galen emphasized the physical involvement of the brain in epilepsy.

The disease epilepsy is characterized by a sudden and recurrent malfunction of the brain which manifests itself as an epileptic seizure. These seizures are fundamentally divided into two main classes - *generalized* and *partial*. While generalized seizures involve almost the entire brain, partial (focal) seizures have clinical or electroencephalographic evidence of a localized onset and usually stay confined to one hemisphere [58]. The origin of focal seizures is usually called *epileptic focus*.

Approximately $0.5\% - 0.8\%$ of the world's population suffers from epilepsy [10]. With today's available antiepileptic drugs, seizures can be controlled satisfactorily in about 67% of the affected individuals, another 8% may profit from epilepsy surgery. The remaining 25% of epilepsy patients cannot be treated sufficiently by any available therapy. Successful surgical treatment of focal epilepsies requires exact localization of the epileptic focus and its delineation from functionally relevant areas. For this purpose, different presurgical evaluation methods are currently in use [26]. Neurological and neuropsychological examinations

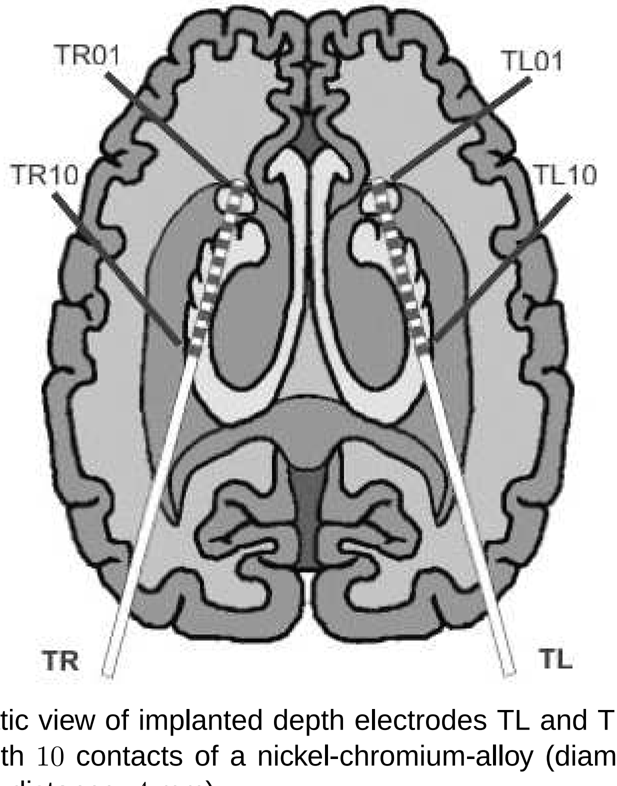

Figure 4.1: Schematic view of implanted depth electrodes TL and TR. Each depth electrode is equipped with 10 contacts of a nickel-chromium-alloy (diameter: 1 mm, length: 2.5 mm, inter-contact distance: 4 mm).

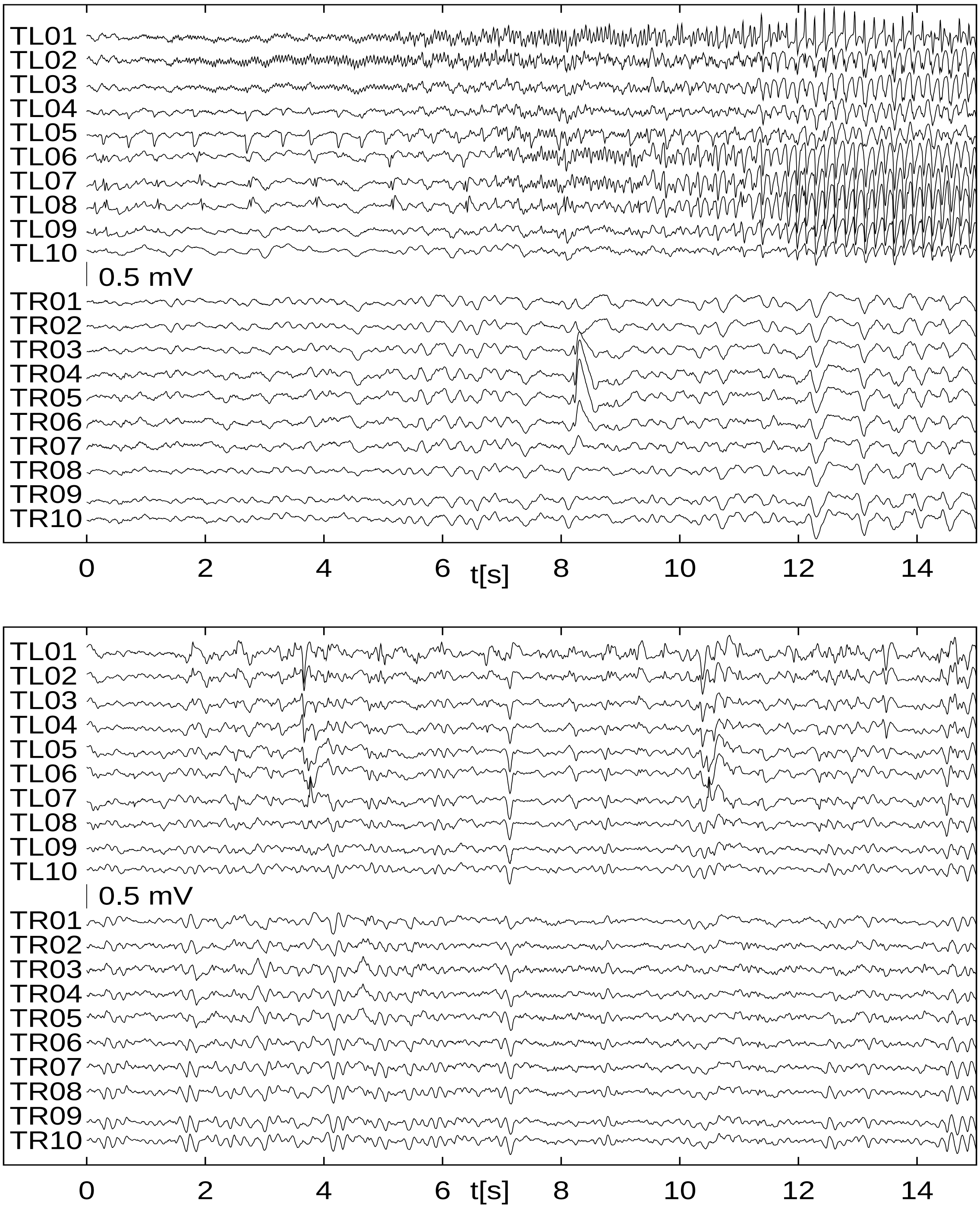

Figure 4.2: Two exemplary multichannel SEEG recordings of a patient with mesial temporal lobe epilepsy originating in the left hippocampal formation. Top: Onset of a seizure and beginning of ictal activity. The seizure starts about 1 s after the start of the recording with low amplitude high frequency activity in channels TL01-TL03. Bottom: Interictal activity.

are complemented by noninvasive techniques such as multichannel magnetoencephalography (MEG) and by neuroimaging techniques such as magnetic resonance imaging (MRI). Single photon emission computed tomography (SPECT) also plays an important role in the epilepsy diagnostics.

One of the most important diagnostic tools in epileptology is the EEG. As reported by Caton in 1875, the first recordings of electrical activity of the brain were performed in exposed brains of rabbits and monkeys. In 1929 Hans Berger performed the first measurements of brain electrical activity in humans [16]. Nowadays, the EEG is of particularly high value for the localization of the epileptic focus.

The standard method for epileptic focus localization is to record the patient's spontaneous habitual seizure. Depending on the individual occurrence of seizures, this task requires long-lasting and continuous recordings of the EEG. In case of ambiguous scalp EEG findings, invasive recordings of the electrocorticogram (ECoG) or the stereo-EEG (SEEG) via chronically implanted intracranial electrodes are applied. A typical implantation scheme of these electrodes is presented in Fig. 4.1. The exceptional signal to noise ratio and the excellent temporal and spatial resolution yielded by chronically implanted electrodes allow a substantially increased precision in the design of the surgical intervention justifying the high degree of invasiveness. To avoid the necessity of recording seizure activity, reliable EEG analysis techniques capable of localizing and demarcating the epileptic focus even during the *seizure-free* (also called *interictal*) interval would be of great value [61, 9, 5]. The upper panel corresponds to the onset of the seizure which is observed in the left hemisphere of the brain (channels TL01-TL03). In Fig. 4.2 (bottom) a typical interictal EEG is shown. In this example one can see some differences between the activity in the focal and in the non-focal hemisphere, but in general this so called *interictal epileptiform activity* does not allow to reliably localize the epileptic focus.

The conventional surgical treatment of mesial temporal lobe epilepsy consists of a complete resection of the hippocampal formation in the focal hemisphere of the brain. That is why precise (up to one or two contacts) localization of epileptic focus can not be confirmed by post-operative seizure control. This is in contrast to the case for which the focus is located in the neocortex. Defining only the focal hemisphere of the brain is usually called lateralization. Nonetheless, in this Chapter we will use the term localization rather than lateralization keeping in mind the distinction between them.

4.2 Localization of the epileptic focus

Already in the 1960s, it was proposed that high voltages in the EEG recorded during the seizure must represent "hypersynchrony" of individual neurons. Although *desynchronization* processes during epileptic seizures have recently been discussed [78] the fact of highly synchronous ictal activity of large populations of neurons in the brain is well accepted in the EEG analysis community [26, 79]. This fact along with the findings from a careful visual inspection of the interictal EEG, e.g., synchronous occurrence of spikes, suggests the following working hypothesis for determining the focal hemisphere of the epileptic brain.

> During the interictal period the average synchronization in the focal hemisphere of the brain of epileptic patients is higher than in the non-focal hemisphere.

To test this hypothesis, which has also been tested in Refs. [77, 12], EEG recordings from 29 patients were analyzed retrospectively (Tab. 4.1). These patients were undergoing presurgical diagnostics at the Department of Epileptology, University of Bonn, Germany.

All patients achieved complete post-operative seizure control (cf. Ref. [25]) after resection of the brain area which was correctly assumed to be the epileptic focus. For our study this means that for all patients the correct hemisphere of the epileptic focus location was known exactly. In 18 patients the focus was localized in the left brain hemisphere while in 11 patients the focus was in the right hemisphere. In the following, the brain hemisphere containing the epileptic focus will be referred to as the focal side, whereas the opposite hemisphere will be referred to as the non-focal side. From these 29 patients, 83 artifact-free interictal EEG recordings (mean duration: 47 minutes) were evaluated by applying a moving window technique to all combinations of channels of the same side.

The EEG recordings were performed under video control using chronically implanted intra-hippocampal depth electrodes (see Fig. 4.1). After the implantation the correct placement of the electrodes was verified by magnetic resonance imaging. The recording was

Patient-ID	Focus Side	Length [min]	Patient-ID	Focus Side	Length [min]
A	R	21	P	R	79
B	R	621	Q	R	121
C	L	223	R	R	79
D	L	36	S	L	65
E	L	150	T	L	64
F	R	191	U	L	70
G	L	137	V	L	418
H	R	34	W	L	214
I	L	88	X	L	26
J	L	28	Y	L	68
K	L	29	Z	L	122
L	R	21	a	l	21
M	L	21	b	L	52
N	R	593	c	R	70
O	R	226	$\sum$	R11/L18	3886

Table 4.1: Patient characteristics. Depicted are patient-ID, the focal side, and the length of the recordings.

carried out on a 128-channel amplifier system with band-pass filter settings of $0.5 - 85$ Hz (12 dB/oct.) using a common average reference. The sampling rate was 173.61 Hz and the resolution of the AD converter was 12 bit.

4.2.1 Performance of synchronization measures

4.2.1.1 Methods

For each patient all recordings were analyzed using a moving window technique, which represents a common way of handling long term EEG recordings [14]. EEG signals were divided into segments of 4096 sampling points each, corresponding to a window length of 23.6 seconds at the given sampling rate. This window length can be regarded as a compromise between the required statistical accuracy for the calculation of all measures used in this thesis and the approximate stationarity within a window length [17].

For the localization analysis we used in total four different classes of measures. As measures of generalized synchronization we used the nonlinear interdependencies H, S, M, N described in Sec. 3.5. Time-delay embedding parameters were set to the following values: dimension $d = 10$, time delay $\tau = 5$. The number of nearest neighbors k and a Theiler-correction T to exclude temporally correlated neighbors [121] were: $k = 10$ and $T = 10$. Whenever the values of H, M, and N became negative, which can happen especially for the surrogate data, they were replaced by 0.

Phase synchronization was quantified using the two indices of phase synchronization σ and γ introduced in Sec. 3.4.5. These indices were calculated for the phase differences where each phase was extracted using both the Hilbert transform (in the following denoted as σH and γH), and the wavelet transform. The phase synchronization analysis using wavelet transform allows to concentrate on a specific frequency range by choosing the main frequency of the wavelet ω_0. We have selected several physiologically interesting frequencies. In the following Figures these indices are denoted by $\sigma\omega\omega_0$ and $\gamma\omega\omega_0$, where ω_0 stands for the actual value of the main frequency of the corrected complex Morlet wavelet.

As already mentioned in Sec. 3.4.1, the easiest way to compute the HT is to perform a fast FT (FFT) of the original time series, shift the phase of every frequency component by $-\pi/2$ and to apply the inverse FFT. The phase shift can conveniently be implemented by swapping the imaginary and the real parts of the FT. To reduce boundary effects one can exclude several periods at the beginning and at the end of the resulting analytical signal. In our calculations we proceeded in a similar way. We used a moving window technique with 50% overlap, and discarded 25% of the points at the beginning and at the end of the resulting analytical signal. The wavelet transform was implemented in the frequency domain using the convolution theorem.

As we suggested in Sec. 3.4.5, we calculated both indices of phase synchronization for rank ordered phases. We checked hereby whether non-uniformity in the evolution of each

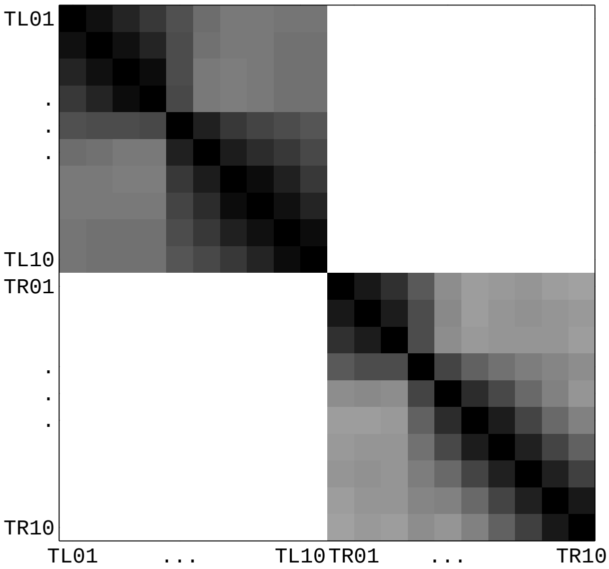

Figure 4.3: Exemplary color-coded synchronization matrix obtained for the index based on circular variance. The phases were extracted with the wavelet transform, $\omega_0 = 18$. Dark colors correspond to high values of the index, white corresponds to 0. For the example shown here the focus was in the left side.

phase has an influence on the values of the phase synchronization indices. In the following Figures these indices are denoted by σHI and γHI.

Mutual information was estimated for pairwise combinations of channels of the same side using our new estimators introduced in Sec. 3.2.2. In the following Figures we denote the "cubic" estimator (Eq.(3.34)) as ICk and the "rectangular" estimator (Eq.(3.40)) as IRk, where k stands for the number of the nearest neighbors used in the calculations.

The two linear measures of synchronization described in Sec. 3.1 were also calculated and in the Figures they are denoted by C_0 and C_{max}.

For all measures and each patient we calculated a synchronization matrix for each window of the respective recording. The synchronization matrix contains the values of the synchronization measure for all combinations of channels of the same side in the respective window. In Fig. 4.3 an exemplary synchronization matrix is presented. We can see that high values are found predominantly in the upper left sub-matrix. This indicates that in the case presented here synchronization in the focal hemisphere is higher than in the non-focal hemisphere.

To test the hypothesis proposed in the beginning of this Section we introduce a localization index, which is defined just as a difference between average synchronization in the focal and non-focal hemisphere. Suppose we are calculating the measure M, then the localization index L is given by

$$L = \langle M_{\text{foc}} \rangle - \langle M_{\text{nonfoc}} \rangle. \tag{4.1}$$

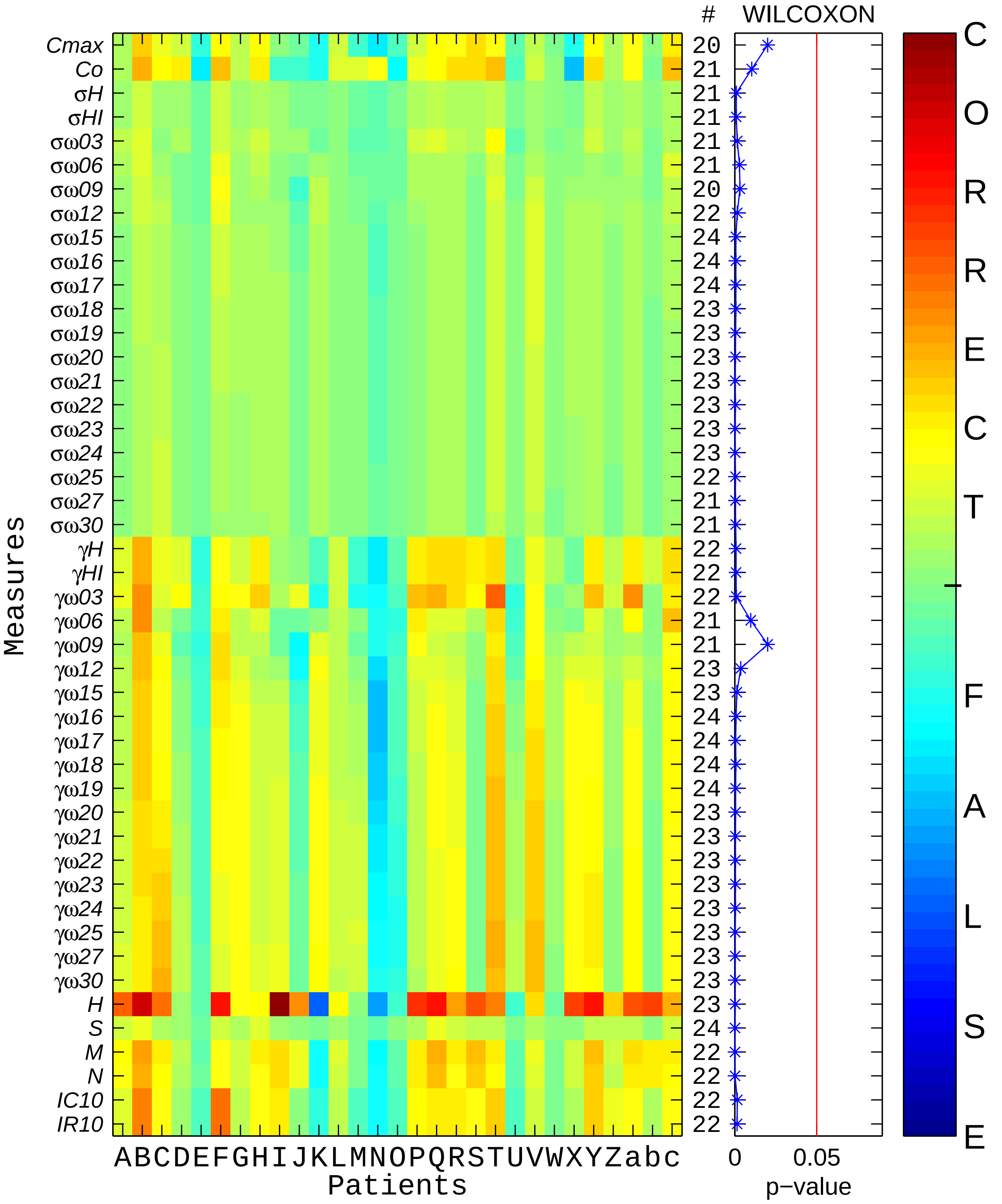

Figure 4.4: Color-coded localization index for each patient and each measure. Warm colors denote a correct localization whereas chilled colors mark a false one.

To get the localization index for one window the averaging should be done over all channel combinations in the focal and the non-focal hemisphere, respectively. For the whole recording of one patients the localization index is obtained by additional averaging over all windows. The order of the two averaging steps is not important since L linearly depends on a measure. According to our hypothesis, a positive value of the localization index corresponds to a correct localization, and a negative one to a false localization.

The statistical significance of the obtained results is checked by means of the Wilcoxon signed-rank test. It is a nonparametric procedure used to test the null hypothesis that two variables have the same distribution. The test makes no assumptions about the shapes of the distributions of the two variables. However, it takes into account information about the magnitude of differences within pairs and gives more weight to pairs that show large differences than to pairs that show small differences. This test statistic is based on the ranks of the absolute values of the differences between the two variables. An observed significance level is often called p value. This value is the basis for deciding whether or not to reject the null hypothesis. It is the probability that a statistical result as extreme as the one observed would occur if the null hypothesis were true. If the observed significance level is small enough, usually less than 0.05 or 0.01, the null hypothesis is rejected.

Note that the assumption of higher values of synchronization in the focal hemisphere does not bias our analysis scheme. The opposite difference would be detected automatically.

4.2.1.2 Results

The results of the localization analysis are presented in Fig. 4.4. Each cell of this color-coded matrix represents the value of the localization index for the corresponding patient and measure. The numbers at the right of the Figure represent the number of correctly classified patients (positive localization index). This number varies from 20 to 24 depending on the measure. For several measures it is not easy to read from the Figure whether the localization is correct or false because the absolute values of the localization index are very small. A question arises if these small values are significant enough to be judged as a success or a failure of the classification procedure. One possible way to answer this question is to perform a statistical analysis using standard hypothesis tests.

4.2.1.3 Statistical validation

The distribution which we first put to the Wilcoxon signed-rank statistical test are the distributions of synchronization values in the focal and in the non-focal side of the brain, respectively. Each distribution is across patients.

The test retrieved a significant ($p < 0.05$) difference between the distributions of focal and non-focal average synchronization values for all measures. All p values of the Wilcoxon test are presented in the second subplot of Fig. 4.4.

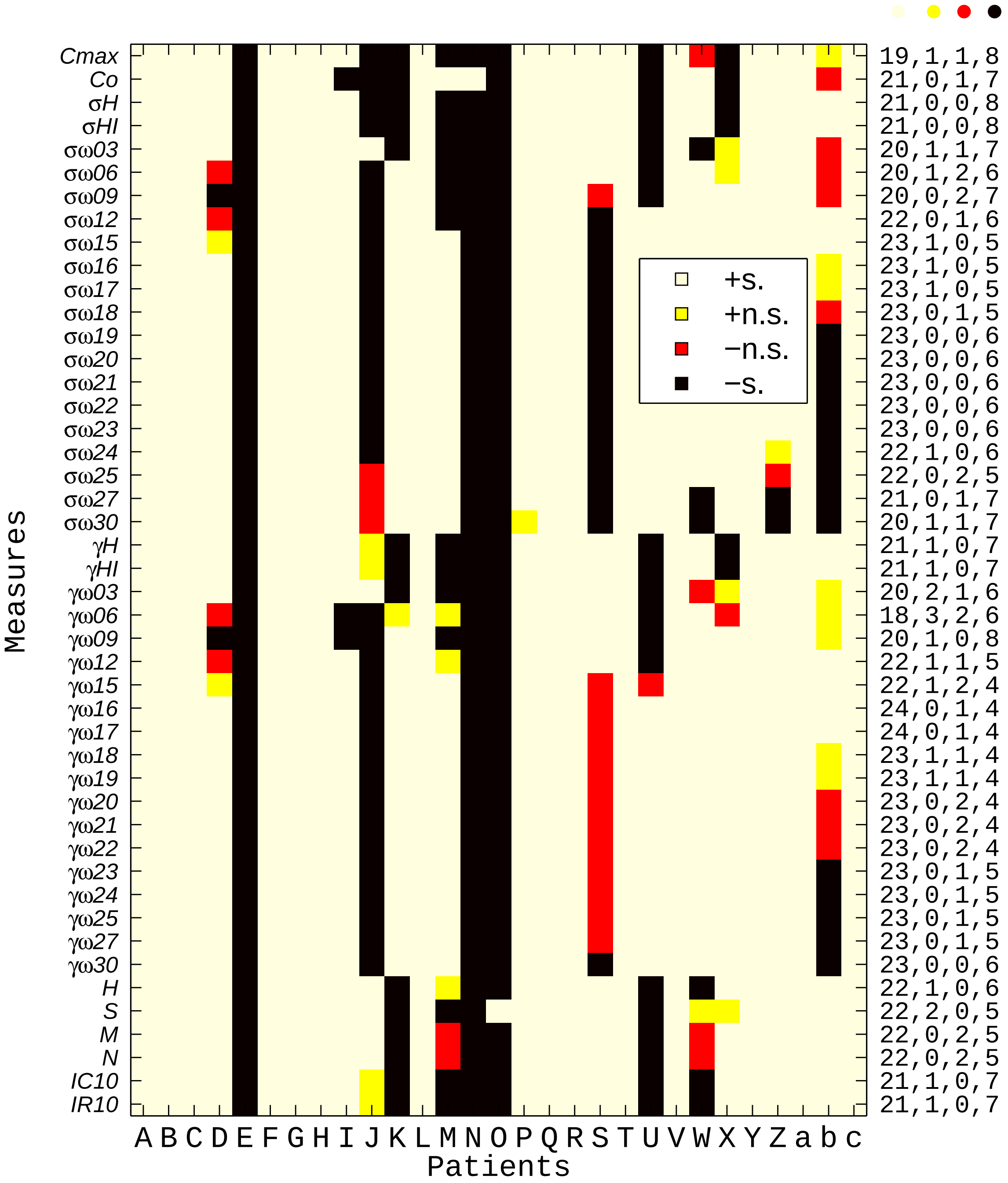

Figure 4.5: Results of the localization analysis in combination with the Wilcoxon signed-rank test. The beige color (+s.) corresponds to a correct and significant localization according to the Wilcoxon test. The yellow color (+n.s.) indicates the still positive but non-significant values of the localization index. The red color (-n.s.) is used for non-significant but negative values of the localization index and the black one (-s.) denotes significantly false localizations. The four columns of numbers to the right of the Figure denote the number of each classification for each measure.

The Wilcoxon signed-rank test can be applied not only to the average over all windows but also to each patient and each measure separately. The tested distributions in this case are the distributions of the synchronization values for the focal and non-focal side in each window. The results are shown in Fig. 4.5. The majority of the cells in this color-coded plot indicate a correct and significant localization according to the Wilcoxon test. There are also cells with positive values of the localization index, and several cells with negative values which are found to be non-significant. Already from this plot one can see a high correlation in the results for different measures. The number of patients which are either correctly or falsely localized by all measures is considerably higher then the number of questionable localizations with regard to the choice of the measure. The smallest number of correctly classified cases (18 patients) is found for the index based on circular variance calculated for the phase difference obtained by the wavelet transform with the central frequency of 6 Hz. The best results (24 patients) are found for the same measure but with central frequency of the wavelet in a range $16 - 17$ Hz. In a wider range from 16 Hz to 23 Hz the results for both indices have just one significant correctly localized case less. The results for rank ordered phases are the same as the results obtained for original phases. The difference in the performance of all four measures of generalized synchronization disappears after significance analysis. This is an expectable result because the difference between them is only in the normalization.

4.2.1.4 Correlations between the different measures

In Fig. 4.6 the pairwise correlation coefficients between all[1] average (over all channel combinations) synchronization values are depicted based on the entire database analyzed in this Section.

This plot confirms what we have already mentioned in the previous Section, namely that all the measures correlate well with each other. The minimal correlation values are observed between the measure of generalized synchronization H and the majority of phase synchronization measures based on wavelet transform, but still the value of correlation coefficient is about 0.5. One can roughly distinguish clusters of measures corresponding to different classes, namely to linear cross-correlation, phase synchronization based on the Hilbert and wavelet transform, generalized synchronization and mutual information. The highest correlation is observed within the following three small clusters. The first one consists of the two measures of phase synchronization based on the Hilbert transform for original and rank ordered phases. The second one consists of two variants of the generalized synchronization measure, namely M and N, and both estimators of mutual information also have high correlation. The biggest cluster consists of phase synchronization measures based on wavelet transform for the frequency range $12-30$ Hz, the other frequency range $3-9$ Hz is

[1] In general both indices of phase synchronization are strongly correlated. Therefore, for the sake of clearness of the picture we omit indices based on Shannon entropy.

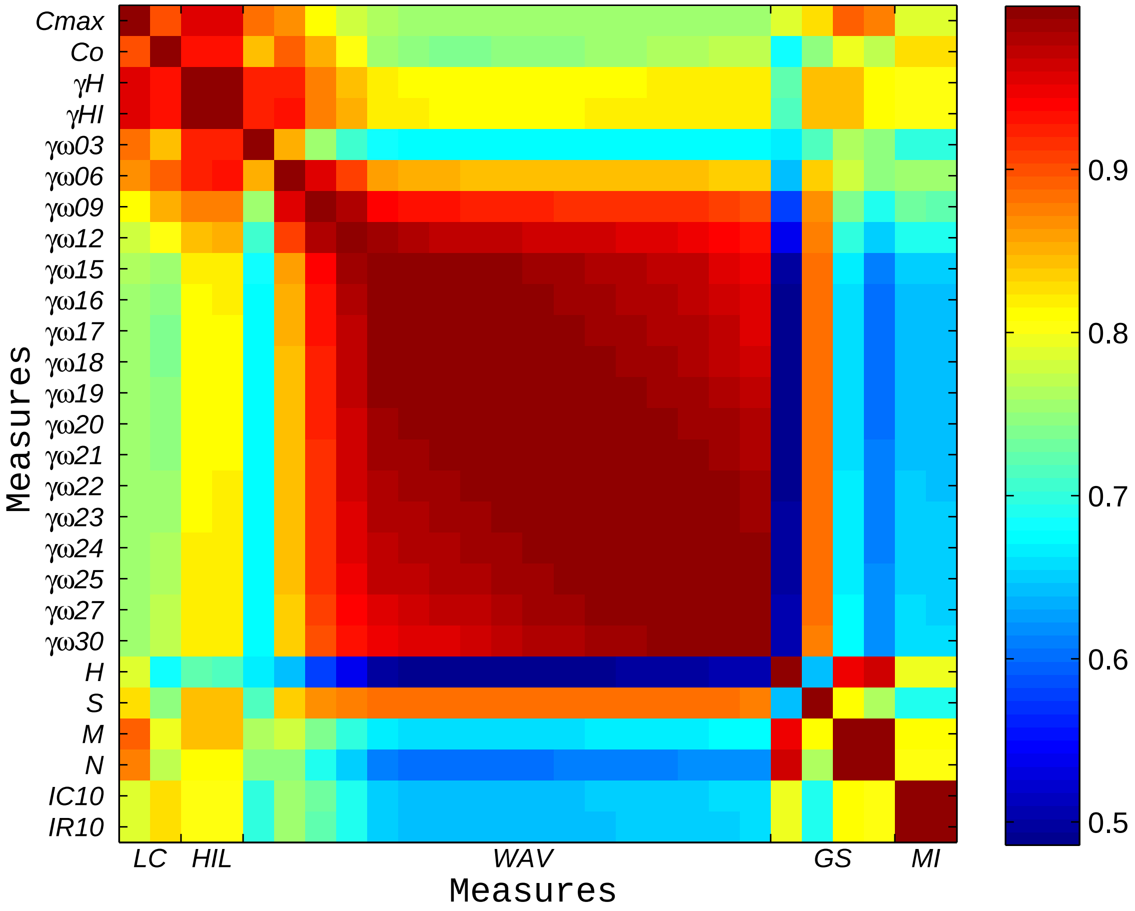

Figure 4.6: Correlation coefficients between the average (over all channel combinations) synchronization values. Coefficients are determined using all 9691 windows from all 29 patients. Different classes of measures are denoted on abscissa axis by *LC* for linear correlation, *HIL* and *WAV* for phase synchronization based on the Hilbert and the wavelet transform, respectively, *GS* for generalized synchronization, and *MI* for mutual information.

not highly correlated with the previous one. This is in correspondence with the localization performance, where the worst results were obtained for the frequency 6 Hz. The difference between S and all other measures of generalized synchronization is also expectable since S is the only measure of generalized synchronization which uses a local neighborhood (see Sec. 3.5).

4.2.2 Bivariate surrogate data analysis

Using the Wilcoxon test for the majority of patiens/measures we found a significant difference between average values of synchronization in the focal and the non-focal hemisphere of the epileptic brain. This difference has also been reported for linear and nonlinear univariate measures which obviously are not sensitive to possible synchronization between different areas of the brain [9, 7]. Therefore, the question arises whether the difference in synchronization values is really caused by synchronization phenomena or rather by some other overall properties of the EEG. This question can be addressed using the concept of surrogate data. The method of surrogate data allows to test the results, for which analytical estimation can not be given, against a specified null-hypothesis. For our purpose, the null hypothesis can be formulated as follows.

> $\mathcal{H}_0$: Any potential difference in the average values of the synchronization measure found for the focal and the non-focal hemisphere can be explained by some differences in properties of the overall dynamics of the focal and the non-focal hemisphere not related to synchronization.

In the framework of nonlinear time series analysis the concept of surrogates was at first used as a test for nonlinearity [122, 112, 113]. In general, an ensemble of surrogate time series is constructed from the original time series in such a way that they have all properties consistent with the specified null hypothesis in common with the original, but are otherwise random. A discriminating statistics, which has to be sensitive to at least one property that is inconsistent with the null hypothesis, is calculated for both the original time series and the surrogates. If the result for the original falls outside the range of values obtained for the surrogates, the null hypothesis can be rejected.

In the next paragraphs, we will summarize some aspects of our extensive study of the necessity, strength and caveats of bivariate surrogate data analysis [6]. In this paper a hierarchy of null-hypotheses along with different algorithms for their testing have been discussed.

The main purpose of bivariate surrogate data analysis is to verify the interdependence between processes under consideration. The simplest null hypothesis for this analysis is that they are independent linear stochastic processes. To test against this null-hypothesis one should preserve only the linear properties of each process. But this is not the most interesting case. Actually, from the algorithmic point of view a procedure for generating this type of surrogates is exactly the same as for univariate surrogates. It should simply be applied to time series of each process separately. Moreover, the rejection of this null hypothesis does not necessarily mean that the processes are dependent, they could for example contain some structure which is not consistent with the assumption of a linear stochastic process.

A natural extension of the previous null hypothesis is to take linear interactions into account. The processes are still assumed to be linear stochastic ones. Under linear interaction we understand all interactions which can be completely described in terms of the linear cross-correlation function or in terms of its equivalent in the frequency domain, the coherence function. To test against this extended null hypothesis one has to preserve the cross-correlation between the processes and the two autocorrelation functions. An algorithm for generating an ensemble of surrogates has to work for both time series simultaneously. In the simplest case of phase randomized surrogates the phase difference between corresponding components in Fourier transform should be preserved. The rejection of this null hypothesis can indicate nonlinear synchronization or, e.g., nonlinear structure in independent processes.

To take a nonlinear structure into consideration it is possible to assume as a null hypothesis two processes with arbitrary structure but without nonlinear interdependence and without significant linear cross- correlation. The easiest way to fulfill this is to take different parts (e.g., at different instances of time) of a time series generated by the same process. Under the stationarity assumption the shifted versions of the same time series would have statistically the same structure and autocorrelation. Assuming ergodicity, they should be independent.

A further generalization of this null hypothesis includes the linear cross-correlation as well. It is, however, impossible to generate corresponding surrogates, because preserving the structure along with the autocorrelation and the cross-correlation at the same time completely specifies both time series and does not leave any degree of freedom for randomization. As a way out of this dilemma, Schreiber proposed to preserve only a part of the cross-correlation and/or the autocorrelation functions (up to some time lag τ) [110]. This method uses a simulated annealing technique for generating an ensemble of surrogates and is therefore very time consuming.

To answer the question we have put in the beginning of this Section we will use time-shifted surrogates. This type of surrogates is perfectly suited to test against the null-hypothesis $\mathcal{H}_0$.

4.2.2.1 Methods

The method of time-shifted surrogates was introduced in Ref. [94]. Later it was used in Ref. [78], and compared with other bivariate surrogate data methods in Ref. [6]. In contrast to these studies, we use randomly chosen windows as time shifted surrogates. This allows us to avoid unwanted high linear cross-correlation between surrogates.

Time-shifted surrogate data analysis was performed for the entire localization database of patients (cf. Tab. 4.1) and for each synchronization measure. We proceeded in a slightly different manner for symmetric measures (linear cross-correlation, phase synchronization) and for asymmetric measures (generalized synchronization). For symmetric measures the synchronization matrix (see Fig. 4.3) is symmetric but the surrogate synchronization matrix is in general asymmetric. The first line of the surrogate synchronization matrix for a window i (see Tab. 4.2) contains the value of the synchronization measure between the channel A in the window i and all channels ($B^*, \ldots, Z^*$) from the surrogate window i^*. The second line contains the value of the synchronization measure between the channel B in the window i and all channels except B^* from the surrogate window i^*. In general $M_{AB^*} \neq M_{BA^*}$. For asymmetric measures the original synchronization matrix is already asymmetric. To calculate the surrogate synchronization matrix we proceed essentially as in the previous case. We use the indices found for the original time series of channel A in

Channel-ID	A	B	$\dots$	Z
A		M_{AB^*}	$\dots$	M_{AZ^*}
B	M_{BA^*}		$\dots$	M_{BZ^*}
$\dots$	$\dots$	$\dots$		$\dots$
Z	M_{ZA^*}	M_{ZB^*}	$\dots$	

Table 4.2: Schematic representation of the surrogate synchronization matrix.

the window i to calculate the conditional radius (see Sec. 3.5) in the surrogate time series of channel B from the surrogate window i^*.

For each window we randomly chose 19 surrogate windows from the recordings of the same patient and calculated 19 surrogate synchronization matrices.

4.2.2.2 Results

In Fig. 4.7 the results of the localization analysis in combination with the time-shifted surrogate data test are shown. As in Fig. 4.5 each cell corresponds to one patient and one measure. Again the majority of the cells indicate correct and significant localization with regard to the time-shifted surrogate test. It means that the average localization index calculated for the original recording lies outside (in this case higher than) any of 19 average localization indices calculated for the surrogate localization matrix. In comparison with the results presented in Fig. 4.5, the number of nonsignificant cases has increased, i.e., significant results with regard to Wilcoxon statistical test turned out to be nonsignificant with regard to the time-shifted surrogates test. This means that the difference in the distributions of the average synchronization values in the focal and the non-focal side is due to some overall properties of the EEG recordings but not due to synchronization. There are a few patients/measures for which the opposite is true, namely the results are significant with regard to the time-shifted surrogates test but nonsignificant with regard to Wilcoxon statistical test. It might mean that the difference in the average synchronization values in the focal and non-focal side is so small that it can not be distinguished by this statistical test. But using the time-shifted surrogate data analysis which takes into account additional information, e.g., about the hidden structure, allows to retrieve a significant distinction.

While the time-shifted surrogate test is applied to each window separately, it is possible to check the significance of the localization index for each window and to exclude the nonsignificant windows from the analysis. It will not automatically lead to significant results after the averaging over all remaining windows, since both correctly and falsely localized windows are still included in the analysis. In Fig. 4.8 the results of this analysis are shown. The number of nonsignificant cases decreases in comparison with the previous analysis (cf. Fig. 4.7). Moreover, the number of correctly localized patients increases on average by one patient for each measure. A small difference between the localization results in combination with the Wilcoxon statistical test and the time-shifted surrogate data test is found for the phase synchronization measures based on the wavelet transform and

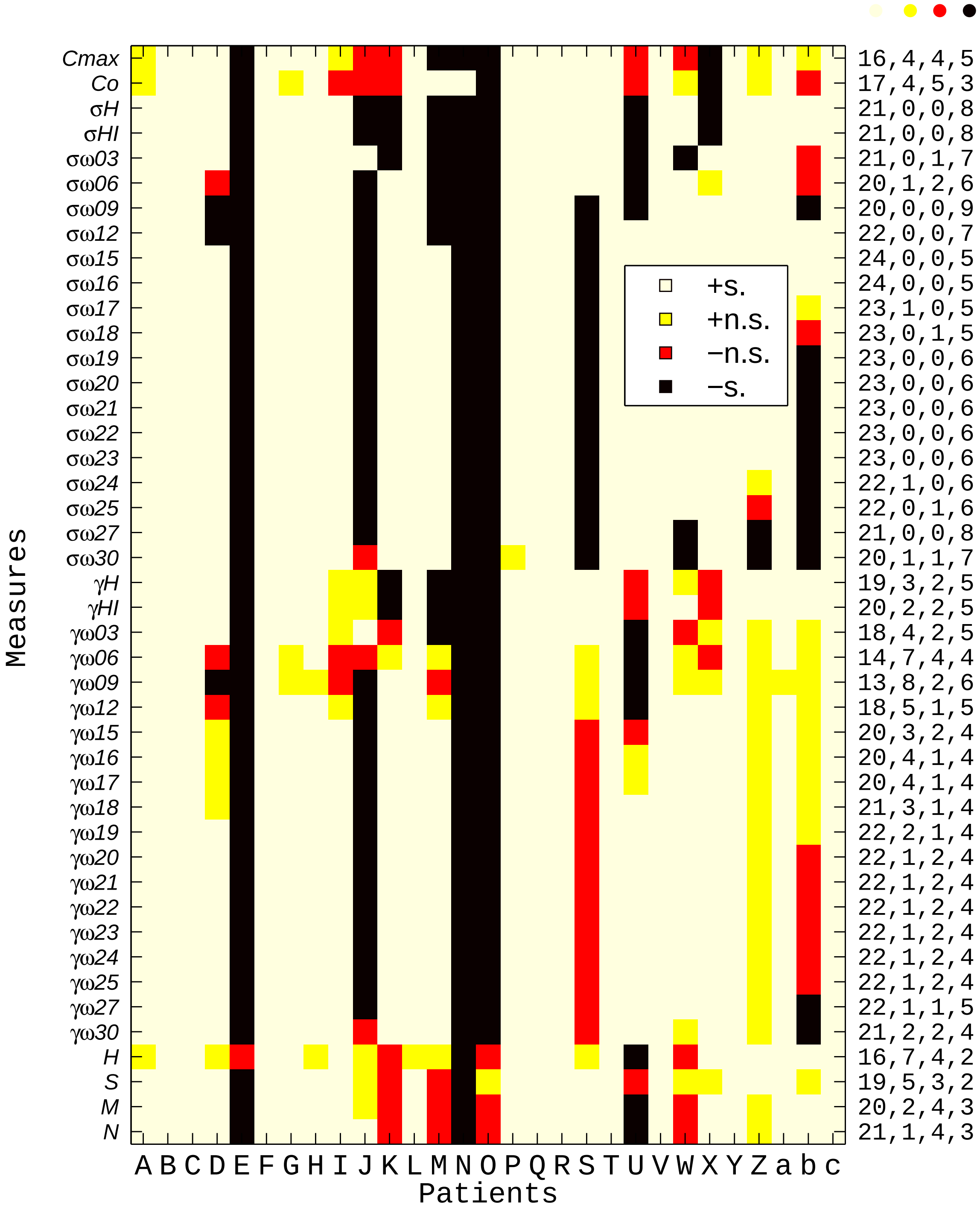

Figure 4.7: Results of the localization analysis in combination with the time-shifted surrogate data test. All notations as in the Fig. 4.5.

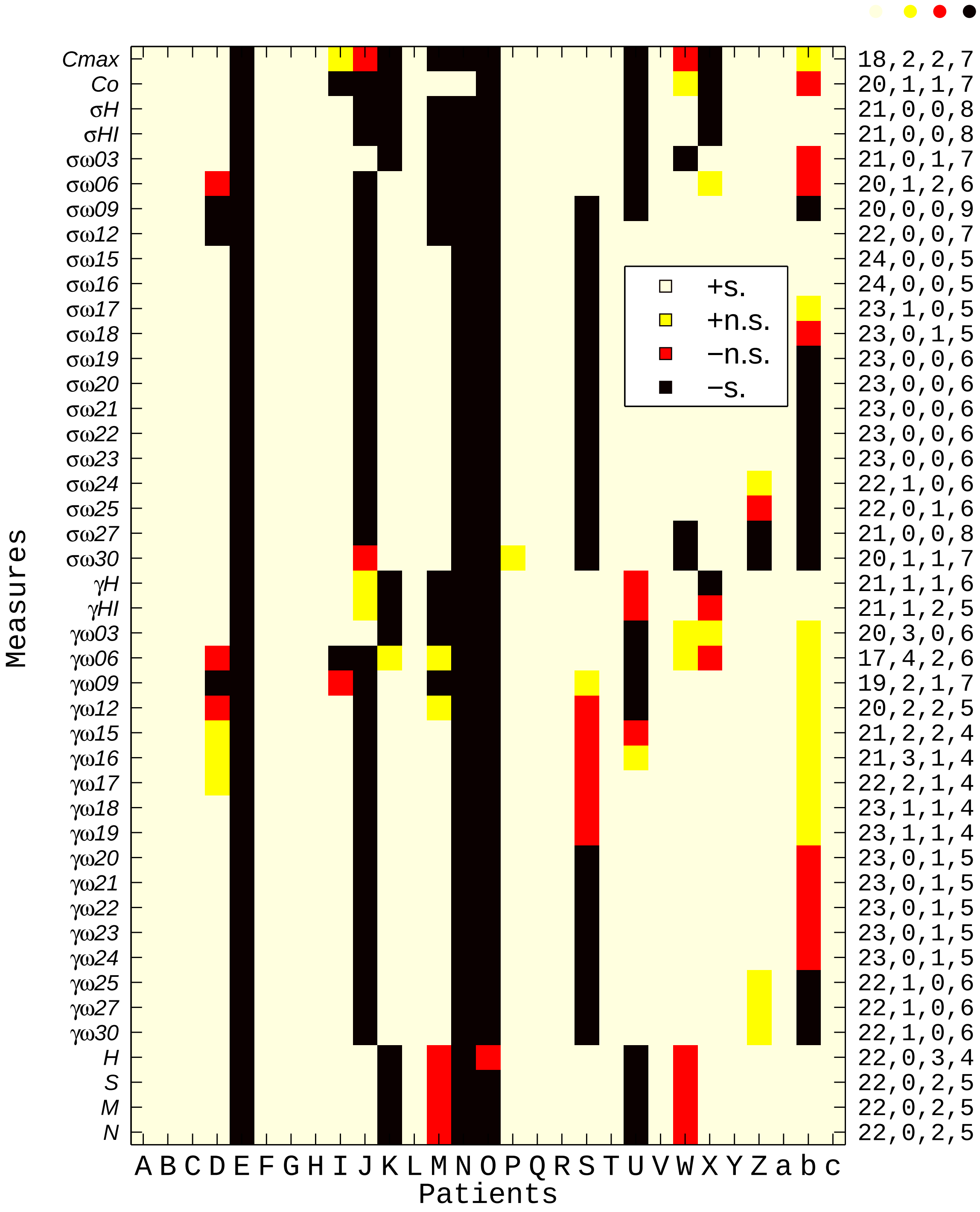

Figure 4.8: Results of the localization analysis in combination with the time-shifted surrogate data test. Only significant windows are included in averaging. All notations as in the Fig. 4.5.

for linear cross-correlation measures. The smallest number of correctly classified cases (17 patients) is again found for $\gamma\omega 6$. The highest number (24 patients) is obtained for the index based on Shannon entropy in the frequency range $15 - 16$ Hz. For higher frequencies of $18 - 23$ Hz the results for both indices have just one significant correctly localized case less. The difference in the performance of all four measures of generalized synchronization which can be seen in Fig. 4.7 disappears after excluding all nonsignificant windows from the average.

4.3 Discussion

In this study we addressed the question whether it is possible to identify the location of an epileptic focus from an intracranial EEG recorded during interictal (seizure-free) intervals. Based on theoretical considerations and motivated by earlier studies we tested the hypothesis that during the interictal period the average synchronization in the focal hemisphere of the brain of epileptic patients is higher than in the non-focal hemisphere. Our comparative study of many different bivariate synchronization measures along with a standard statistical validation confirmed this hypothesis. Moreover, the results obtained with time-shifted surrogate analysis showed that our results are indeed caused by synchronization phenomena rather than by some other overall properties of the dynamics.

A comparative study showed high correlations between all synchronization measures. Comparatively low correlation between one variant of the generalized synchronization H measure and phase synchronization measures based on the wavelet transform in combination with their correct localization of the epileptic focus for different patients could be used for a discriminant analysis using both measures.

The obtained results could not compete in their performance with EEG expert readers who perform a visual inspection of ictal-EEG recordings, i.e., recordings containing seizure. However, our results were obtained without the necessity of observing seizure activity. On the other hand, we have to admit that the results obtained for a partly overlapped data base of patients using univariate analysis in combination with surrogate data correction show better performance [9, 7].

Beyond any doubts, the prediction of epileptic seizures is another challenging task in the analysis of EEG recordings from epilepsy patients (for an overview see Refs. [67, 68, 63]). Recently it was found that bivariate techniques are superior to univariate techniques for this task [75, 74, 76, 73, 54]. One of the distinctive features of these studies along with comparison of many univariate and bivariate measures is a careful statistical validation of the results. This validation is also based on the concept of surrogate data, namely seizure time surrogates [8] and measure profile surrogates [55, 53].

Chapter 5

Clustering of data

In this Chapter a new algorithm for data clustering is presented [50, 51]. We start with a short introduction to clustering in general.

Classification or organizing of data is very important in many scientific disciplines. It is one of the most fundamental mechanism of understanding and learning [43, 27]. Depending on the problem, classification can be exclusive or overlapping, supervised or unsupervised. In the following we will be interested only in exclusive unsupervised classification. This type of classification is usually called clustering or cluster analysis.

An instance of a clustering problem consists of a set of objects and a set of properties (called characteristic vector) for each object. The main goal of clustering is the division of objects into groups using only the characteristic vectors. Cluster analysis organizes data either as a single grouping of individuals into non-overlapping clusters or as a hierarchy of nested partitions. The first approach is called partitional clustering (PC), the second one is the hierarchical clustering (HC). One of the main features of HC methods is the visual impact of the *dendrogram* which enables one to see how objects are merged into clusters. From any HC one can obtain a PC by restricting oneself to a “horizontal” cut through the dendrogram, while one cannot go in the other direction and obtain a full hierarchy from a single PC. Because of the wide spread of their applications, there are a large variety of different clustering methods in usage, see, e.g., Refs. [43, 27] for an overview.

The crucial point of all clustering algorithms is the choice of a *proximity measure*. This is obtained from the characteristic vectors and can be either an indicator for similarity (i.e., large for similar and small for dissimilar objects), or a distance-like quantity. In the latter case it is convenient but not obligatory if it satisfies the standard axioms of a metric (positivity, symmetry, and triangle inequality). A matrix of all pairwise proximities is called proximity matrix. Among HC methods one should distinguish between those where one uses the characteristic vectors only at the first level of the hierarchy and derives the proximities between clusters from the proximities of their constituents, and methods where the proximities are calculated each time from their characteristic vectors. The latter strategy (which is used also in the present Chapter) allows of course for more flexibility but might also be computationally costlier.

Quite generally, the “objects” to be clustered can be either single (finite) patterns (e.g., DNA sequences) or random variables, i.e., *probability distributions*. In the latter case the data are usually supplied in form of a statistical sample, and one of the simplest and most widely used similarity measures is the linear (Pearson) correlation coefficient. But this is not sensitive to nonlinear dependencies which do not manifest themselves in the covariance and can thus miss important features. This is in contrast to mutual information (MI) described in Sec. 3.2.

This Chapter is organized as follows. In Sec. 5.1 the relation of MI in Shannon sense to the MI in Kolmogorov theory is described and the MI clustering algorithm, called MIC, is presented. In Sec. 5.2 MI distance measure for the Shannon case is introduced and in Section 5.3 two applications of the new clustering algorithm are discussed. In the last Section 5.4 we present a short discussion about the place of MIC algorithm among other clustering algorithms.

5.1 New clustering algorithm

One of the important features of MI is that it has also an “algorithmic” cousin, defined within algorithmic (Kolmogorov) information theory [66] which measures the similarity between individual objects.

Let us recall some important facts of algorithmic information theory. In contrast to Shannon theory where the basic objects are random variables and entropies are *average* informations, algorithmic information theory deals with individual symbol strings and with the actual information needed to specify them. To “specify” a sequence X means here to give the necessary input to a universal computer U, such that U prints X on its output and stops. The analogon to entropy, called here usually the *complexity* $K(X)$ of X, is the minimal length of an input which leads to the output X, for fixed U. It depends on U, but it can be shown that this dependence is weak and can be neglected in the limit when $K(X)$ is large [66].

Let us denote the concatenation of two strings X and Y as XY. Its complexity is $K(XY)$. It is intuitively clear that $K(XY)$ should be larger than $K(X)$ but cannot be larger than the sum $K(X) + K(Y)$. Finally, one expects that $K(X|Y)$, defined as the minimal length of a program printing X when Y is furnished as an auxiliary input, is related to $K(XY) - K(Y)$. Indeed, one can show [66] (again within correction terms which become irrelevant asymptotically) that

$$0 \leq K(X|Y) \simeq K(XY) - K(Y) \leq K(X). \tag{5.1}$$

Notice the close similarity with Shannon entropy.

The algorithmic information in Y about X is finally defined as

$$I_{\rm alg}(X,Y) = K(X) - K(X|Y) \simeq K(X) + K(Y) - K(XY). \tag{5.2}$$

Within the same additive correction terms, one shows that it is symmetric, $I_{\rm alg}(X,Y) = I_{\rm alg}(Y,X)$, and can thus serve as an analogon to mutual information.

Using Turing's proof of the halting theorem one can show that $K(X)$ is in general not computable. But one can easily give upper bounds. Indeed, the length of any input which produces X (e.g., by spelling it out verbatim) is an upper bound. Improved upper bounds are provided by any file compression algorithm such as gnuzip or UNIX "compress". Good compression algorithms will give good approximations to $K(X)$, and algorithms whose performance does not depend on the input file length (in particular since they do not segment the file during compression) will be crucial for the following.

Another feature of MI which is essential for the present application is the *grouping property*: The MI between three objects (distributions) $X, Y,$ and Z is equal to the sum of the MI between X and Y, plus the MI between Z and the combined object (joint distribution) (XY),

$$I(X,Y,Z) = I(X,Y) + I((X,Y),Z). \tag{5.3}$$

Within Shannon information theory this is an exact theorem (see below), while it is true in the algorithmic version up to the usual logarithmic correction terms [66]. Since $X, Y,$ and Z can be themselves composite, Eq.(5.3) can be used recursively for a cluster decomposition of MI. This motivates the main idea of our clustering method: instead of using, e.g., centers of masses in order to treat clusters like individual objects in an approximative way only, we treat them exactly like individual objects when using MI as proximity measure.

More precisely, we propose the following scheme for clustering n objects with MIC:

(1) Compute a proximity matrix based on pairwise mutual informations; assign n clusters such that each cluster contains exactly one object;

(2) find the two closest clusters i and j;

(3) create a new cluster (ij) by combining i and j;

(4) delete the lines and columns with indices i and j from the proximity matrix, and add one line/column containing the proximities between cluster (ij) and all other clusters;

(5) if the number of clusters is still > 2, goto (2); else join the two clusters and stop.

5.2 Mutual information distance measure

Mutual information itself is a similarity measure in the sense that small values imply large "distances" in a loose sense. But it would be useful to modify it such that the resulting quantity is a metric in the strict sense, i.e., satisfies the triangle inequality. Indeed, the first such metric is well known (see e.g the problem 15 of Chapter 2 in Ref. [21]): The quantity

$$d(X,Y) = H(X|Y) + H(Y|X) = H(X,Y) - I(X,Y) \tag{5.4}$$

satisfies the triangle inequality, in addition to being non-negative and symmetric and to satisfying $d(X,X) = 0$. The proof proceeds by first showing that for any Z

$$H(X|Y) \leq H(X,Z|Y) \leq H(X|Z) + H(Z|Y). \tag{5.5}$$

But $d(X,Y)$ is not optimal for our purposes. Since we want to compare the proximity between two single objects and that between two clusters containing maybe many objects, one would like the distance measure to be unbiased by the sizes of the clusters. As argued forcefully in Refs. [64, 65], this is not true for $I_{\text{alg}}(X,Y)$, and for the same reasons it is not true for $I(X,Y)$ or $d(X,Y)$ either: A mutual information of 1000 bits should be considered as large, if X and Y themselves are just 1000 bits long, but it should be considered as very small, if X and Y are huge.

As shown in Refs. [64, 65] within the algorithmic framework, one can form two different distances which measure *relative* distance, i.e., which are normalized by dividing by a total entropy. We sketch here only the theorems and proofs for the Shannon version, they are indeed very similar to their algorithmic analoga in Refs. [64, 65][1].

THEOREM 1: The quantity

$$D(X,Y) = 1 - \frac{I(X,Y)}{H(X,Y)} = \frac{d(X,Y)}{H(X,Y)} \tag{5.6}$$

is a metric, with $D(X,X) = 0$ and $D(X,Y) \leq 1$ for all pairs (X,Y).

PROOF: Symmetry, positivity and boundedness are obvious. Since $D(X,Y)$ can be written as

$$D(X,Y) = \frac{H(X|Y)}{H(X,Y)} + \frac{H(Y|X)}{H(Y,X)}, \tag{5.7}$$

it is sufficient for the proof of the triangle inequality to show that each of the two terms on the r.h.s. is bounded by an analogous inequality, i.e.,

$$\frac{H(X|Y)}{H(X,Y)} \leq \frac{H(X|Z)}{H(X,Z)} + \frac{H(Z|Y)}{H(Z,Y)} \tag{5.8}$$

and similarly for the second term. Eq.(5.8) is proven straightforwardly, using Eq.(5.5) and the basic inequalities $H(X) \geq 0$, $H(X,Y) \leq H(X,Y,Z)$ and $H(X|Z) \geq 0$:

$$\frac{H(X|Y)}{H(X,Y)} \;=\; \frac{H(X|Y)}{H(Y)+H(X|Y)} \leq \frac{H(X|Z)+H(Z|Y)}{H(Y)+H(X|Z)+H(Z|Y)}$$

[1] We recently found out that this was first proven in Ref. [70].

$$
\begin{aligned}
&= \frac{H(X|Z)+H(Z|Y)}{H(X|Z)+H(Y,Z)} \leq \frac{H(X|Z)}{H(X|Z)+H(Z)} + \frac{H(Z|Y)}{H(Y,Z)} \\
&= \frac{H(X|Z)}{H(X,Z)} + \frac{H(Z|Y)}{H(Z,Y)}. \qquad (5.9)
\end{aligned}
$$

THEOREM 2: The quantity

$$
D'(X,Y) = 1 - \frac{I(X,Y)}{\max\{H(X),H(Y)\}} = \frac{\max\{H(X|Y),H(Y|X)\}}{\max\{H(X),H(Y)\}} \qquad (5.10)
$$

is also a metric, also with $D'(X,X) = 0$ and $D'(X,Y) \leq 1$ for all pairs (X,Y). It is sharper than D, i.e., $D'(X,Y) \leq D(X,Y)$.

PROOF: Again we have only to prove the triangle inequality, the other parts being trivial. For this we have to distinguish different cases [65].
Case 1: $\max\{H(Z),H(Y)\} \leq H(X)$. Using Eq.(5.5) we obtain

$$
\begin{aligned}
D'(X,Y) &= \frac{H(X|Y)}{H(X)} \leq \frac{H(X|Z)}{H(X)} + \frac{H(Z|Y)}{H(Y)} \\
&= D'(X,Z) + D'(Z,Y). \qquad (5.11)
\end{aligned}
$$

Case 2: $\max\{H(Z),H(X)\} \leq H(Y)$. This is completely analogous.
Case 3: $H(X) \leq H(Y) < H(Z)$. We now have to show that

$$
\begin{aligned}
D'(X,Y) &= \frac{H(Y|X)}{H(Y)} \leq \frac{H(Y|Z)+H(Z|X)}{H(Y)} \qquad (5.12) \\
&\overset{?}{\leq} D'(X,Z) + D'(Z,Y) = \frac{H(Z|X)}{H(Z)} + \frac{H(Z|Y)}{H(Z)}.
\end{aligned}
$$

Indeed, if the r.h.s. of the first line is less than 1, then

$$
\begin{aligned}
\frac{H(Y|X)}{H(Y)} &\leq \frac{H(Y|Z)+H(Z|X)}{H(Y)} \\
&\leq \frac{H(Y|Z)+H(Z|X)+H(Z)-H(Y)}{H(Z)} \\
&= \frac{H(Z|Y)+H(Z|X)}{H(Z)}, \qquad (5.13)
\end{aligned}
$$

and Eq.(5.12) holds. If it is larger than 1, then also $(H(Z|Y)+H(Z|X))/H(Z) \geq 1$. Eq.(5.12) must now also hold, since $H(Y|X)/H(Y) \leq 1$.
Case 4: $H(Y) \leq H(X) < H(Z)$. This is completely analogous to case 3.

Apart from scaling correctly with the total information, in contrast to $d(X,Y)$, the algorithmic analog to $D'(X,Y)$ is also *universal* [65], while $D(X,Y)$ is universal up to a factor 2.

Essentially this means that if $X \approx Y$ according to any non-trivial distance measure, then $X \approx Y$ also according to D and D'. In contrast to the other properties of D and D', this is not easy to carry over from algorithmic to Shannon theory. The proof in Ref. [65] depends on X and Y being discrete, which is obviously not true for probability distributions. Based on the universality argument, it was argued in Ref. [65] that D' should be superior to D, but the numerical studies shown in that reference did not show a clear difference between them. In the following we shall therefore use primarily D for simplicity, but we checked that using D' did not give systematically better results.

A major difficulty appears in the Shannon framework, if we deal with continuous random variables. As we mentioned in Sec. 3.2, Shannon informations are only finite for coarse-grained variables, while they diverge if the resolution tends to zero. This means that dividing MI by entropy as in the definitions of D and D' becomes problematic. One has essentially two alternative possibilities. The first is to actually introduce some coarse-graining, although it would have not been necessary for the definition of $I(X, Y)$, and divide by the coarse-grained entropies. This introduces an arbitrariness, since the scale Δ is completely ad hoc, unless it can be fixed by some independent arguments. We have found no such arguments, and thus we propose the second alternative. There we take $\Delta \to 0$. In this case $H(X) \sim m_x \log \Delta$, with m_x being the dimension of X. In this limit D and D' would tend to 1. But using similarity measures

$$S(X,Y) = (1 - D(X,Y)) \log(1/\Delta), \tag{5.14}$$

$$S'(X,Y) = (1 - D'(X,Y)) \log(1/\Delta) \tag{5.15}$$

instead of D and D' gives *exactly* the same results in MIC, and

$$S(X,Y) = \frac{I(X,Y)}{m_x + m_y}, \quad S'(X,Y) = \frac{I(X,Y)}{\max\{m_x, m_y\}}. \tag{5.16}$$

Thus, when dealing with continuous variables, we divide the MI either by the sum or by the maximum of the dimensions. When starting with scalar variables and when X is a cluster variable obtained by joining m elementary variables, then its dimension is just $m_x = m$.

5.3 Applications of MIC

5.3.1 Mitochondrial DNA

As a first application, we study the mitochondrial DNA of a group of 34 mammals (see Fig. 5.1). Exactly the same data [1] had previously been analyzed in Refs. [64, 96]. This group includes among others[2] some rodents and related species[3], ferungulates[4], and primates[5]. It had been chosen in Ref. [64] because of doubts about the relative order among these three groups [19, 96].

Obviously, we are here dealing with the algorithmic version of information theory, and informations are estimated by lossless data compression. For constructing the proximity matrix between individual taxa, we proceed essentially as in Ref. [64] using the special compression program GenCompress [2] to estimate the complexity.

In Ref. [64], this proximity matrix was then used as the input to a standard HC algorithm (neighbor-joining and hypercleaning) to produce an evolutionary tree. It is here where our treatment deviates crucially. We used the MIC algorithm described in the beginning of this section, with distance $D(X,Y)$. The joining of two clusters (the third step in the MIC algorithm) is obtained by simply concatenating the DNA sequences. There is of course an arbitrariness in the order of concatenation. For sequence length $\to \infty$, sequences XY and YX would have the same information, but in reality they will in general give compressed sequences of different lengths. As more and more sequences are concatenated in larger clusters, this problem becomes more and more severe. In a pre-analysis step we could rule out any influence of this effect on our results. The resulting evolutionary tree obtained with Gencompress is shown in Fig. 5.1.

The overall structure of this tree closely resembles the one shown in Ref. [96]. All primates are correctly clustered and also the relative order of the ferungulates is in accordance with Ref. [96]. On the other hand, there are a number of connections which obviously do not reflect the true evolutionary tree, see for example the guinea pig with bat, and elephant with platypus. But the latter two, inspite of being joined together, have a very large distance from each other, thus their clustering just reflects the fact that neither the platypus nor the elephant have other close relatives in the sample. Thus we expect that our method would work better with a larger sample where families are represented by more species and thus better defined. All in all, however, already the results shown in Fig. 5.1 capture surprisingly well the overall structure shown in Ref. [96]. Notice that dividing MI by the total information is essential for this success. If we had used instead $I_{\rm alg}(X,Y)$ itself or the non-normalized distance $d(X,Y)$ defined in Eq.(5.4), the clustering algorithm used in Ref. [64] would not have changed much, since all 34 DNA sequences have roughly the

[2] opossum (*Didelphis virginiana*), wallaroo (*Macropus robustus*), and platypus (*Ornithorhyncus anatinus*)

[3] rabbit (*Oryctolagus cuniculus*), guinea pig (*Cavia porcellus*), fat dormouse (*Glis glis*), rat (*Rattus norvegicus*), squirrel (Scuirus vulgaris), and mouse (*Mus musculus*)

[4] horse (*Equu caballus*), donkey (*Equus asinus*), Indian rhinoceros (*Rhinoceros unicornis*), white rhinoceros (*Ceratotherium simum*), harbor seal (*Phoca vitulina*), grey seal (*Halichoerus grypus*), cat (*Felis catus*), dog (*Canis familiaris*), fin whale (*Balenoptera physalus*), blue whale (*Balenoptera musculus*), cow (*Bos taurus*), sheep (*Ovis aries*), pig (*Sus scrofa*), hippopotamus (*Hippopotamus amphibius*), neotropical fruit bat (*Artibeus jamaicensis*), African elephant (*Loxodonta africana*), aardvark (*Orycteropus afer*), and armadillo (*Dasypus novemcintus*)

[5] human (*Homo sapiens*), common chimpanzee (*Pan troglodytes*), pigmy chimpanzee (*Pan paniscus*), gorilla (*Gorilla gorilla*), orangutan (*Pongo pygmaeus*), gibbon (*Hylobates lar*), and baboon (*Papio hamadryas*)

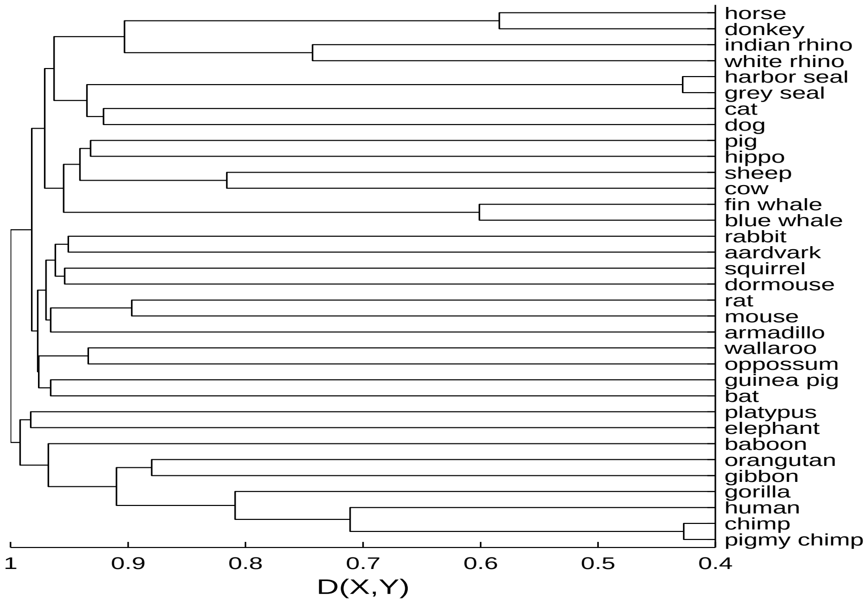

Figure 5.1: Phylogenetic tree for 34 mammals (31 eutherians plus 3 non-placenta mammals). In contrast to Fig. 5.4, the heights of nodes are equal to the distances between the joining daughter clusters.

same length and the same information content. But our MIC algorithm would have been completely screwed up: After the first cluster formation, we have DNA sequences of very different lengths to compare with. If we use MI itself as a similarity measure, we would mainly join large clusters (since they tend to have large MI). If we would use $d(X, Y)$, we would mainly join small clusters since they have smaller distances.

A heuristic reasoning for the use of MIC for the reconstruction of an evolutionary tree might be given as follows: Suppose that a proximity matrix has been calculated for a set of DNA sequences and the smallest distance is found for the pair (X, Y). Ideally, one would remove the sequences X and Y, replace them by the sequence of the common ancestor (say Z) of the two species, update the proximity matrix to find the smallest entry in the reduced set of species, and so on. But the DNA sequence of the common ancestor is not available. One solution might be that one tries to reconstruct it by making some compromise between the sequences X and Y. Instead, we essentially propose to concatenate the sequences X and Y. This will of course not lead to a plausible sequence of the common ancestor, but it will *optimally represent the information* about the common ancestor. During the evolution since the time of the ancestor Z, some parts of its genome might have changed both in X and in Y. These parts are of little use in constructing any phylogenetic tree. Other parts might not have changed in either. They are recognized anyhow by any sensible algorithm. Finally, some parts of its genome will have mutated significantly in X but not in Y, and vice versa. This information is essential to find the correct way through higher hierarchy levels of the evolutionary tree, and it is preserved in concatenating.

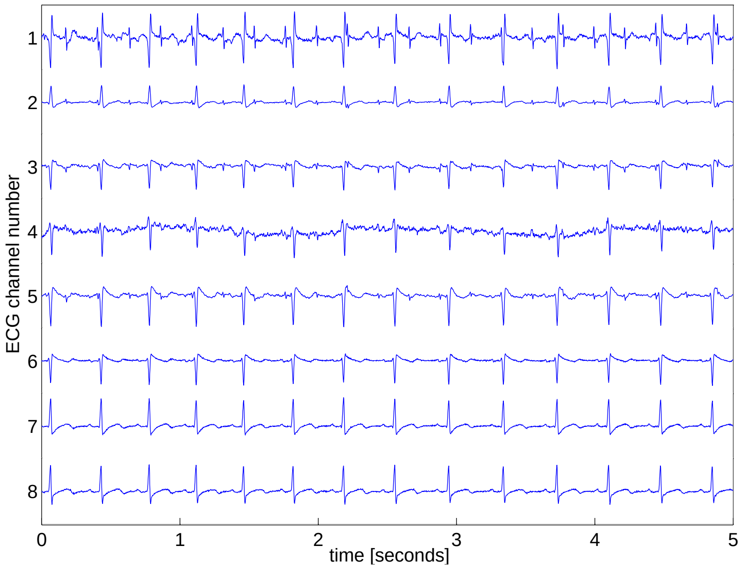

Figure 5.2: ECG of a pregnant woman.

5.3.2 Minimally dependent components in electrocardiograms

As our second application we choose a case where Shannon theory is the proper setting. We show in Fig. 5.2 an electrocardiogram (ECG) recorded from the abdomen and thorax of a pregnant woman [24] (8 channels, sampling rate 500 Hz, 5 s total). It is already seen from this graph that there are at least two important components in this ECG: the heartbeat of the mother, with a frequency of ≈ 3 beat/s, and the heartbeat of the fetus with roughly twice this frequency. Both are not synchronized. In addition there is noise from various sources (muscle activity, measurement noise, etc.). While it is easy to detect anomalies in the mother's ECG from such a recording, it would be difficult to detect them in the fetal ECG.

As a first approximation we can assume that the total ECG is a linear superposition of several independent sources (mother, child, noise$_1$, noise$_2$,...). A standard method to disentangle such superpositions is *independent component analysis* (ICA) [42]. In the simplest case one has n independent sources $s_i(t),\ i = 1 \ldots n$ and n measured channels $x_i(t)$ obtained by instantaneous superpositions with a time independent non-singular matrix $\mathbf{A}$,

$$x_i(t) = \sum_{j=1}^{n} A_{ij} s_j(t) \ . \qquad (5.17)$$

In this case the sources can be reconstructed by applying the inverse transformation $\mathbf{W} = \mathbf{A}^{-1}$ which is obtained by minimizing the (estimated) mutual informations between the transformed components $y_i(t) = \sum_{j=1}^{n} W_{ij} x_j(t)$. If some of the sources are Gaussian, this leads to ambiguities [42], but it gives a unique solution if the sources have more structure.

In reality things are not so simple. For instance, the sources might not be independent, the number of sources (including noise sources!) might be different from the number of channels, and the mixing might involve delays. For the present case this implies that the heartbeat of the mother is seen in several reconstructed components y_i, and that the "independent" components are not independent at all. In particular, all components y_i which have large contributions from the mother form a cluster with large intra-cluster MIs and small inter-cluster MIs. The same is true for the fetal ECG, albeit less pronounced. It is thus our aim to

1) optimally decompose the signals into least dependent components;
2) cluster these components hierarchically such that the most dependent ones are grouped together;
3) decide on an optimal level of the hierarchy, such that the clusters make most sense physiologically;
4) project onto these clusters and apply the inverse transformations to obtain cleaned signals for the sources of interest.

Technically we proceeded as follows [117]:

Since we expect different delays in the different channels, we first used Takens delay embedding [118] with time delay 0.002 s and embedding dimension 3, resulting in 24 channels. We then formed 24 linear combinations $y_i(t)$ and determined the de-mixing coefficients W_{ij} by minimizing the overall mutual information between them, using the MI estimator described in Sec. 3.2.2. There, two classes of estimators were introduced, one with square and the other with rectangular neighborhoods. Within each class, one can use the number of neighbors, called k in the following, on which the estimate is based. Small values of k lead to a small bias but to large statistical errors, while the opposite is true for large k. But even for very large k the bias is zero when the true MI is zero, and it is systematically such that absolute values of the MI are underestimated. Therefore, this bias affects neither the ranking of the pairwise MIs nor the determination of the optimal de-mixing matrix. But it depends on the dimension of the random variables, therefore large values of k are not suitable for the clustering. We thus proceeded as follows: We first used $k = 100$ and square neighborhoods to obtain the least dependent components $y_i(t)$, and then used $k = 3$ with rectangular neighborhoods for the clustering. The resulting least dependent components are shown in Fig. 5.3. They are sorted such that the first components $(1 - 5)$ are dominated by the mother's ECG, while the next three contain large contributions from the fetus. The rest contains mostly noise, although some seem to be still mixed.

These results obtained by visual inspection are fully supported by the cluster analysis. The dendrogram is shown in Fig. 5.4. In constructing it we used $S(X, Y)$ (Eq.(5.16)) as similarity measure to find the correct topology. Again we would have obtained much worse results if we had not normalized it by dividing MI by $m_X + m_Y$. In plotting the actual dendrogram, however, we used the MI of the cluster to determine the height at which the

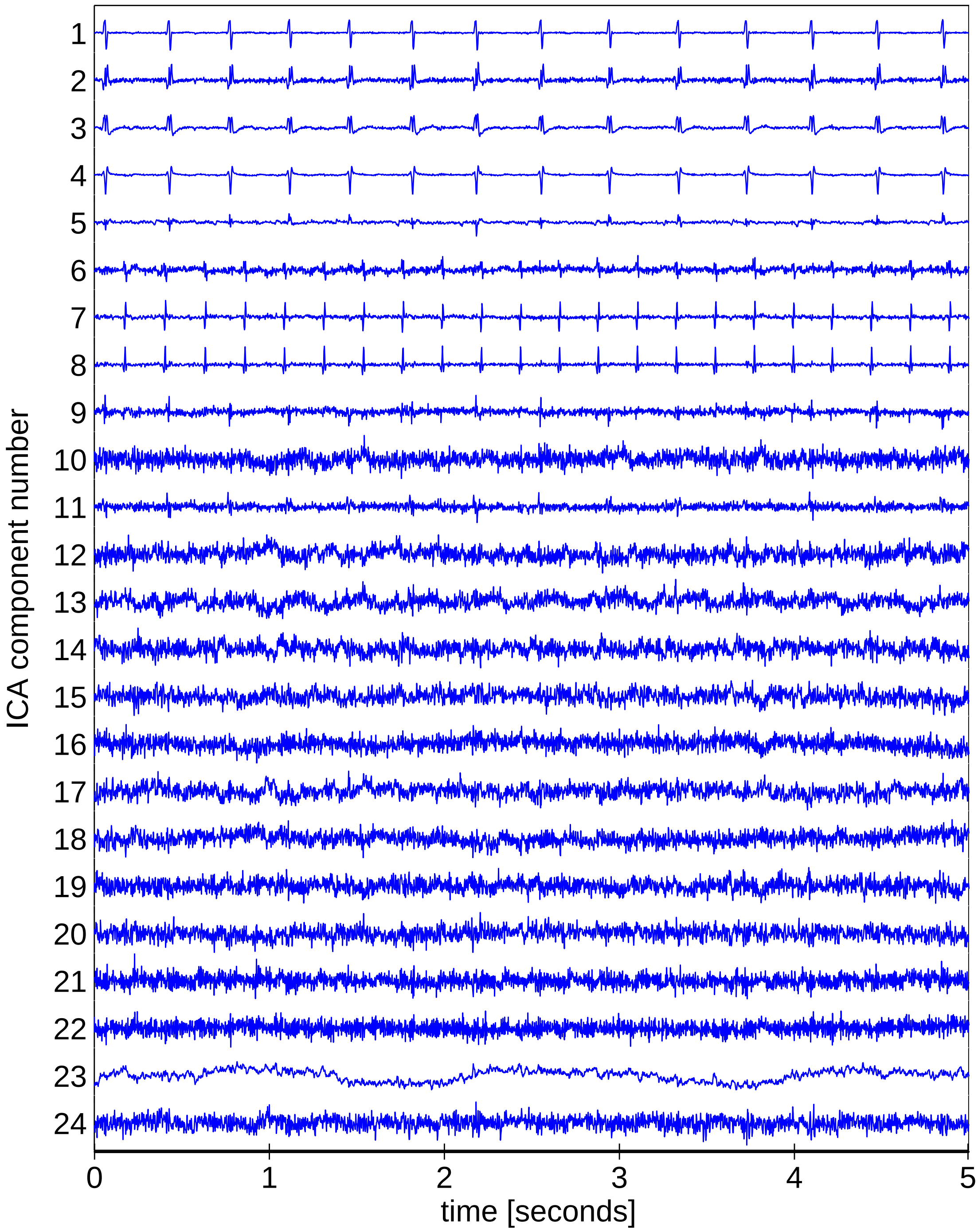

Figure 5.3: Least dependent components of the ECG shown in Fig. 5.2, after increasing the number of channels by delay embedding.

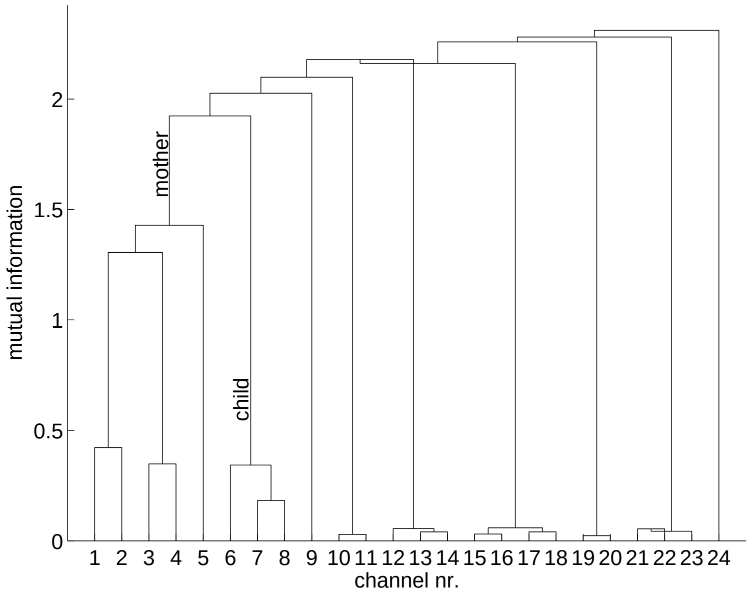

Figure 5.4: Dendrogram for least dependent components. The height where the two branches of a cluster join corresponds to the MI of the cluster.

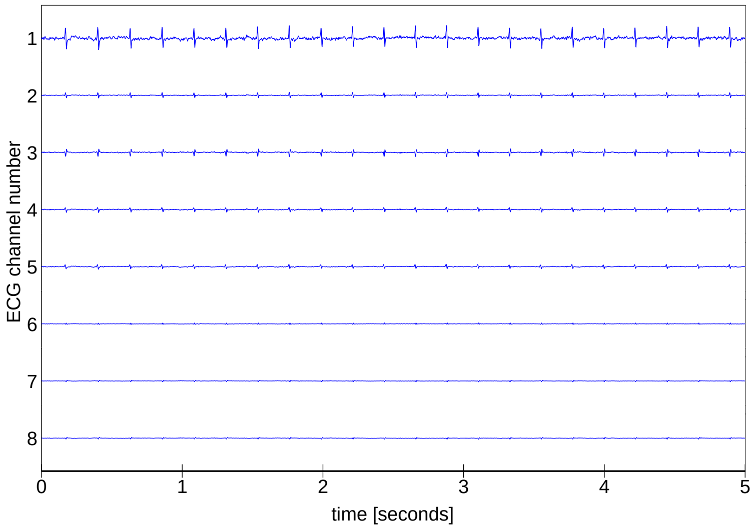

Figure 5.5: Original ECG where all contributions except those of the child cluster have been removed.

two daughters join. The MI of the first five channels, e.g., is ≈ 1.44, while that of channels 6 to 8 is ≈ 0.3. For any two clusters (tuples) $X = X_1 \dots X_n$ and $Y = Y_1 \dots Y_m$ one has $I(X,Y) \geq I(X) + I(Y)$. This guarantees, if the MI is estimated correctly, that the tree is drawn properly. The two slight glitches (when clusters $(1-14)$ and $(15-18)$ join, and when $(21-22)$ is joined with 23) result from small errors in estimating MI. They do in no way effect our conclusions.

In Fig. 5.4 one can clearly see two big clusters corresponding to the mother and to the child. There are also some small clusters which should be considered as noise. For reconstructing the mother and child contributions to Fig. 5.2, we have to decide on one specific clustering from the entire hierarchy. We decided to make the cut at inter-cluster MI equal to 0.1, i.e., two clusters X and Y are joined whenever $I((X),(Y)) \equiv I(X,Y) - I(X) - I(Y) \geq 0.1$. The resulting mother and child clusters are indicated in Fig. 5.4 and were already anticipated in sorting the channels. Reconstructing the original ECG from the child components only, we obtain Fig. 5.5.

5.3.3 One more application of MIC

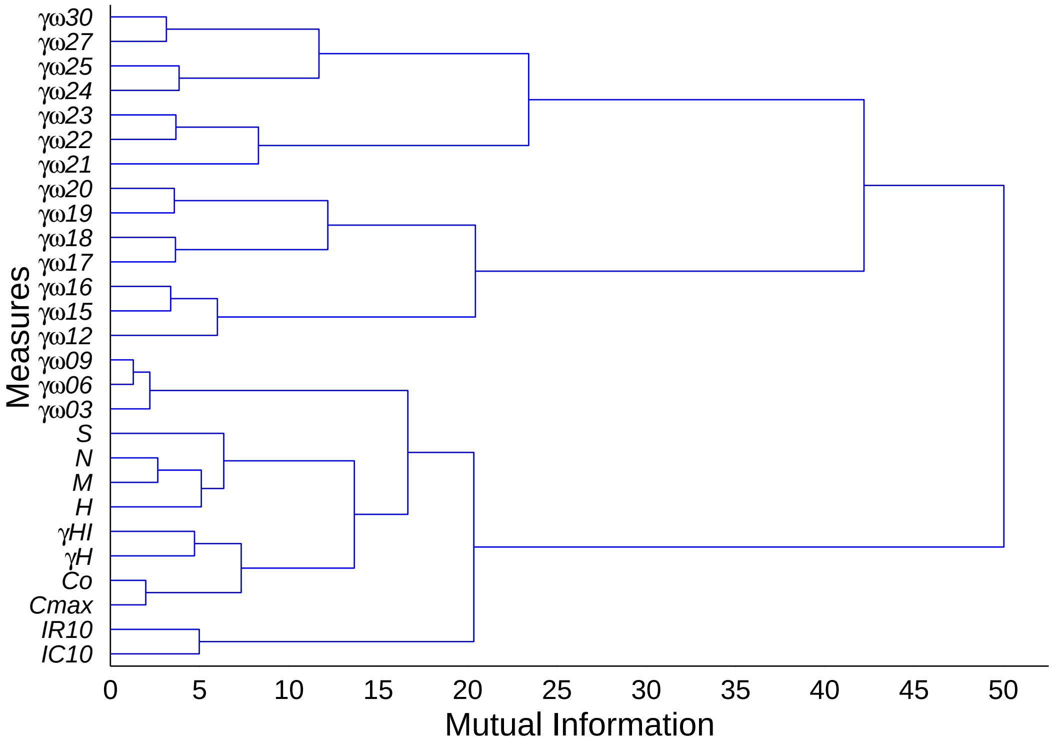

Figure 5.6: Dendrogram for different synchronization measures.

A small application triggered by this and previous Chapters is presented here. In Sec. 4.2.1.4 we investigated the correlations between different synchronization measures using correlation coefficients between average synchronization values (see Fig. 4.6). We drew some conclusions about closeness of these measures based only on correlation matrix, but these conclusions were only of qualitative nature. Here we apply the hierarchical clustering algorithm described in this Chapter to the same data. A tree obtained with $k = 5$ nearest neighbors and the “rectangular” algorithm is presented in Fig. 5.6. Two estimators of mutual information ($IC10$ and $IR10$) constitute the first cluster, the next one contains the two indices of phase synchronization based on the Hilbert transform and calculated for original and rank ordered phase. Then, the measures of phase synchronization based on the wavelet transform start to merge in clusters. The measures of generalized synchronization constitute their own cluster which starts with merging N and M joint subsequently by H and S. It is interesting that the phase synchronization indices based on wavelet transform for the frequency range $3 - 9$ Hz ($\gamma\omega 3 - \gamma\omega 9$) merge not to the “wavelet” but rather to the cluster of all other measures. This fact illustrates once again the poor performance in this frequency range. It is remarkable that measures $\gamma\omega 12$ to $\gamma\omega 20$ which gave the best performance (see the previous Chapter) form one cluster.

5.4 Discussion

We have shown that MI can not only be used as a proximity measure in clustering, but that it also suggests a conceptually very simple and natural hierarchical clustering algorithm. We do not claim that this algorithm, called *mutual information clustering* (MIC), is always superior to other algorithms. Indeed, MI is in general not easy to estimate. Obviously, when only crude estimates are possible, MIC will not give very good results too. But as MI estimates become better, the results of MIC should also improve. The present work was partly triggered by our investigations of a new class of MI estimators for continuous random variables which have very small bias and also rather small variances (see Sec. 3.2.2).

We have illustrated our method with several applications, one from genetics and one from cardiology. For neither application MIC might give the very best clustering, but it seems interesting that one common method gives decent results for both, although they are very different.

The results of MIC should improve, if more data become available. This is trivial, if we mean by that longer time sequences in the application to ECG, and longer parts of the genome in the application of Sec. 5.3.1. It is less trivial that we expect MIC to make fewer mistakes in a phylogenetic tree, when more species are included. The reason is that close-by species will be correctly joined anyhow, and families – which now are represented only by single species and thus are poorly characterized – will be much better described by the concatenated genomes if more species are included.

There are two versions of information theory, algorithmic and probabilistic, and therefore there are also two variants of both MI and MIC. We have discussed in detail one application of each, and shown that indeed common concepts were involved in both. In particular it was crucial to normalize MI properly, so that it is essentially the *relative* MI which is used as proximity measure. For conventional clustering algorithms using algorithmic MI as proximity measure this had already been stressed in Refs. [64, 65], but it is even more important for MIC, both in the algorithmic and in the probabilistic version.

In the probabilistic version, one studies the clustering of probability distributions. However, usually distributions are not provided as such, but are given implicitly by finite random samples drawn (more or less) independently from them. On the other hand, the full power of algorithmic information theory is only reached for infinitely long sequences, and in this limit any individual sequence defines a sequence of probability measures on finite subsequences. Thus the strict distinction between the two theories is somewhat blurred in practice. Nevertheless, one should not confuse the similarity between two sequences (two English books, say) and that between their subsequence statistics. Whereas two sequences are maximally different if they are completely random, their statistics for short subsequences is then identical (all subsequences appear in both with equal probabilities). Thus one should always be aware of what similarities or independencies one is looking for. The fact that MI can be used in similar ways for all these problems is not trivial.

Chapter 6

Summary and Outlook

Synchronization is among the most important phenomena in many branches of natural sciences, engineering and life sciences [89]. Very often the only information to investigate this phenomenon is available in form of time series measured from the systems under consideration. The analysis of biological systems is a prominent example of such a setting. Therefore, to develop and to improve measures which are able to retrieve reliable information about synchronization from time series is of great importance for the understanding of synchronization phenomena. In this thesis four different classes of the synchronization measures were compared with each other. These measures comprised the linear cross-correlation, measures with information theoretic background such as mutual information and transfer entropy, phase synchronization measures based on either Hilbert or wavelet transform, and measures of generalized synchronization.

In the first part of this thesis we introduced different measures of synchronization. For mutual information and transfer entropy a new family of estimators was developed. Their major advantage lies in vastly reduced systematic errors, when compared to previous estimators. This allows to use them on very small data sets. It also makes possible their use in independent component analysis to estimate absolute values of mutual dependencies.

A theoretical comparison of the two phase extraction methods based on Hilbert and wavelet transform was derived in the second part of this thesis. Although the notion of phase plays an important role in oscillation theory and especially for synchronization phenomena the comparison of different phase extraction methods was still missing. Moreover, an extended discussion about the ambiguity of phase definition was presented there.

Since pathological processes such as epilepsy are considered to be related to synchronization phenomena, all the measures of synchronization were applied in the third part of this thesis to the analysis of intracranial EEG recordings from epilepsy patients undergoing pre-surgical diagnostics. In this study we addressed the question whether it is possible to identify the location of an epileptic focus from an EEG recorded during the seizure-free interval. The performance of different measures of synchronization with this respect was compared. The results of the localization analysis for all measures and the majority of patients were found significant with respect to the Wilcoxon statistical test. Another important question whether the obtained results were specific for the synchronization measures was addressed using bivariate surrogate data technique. These results confirm the hypothesis about synchronization as one of the factors responsible for the difference in the focal and the non-focal hemispheres of patients suffering from focal epilepsy.

More generally, synchronization is just one way how systems can show dependencies. Finding dependencies between different (sub-)systems and classifying these systems based on the levels of dependencies among them is an important problem surpassing synchronization. Therefore, studying general dependencies and clustering data based on them constituted another part of the thesis. In this part we introduced a new, conceptually very simple and natural, hierarchical clustering algorithm, called *mutual information clustering* (MIC). We illustrated our method with several applications. Among them are clustering of DNA sequences of mammals and clustering of minimally dependent components of the ECG of a pregnant woman. For these applications MIC might not give the best clustering, but it appears interesting that one common method gives decent results for both, although they are very different.

Finally, we applied MIC to cluster the synchronization/interdependencies measures used in EEG analysis. Using the sequence of average synchronization values as an input we clustered different synchronization measures. These results can be used to optimize the choice of measures for localization of the epileptic focus. The possible extensions of this application to EEG analysis can be, e.g., grouping the different channels for a more precise localization of the epileptic focus or classification of intervals preceding an epileptic seizure and intervals far away from any seizure activity. We believe that these directions will be interesting in epilepsy research.

Bibliography

[1] http://www.ncbi.nlm.nih.gov/.

[2] http://www.cs.ucsb.edu/ mli/Bioinf/software/index.html.

[3] H. D. I. Abarbanel, N. Rulkov, and M. Sushchik. Generalized synchronization of chaos: The auxiliary system approach. *Phys. Rev. E*, 53(5):4528–4535, 1996.

[4] V. S. Afraimovich, N. N. Verichev, and M. I. Rabinovich. Stochastic synchronization of oscillations in dissipative systems. *Radiophys. Quantum Electron.*, 29:795, 1986.

[5] R. G. Andrzejak. *Epilepsie als eine nichtlinear deterministische Dynamik: Eine Untersuchung hirnelektrischer Aktivität mit Methoden der linearen und nichtlinearen Zeitreihenanalyse*. PhD thesis, Dissertation in Physics, University of Bonn, Germany, 2001.

[6] R. G. Andrzejak, A. Kraskov, H. Stögbauer, F. Mormann, and T. Kreuz. Bivariate surrogate techniques: Necessity, strengths, and caveats. *Phys. Rev. E*, 68:066202, 2003.

[7] R. G. Andrzejak, T. Kreuz, F. Mormann, G. Widmann, C. E. Elger, and K. Lehnertz. Improved characterization of neuronal dynamics by focusing on nonlinearity. *(submitted)*, 2003.

[8] R. G. Andrzejak, F. Mormann, T. Kreuz, C. Rieke, A. Kraskov, C. E. Elger, and K. Lehnertz. Testing the null hypothesis of the nonexistence of a preseizure state. *Phys. Rev. E*, 67:010901, 2003.

[9] R. G. Andrzejak, G. Widman, K. Lehnertz, P. David, and C. E. Elger. The epileptic process as nonlinear deterministic dynamics in a stochastic environment: An evaluation on mesial temporal lobe epilepsy. *Epilepsy Res.*, 44:129, 2001.

[10] J. F. Annegers. The epidemiology of epilepsy. In E. Wyllie, editor, *The treatment of epilepsy: Principles and practice*, page 165. Williams and Wilkins, Baltimore, 1996.

[11] E. V. Appleton. The automatic synchronization of triode oscillator. *Proc. Cambridge Phil. Soc. (Math. and Phys. Sci.)*, 21:231, 1922.

[12] J. Arnhold. *Nichtlineare Analyse raum-zeitlicher Aspekte der hirnelektrischen Aktivität von Epilepsiepatienten*. PhD thesis, Dissertation in Physics, University of Wuppertal, Germany, 2000.

[13] J. Arnhold, K. Lehnertz, P. Grassberger, and C. E. Elger. A robust method for detecting interdependences: Application to intracranially recorded EEG. *Physica D*, 134:419, 1999.

[14] J. S. Barlow. Methods of analysis of nonstationary EEGs with emphasis on segmentation techniques: A comparative review. *J. Clin. Neurophysiol.*, 2:267, 1985.

[15] N. N. Bartlett. On the theoretical specification and sampling properties of autocorrelated time-series. *J. R. Stat. Soc.*, B8:27, 1946.

[16] H. Berger. Über das Elektroenkephalogramm des Menschen. *Arch Psychiat Nervenkrankh,* 87:527, 1929.

[17] S. Blanco, H. Garcia, R. Quian Quiroga, L. Romanelli, and O. A. Rosso. Stationarity of the EEG series. *IEEE Eng. Med. Biol.*, 4:395, 1995.

[18] G. E. P. Box and G. M. Jenkins. *Time series analysis: Forecasting and control. Revised Ed.* Holden-Day, San Francisco, 1993.

[19] Y. Cao, A. Janke, P.J. Waddell, M. Westerman, O. Takenaka, S. Murata, N. Okada, S. Pääbo, and M. Hasegawa. Conflict among individual mitochondrial proteins in resolving the phylogeny of eutherian orders. *J. Molec. Evol.*, 47(3):307–322, 1998.

[20] T. L. Carroll and L. M. Pecora. Cascading synchronized chaotic systems. *Physica D*, 67:126, 1993.

[21] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley and Sons, New York, 1991.

[22] G. A. Darbellay. Statistical dependencies in $\mathbf{R}^d$: An information-theoretic approach. In *3rd IEEE European Workshop on Computer-intensive Methods in Control and Data Processing*, pages 83–88, 1998.

[23] G. A. Darbellay and I. Vajda. Estimation of the information by an adaptive partitioning of the observation space. *IEEE Trans. Inform. Th.*, 45:1315, 1999.

[24] B. L. R. De Moor. Daisy: Database for the identification of systems, 1997. www.esat.kuleuven.ac.be/sista/daisy.

[25] J. Engel Jr, P. C. Van Ness, T. B. Rasmussen, and L. M. Ojemann. Outcome with respect to epileptic seizures. In J. Engel Jr, editor, *Surgical Treatment of the Epilepsies,* page 609. Raven Press, New York, 1993.

[26] J. Engel Jr and T. A. Pedley, editors. *Epilepsy: A Comprehensive Textbook*. Lippincott-Raven, Philadelphia, 1997.

[27] B. S. Everitt. *Cluster analysis*. John Wiley & Sons Inc,, New York, 1993.

[28] L. Fabiny, P. Colet, and R. Roy. Coherence and phase dynamics of spatially coupled solid-state lasers. *Phys. Rev. A*, 47:4287, 1993.

[29] J. Fell, P. Klaver, K. Lehnertz, T. Grunwald, C. Schaller, C. E. Elger, and G. Fernández. Human memory formation is accompanied by rhinal-hippocampal coupling and decoupling. *Nature Neurosci.*, 4:1259, 2001.

[30] A. Fraser and H. Swinney. Independent coordinates for strange attractors from mutual information. *Phys. Rev. A*, 33:1134, 1986.

[31] H. Fujisaka and T. Yamada. Stability theory of synchronized motion in coupled dynamical systems. *Prog. Theor. Phys.*, 69(1):32–47, 1983.

[32] D. Gabor. Theory of communication. *Proc. IEEE London*, 93:429, 1946.

[33] P. Grassberger. Entropy estimation from insufficient samplings. E-print, arXiv.org/physics/0307138.

[34] P. Grassberger. Generalizations of the Hausdorff dimension of fractal measures. *Phys. Lett. A*, 107(3):101–105, 1985.

[35] P. Grassberger. Finite sample corrections to entropy and dimension estimates. *Phys. Lett. A*, 128:369, 1988.

[36] P. Grassberger. An optimized box-assisted algorithm for fractal dimensions. *Phys. Lett. A*, 148(1–2):63–68, 1990.

[37] R. Gray. *Entropy and Information Theory*. Springer Verlag, New York, 1990.

[38] A. Grossmann, R. Kronland-Martinet, and J. Morlet. Reading and understanding continuous wavelet transform. In J.M. Combes, A. Grossmann, and Ph.Tchamitchian, editors, *Wavelets: Time-Frequency methods and phase space.*, pages 2–20. Springer, Berlin, 1989.

[39] J. F. Heagy, T. L. Caroll, and L. M. Pecora. Synchronous chaos in coupled oscillator systems. *Phys. Rev. E*, 50:1874, 1994.

[40] B. R. Hunt, E. Ott, and J.A. Yorke. Differential generalized synchronization of chaos. *Phys. Rev. E*, 55(4):4029–4034, 1997.

[41] C. Huygens. *Horoloquium Oscilatorium*. Paris, 1673.

[42] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent component analysis*. John Wiley and Sons, New York, 2001.

[43] A. K. Jain and R. C. Dubes. *Algorithms for Clustering Data*. Prentice Hall, Englewood Cliffs, NJ, 1988.

[44] A. Kaiser and T. Schreiber. Information transfer in continuous processes. *Physica D*, 166:43, 2002.

[45] L. Kocarev and U. Parlitz. General approach for chaotic synchronization with applications to communication. *Phys. Rev. Lett.*, 74:5028, 1995.

[46] L. F. Kozachenko and N. N. Leonenko. *Problm. Inf. Transm.*, 23:95, 1987.

[47] A. Kraskov, T. Kreuz, R. G. Andrzejak, H. Stögbauer, W. Nadler, and P. Grassberger. Extracting phases from aperiodic signals. *Phys. Rev. E (submitted)*, 2003.

[48] A. Kraskov, T. Kreuz, R. Quian Quiroga, P. Grassberger, F. Mormann, K. Lehnertz, and C. E. Elger. Phase synchronization using continuous wavelet transform of the EEG for interictal focus localization in mesial temporal lobe epilepsy. *Epilepsia*, 42(7):43, 2001.

[49] A. Kraskov, T. Kreuz, R. Quian Quiroga, P. Grassberger, F. Mormann, K. Lehnertz, and C. E. Elger. Comparison of two phase synchronization analysis techniques for interictal focus lateralization in mesial temporal lobe epilepsy. *Epilepsia*, 43(7):48, 2002.

[50] A. Kraskov, H. Stögbauer, R. G. Andrzejak, and P. Grassberger. Hierarchical clustering based on mutual information. *Bioinformatics (submitted)*, 2003. E-print, arXiv.org/q-bio/0311039.

[51] A. Kraskov, H. Stögbauer, R. G. Andrzejak, and P. Grassberger. Hierarchical clustering using mutual information. *Phys. Rev. Lett. (submitted)*, 2003. E-print, arXiv.org/q-bio/0311037.

[52] A. Kraskov, H. Stögbauer, and P. Grassberger. Estimating mutual information. *Phys. Rev. E (accepted)*, 2003. E-print, arXiv.org/cond-mat/0305641.

[53] T. Kreuz. *Measuring Synchronization in Model Systems and Electroencephalographic Time Series from Epilepsy Patients*. PhD thesis, Dissertation in Physics, University of Wuppertal, Germany, 2003.

[54] T. Kreuz, R. G. Andrzejak, A. Kraskov, F. Mormann, H. Stögbauer, C. E. Elger, P. Grassberger, and K. Lehnertz. Time profile surrogates: A new method to validate the performance of seizure prediction algorithms. *Epilepsia*, 44(9):231, 2003.

[55] T. Kreuz, R. G. Andrzejak, F. Mormann, A. Kraskov, H. Stögbauer, C. E. Elger, K. Lehnertz, and P. Grassberger. Measure profile surrogates: A new method to validate the performance of epileptic seizure prediction algorithms. *Phys. Rev. E (accepted)*, 2003.

[56] J. P. Lachaux, E. Rodriguez, M. Le Van Quyen, A. Lutz, J. Martinerie, and F. J. Varela. Studying single-trials of phase-synchronous activity in brain. *Int. J. Bifurcation Chaos Appl. Sci. Eng.*, 10(10):2429, 2000.

[57] J. P. Lachaux, E. Rodriguez, J. Martinerie, and F. J. Varela. Measuring phase synchrony in brain signals. *Hum. Brain Mapp.*, 8:194, 1999.

[58] J. Laidlaw, A. Richens, and J. Oxley. *A Textbook of Epilepsy*. Churchill Livingstone, New York, 1988.

[59] M. Le Van Quyen, J. Martinerie, C. Adam, and F. J. Varela. Nonlinear analysis of interictal EEG map the brain interdepences in human focal epilepsy. *Physica D*, 127:250, 1999.

[60] Dae-Sic Lee, Won-Ho Kye, Sunghwan Rim, Tae-Yoon Kwon, and Chil-Min Kim. Generalized phase synchronization in unidirectionally coupled chaotic oscillators. *Phys. Rev. E*, 67:045201, 2003.

[61] K. Lehnertz, R. G. Andrzejak, J. Arnhold, T. Kreuz, F. Mormann, C. Rieke, G. Widman, and C. E. Elger. Nonlinear EEG analysis in epilepsy: Its possible use for interictal focus localization, seizure anticipation, and prevention. *J. Clin. Neurophysiol.*, 18:209, 2001.

[62] K. Lehnertz, J. Arnhold, P. Grassberger, and C. E. Elger. *Chaos in Brain?* World Scientific, Singapore, 2000.

[63] K. Lehnertz, F. Mormann, T. Kreuz, R. G. Andrzejak, C. Rieke, P. David, and C. E. Elger. Seizure prediction by nonlinear EEG analysis. *IEEE Eng. Med. Biol.*, 22:57, 2003.

[64] M. Li, J. H. Badger, X. Chen, S. Kwong, P. Kearney, and H. Zhang. An information-based sequence distance and its application to whole mitochondrial genome phylogeny. *Bioinformatics*, 17(2):149–154, 2001.

[65] M. Li, X. Chen, X. Li, B. Ma, and P. Vitanyi. The similarity metric. 2002. E-print, arxiv.org/cs.CC/0111054.

[66] M. Li and P. Vitanyi. *An introduction to Kolmogorov complexity and its applications*. Springer, New York, 1997.

[67] B. Litt and J. Echaux. Prediction of epileptic seizures. *Lancet Neurol*, 1:22, 2002.

[68] B. Litt and K. Lehnertz. Seizure prediction and the preseizure period. *Curr Opin Neurol*, 15:173, 2002.

[69] F. H. Lopes da Silva. EEG analysis: Theory and practice. In E. Niedermayer and F. H. Lopes da Silva, editors, *Electroencephalography, basic principles, clinical applications and related fields*, page 1097. Urban and Schwarzenberg, 3rd Ed. (Williams and Wilkins, Baltimore), 1993.

[70] F. H. López de Màntaras. A distance based attribute selection for decision tree induction. *Machine Learning*, 6(1):81–92, 1991.

[71] E. N. Lorenz. Deterministic non-periodic flow. *J. Atmos. Sci.*, 20:130, 1963.

[72] F. Mormann. Synchronisationsphänomene in synthetischen Zeitreihen und Zeitreihen hirnelektrischer Aktivität. Master's thesis, Department of Physics, University of Bonn, Germany, 1998.

[73] F. Mormann. *Synchronization phenomena in the human epileptic brain*. PhD thesis, Dissertation in Physics, University of Bonn, Germany, 2003.

[74] F. Mormann, R. G. Andrzejak, T. Kreuz, C. Rieke, P. David, C. E. Elger, and K. Lehnertz. Automated preictal state detection based on a decrease in synchronization in intracranial electroencephalography recordings from epilepsy patients. *Phys. Rev. E*, 67:021912, 2003.

[75] F. Mormann, T. Kreuz, R. G. Andrzejak, P. David, K. Lehnertz, and C. E. Elger. Epileptic seizures are preceded by a decrease in synchronization. *Epilepsy Res.*, 53:173, 2003.

[76] F. Mormann, T. Kreuz, C. Rieke, R. G. Andrzejak, A. Kraskov, P. David, C. E. Elger, and K. Lehnertz. On the predictability of epileptic seizures. *Clin. Neurophysiol. (submitted)*, 2003.

[77] F. Mormann, K. Lehnertz, P. David, and C. E. Elger. Mean phase coherence as a measure for phase synchronization and its application to the EEG of epilepsy patients. *Physica D*, 144:358, 2000.

[78] T. I. Netoff and S. J. Schiff. Decreased neuronal synchronization during experimental seizures. *J. Neurosci.*, 22:7297–307, 2002.

[79] E. Niedermeyer. *The Epilepsies - Diagnosis and Management*. Urban and Schwarzenberg, Baltimore, 1990.

[80] G. Osipov, A. Pikovsky, M. Rosenblum, and J. Kurths. Phase synchronization effects in a lattice of nonidentical Rössler oscillators. *Phys. Rev. E*, 55:2353, 1997.

[81] M. Paluš and A. Stefanovska. Direction of coupling from phases of interacting oscillators: An information-theoretic approach. *Phys. Rev. E*, 67:055201, 2003.

[82] P. Panter. *Modulation, noise, and spectral analysis*. McGraw-Hill, New York, 1965.

[83] U. Parlitz, L. Junge, W. Lauterborn, and L. Kocarev. Experimental observation of phase synchronization. *Phys. Rev. E*, 54:2115, 1996.

[84] L. M. Pecora and T. L. Carroll. Synchronization in chaotic systems. *Phys. Rev. Lett.*, 64:821, 1990.

[85] D. W. Peterman, M. Ye, and P. E. Wigen. High frequency synchronization of chaos. *Phys. Rev. Lett.*, 74:1740, 1995.

[86] A. S. Pikovsky. On the interaction of strange attractors. *Z. Phys. B: Condens Matter*, 55(2):149, 1984.

[87] A. S. Pikovsky. Phase synchronization of chaotic oscillations by a periodic external field. *Sov. J. Commun. Technol. Electron.*, 30(10):1970, 1985.

[88] A. S. Pikovsky, M. G. Rosenblum, and J. Kurths. Synchronization in a population of globally coupled chaotic oscillators. *Europhys. Lett.*, 34:165, 1996.

[89] A. S. Pikovsky, M. G. Rosenblum, and J. Kurths. *Synchronization. A universal concept in nonlinear sciences*. Cambridge Univ. Press, Cambridge, UK, 2001.

[90] W. H. Press, B. Flannery, S. Teukolsky, and W. Vetterling. *Numerical recipes in C: The art of scientific computing*. Cambridge Univ. Press, Cambridge, UK, second edition, 1992.

[91] K. Pyragas. Continuous control of chaos by self-controlling feedback. *Phys. Lett. A*, 170:421, 1992.

[92] K. Pyragas. Weak and strong synchronization of chaos. *Phys. Rev. E*, 54:4508, 1996.

[93] R. Quian Quiroga, J. Arnhold, and P. Grassberger. Learning driver-response relationships from synchronization patterns. *Phys. Rev. E*, 61:5142, 2000.

[94] R. Quian Quiroga, A. Kraskov, T. Kreuz, and P. Grassberger. Performance of different synchronization measures in real data: A case study on electroencephalographic signals. *Phys. Rev. E*, 65:041903, 2002.

[95] A. Renyi. *Probability Theory*. North Holland, Amsterdam, 1971.

[96] A. Reyes, C. Gissi, G. Pesole, F. M. Catzeflis, and C. Saccone. Where do rodents fit? Evidence from the complete mitochondrial genome of sciurus vulgaris. *Mol. Biol. Evol.*, 17:979–983, 2000.

[97] E. Rosa Jr., W. B. Pardo, C. M. Ticos, J. A. Walkenstein, and M. Monti. Phase synchronization of chaos in a plasma discharge tube. *Int. J. Bifurc. Chaos*, 10:2551, 2000.

[98] M. Rosenblum, A. Pikovsky, J. Kurths, G. Osipov, I. Kiss, and J. Hudson. Locking-based frequency measurement and synchronization of chaotic oscillators with complex dynamics. *Phys Rev Lett*, 89:264102, 2002.

[99] M. G. Rosenblum, L. Cimponeriu, A. Bezerianos, A. Patzak, and R. Mrowka. Identification of coupling direction: Application to cardiorespiratory interaction. *Phys Rev E*, 65:041909, 2002.

[100] M. G. Rosenblum and A. S. Pikovsky. Detecting direction of coupling in interacting oscillators. *Phys Rev E*, 64:045202, 2001.

[101] M. G. Rosenblum, A. S. Pikovsky, and J. Kurths. Phase synchronization of chaotic oscillators. *Phys. Rev. Lett.*, 76(11):1804, 1996.

[102] M. G. Rosenblum, A. S. Pikovsky, J. Kurths, C. Schaefer, and P. A. Tass. Phase synchronization: from theory to data analysis. In F. Moss and S. Gielen, editors, *Handbook of biological physics*, page 297. Elsevier Science, Amsterdam, 2001.

[103] A. G. Rossberg, K. Bartholom, and J. Timmer. Data driven optimal filtering for phase and frequency of noisy oscillations: Application to vortex flowmetering. 2003. E-print, arxiv.org/nlin.CD/0305039.

[104] O. E. Rössler. An equation for continuous chaos. *Phys. Lett. A*, 57:397, 1976.

[105] R. Roy and K. S. Thornburg. Experimental synchronization on chaotic lasers. *Phys. Rev. Lett.*, 72:2009, 1994.

[106] N. F. Rulkov, M. M. Sushchik, L.S. Tsimring, and H. D. I. Abarbanel. Generalized synchronization of chaos in directionally coupled chaotic systems. *Phys. Rev. E*, 51(2):980–994, 1995.

[107] N. F. Rulkov, L.S. Tsimring, and H. D. I. Abarbanel. Tracking unstable orbits in chaos using dissipative feedback control. *Phys. Rev. E*, 50:314, 1994.

[108] S.J. Schiff, P. So, T. Chang, R. E. Burke, and T. Sauer. Detecting dynamical interdependence and generalized synchrony through mutual prediction in a neural ensemble. *Phys. Rev. E*, 54:6708, 1996.

[109] A. Schmitz. Measuring statistical dependence and coupling of subsystems. *Phys. Rev. E*, 62:7508, 2000.

[110] T. Schreiber. Constrained randomization of time series data. *Phys. Rev. Lett.*, 80(10):2105, 1998.

[111] T. Schreiber. Measuring information transfer. *Phys. Rev. Lett.*, 85:461, 2000.

[112] T. Schreiber and A. Schmitz. Improved surrogate data for nonlinearity tests. *Phys. Rev. Lett.*, 77(4):635, 1996.

[113] T. Schreiber and A. Schmitz. Surrogate time series. *Physica D*, 142:346, 2000.

[114] B. W. Silverman. *Density estimation for statistics and data analysis*, volume 26 of *Monographs on Statistical and Applied Probability*. Chapman & Hall, London, 1986.

[115] D. A. Smirnov and B. P. Bezruchko. Estimation of interaction strength and direction from short and noisy time series. *Phys. Rev. E*, 68:046209, 2003.

[116] C. J. Stam and B. W. van Dijk. Synchronization likelihood: An unbiased measure of generalized synchronization in multivariate data sets. *Physica D*, 163:236, 2002.

[117] H. Stögbauer, A. Kraskov, and P. Grassberger. Potential of mutual information in application to ICA. To be published.

[118] F. Takens. Detecting strange attractors in turbulence. In D. A. Rand and L. S. Young, editors, *Dynamical Systems and Turbulence*, volume 898 of *Lecture Notes in Mathematics*, page 366. Springer-Verlag, Berlin, 1980.

[119] D. Y. Tang, R. Dykstra, M. W. Hamilton, and N. R. Heckenberg. Experimental evidence of frequency entrainment between coupled chaotic oscillations. *Phys. Rev. E*, 57(3):3649, 1998.

[120] P. A. Tass, M. G. Rosenblum, J. Weule, J. Kurths, A. Pikovsky, J. Volkmann, A. Schnitzler, and H. J. Freund. Detection of n:m phase locking from noisy data: Application to magnetoencephalography. *Phys. Rev. Lett.*, 81(15):3291, 1998.

[121] J. Theiler. Spurious dimensions from correlation algorithms applied to limited time-series data. *Phys. Rev. A*, 34:2427, 1986.

[122] 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*, 58:77, 1992.

[123] D. E. Vakman and L. A. Vainshtein. Amplitude, phase, frequency — fundamental concepts of oscillation theory. *Sov. Phys. Usp.*, 20(12):1002, 1977.

[124] B. van der Pol. Forced oscillations in a circuit with non-linear resistance. *Phil. Mag.*, 3:64, 1927.

[125] B. van der Pol and J. van der Mark. The heartbeat considered as a relaxation oscillation, and an electrical model of the heart. *Phil. Mag.*, 6:763, 1928.

[126] F. J. Varela. Resonant cell assemblies: A new approach to cognitive functions and neuronal synchrony. *Biol. Res.*, 28:81, 1995.

[127] F. J. Varela, J. P. Lachaux, E. Rodriguez, and J. Martinerie. The brain web: Phase synchronization and large-scale integration. *Nature Rev. Neurosci.*, 2:229, 2001.

[128] J. A. Vastano and H. L. Swinney. Information transport in spatiotemporal systems. *Phys. Rev. Lett.*, 60:1773, 1988.

[129] J. D. Victor. Binless strategies for estimation of information from neural data. *Phys. Rev. E*, 66:051903–1, 2002.

[130] Z. Zheng and G. Hu. Generalized synchronization versus phase synchronization. *Phys. Rev. E*, 62:7882, 2000.

[131] Z. Zheng, X. Wang, and M. C. Cross. Transitions from partial to complete generalized synchronizations in bidirectionally coupled chaotic oscillators. *Phys. Rev. E*, 65:56211, 2002.

Danksagung

Zunächst möchte ich meinem Doktorvater Prof. Dr. P. Grassberger, dafür danken, daß er es mir ermöglicht hat, diese interdisziplinäre Doktorarbeit am John von Neumann Institute for Computing im Forschungszentrum Jülich, Deutschland durchzuführen. Besonders danke ich ihm für sein ständiges Interesse, die vielen lehrreichen Gespräche und die spannenden Diskussionen, die zum Gelingen dieser Doktorarbeit maßgeblich beigetragen haben, sowie die stets einladend offene Bürotür. Außerdem bin ich ihm dankbar dafür, dass er es mir ermöglicht hat, meine Forschungsergebnisse auf zahlreichen nationalen und internationalen Konferenzen zu präsentieren und dort Kontakte zu vielen interessanten Wissenschaftlern zu knüpfen.

Den Kollegen am John von Neumann Institute for Computing Dr. Ralph G. Andrzejak, Dr. Hsiao-Ping Hsu, Dr. Thomas Kreuz, Dr. Walter Nadler, Harald Stögbauer und Dr. Rodrigo Quian Quiroga danke ich für das ausgezeichnete Arbeitsklima, sowie für viele interessante und weiterführende Diskussionen. Bei Dr. Ralph G. Andrzejak, Dr. Thomas Kreuz und Harald Stögbauer möchte ich mich sehr für ihre kollegiale Zusammenarbeit und ständige Hilfsbereitschaft und für das aufmerksame Korrekturlesen vorliegender Arbeit bedanken. Außerdem bin ich Priv. Doz. K. Lehnertz und seiner Forschungsgruppe in der Klinik für Epileptologie der Universität Bonn für die produktive Zusammenarbeit sehr dankbar.

Außerdem möchte ich die Hilfsbereitschaft von Frau Helga Frank betonen, die für mich als Ausländer besonders wichtig war.

Mein besonderer Dank gilt Prof. Dr. B.P. Bezruchko, Saratow Universität, Rußland, der mich am Anfang meiner wissenschflichen Laufbahn betreut hat und bei mir das Interesse für die Forschung im Grenzgebiet zwischen Physik und Medizin geweckt hat.

Beonders möchte ich meine Mutter Prof. Dr. S.P. Mushtakova hervorheben, die mich für das wissenschaftliche Arbeiten inspiriert hat und mir als Vorbild dient.

Bei meiner Frau Nadja Vidro bedanke ich mich sehr für ihre Geduld und ständige Unterstützung in allen Lebenslagen. Ohne Sie wäre dies alles gar nicht möglich gewesen.

Aachen, im Januar 2003

Alexander Kraskov

Already published:

Modern Methods and Algorithms of Quantum Chemistry -
Proceedings
Johannes Grotendorst (Editor)
Winterschool, 21 - 25 February 2000, Forschungszentrum Jülich
NIC Series Volume 1
ISBN 3-00-005618-1, February 2000, 562 pages
out of print

Modern Methods and Algorithms of Quantum Chemistry -
Poster Presentations
Johannes Grotendorst (Editor)
Winterschool, 21 - 25 February 2000, Forschungszentrum Jülich
NIC Series Volume 2
ISBN 3-00-005746-3, February 2000, 77 pages
out of print

Modern Methods and Algorithms of Quantum Chemistry -
Proceedings, Second Edition
Johannes Grotendorst (Editor)
Winterschool, 21 - 25 February 2000, Forschungszentrum Jülich
NIC Series Volume 3
ISBN 3-00-005834-6, December 2000, 638 pages

Nichtlineare Analyse raum-zeitlicher Aspekte der
hirnelektrischen Aktivität von Epilepsiepatienten
Jochen Arnold
NIC Series Volume 4
ISBN 3-00-006221-1, September 2000, 120 pages

Elektron-Elektron-Wechselwirkung in Halbleitern:
Von hochkorrelierten kohärenten Anfangszuständen
zu inkohärentem Transport
Reinhold Lövenich
NIC Series Volume 5
ISBN 3-00-006329-3, August 2000, 146 pages

Erkennung von Nichtlinearitäten und
wechselseitigen Abhängigkeiten in Zeitreihen
Andreas Schmitz
NIC Series Volume 6
ISBN 3-00-007871-1, May 2001, 142 pages

Multiparadigm Programming with Object-Oriented Languages -
Proceedings
Kei Davis, Yannis Smaragdakis, Jörg Striegnitz (Editors)
Workshop MPOOL, 18 May 2001, Budapest
NIC Series Volume 7
ISBN 3-00-007968-8, June 2001, 160 pages

Europhysics Conference on Computational Physics -
Book of Abstracts
Friedel Hossfeld, Kurt Binder (Editors)
Conference, 5 - 8 September 2001, Aachen
NIC Series Volume 8
ISBN 3-00-008236-0, September 2001, 500 pages

NIC Symposium 2001 - Proceedings
Horst Rollnik, Dietrich Wolf (Editors)
Symposium, 5 - 6 December 2001, Forschungszentrum Jülich
NIC Series Volume 9
ISBN 3-00-009055-X, May 2002, 514 pages

Quantum Simulations of Complex Many-Body Systems:
From Theory to Algorithms - Lecture Notes
Johannes Grotendorst, Dominik Marx, Alejandro Muramatsu (Editors)
Winter School, 25 February - 1 March 2002, Rolduc Conference Centre,
Kerkrade, The Netherlands
NIC Series Volume 10
ISBN 3-00-009057-6, February 2002, 548 pages

Quantum Simulations of Complex Many-Body Systems:
From Theory to Algorithms- Poster Presentations
Johannes Grotendorst, Dominik Marx, Alejandro Muramatsu (Editors)
Winter School, 25 February - 1 March 2002, Rolduc Conference Centre,
Kerkrade, The Netherlands
NIC Series Volume 11
ISBN 3-00-009058-4, February 2002, 194 pages

**Strongly Disordered Quantum Spin Systems in Low Dimensions:
Numerical Study of Spin Chains, Spin Ladders and
Two-Dimensional Systems**
Yu-cheng Lin
NIC Series Volume 12
ISBN 3-00-009056-8, May 2002, 146 pages

**Multiparadigm Programming with Object-Oriented Languages -
Proceedings**
Jörg Striegnitz, Kei Davis, Yannis Smaragdakis (Editors)
Workshop MPOOL 2002, 11 June 2002, Malaga
NIC Series Volume 13
ISBN 3-00-009099-1, June 2002, 132 pages

**Quantum Simulations of Complex Many-Body Systems:
From Theory to Algorithms - Audio-Visual Lecture Notes**
Johannes Grotendorst, Dominik Marx, Alejandro Muramatsu (Editors)
Winter School, 25 February - 1 March 2002, Rolduc Conference Centre,
Kerkrade, The Netherlands
NIC Series Volume 14
ISBN 3-00-010000-8, November 2002, DVD

Numerical Methods for Limit and Shakedown Analysis
Manfred Staat, Michael Heitzer (Eds.)
NIC Series Volume 15
ISBN 3-00-010001-6, February 2003, 306 pages

**Design and Evaluation of a Bandwidth Broker that Provides
Network Quality of Service for Grid Applications**
Volker Sander
NIC Series Volume 16
ISBN 3-00-010002-4, February 2003, 208 pages

**Automatic Performance Analysis on Parallel Computers with
SMP Nodes**
Felix Wolf
NIC Series Volume 17
ISBN 3-00-010003-2, February 2003, 168 pages

Haptisches Rendern zum Einpassen von hochaufgelösten Molekülstrukturdaten in niedrigaufgelöste Elektronenmikroskopie-Dichteverteilungen
Stefan Birmanns
NIC Series Volume 18
ISBN 3-00-010004-0, September 2003, 178 pages

Auswirkungen der Virtualisierung auf den IT-Betrieb
Wolfgang Gürich (Editor)
GI Conference, 4 - 5 November 2003, Forschungszentrum Jülich
NIC Series Volume 19
ISBN 3-00-009100-9, October 2003, 126 pages

NIC Symposium 2004
Dietrich Wolf, Gernot Münster, Manfred Kremer (Editors)
Symposium, 17 - 18 February 2004, Forschungszentrum Jülich
NIC Series Volume 20
ISBN 3-00-012372-5, February 2004, 482 pages

Measuring Synchronization in Model Systems and Electroencephalographic Time Series from Epilepsy Patients
Thomas Kreutz
NIC Series Volume 21
ISBN 3-00-012373-3, February 2004, 138 pages

Computational Soft Matter: From Synthetic Polymers to Proteins - Poster Abstracts
Norbert Attig, Kurt Binder, Helmut Grubmüller, Kurt Kremer (Editors)
Winterschule, 29 February - 6 March 2004, Gustav-Stresemann-Institut Bonn
NIC Series Volume 22
ISBN 3-00-012374-1, February 2004, 120 pages

Computational Soft Matter: From Synthetic Polymers to Proteins - Lecture Notes
Norbert Attig, Kurt Binder, Helmut Grubmüller, Kurt Kremer (Editors)
Winterschule, 29 February - 6 March 2004, Gustav-Stresemann-Institut Bonn
NIC Series Volume 23
ISBN 3-00-012641-4, February 2004, 440 pages

All volumes are available online at http://www.fz-juelich.de/nic-series/.