ARISTOTLE UNIVERSITY OF THESSALONIKI
DEPARTMENT OF MATHEMATICS, PHYSICS
and COMPUTATIONAL SCIENCES
FACULTY OF TECHNOLOGY

ANGELIKI D. PAPANA
B.Sc. in Mathematics

NONLINEAR STATISTICAL ANALYSIS OF BIOLOGICAL TIME SERIES

PhD THESIS

Thessaloniki, 2009

ANGELIKI D. PAPANA

NONLINEAR STATISTICAL ANALYSIS OF BIOLOGICAL TIME SERIES

PhD THESIS

Submitted to Department of Mathematics,
Physics and Computational Sciences,
Faculty of Technology
Day of oral examination: 24 June, 2009

Adjudicate committee

Ass. Professor D. Kugiumtzis, Supervisor
Ass. Professor G. Zioutas, Member of the consultive committee
Professor A. Bountis, Member of the consultive committee

Researcher A' A. Provata, Examiner
Professor C. Moyssiadis, Examiner
Asso. Professor N. Maglaveras, Examiner
Ass. Professor I. Rekanos, Examiner

Acknowledgements

This research project is implemented within the framework of the ”Reinforcement Programme of Human Research Manpower” (PENED) and is co-financed at 90% jointly by E.U.-European Social Fund (75%) and the Greek Ministry of Development-GSRT (25%) and at 10% by Rikshospitalet, Norway.

NONLINEAR STATISTICAL ANALYSIS OF BIOLOGICAL TIME SERIES

ISBN

Contents

Abstract

In the present thesis, methods of non-linear time series analysis and dynamical systems analysis have been combined with statistical methods, and statistical measures with high discriminating power directly estimated on time series have been studied and developed. Specifically, existing univariate and multivariate linear and nonlinear measures have been thoroughly reviewed, new nonlinear measures have been defined here and existing methods have been extended or modified to gain statistical significance and be more effective on different applications. All the investigated measures were first tested on simulation studies and were then applied on real data. Electroencephalogram (EEG) recordings from epileptic patients have been considered in order to evaluate the effectiveness of the measures to detect dynamical changes in the brain activity of the patients at different states, e.g. in order to discriminate between EEG records many hours before the seizure onset and EEG records from one hour before the seizure onset. The prediction of the onset of an epileptic seizure and the detection of the dynamical changes of the brain dynamics just before the seizure onset were investigated in the frame of univariate time series analysis. The performance of the univariate measures was quite promising and therefore the study was extended beyond the univariate case. Thus, the existence of interdependencies between different brain areas and the identification of the direction of the causal effects among the brain areas was investigated in the frame of multivariate time series analysis.

The study focused on information measures, as they are model-free, computationally efficient and do not require any prior knowledge of the distribution of the data. The evaluation of the information measures was first assessed by Monte Carlo simulations on well-known dynamical systems in order to examine their discriminating power and significance. Dynamical characteristics of the systems varied by changing the parameters of the equation of the systems. The discriminating ability of the measures was also assessed in terms of the time series length. The stochasticity of the systems was also a factor that was examined as it can be controlled by variation of the level of noise added in the observations of the systems. The variation of the complexity or stochasticity of a system is considered to simulate the different states of the brain activity and thus of the EEG signal.

Mutual information is an essential tool in nonlinear time series analysis and therefore was thoroughly reviewed and tested on different applications, e.g. in detecting dynamical changes of systems. Different estimators of mutual information have been compared and the selection of their parameters was investigated in order to be optimized. The promising performance of mutual information led to a more comprehensive study which included also entropy measures (e.g. Shannon, Tsallis, Permutation). This investigation has led to the extension and modification of many existing measures. The statistical significance and power of the measures in discrimination tasks were assessed using different statistical tests (t-test, Wilcoxon

rank sum test, ROC curves, surrogate data test). It was also examined whether the discriminating power of the measures can be improved if the time series were first transformed to have uniform or normal marginal distribution.

The results from the simulation studies on known dynamical systems were required in order to interpret the values of the estimated measures on the EEG recordings in terms of their stochasticity and complexity. Thus, the measures were evaluated in discriminating EEG signal from preictal and interictal stages. The effectiveness of the nonlinear measures was compared to that of some linear ones, as there are contradictory studies on this matter. Based on the frame of univariate time series analysis, EEG recordings from different channels were used for the analysis; channels were chosen based on knowledge of the epileptic focus area or were randomly selected in cases of generalized seizures to cover all parts of the brain.

EEG recordings are multivariate signals, and therefore one should also test the interdependencies between the recordings from the different channels, i.e. the existence of interactions among the different parts of the brain should be examined. Here, the direction of the information flow was investigated using bivariate nonlinear measures. The study focused again on information measures detecting the direction of interdependencies between interacting systems, however also other types of causal measures were included for comparative studies, e.g. state-space and synchronization measures. The evaluation of the causality measures was initially assessed on well-known nonlinear systems and results were considered in order to apply the measures on EEG signals. This study led us again in the modification of the existing measures and the extraction of measures that seem to improve their performance.

There exist measures that can discriminate among different dynamical systems and detect the dynamical changes of a systems and some of the measures that have been defined here may improved the discriminating power of the original measures when tested on synthetic data, however even these improved methods do not perform dramatically better on EEG. EEG signal is very complex and the finite length of the EEG signal, the measurement artifacts and the interactions between the different brain parts render the problem of the EEG analysis to be even harder. There is still a long way to go for the use of any measure in clinical practice. The main contribution of this work is to assess the statistical significance of the information measures for univariate and bivariate time series, propose modifications to attain better significance and show in an objective clinical setting the shortcomings and possible limited success of many existing measures for the problem of seizure prediction.

Publications

The findings of this thesis have been presented in conferences, have been published or have been submitted to international journals in the field of nonlinear dynamics, time series analysis or statistical analysis. The respective references of the papers in proceedings and journals are presented here.

P1 A. Papana and D. Kugiumtzis. Cumulative Mutual Information Function as a Statistic for the Test for Nonlinearity on Time Series (in Greek). Proceedings of the 18th Greek Statistical Conference, Rhodes, pp. 315–325, 2005.

P2 D. Kugiumtzis, A. Papana, I. Vlachos and P.G. Larsson. Time series feature evaluation in discriminating preictal EEG states. *Lecture Notes in Computer Science*, Vol. **4345**, pp. 298–310, 2006.

P3 A. Papana and D. Kugiumtzis. Linear and nonlinear correlation measures of time series for seizure prediction (in Greek). Proceedings of the 19th Greek Statistical Conference, Kastoria, pp. 415–423, 2006.

P4 D. Kugiumtzis, I. Vlachos, A. Papana and P.G. Larsson. Assessment of Measures of Scalar Time Series Analysis in Discriminating Preictal States. NEUROMATH Workgroup meeting and Workshop, Rome, 2–4 December, 2007.

P5 D. Kugiumtzis, I. Vlachos, A. Papana and P.G. Larsson. Assessment of Measures of Scalar Time Series Analysis in Discriminating Preictal States. *International Journal of Bioelectromagnetism*, Vol. **9** (3), pp. 134–145, 2007.

P6 A. Papana and D. Kugiumtzis. Evaluation of Histogram-Based Estimators of Mutual Information in Time Series (in Greek). Proceedings of the 20th Greek Statistical Conference, Nicosia, Cyprus, pp. 329–336, 2007.

P7 A. Papana and D. Kugiumtzis. Evaluation of Mutual Information Estimators on Nonlinear Dynamic Systems. *Nonlinear Phenomena in Complex Systems*, Vol. **11** (2), pp. 225–232, 2008.

P8 A. Papana and D. Kugiumtzis. Detection of Directionality of Information Transfer in Nonlinear Dynamical Systems. Topics on Chaotic Systems, Selected Papers from CHAOS 2008 International Conference, 3-6 June, Chania, Crete, pp. 251–264, World Scientific.

P9 A. Papana and D. Kugiumtzis. Detection of Dynamical Changes of Systems using Information Measures (in Greek). Proceedings of the 21rst Greek Statistical Conference, Samos, Chania, Crete (to be published).

P10 A. Papana and D. Kugiumtzis. Evaluation of Mutual Information Estimators for Time Series. *International Journal Bifurcation and Chaos* (to be published).

P11 A. Papana and D. Kugiumtzis. Improving Statistical Significance of Information Causality Measures. Proceedings of International Conference *CHAOS 2009*, 1-5 June, Chania, Crete.

Poster Presentations

PP1 A. Papana, D. Kugiumtzis 17th Panhellenic Conference and Summer School, *Complexity in Science and Society*, 14–26 July, Patra and Ancient Olympia, 2004. Poster title: Statistical Detection of Changes in the Underlying Dynamics of a System.

PP2 A. Papana, D. Kugiumtzis 19th Panhellenic Conference and Summer School, *Complexity and Nonlinear Dynamics*, 10–22 July, Thessaloniki, 2006. Poster title: Discrimination of Regimes of Dynamical Systems using Measures of Correlation.

PP3 D. Kugiumtzis, I. Vlachos, A. Papana, P.G. Larsson 4th International Workshop on Seizure Prediction, 4–7 June, Kansas, 2009. Poster title: Search for Optimal Time Series Measures to Discriminate Preictal States.

Chapter 1

Introduction

Epilepsy is a neurological disorder that affects $1\% - 2\%$ of the world population. The etiology of epilepsy is manifold; inherited, after trauma or after infections. Epilepsy is characterized by recurrent unprovoked seizures which occur when there is a sudden change in the normal way that the brain cells communicate through electrical signals (Bruce, 1980; Porter, 1993). The normal pattern of neuronal activity becomes disturbed and causes strange sensations, emotions and behavior or sometimes convulsions, muscle spasms and loss of consciousness. These physical changes are called epileptic seizures. Epilepsy should not be understood as a single disorder, but rather as a group of syndromes with vastly divergent symptoms, which involve episodic abnormal electrical activity in the brain. Seizure types are organized mainly according to whether the source of the seizure within the brain is localized (partial or focal onset seizures) or distributed (generalized seizures). There are over 40 different types of epilepsy that may appear at different ages, need different treatment and have different prognosis.

Neurons are electrically active cells which are primarily responsible for carrying out the functions of the brain. The cause of an epileptic seizure is the abnormal synchronous neural firing in one or several cortical areas. The resulting ictal spikes in the EEG (electroencephalogram) are the result of this hyper-synchronized cortical activity. The epileptic focus area, which is the brain area the seizure starts, may be localized to a small area or generalized due to the coupling between spatially disparate neuron pools.

Most epileptic patients experience the onset of a seizure as a sudden and unexpected event. However, the transition to the seizure (ictal state) might not be an abrupt phenomenon but rather evolves via a temporally extended preictal state (Viglione and Walsh, 1975; Iasemidis et al., 1990; Andrzejak et al., 2003). Thus, a preictal state is the state of brain just before the seizure onset. Provided that such detection of a preictal state could be achieved with a sufficient sensitivity and specificity, that would significantly improve the therapeutic possibilities and thereby the quality of life of epileptic patients. If this transition implies or evolves a gradual change in the dynamics of the brain then it can be theoretically detected. Evidence

that a preictal state exists are based on clinical findings including an increase in cerebral blood flow (Weinand et al., 1997), oxygen availability (Adelson et al., 1999) and blood oxygen-level dependent signal (Federico et al., 2005) as well as changes in heart rate (Delamont et al., 1999) before a seizure occurs. However, there is no clear evidence of the duration of a preictal state, and it may vary from patient to patient depending on the type of epilepsy.

As the neurons of the brain act through electric activity, the most common diagnostic procedure is to record the electric activity of the brain. EEG is the recording of electrical activity along the scalp produced by the firing of neurons within the brain. EEG is an essential component in the evaluation of the epilepsy (Lehnertz et al., 2001; Shiau et al., 2003; Mormann et al., 2007) as epileptic activity can create clear abnormalities on a standard EEG. EEG has been used for the detection of the epileptic focus area and aims to give insight about the prediction of epileptic seizures. The mechanisms of icto-genesis remain largely unknown except for certain distinct types of epilepsies such as reflex epilepsies (Kalitzin et al., 2002). The gradual changes that might occur in the EEG before a seizure and the best methods to detect them could thus vary considerably from patient to patient. In Ch. 2 the basic information about the EEG is given, e.g. EEG recording and the types of EEG.

The study of the EEG is assessed in terms of time series analysis. Time series (e.g. EEG) are measurements at successive times, spaced at uniform time intervals. Time series analysis aims to identify the nature of the phenomenon represented by the sequence of the observations, understand the underlying system of the time series, model the time series and make forecasts. The generating mechanism, i.e. the dynamics of the time series, are usually unknown. Although one might have no profound knowledge of the system, under some conditions its dynamics can be reconstructed from a single time series only (Takens, 1980). Multi-channel EEG recordings are thought to comprise significant information about the brain dynamics.

Linear methods of time series analysis interpret all regular structure of a time series. The intrinsic dynamics of linear systems are governed by the rule that small causes lead to small effects. Linear time series analysis can thoroughly characterize time series with constant mean (first moment), variance and covariance (second moments). However, in some systems irregular behavior can be observed, and this should not be solely attributed to some random external input. Chaos theory accounts for irregular behavior from nonlinear, chaotic dynamical systems with purely deterministic equations of motion. Dynamical systems are systems that can be defined by a set of variables whose values change with time. Nonlinear time series analysis is a powerful tool that enables the extraction of characteristic quantities of a nonlinear dynamical system solely by analyzing the time course of one of its variables. EEG signals present irregular behavior and therefore one should use methods from nonlinear time series analysis and chaos in order to investigate its characteristics, at least as part of the analysis (Palus, 1996). An extensive introduction on nonlinear time series and chaos analysis can be found at e.g. Grassberger

et al. (1991); Kantz and Schreiber (1997). The first part of Ch. 3 gives some fundamental aspects of nonlinear time series analysis. As the main aim of this thesis is to study, evaluate, develop and apply methods based on nonlinear dynamics and chaos theory in order to characterize EEG time series, detect any changes in the dynamics of the EEG signal before a seizure and discriminate EEG records from different stages (interictal, preictal, ictal), statistical tools are needed to infer for the statistical significance of the results. In the second part of Ch. 3, the statistical tools that have been used, are presented.

Information theory is a relatively new branch of mathematics developed in the 1940s and deals with the study of problems concerning any system, e.g. of information processing and decision making. Information theory was established by Shannon (1948) and this paper marked its development. Information theoretical measures (Shannon, 1948; Renyi, 1961; Kullback and Leibler, 1951) are widely used to analyze correlation and interdependency, characterize the degree of randomness of a system or time series, quantify the difference between probability distributions (Cover and Thomas, 1991; Gray, 1990) and therefore characterize dynamical systems. Information theory is based on probability theory and statistics and it is also a branch of ergodic theory (Birkhoff, 1931; Hoph, 1937) which deals with invariant transformations, assuming time series stationarity. The most important information measures developed are the entropy, and the mutual information. Entropy measures the information in a random variable and is a measure of the amount of uncertainty associated with the values of the variable. Mutual information measures the amount of information that can be obtained about one random variable by observing another. Mutual information is suitably defined for time series and can be considered as a generalization of the autocorrelation function, as it is used to measure any form of correlation (linear and nonlinear) in a time series.

Mutual information is presented thoroughly (definition, estimation and applications) in Ch. 4. Mutual information is used here for the investigation of the independence and nonlinear dependence of time series measurements and the detection of dynamical changes on systems. The most commonly used estimators of mutual information are evaluated, and the proper selection of their parameters is studied including many existing criteria in order to be optimized. The evaluation of the mutual information estimators is assessed on synthetic data from various linear and nonlinear systems.

In Ch. 5, the study of mutual information is extended in order to include more correlation measures. New univariate correlation measures are also defined here and are evaluated. "Cumulative mutual information" is a naive extension of mutual information which measures the overall linear and nonlinear correlations of a time series, relaxing the dependence on the delays. "Generalized decorrelation time" is a measure of the "memory" of a system and is an extension of the linear decorrelation time which is based on the autocorrelation function. Another two features are extracted that attempt to measure the deviation from linearity. The first measure is based on the surrogate data test (p-value from the test), and the second one is based

on the difference of the estimated mutual information to the "Gaussian" mutual information, i.e. the mutual information under the assumption that the process is Gaussian. The "cumulative declination from linearity" is extracted again in order to treat the dependence on the delay.

Various existing information measures, such as Tsallis entropy and Permutation entropy, were also included in an overall evaluation of correlation measures. The evaluation of the measures was assessed in cases of well-known dynamical systems, and their discriminating power and significance was examined. Linear and nonlinear/chaotic systems, both maps and flows, were used and it was examined whether nonlinear measures are more suitable than linear ones in identifying dynamical changes. By changing some parameters of the equations of the nonlinear systems, their dynamical characteristics and complexity can be controlled. Addition of noise in the observation of the systems allowed the examination of the effect of stochasticity on the measures. The complexity and stochasticity of the systems were controlled in order to simulate the different states of the brain activity. Thus, the discrimination ability of the measures was assessed statistically in terms of time series lengths, noise level and complexity of the systems. All measures were also estimated on transformed time series that had Gaussian or uniform amplitudes, in order to investigate a possible improvement of the discriminating power of the measures. The methods of nonlinear dynamical analysis have been combined with statistical methods in order to define measures with a high discriminating power.

Much research has been done on the prediction of epileptic seizure and current studies have shown that a number of measures derived from both linear and nonlinear signal processing are to some extent capable of extracting information from EEG. Typically, a study on the predictability of epileptic seizures involves the calculation of measures from the multi-channel EEG using a moving data window technique. The resulting measure profiles are then scanned for prominent features which can be related to consecutive stages of the brain activity. These features might be drops or peaks or any other distinct pattern in the measure profile. The ability of the measures to discriminate the preictal from the interictal interval is evaluated with test statistics quantifying the occurrence of these features relative to the number of seizure.

To this respect, the measures were estimated on consecutive EEG segments from all the stages and the measure profiles were compared using statistical tests. The first aim was to discriminate between the different states and the second one was to detect the dynamical changes on EEG and how often these changes can be observed. The measures that have already been thoroughly studied on synthetic data, were tested also on EEG recordings from epileptic patients. Again, it was examined whether nonlinear measures are more suitable than linear ones in identifying dynamical changes when studying the EEG signal. The results from the simulation study helped in the interpretation of the results on EEG in terms of their stochasticity and complexity. The set-up and the results of the application on the EEG are presented in Ch. 6.

The study on mutual information and the other nonlinear correlation measures was extended in order to include also causal effects. The study focused on information causality measures detecting the direction of interdependencies between interacting systems. However, it was extended in order to include also other types of causality measures, i.e. state-space and synchronization measures, for comparative purposes. The evaluation of the causal measures was assessed initially on nonlinear interacting systems from synthetic data (well-known coupled systems for a range of coupling strengths). This study has led in the modification of some existing measures which are mainly based on the frame of surrogates or on the estimation scheme of the measures. The corrected measures aim to reduce the bias of estimation and specifically to preserve the values of the measures at the zero level in case of no causal effect. Along with the study of the performance of the causal measures and their modifications, it was also investigated the influence of the embedding dimension selection for the reconstruction of the state space, time series length and noise level in the performance of the measures. The proposed corrected measures seem to improve the statistical significance of the original measures and reduce the bias in case of no causality.

As EEG recordings are multivariate signals, one should also examine whether interdependencies between the EEG recordings exist and investigate the direction of the information flow between the different brain areas. Thus, the findings from the simulation study on the causality measures were tested again on EEG recordings in order to determine the direction of information flow between different brain areas. In Ch. 7, the original and modified causality measures are presented along with the evaluation scheme and the applications on EEG data.

Completing this study, it is obvious that detection of dynamical changes and information flow on real data is not an easy task. The results from the simulation studies are indicative of the effectiveness of a measure on a certain application, however it not always true when real, noisy and complex systems are examined. It was observed that nonlinear measures do not always perform better than linear ones on the discrimination among different dynamical systems or brain states. The extracted modified measures that were developed as part of this work may have improved the discriminating power and statistical significance of some existing measures, but still they do not perform dramatically better on EEG. The overall conclusions of this study are drawn in Ch. 8.

Chapter 2

Physiology

In this chapter, the background about the physiology of the brain and electroencephalogram (EEG) is given. It is explained how the brain activity can lead to the seizure onset and what is the role of the neurons in this phenomenon. EEG types and bands, EEG recording methods and types of electrodes and EEG artifacts are also discussed here. Finally, preprocessing methods of EEG are also discussed.

Brain is divided into four main areas due to their functions. The two hemispheres of the brain are constituted from four main parts. The frontal part of the brain (left (LF) and right (RF)) contains the motor area, the temporal (left (LT) and right (RT)) lobe placed at the two sides of the brain designates the hearing and the memory, the occipital (OC) lobe placed at the back of the head designates the vision, and the parietal or middle (MI) part of the brain designates the sensory area.

2.1 EEG

EEG records measure the electrical activity of the brain, which is small and measured in microvolts (mV). Figure 2.1 shows one second of EEG signal from normal activity. There are two main types of EEG in terms of the recording method; ex-

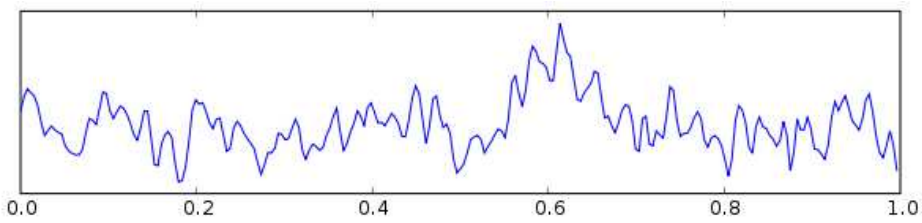

Figure 2.1: One second of EEG signal.

tracranial or scalp and intracranial EEG.

2.1.1 Extracranial EEG

Scalp potentials are measured with electrodes placed on the scalp at specified locations. An EEG voltage signal represents a difference between the voltages at two electrodes and the display of the EEG may be set up in several ways. The are four types of representation of the EEG channels which are called montages. In bipolar montage, measurements from each channel represent the difference between two adjacent electrodes. In referential montage, each channel represents the difference between a certain electrode and a designated reference electrode. In the average reference montage, the outputs of all the amplifiers are summed and averaged, and this averaged signal is used as the common reference for each channel. Finally, in the Laplacian montage each channel represents the difference between an electrode and a weighted average of the surrounding electrodes.

The electrodes are attached to an individual wire or use caps into which electrodes are embedded. The electrode placement is usually done according to the 10–20 International System where electrodes form a regular grid covering the frontal, temporal, parietal and occipital areas of the scalp (Jasper, 1958). The original 10–20 system included only 20 electrodes, however high-density arrays were developed which contain up to 256 electrodes evenly spaced around the scalp, e.g. high density electrode recordings of 64 electrodes (Tucker, 1993).

2.1.2 Intracranial EEG

Intracranial EEG is measured by electrodes placed near the surface of the brain in order to observe deeper brain activity. The EEG electrodes are placed under the surface of the scalp and skull directly on the brain surface. These electrodes provide high-quality and detailed measurements of the brains electrical activity. There are two categories of intracranial EEG. The first one, called intraoperative electrocorticography (ECoG), has the electrodes placed on the brain in relation to a surgical procedure. The second category for extracting intracranial EEG is called invasive monitoring or extra-operative monitoring and the electrodes are placed in or over the brain at the epileptic focus area.

There are three main types of electrodes used for intracranial EEG; strip, grid, and depth electrodes. Strip electrodes are multiple electrodes (usually eight) attached to a rectangular grid and implanted in the subdural cortical layer of the brain to record electrical activity. They are used when the examined brain region is small and usually consist of eight electrodes. Grid electrodes are multiple electrodes attached to a rectangular grid and implanted in the subdural cortical layer of the brain to record electrical activity. Grids require the most extensive procedure; sheets of electrodes are placed over the surface of the brain via a craniotomy. Grids are very useful when the general area of seizure onset is known, but fine detail is needed as the electrodes are closely spaced. Depth electrodes, which look like a single thin wire, may be used to access structures deep within the brain, such as the amygdala and the hippocampus. These are the only EEG electrodes that actu-

ally record from within the substance of the brain. The main advantage of depth electrodes is the ability to record from parts of the brain that are not on the surface, where grids and strips record from.

2.1.3 Artifacts in EEG

Although EEG is designed to record cerebral activity, it also records electrical activities arising from sites other than the brain. The recorded activity that is not of cerebral origin is termed artifact and can be divided into physiologic and extra-physiologic artifacts (mechanical). Artifacts are mostly observed on extracranial EEG. Physiologic artifacts are generated from the patient and arise from sources other than the brain. Physiologic artifacts stem from eye blinks, eye movements or muscular movements. The EEG frequency refers to the rhythmic repetitive activity and is given in Hz. Eye blinks result in a slow signal (with a frequency less than 4 Hz) that appears mainly on the frontal area and is symmetric between the two hemispheres (see Fig.2.2a). Artifacts from eye movement also appear mainly on frontal

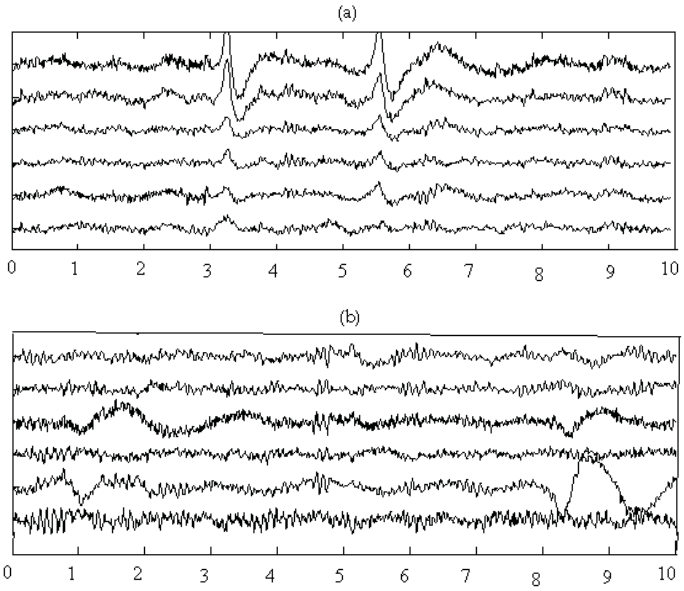

Figure 2.2: EEG signal with artifacts from (a) eye blink and (b) electrode movement.

and temporal area and are more propagated than eye blinks. In this case the signal is anti-symmetric between the two hemispheres. Muscular activity results in high frequency signals (frequency larger than 13 Hz), often much higher than cerebral signals. The main head muscle is the jaw which can create an important artifact in the temporal area (lasting up to 5 sec). Frontal muscles can appear as well since they are located just under the electrodes. Extra-physiologic artifacts arise from outside the body, i.e. electrode movement. Wire or electrode movements create

a low frequency artifact (frequency less than 2 Hz) on that electrode, where the signal has often high amplitude (see Fig.2.2b). Often a mechanical artifact appears after a head movement.

EEG in clinical settings always contains artifacts. There are different ways to handle the artifacts. Methods for artifact removal have been developed, however these procedures may influence the signal. Some certain artifacts, as eye movements, may be detected and extracted from the signal. However, explicit detection of artifacts is too hard to materialize as there is a very wide variation of artifacts. Some robust methods also exist that are not sensible to artifacts, but these are effective to a small degree.

2.1.4 Extracranial vs intracranial EEG

Scalp EEG recordings have an excellent temporal resolution, however they do not provide enough detail in order to detect the epileptic focus area. The scalp is relatively far from the brain tissue as it is separated by the skull, skin and the fluid. All of these barriers degrade the EEG signal. Intracranial electrodes provide more detailed EEG recordings as they are placed directly on the surface of the brain or within the brain and thus give a more accurate reading, have a better spatial resolution and less artifacts than scalp EEG. However, the recording procedure of intracranial EEG is more complicated and invasive compared to the scalp EEG. Moreover, the number of intracranial electrodes implanted in the brains is very small compared to anything approaching full spatial coverage, even for recordings at intermediate spatial scales. Recording from intracranial electrodes showed reproducible results in clinical applications, whereas the scalp electrodes showed much more varying results. Seizure-prediction studies to date have been carried out on both intracranial, e.g. Mormann et al. (2005); Iasemidis et al. (2005); Osterhage et al. (2008) and scalp recordings, e.g. Hively and Protopopescu (2003); Li (2006); Osowski et al. (2007).

2.1.5 EEG bands

EEG is typically described in terms of rhythmic activity and transients. The rhythmic activity is divided into bands of frequency. EEG activity shows oscillations at a variety of frequencies. Several of these oscillations have characteristic frequency ranges, spatial distributions and are associated with different states of brain functioning. Specifically, up to 3 Hz is the δ band (adults at deep sleep, babies), θ band varies in the range of 3–7 Hz (children up to 13 years old), α band at 7–12 Hz (relaxed or closing the eyes), β band at 12–30 Hz (thinking, active concentration) and γ band at 30-100Hz or more Hz (certain cognitive or motor functions). All bands are characterized by certain forms of electromagnetic oscillations. An example of EEG signal filtered to present only the α and respectively the δ waves is given in Fig.2.3.

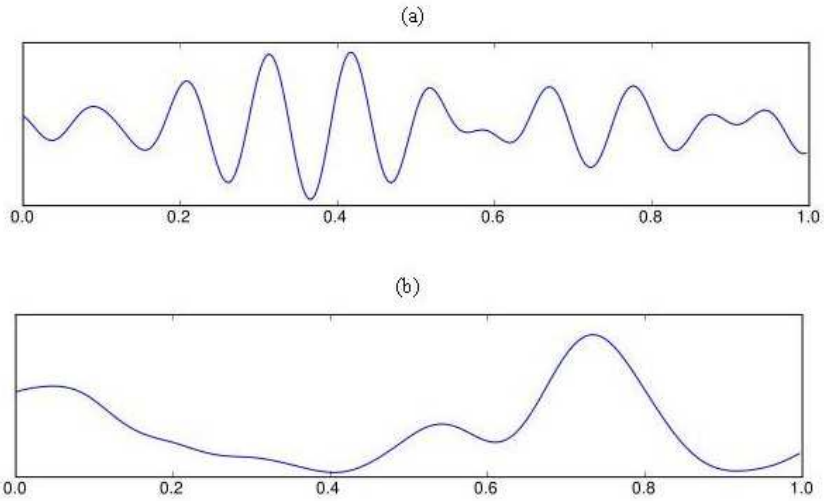

Figure 2.3: One second sample EEG filtered to present only (a) α and (b) δ waves, respectively.

2.1.6 EEG preprocessing

EEG activity is quite small, measured in microvolts (mV) and thus needs to be amplified. The EEG is digitized from the amplifiers with analog-to-digital sampling techniques. Bandpass filters are applied when interested on a certain band of the signal activity. For temporal filtering a fixed-frequency bandpass filter is usually used. Bandpass filtering has the disadvantage that the range of explored frequencies must be arbitrarily defined prior to performing the analysis and does not allow the simultaneous exploration of the whole range of the EEG frequency spectrum.

Due to the diffusion of the skull and skin, a scalp EEG electrode actually measures a mixture of signals from several neuronal clusters. Therefore, spatial and temporal filters are commonly applied before any further analysis of the EEG signal (Perrin et al., 1989; Muller-Gerking et al., 1999; Dornhege et al., 2006). Spatial filters, such as a Laplacian filter (Srinivasan et al., 1998; Song and Epps, 1998), are usually applied to concentrate the signals to a single neuronal cluster. Morphological filters are efficient tools that decompose raw EEG signal into several physical parts. Background activity and spike component can be separated and the main morphological characteristic of spikes can be retained.

Mathematical transformations are applied to signals to obtain further information that is not readily available in the raw signal. The Fourier transform is one of the most popular transformations and is used to express the oscillations of the signal across the entire frequency spectrum. Gabor transform technique (Yao, 1993) is a special case of Fourier transform that maximizes the relationship between the time and the frequency that has been applied to analyze the main characteristics of human sleep EEG (Schonwald et al., 2003). A wavelet transform (Chui, 1992)

is a transform that can map a function of time into a space of frequencies. Its advantage compared to the Fourier transform is that the wavelet transform is capable of providing the time and frequency information simultaneously, hence giving a time-frequency representation of the signal. Wavelet transforms are efficient time-frequency decomposition methods, also used on EEG signals (Schiff et al., 1994; Rosso et al., 2002).

Signal processing methods both in the time and the frequency domain as averaging and filtering have been used on EEG data. EEGLAB software toolbox for Matlab, available from http://www.sccn.ucsd.edu/eeglab/, contains standard data analysis functions, such as data filtering, data epoch extraction, artifact removal, average reference conversion and data resampling. BESA (Brain Electrical Source Analysis) available from http://www.besa.de/, provides a software for research and clinical applications in the fields of physiology, neuroimaging and neuroscience, such as processing of EEG and MEG. There are also other commercial or freeware packages. In this work, these packages were not used and all the programs were built by our research group.

2.2 MEG

Magnetoencephalography (MEG) is an imaging technique used to measure the magnetic fields produced by electrical activity in the brain. Current flow within cortical neurons produces a surrounding neuromagnetic field that can be measured above the head surface by specialized conducting detectors. MEG is a non-invasive method that allows the localization of the electrical activity of nerve cells within the brain with a few mm accuracy and fast time resolution. The magnetic field generated by intracellular currents from an individual neuron is extremely small to be detected outside the head. Several thousand synchronously active neurons are needed to generate the field measured by the MEG. MEG measures is the very weak signal from the magnetic fields of the brain, while simultaneously discriminating against interference from strong background noise. Thus, MEG is a noninvasive functional imaging technique in which the weak magnetic forces associated with the electrical activity of the brain are recorded externally on the scalp. One of the clinical applications of MEG is in detecting and localizing epileptiform spiking activity in epileptic patients.

EEG measures the electric field of brain and MEG measures the magnetic field. Although they are generated by the same neurophysiologic processes, however there are important differences concerning the neurogenesis of MEG and EEG (Cohen and Cuffin, 1983). MEG has a better spatial resolution than EEG, as magnetic fields are less distorted by the resistive properties of the skull and scalp and is more sensitive to superficial cortical activity. MEG has the advantage over scalp EEG that is reference-free. It is not fully elucidated how much area of the spike activity contributes to the magnetic field of the scalp, and whether MEG can detect spike activities generated in a brain area.

2.3 Brain States

The period of time between two seizures in epilepsy, where the brain is considered to be in normal activity, is called interictal state. The interictal state usually corresponds to more than 99% of the life of epileptic patients. The state of the brain before a seizure occurs is called preictal state. Ictal state refers to the seizure period. Finally, postictal state refers to the state shortly after the seizure.

2.4 EEG Data for the Evaluation of the Nonlinear Measures

The present study aims to assess the discriminating power of a set of nonlinear measures, in order to apply the findings for the discrimination of different brain states of epileptic patients. The dynamical changes in the brain dynamics before the seizure are investigated at the conditions of typical clinical practice, where EEG recordings are delivered without preprocessing (e.g. artifact removal).

The EEG data were recorded on a routine basis at the Department of Neurodiagnostics, Rikshospitalet University Hospital, Norway. The methods considered in this study do not rely on very high frequency components, and therefore the data were high-pass filtered at 0.3 Hz and low-pass filtered at 40 Hz. Data were sub-sampled at 100 Hz after smooth interpolation. Patients seizures were mainly generalized. Extracranial EEG records have been used with either 25 or 63 electrodes (channels), as in Fig.2.4. EEG with (28) intracranial electrodes from one

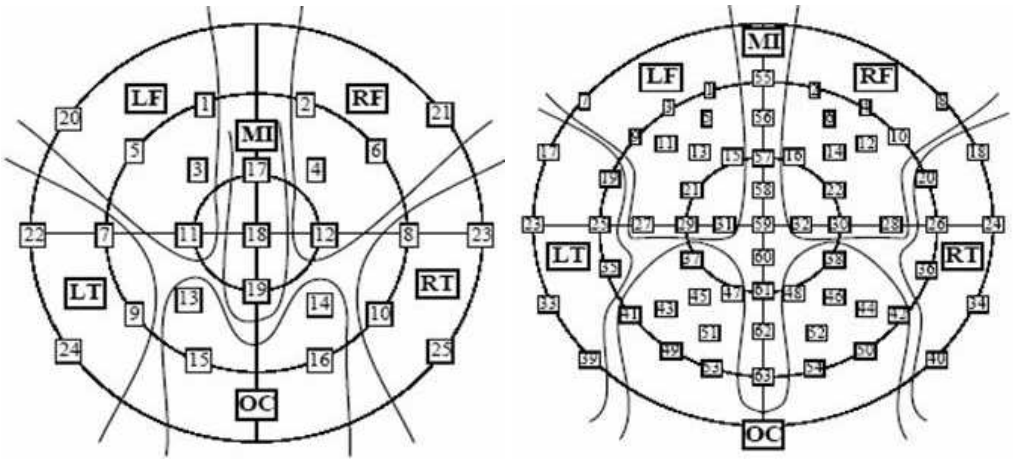

Figure 2.4: Extracranial EEG recording system with (a) 25 electrodes and (b) 64 electrodes.

patient was also examined for comparative studies.

EEG is a highly complex and dynamic signal. It underlies different rhythms (circadian and ultradian) and changes due to changes in arousal, e.g. sleep changes

the characteristics of the EEG, so that EEG is prone to artifacts. The EEG signals depend on the recording setup. Different recording montages give different spatial filtering and hence different information in the EEG (Nunez, 1995). The reported findings in the applications, that will be discussed in the following chapters, are mainly based on EEG measurements with a calculated average reference. However, changing the spatial filtering parameters might have led to different results.

Chapter 3

Methodology

The first section of this chapter introduces the basic concepts of dynamical systems, chaos theory, fractal theory and nonlinear time series analysis and how these concepts are connected. Definition and the main properties of dynamical/ chaotic systems are presented and the main tools of nonlinear time series analysis are described. In the second part of the chapter, the statistical analysis tools that have been used are presented, e.g. the concept of statistical testing.

3.1 Dynamical Systems and Chaos Theory

3.1.1 Dynamical systems

Dynamics is the study of change. A dynamical system is a system of variables which change with time and may interact. Dynamical systems are divided in two types; deterministic and stochastic. A dynamical system is called deterministic when its controlling mechanism is fully understood, whereas a systems that involves randomness in its mechanism is called stochastic. The variables of the system as a function of time determine the course of the system's behavior and describe the state of the system at any given time. The state space of a system is a mathematical construction, defined by the active system variables. The state of a dynamical system is determined by a point and the evolution of the system is given by a trajectory in the state space forming a geometrical shape called attractor.

Chaos theory focuses on the study of nonlinear dynamical systems, i.e. systems whose output is not proportional to its input, and describes their behavior. A deterministic nonlinear system is called chaotic system if its trajectory shows sensitivity to initial conditions, i.e. two trajectories starting at nearby points diverge in the course of time. Thus a dynamical system must involve some form of nonlinearity in order to display chaotic behavior. Since nonlinear systems can exhibit chaos, this is a natural starting point when irregularity is present in a signal. The theory of fractals was developed in the same period with chaos theory and the two theories are directly related. Fractals (Mandelbrot, 1974) are rough or frag-

mented geometric shapes that are self similar at different scales. Strange attractor are the complicated geometrical shapes to which chaotic systems evolve and are often fractals.

The two main properties of chaos is the sensitivity to the initial conditions and the self-similarity. The sensitivity to the initial conditions reflects the fact that small changes in the starting conditions of a chaotic system will drastically change the long-term behavior of the system. Because of the extreme dependence on the initial conditions, one cannot create a model that will accurately predict future values of the system. A common source of such sensitivity to initial conditions is that the dynamics involve a repeated stretching and folding. The stretching operation tends to quickly separate nearby points, while the folding operation ensures that all points will remain bounded in some region of the state space. The competitive effect of repeated stretching and folding produces very complex and irregular structures in state space. The second property of chaotic systems is that strange attractors are self-similar, i.e. similar to themselves at all scales.

Topological Invariants

The two prominent properties of chaotic systems can be quantified by characteristic measures on the attractors, i.e. the trajectories of the systems. These nonlinear characteristics of a dynamical system in the framework of nonlinear dynamical systems and chaos are defined as invariant measures of the systems, i.e. constant magnitudes that are not affected by the evolution of the system. A topological invariant of a space is a property that depends only on the topology of the space. Topological invariants are based on the ”stretching” and ”folding” mechanisms present in the dynamics of the system, are stable under parametric variations and allow system features to be verified independently. Static invariants may involve the statistical characterization of the distribution of points in the phase space, e.g. fractal dimension (Feder, 1988), correlation dimension (Grassberger and Procaccia, 1983). Dynamical invariants are based on the evolution properties of the trajectories and study invariant system properties under coordinate transformations, such as Lyapunov exponents (Abarbanel et al., 1993), and entropy. Only definitions of some of the most important invariant measures are given here, as their estimation is based on the framework of nonlinear time series analysis, which is discussed in the next section of this chapter. For the estimation of these topological invariants, but also for many other linear and nonlinear measures, many software packages have been developed, e.g. TISEAN (http://www.mpipks-dresden.mpg.de/ tisean).

Dimensions

The dimension of an attractor is clearly the first level of knowledge necessary to characterize its properties and it is an indicator of the degree of complexity of the system. The complexity of a system is indicated from the effective (or not) prediction of future outputs or states of the system. Euclidean dimension is given

by the number of coordinates required to specify a point in the state space. The topological dimension is an expansion of Euclidean dimension. The topological dimension reflects the number of real parameters necessary to describe different points in the state space. A single point therefore has a topological dimension equal to zero, a curve has dimension one, a surface has dimension two and so on.

Fractal dimension

Fractal dimension or box counting dimension is based on scaling the mass of the attractor with size. The general formula to compute the fractal dimension is

$$D_0 = \lim_{r \to 0} \frac{\log(\frac{1}{M(r)})}{\log(r)} \tag{3.1}$$

where $M(r)$ is the number of hypercubes of a given dimension with side length r required to cover the attractor. This definition has many practical limitations for its estimation. In fractal dimension only the geometrical structure of the attractor is taken into account without considering the distribution of the points on the attractor.

Information dimension

The information dimension is a generalization of the fractal dimension that takes into account the relative probability of the hypercubes used to cover the set (Balatoni and Renyi, 1956). The information dimension is given by

$$D_1 = -\lim_{\epsilon \to 0} \frac{H(\epsilon)}{\log \epsilon} \tag{3.2}$$

where

$$H(\epsilon) = \sum_{i=1}^{N(\epsilon)} P_i \log \frac{1}{P_i} \tag{3.3}$$

is the Shannon entropy of the state variable discretized by the hypercubes and P_i is the probability contained within the i-th hypercube. The entropy $H(\epsilon)$ is the amount of information necessary to specify the state of the system within an accuracy ϵ. D_1 shows how fast the information necessary to specify a point on the attractor increases as ϵ decreases.

Correlation dimension

Correlation dimension is a measure of the dimensionality of the space occupied by the attractor and gives an estimate of the fractal dimension of the attractor. For a set of N points in an m-dimensional space, the correlation integral is defined as

$$C(\epsilon) = \frac{g}{N^2} \tag{3.4}$$

where g is the total number of pairs of points which have a distance between them that is less than a small positive constant ϵ. If $C(\epsilon)$ increases as a power law function of ϵ, then $C(\epsilon)$ versus ϵ on a log-log plot should be a straight line, and the slope of this line is the estimate of the correlation dimension D_2. The correlation dimension D_2 is defined as

$$D_2 = \lim_{\epsilon \to 0, N \to \infty} \frac{\log C(\epsilon)}{\log \epsilon} \tag{3.5}$$

The dimensions D_0 and D_1 can be estimated in the same way.

Lyapunov exponents

The property of chaotic systems to be sensitive on initial conditions has as a result that trajectories that are very close, deviate exponentially. Thus, Lyapunov exponents of a dynamical system are quantities that characterize the rate of separation of infinitesimally close trajectories. A negative Lyapunov exponent indicates a local average rate of contraction while a positive one indicates a local average rate of expansion. The mean exponent that corresponds to the direction of the largest deviation of the trajectories is the maximum Lyapunov exponent. For flows, at least one Lyapunov exponent is zero and corresponds to the direction of the motion of the trajectory. If the dynamical system is not chaotic, then the system has no positive Lyapunov exponents. For dissipative systems the sum of Lyapunov exponents should be negative. The Lyapunov spectrum (the set of Lyapunov exponents) can be used to give an estimate of the fractal dimension of the considered dynamical system according to the KaplanYorke conjecture (Kaplan and Yorke, 1979)

$$d_L = K + \frac{\sum_{i=1}^{K} \lambda_i}{\lambda_{K+1}} \tag{3.6}$$

where λ_i are the Lyapunov exponents and K is the maximum integer such that the sum of the K largest exponents is still non-negative and gets negative when the λ_{K+1} is added.

3.1.2 Nonlinear time series analysis

Nonlinear dynamical systems with complex and irregular behavior can be characterized by means of nonlinear time series analysis. The aim of nonlinear time series analysis is to study characteristics of the deterministic dynamical systems, describe them with a proper model and predict future values. The nonlinear time series analysis methods are based on the theory of dynamical systems. A time series showing irregular behavior may be analyzed under the hypothesis that it is derived from a nonlinear dynamical system that may be contaminated by noise. There are two types of noise; observational or measurement noise and dynamical or system noise. Observational noise does not effect the future evolution of the system and is the error between the true value in a system and its observed value due

to imprecision in measurement. In contrast, dynamical noise does effect the future evolution of the system. In this case, the output of a dynamical system becomes corrupted with noise, and the noisy point is used as input during the next iteration.

Reconstruction theorem

The main tools to study time series from nonlinear dynamical systems is the reconstruction of the state space and the estimation of invariant measures of the underlying dynamical system. The state space of a dynamical system is a space in which all possible states of the system are represented with points in an d-dimensional space. Each possible state of the system at a given time corresponds to a unique point in the state space. The embedding theorem which enables the reconstruction of the asymptotic dynamics of a system from a measured time series is introduced here.

Let $\{x_t\}$, $t = 1, \ldots, n$ be a measured univariate time series from a dynamical system, whereas its original state space is an unknown manifold of some dimension d. If t_s is the sampling interval then $x_t = x(kt_s)$, for $k = 1, \ldots, n$. The evolution of the system at discrete time steps is determined by $\mathbf{s}_{t+1} = \mathbf{f}(\mathbf{s}_t)$, where $\mathbf{s}_t$ is the state at time t and $\mathbf{f}$ is the the system function defined on the original manifold. The time series $\{x_t\}$ is derived from the equation $x_t = h(\mathbf{s}_t)$, where h is a measurement function. The aim is the reconstruction of the original state space, when only time series x_t is known and the system dynamics on the reconstructed state space should be homeomorphic to the dynamics of the original attractor. For noise-free infinite time series, Taken's theorem (Takens, 1980) assures that all dynamical properties of the original system can be preserved in a reconstructed state space $\Re^m$, as long as $m \geq 2d+1$, whereas d is the dimension of the attractor and m is called embedding dimension.

The most common reconstruction method is the method of delays, whereas points in $\Re^m$ are formed from the measured time series as $\mathbf{x}_t = (x_t, x_{t-\tau}, \ldots, x_{(m-1)\tau})'$, where τ is called the delay time. The two parameters of the method of delays define the time window length $\tau_w = (m-1)\tau$, which is usually chosen around the mean orbital period for flows, and thus depends on the form of the oscillations of the time series (Kugiumtzis, 1996), while for maps τ is set to 1. Selection of m and τ is also important and usually the selection of the one parameter is often not independent of the other, although it has been shown that their selection should not be independent of each other (Kugiumtzis, 1996; Olbrich and Kantz, 1997).

Selection of the embedding dimension

The most common method to obtain the optimum embedding dimension m for phase space reconstruction is the false nearest neighbors (FNN) procedure (Kennel et al., 1992). The main idea of this method is to increase gradually the embedding dimension of the state space and check the neighborhoods of points embedded in state spaces of increasing dimensions. Points apparently lying close together due

to the projection may be separated in higher embedding dimensions. For an insufficient choice of the embedding dimension, the topological structure will no longer be preserved and points will be projected into neighborhoods of other points to which they would not belong in higher dimensions. These points are called false neighbors. False neighbors are not typically mapped into the image of the neighborhood, but somewhere else, so that the average radius of the neighborhoods becomes quite large. Specifically, for each point $\mathbf{x}_i$ in the time series one should look for its nearest neighbor in an m-dimensional space and measure the percentage of false neighbors as m increases. The criterion that the embedding dimension is high enough is that the fraction of points with false neighbors is sufficiently small. The smallest embedding dimension in order to start looking for false neighbors is usually set to be 1.

Selection of the delay time

Various methods for choosing an adequate delay time τ have been proposed. The time delay τ has to be large enough so that the information from measuring the value of variable x at time $i + \tau$ will be relevant and significantly different from the information from knowing the value at time i. However, τ should not be larger than the typical time in which the system loses memory of its initial state. If τ is chosen larger than it should, then the attractor will be folded and points will look more or less random since they will be uncorrelated.

Schuster (1988) suggested choosing as the delay time, the first zero of autocorrelation function to ensure the linear independence of the coordinates. The sample autocorrelation function of the time series $\{x_t\}, t = 1, \ldots, n$ indicates the degree of correlation between values of the series separated by τ lags and is given as

$$r(\tau) = \frac{\sum_{t=1}^{n-\tau}(x_t - \overline{x})(x_{t+\tau} - \overline{x})}{\sum_{t=1}^{n-\tau}(x_t - \overline{x})^2}, \tag{3.7}$$

where $\overline{x}$ is the mean of the time series.

The most common method for selecting the time delay originates from Fraser and Swinney (1986) and is the lag of the first local minimum of mutual information. Mutual information (MI) is a measure of mutual dependence between two random variables quantifying the amount of uncertainty about one variable reduced when knowing the other. MI of two continuous random variables X and Y is defined as

$$\mathcal{I}(X, Y) = \int_X \int_Y f_{X,Y}(x, y) \log_a \frac{f_{X,Y}(x, y)}{f_X(x) f_Y(y)} \mathrm{d}x\mathrm{d}y, \tag{3.8}$$

where $f_{X,Y}(x, y)$ is the joint probability density function (pdf) of X and Y, $f_X(x)$ and $f_Y(y)$ are the marginal pdfs of X and Y, respectively. For a time series $\{x_t\}$, $t = 1, \ldots, n$, sampled at fixed times t_s, MI is defined as a function of the delay τ assuming the two variables $X = X_t$ and $Y = X_{t-\tau}$, i.e.

$$\mathcal{I}(\tau) = \mathcal{I}(X_t, X_{t-\tau}). \tag{3.9}$$

The different estimation methods of mutual information are described in the next chapter.

The reconstruction of the state space is the prerequisite step for the estimation of most nonlinear system characteristics, such as Lyapunov exponents, correlation dimensions and entropies. Many of the nonlinear measures that have been used in this work, estimate such system characteristics and use the embedding theorem in their estimation.

3.2 Statistical analysis

The statistical tools that were used in the study are presented in this section. Descriptive statistics, such as the mean and standard deviation, can be used to summarize the data and to describe the sample. Inferential statistics is used to model patterns in the data and account for randomness and draw inferences about different samples that may stem from a hypothesis testing, from the investigation of the estimates of numerical characteristics, from descriptions of the association of the samples or from the modelling of the relationships of the samples (e.g. regression, analysis of variance).

A statistical test is a framework for searching answers to questions about experimental data. The measured time series of an observable quantity reflects the behavior of the system or the process it originates from. By exploring the underlying mechanism of the time series, one can realize whether the system is random or deterministic and predictable. If the underlying mechanism can be inferred from the data then a mathematical model may be formed in order to make prediction.

A property of the examined data is quantified by some real function or statistic $\theta(x)$, which is called discriminating statistic. Let consider that the null hypothesis H_0 is set $\theta = \theta_0$. In the framework of the Fisher's significance testing, the probability of an outcome of magnitude $\theta(x)$ or larger is investigated, provided that the null hypothesis is valid (Neyman and Pearson, 1933; Fisher, 1935, 1973). A result is called statistically significant if it is unlikely to have occurred by chance. The p-value of a test is the probability of obtaining a result at least as extreme as the one that was actually observed, assuming that the null hypothesis is true. The significance and the p-value are estimated, under the condition of the validity of H_0. Thus, p-value is a measure that indicates the validity of H_0; for small p-values the H_0 is rejected. The probability of the false rejection of H_0 is called Type I error and is denoted by α and the probability of the false acceptance of H_0 is called Type II error and is denoted by β. The power of a statistical test is the probability that the test will reject a false null hypothesis, and thus the power is equal to $1 - \beta$. The significance level α is chosen before the test is done.

In the classical approach to significance and hypothesis testing, the discriminating statistic is carefully tailored to be sensitive to deviations of H_0. It is usually preferred to choose discriminating statistics with known distributions for a given null hypothesis. However, the distribution of any statistic and its confidence range for a

given α can be accurately estimated by Monte-Carlo simulation method (Barnard, 1963; Hope, 1968). The main idea of the Monte-Carlo simulation method is to compute values of the statistic for many different realizations of the null hypothesis and to empirically estimate the distribution of it from these values. Generation of the realizations depends on the H_0, i.e. a random permutation of the original data generates independent and identically distributed noise (iid). Monte-Carlo methods rely on the computation of random or pseudo-random numbers and probability statistics. The two main Monte Carlo methods for statistical testing are the bootstrap and the randomization. The latter is used particularly in the nonlinear analysis of time series under the term surrogate data test.

3.2.1 Surrogate data test

The surrogate data technique is used when the hypothesis of a particular process type is investigated for a given time series. Surrogates should preserve all the important statistical properties of the original data but the property that are tested for. The most common application of surrogate data test is for testing the existence of nonlinearity in the data. In this case, surrogates are constructed as realizations of a linear stochastic process replicating the linear properties of the original data. The most general null hypothesis when testing for nonlinearity is that the original scalar data are generated by a Gaussian normal process measured through a static and possibly nonlinear transform, where the surrogates should have the same autocorrelation and amplitude distribution as the original time series. Different algorithms generating surrogate data have been developed (Theiler et al., 1992; Schreiber and Schmitz, 1996; Schreiber, 1998; Kugiumtzis, 2002).

The surrogate data test for any null hypothesis H_0 is given here in detail. First, M surrogate data representing H_0 are generated. A discriminating statistic is used, e.g. mutual information that is sensitive to deviations from H_0 for the nonlinearity test because it can detect nonlinear dynamics. The statistic is estimated from the original time series and the surrogates, and let q_0 be the value of the statistic estimated on the original data and $q_1, \ldots, q_M$ on the surrogates. If q_0 does not lie in the distribution of $q_1, \ldots, q_M$ then H_0 can be rejected.

Often the assumption that $q_1, \ldots, q_M$ follow a normal distribution is made and the rejection of H_0 is given parametrically from the significance s and the p-value of the discriminating statistic

$$s = \frac{q_0 - \overline{q}}{s_q}, \;\; p = 2(1 - \Phi(|s|)) \tag{3.10}$$

where $\overline{q}$ and s_q are the mean and the standard deviation of $q_1, \ldots, q_M$, and Φ is the cumulative distribution function of the standard normal distribution. Customarily s is given in units of 'sigmas'. Assuming that the test statistics has essentially a gaussian distribution under H_0, the rejection region at $\alpha = 0.05$ is $|q_0| > 1.96$. Thus, a small p-value (e.g. less than $\alpha = 0.05$) indicates the rejection of H_0.

If one cannot assume that the test statistics has essentially a gaussian distribution than a non parametric approach can be considered using the rank ordering. The values q_0 and $q_1, \ldots, q_M$ are sorted and the rejection of H_0 depends on whether q_0 is ranked first, last or in between. The p-values of the test are estimated then depending on the rank value of q_0; $p = r_0/M$ for one sided test or $p = 2(1 - r_0/M)$ for two sided test, where r_0 is the rank value of q_0. Such a technique demands a large number of surrogates.

3.2.2 Trend detection

Detection of dynamical changes in the dynamics of a system is assessed by a parametric test indicating linear trends in time dependent components. The parametric test for the detection of a linear trend is based on the linear regression of the values of a time series X on time T. Let

$$S_{tt} = \sum_{i=1}^{n} t_i^2 - n\bar{t}^2, \;\; S_{tx} = \sum_{i=1}^{n} t_i x_i - n\overline{tx}, \tag{3.11}$$

and

$$\hat{\beta} = \frac{S_{tx}}{S_{tt}}, \;\; s_\beta = \frac{S_{tt}S_{xx} - S_{tx}}{(n-2)S_{tt}}, \tag{3.12}$$

and the test statistic is defined as

$$T = \frac{\hat{\beta}}{s_\beta}. \tag{3.13}$$

T follows a Student's t-distribution with $n - 2$ degrees of freedom, where n is the sample size. The rejection of H_0 depends on the significance level α. The p-values of the test can be defined as in Eq.(3.10), whereas instead of Φ, the F function of the Student's t distribution is used. The statistical method of the trend detection has been used in applications for the detection of dynamical changes in the brain dynamics of epileptic patients before a seizure.

3.2.3 Statistical tests for parameters

For the discrimination of two populations (dynamical systems or brain states), two statistical tests were used; the t-test for means and the Wilcoxon rank sum test for medians. In both tests it is assumed that there is no time dependence among the values of the two populations.

t-test

The t-test examines the null hypotheses that the means of two samples $\{x_t\}, t = 1, \ldots, n_x$ and $\{y_t\}, t = 1, \ldots, n_y$, are equal. The assumptions of the test is that

each variable follows a normal distribution. For equal variances, the test statistic is defined as

$$t = \frac{\overline{x} - \overline{y}}{s\sqrt{1/n_x + 1/n_y}} \tag{3.14}$$

where $\overline{x}, \overline{y}$ are the means of x and y, respectively and s is the estimate of the common standard variation of the samples given as

$$s = \frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 1} \tag{3.15}$$

where s_x and s_y are the sample variances for x and y. The test statistic t follows a Student's t-distribution with $n_x + n_y - 2$ degrees of freedom.

When the variances of x and y cannot be assumed to be equal , the test statistic is defined as

$$t = \frac{\overline{x} - \overline{y}}{\sqrt{s_x^2/n_x + s_y^2/n_y}}. \tag{3.16}$$

The distribution of the test statistic is again the Student's t distribution but the degrees of freedom are calculated from the expression

$$\frac{s_x^2/n_x + s_y^2/n_y}{(s_x^2/n_x)^2/(n_x - 1) + (s_y^2/n_y)^2/(n_y - 1)}. \tag{3.17}$$

Wilcoxon rank sum test

The second test is the Wilcoxon rank sum test for the null hypotheses that the medians of the two tested populations are equal. Here, the assumption for normal distribution is relaxed and it is only required that the underlying distributions of the two samples have the same general form. In order to perform the test, the observations from both samples are arranged into a single ranked series, i.e. all the observations are ranked without regard to which sample they are from. The ranks for the observations of each sample are added up and let us denote these sum as n_L, n_S where $n_L < n_S$. The test statistic is defined as

$$Z_W = \frac{W - \mu_W}{s_W}, \tag{3.18}$$

where

$$\mu_W = n_L(n_L + n_S + 1)/2, \;\; s_W = \sqrt{n_L n_L(n_L + n_S + 1)/12}. \tag{3.19}$$

Z_W follows the standard normal distribution N(0,1).

3.2.4 ROC curves

The discrimination of two different populations or states can also be evaluated using Receiver Operating Characteristic (ROC) curves. In this study the two populations are two dynamical systems or different states of brain activity. ROC curves were developed in the 1950's in the framework of statistical decision theory (Green and Swets, 1966) and was originally used during World War II for the analysis of radar images. Nowadays they are mostly used in medical decision-making, but also in radiology, psychology, machine learning and data mining.

ROC curves are used in order to evaluate the diagnostic performance or the accuracy of a test aiming to discriminate between two different cases. Naming one state (population) as "positive" and the other as "negative", ROC makes a graphical representation of the trade-off between the sensitivity (true positive rate) and specificity (false negative rate). Graphically, the ROC curves show the dependence of the complementary of specificity on the sensitivity for varying threshold (cut-off) values. A cut-off point indicates the value that defines the two states, above which a state is considered as 'positive' and below which a state is considered as 'negative'. Thus the position of the cut-off point determines the number of true positive, true negatives, false positives and false negatives. For each cut-off the cumulative function of the distribution regarding each state is calculated. An example of a pair of distributions and the ROC curve is shown in Fig.3.1. The curve always

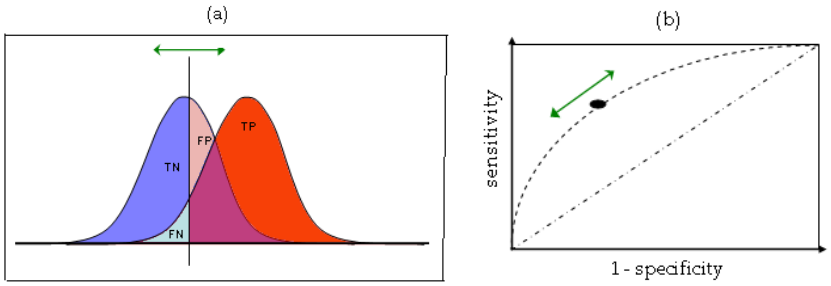

Figure 3.1: (a) A pair of distributions with medium overlap. The vertical line represents a cut-off point (threshold value). TP denotes the true positives, TN the true negatives, FP the false positives and FN the false negatives. (b) The ROC curve for all cut-off points.

goes from point (0,0) to (1,1) varying the threshold cut-off. For the point (0,0), the classifier finds no positives (all values are classified to state "positive"), and for the (1,1) everything is classified as "negative".

The total area under the ROC curve (AUC) is a measure of the performance of the diagnostic test since it reflects the test performance at all possible cut-off levels and thus it is a convenient way of comparing classifiers. The AUC is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one and is equivalent to the Wilcoxon test of

ranks for a corresponding significance level (Fawcett, 2006). A random classifier has an area of 0.5 and there is no distinction of the two states for any cut-off value, while an ideal classifier has an area of 1, i.e. there is absolute distinction for the whole range of cut-off values.

Chapter 4

Mutual Information

This chapter deals with the definition, the properties and the estimation methods of mutual information and presents different applications of mutual information on time series. The most naive estimator was first used in some applications in order to investigate its usefulness. Mutual information was tested as a statistic for the detection of systems' dynamical changes, as well as for independence and nonlinearity test. The effectiveness of mutual information in these applications motivated a further evaluation of mutual information estimators, in order to find the most reliable estimator and also optimize their corresponding free parameters.

All information measures stem from the concept of Shannon entropy (Shannon, 1948). The Shannon entropy is a measure that reflects the randomness or the uncertainty of a variable and is defined in terms of its probability distribution. The Shannon entropy of a variable X is a measure of the amount of uncertainty associated with the values of X. For a discrete variable X, it is defined as

$$\mathcal{H}(X) = -\sum_i p_X(x) \log p_X(x), \tag{4.1}$$

where $p_X(x)$ is the probability mass function of X. If X is a continuous variable, it is called differential entropy and is defined as

$$\mathcal{H}(X) = -\int_X f_X(x) \log f_X(x), \tag{4.2}$$

where $f_X(x)$ is the probability density function of X. The units of the entropy depend on the base of the logarithm. It can be measured in bits (base 2 logarithm) or nats (natural logarithm) depending on the base of the logarithm. The joint entropy of two variables measures how much entropy is contained in a joint system of two random variables and is defined for discrete variables as

$$\mathcal{H}(X,Y) = -\sum_i p_{X,Y}(x,y) \log p_{X,Y}(x,y), \tag{4.3}$$

where $p_{X,Y}(x, y)$ is the joint probability mass function of (X, Y). The conditional entropy of the variable Y given X is defined as

$$\mathcal{H}(X|Y) = -\sum_{x,y} p_{X,Y}(x, y) \ln \frac{p_{X,Y}(x, y)}{p_X(x)}. \tag{4.4}$$

4.1 Mutual Information: Definition and Properties

The mutual information (MI), defined in Eq.(3.9) for continuous random variables X and Y, is an entropy-based measure of both linear and non-linear correlation. MI can be estimated in terms of Shannon entropies as

$$\mathcal{I}(X, Y) = \mathcal{H}(X) + \mathcal{H}(Y) - \mathcal{H}(X, Y). \tag{4.5}$$

Despite of being the sum of entropies, MI is invariant under homeomorphisms. MI is always positive, with equality holding for independent variables, and $\mathcal{I}(X, Y) \leq \mathcal{H}(X) \leq \log_a n$ (Jensen inequality). Zero MI always implies statistical independence, unlike the correlation coefficient, which can be zero for highly dependent non-Gaussian data. For n random variables $X_1, X_2, \ldots, X_n$, the MI is defined as

$$\mathcal{I}(X_1, X_2, \ldots, X_n) = \sum_{k=1}^{n} \mathcal{H}(X_k) - \mathcal{H}(X_1, X_2, \ldots, X_n) \tag{4.6}$$

and is called redundancy. It is known that Shannon entropy cannot be accurately estimated due to finite sample effects (Grassberger, 1988; Kantz and Shürmann, 1996). The estimation of MI in terms of entropies is not discussed here but only in terms of the probability functions, that are present in the entropy expressions.

Assuming a partition of the domain of X and Y, the double integral in Eq.(3.9) becomes a sum over the cells of the two-dimensional partition

$$I(X, Y) = \sum_{i,j} p_{X,Y}(i, j) \log_a \frac{p_{X,Y}(i, j)}{p_X(i) p_Y(j)}, \tag{4.7}$$

where $p_X(i)$, $p_Y(j)$, and $p_{X,Y}(i, j)$ are the marginal and joint probability mass functions over the elements of the one and two-dimensional partition. In the limit of fine partitioning the expression in Eq.(4.7) converges to Eq.(3.9). This may partly justify the abuse of notation of MI for the continuous and the discretized variables. The units of information depends again on the base of the logarithm, as for Shannon entropy. In the presence of a third variable Z, the conditional mutual information $I(X, Y|Z)$ of the variables X, Y given the variable Z is defined as

$$I(X, Y|Z) = H(X|Z) + H(Y|Z) - H(X, Y|Z). \tag{4.8}$$

MI estimated from a time series is defined in Eq.(3.9). In time series analysis, MI is used in order to measure the global correlations, both linear and non-linear,

between their terms with time lag τ. Given two time series $\{x_t\}$ and $\{y_t\}$, $t = 1, \ldots, n$, their mutual information is the average number of bits of X that can be predicted by measuring Y. This relationship is symmetrical, $\mathcal{I}(X, Y) = \mathcal{I}(Y, X)$.

MI is a correlation measure that uses distributions instead of means as autocorrelation function and there is proper transformation of MI in order to satisfy Renyi axioms (Granger and Lin, 1994). Renyi (1959) stated seven properties which should be fulfilled by any dependence measure, $\delta(X, Y)$, between two random variables X, Y defined over the same probability space, and are the following:

- The measure is defined for any pair of random variables X and Y, neither of them being constant with probability one.
- $\delta(X, Y) = \delta(Y, X)$
- $0 \leq \delta(X, Y) \leq 1$
- $\delta(X, Y) = 0$ if and only if X and Y are independent.
- $\delta(X, Y) = 1$ if there is a strict dependence between X and Y, i.e either $X = g(Y)$ or $Y = f(X)$, where $g(\cdot)$ and $f(\cdot)$ are Borel-measurable functions.
- If the Borel-measurable functions $f(\cdot)$ and $g(\cdot)$ map the real line in a one-to-one way into itself, then $\delta(f(X), g(Y)) = \delta(X, Y)$.
- If the joint distribution of (X, Y) is normal, then $\delta(X, Y) = |r(X, Y)|$, where $r(X, Y)$ is the correlation coefficient of X and Y.

This set of axioms has been considered too restrictive by various authors (e.g. (Schweizer and Wolff, 1981; Granger et al., 2004)) and slight modifications of them have proposed.

The distribution of any MI estimator is not known in general. There are generally few analytic results on MI. Theoretical results are obtained for iid variables (Roulston, 1997) or expressions of MI in terms of the correlation coefficient for some known distributions, e.g. Gaussian and Gamma distribution (Pardo, 1995; Hutter and Zaffalon, 2005). Some statistical results on the mean, variance and bias of the MI estimator using fixed partitioning can be found in (Zografos, 1993; Roulston, 1997; Moddemeijer, 1988; Abarbanel et al., 2001). For chaotic systems in particular, the discontinuity of the density function of their variables does not allow for an analytic derivation of the statistics of MI estimators.

4.2 Mutual Information Estimators

The estimation of MI boils down to the estimation of the densities in Eq.(3.9) or probabilities in Eq.(4.7). MI estimation involves one and two dimensional density estimation. Density estimation has been studied extensively and different methods have been suggested and compared in the statistical literature, e.g. see (Scott,

1979; Freedman and Diaconis, 1981; Silverman, 1986), but it is still to be investigated whether these methods and the suggested criteria for the selection of method specific parameters are also suitable for MI estimation. The estimators of MI, denoted $I(\tau)$, differ in the estimation of these marginal and joint probabilities or densities, using binning (Fraser and Swinney, 1986; Darbellay and Vajda, 1999), kernels (Silverman, 1986; Moon et al., 1995) or correlation integrals (Diks and Manzan, 2002), k-nearest neighbors (Paninski, 2003; Kraskov et al., 2004), B-splines (Daub et al., 2004) or the Gram-Charlier polynomial expansion (Blinnikov and Moessner, 1998). All these estimators depend on at least one parameter. Only the most commonly used MI estimators are described here.

4.2.1 Binning estimators

Equidistant estimator

The most common MI estimator is the naive equidistant binning estimator (ED). The main idea of this estimator is to partition the domain of each variable into a finite number b of discrete bins (equidistant partitioning). In general, let $\{x_t\}$ and $\{y_t\}$, $t = 1, \ldots, n$ be samples in pairs. The distributions of X and Y, denoted by $P_X(i)$ and $P_Y(j)$ respectively are approximated by the histograms of b intervals for each variable, that uniformly divide the range of their values, i.e. x_{min}-x_{max} and y_{min}-y_{max}. The number b of the intervals is not required to be the same for each variable. Thus, the two-dimensional XY plane is partitioned so that the values of each variable are divided in b equidistant intervals, generating $b \times b$ cells. Let denote $O_{XY}(i,j)$ the occupancy of the (i,j)-th cell of the partition of the XY plane. Then, the joint probability $P_{XY}(i,j)$ of (X,Y) is determined as $P_{XY}(i,j) = O_{XY}(i,j)/n$, and therefore is comprised of $b \times b$ values. A discrete approximation of MI is given by

$$I(X,Y) = \sum_{i=1}^{b} \sum_{j=1}^{b} P_{XY}(i,j) \log \frac{P_{XY}(i,j)}{P_X(i) P_Y(j)} \tag{4.9}$$

where there is no contribution to the sum if $P_{XY}(i,j)$ is equal to zero.

Equiprobable estimator

A second binning estimator is the equiprobable binning estimator (EP), which is derived by partitioning the domain of each variable in b intervals of the same occupancy (equiprobable partitioning), and therefore of different width (Palus, 1995; Cellucci et al., 2005). This partitioning actually transforms the sample univariate distribution to a discrete uniform with b components minimizing the effect of the univariate distribution on the estimation. This method reduces the sensitivity to any outlying values of X and Y. Let consider again that the domains of X and Y are partitioned into the same number of intervals b, so that there is an equal occupancy in each interval. Thus, the probability $P_X(i)$ of the i-th element of the partition of

X will be $P_X(i) = 1/b$ and respectively for Y, it is $P_Y(i) = 1/b$. Under the null hypothesis of statistical independence, the expected occupancy of the (i, j)-th element of the partition of the XY plane will be $E_{XY}(i, j) = nP_X(i)P_Y(j) = n/b^2$.

Selection of b for the binning estimators For the estimation of MI using the two binning estimators, it is essential to use a suitable number of intervals b for the determination of the histograms. If b is selected too large, then each interval will have small occupancy. If b is too small, with the limiting case $b = 1$, then the structure of the distribution will not be discerned. The problem of the selection of b has been examined by several investigators, e.g. see Bendat and Piersol (1966); Cocatre-Zilgien and Delcomyn (1992); Tukey and Mosteller (1977). For the selection of b for the EP estimator, Cochran (1954) suggested to be the largest possible value that gives $E_{XY}(i, j) \geq 5$ in at least 80% of the intervals. Cellucci et al. (2005) suggested to set as b the largest possible value that gives $E_{XY}(i, j) \geq 5$ for all elements of the XY partition, and therefore $b \leq \sqrt{n/5}$. A thorough investigation on the selection of b has been also conducted within this study and the results are discussed further below.

Adaptive estimators

Fraser-Swinney estimator Fraser and Swinney (1986) suggested a binning estimator using an adaptive partitioning. This method constructs a locally adaptive partition of the two-dimensional plane. It starts with a partition of equiprobable bins for each variable and makes a finer partition in areas where the joint probability density is non-uniform until the joint distribution on the cells is approximately uniform. The final partition is finer in dense regions whereas less occupied regions are covered with larger cells. In particular, the XY plane is partitioned into a number of cells, and the joint probability distribution on any cell is estimated by dividing the number of points of the cell by the total number of points. The partition of this method is nonuniform and adaptive. A finer partition is constructed in regions of the plane where P_{XY} has a detailed structure. A sequence of partitions $G_0, G_1, \ldots, G_m$ is constructed and each partition is a grid of 4^m cells generated by dividing each axis into 2^m equiprobable intervals, so that $P_X = P_Y = 1/2^m$ for each interval of the partition of X and Y respectively. G_0 is the entire XY plane. Let $R_m(K_m)$ denote an interval of the partition G_m and K_m is an index of the cell that takes one of 4^m possible values. The corresponding approximation of MI is given by

$$i_m = 2m + \sum_{K_m} P_{XY}(R_m(K_m)) \log_2 P_{XY}(R_m(K_m)) \tag{4.10}$$

The partitioning continues cell by cell until P_{XY} is uniform on each cell, i.e. P_{XY} values on each subdivision of it are approximately equal as specified by some statistical criterion. The statistical test constructed is a test for uniformity; it uses a x^2 test to examine the structure on both the $m + 1$ and $m + 2$ generation partition of

$R_m(K_m)$. This approach is not discussed any further here as it is in the same spirit but less effective than the next approach of Darbellay and Vajda (1999).

Darbellay-Vajda estimator Darbellay and Vajda (1999) investigated thoroughly the consistency of the estimator and tested the method on a number of probability distributions. In the mathematical and information theoretic literature the method is referred to as the DarbellayVajda algorithm. This adaptive estimator (AD) is based on the fact that MI can be estimated as a supremum over partitions (Dobrushin, 1959). By constructing a sequence of finer and finer partitions, the corresponding sequence of "MI" will monotonically increase and MI will stop increasing when conditional independence is achieved on all cells of the partition.

The histogram is data adaptive and is effective for multivariate cases; $\Re^d$ is partitioned into 2^d cells by dividing each of its d edges into two equiprobable intervals, where the cells of the partition are hyper-rectangles in $\Re^d$. For the two dimensional histogram obviously it is $d = 2$, however the general case for any d is described here. The probability distributions are estimated again by finding the relative frequencies, i.e. by dividing the points in a cell by the total number of points. Subsequently, each cell is partitioned into 2^d subcells and the procedure is repeated. A cell is not subdivided any further if local independence on that cell has been achieved. Local independence is achieved when the following condition is satisfied

$$\frac{P_{X,Y}(C_{kl})}{P_X(C_{kl})P_Y(C_{kl})} = \frac{P_{X,Y}(C_k)}{P_X(C_k)P_Y(C_k)}, \tag{4.11}$$

where C_k is the examined cell of the partition and C_{kl} is a subpartition of that cell. The mutual information is obtained by summing over all independent cells as

$$I(X,Y) = \sum_k P_{X,Y}(C_k) \ln \frac{P_{X,Y}(C_k)}{P_X(C_k)P_Y(C_k)} \tag{4.12}$$

The advantage of this estimator is that it is data-adaptive and does not a priori determine the number of bins in the partition. The AD estimator has a direct dependence on the number of the data or equivalently on the time series length, which determines the roughness of the partitioning in a somehow automatic way. In the abundance of data, the AD estimator reaches a very fine partition that satisfies the independence condition in each cell, so that the total number of cells is very large and analogous to a fixed-partition with a respectively large number of cells (large b). Note that the dependence of AD on the number of the data is not comparable to that of the fixed-bin estimators because it involves a change of partitioning with the number of data.

4.2.2 k-nearest neighbors estimator

Kraskov et al. (2004) proposed an MI estimator (KNN) that uses the distances of k-nearest neighbors of the two-dimensional points of the time series to estimate the

joint and marginal densities. It uses adaptive cubes whose size is locally adapted in the joint space (the XY plane) and then kept equal in the marginal subspaces. For each reference point from the bivariate sample, a distance length is determined so that the k nearest neighbors are within this distance length. Then, the number of points within this distance from the reference point gives the estimate of the joint density at this point and the respective neighbors in one-dimension give the estimate of the marginal density for each variable. The algorithm uses discs (or squares depending on the metric) of a size adapted locally and then uses the corresponding size in the marginal subspaces, so in some sense the estimator is data adaptive.

Let $Z = (X, Y)$ be the joint space of the X and Y and thus $\mathbf{z}_i = (x_i, y_i), i = 1, \ldots, n$ the points on the XY plane, and $\|\mathbf{z}_i - \mathbf{z}_j\| = \max(\|x_i - x_j\|, \|y_i - y_j\|)$ be the maximum norm of $\mathbf{z}_i, \mathbf{z}_j$, $i, j = 1, \ldots, n$. Let also denote $\varepsilon(i)/2$ the distance from z_i to its k-th neighbor, and $\varepsilon_x(i)/2$ and $\varepsilon_y(i)/2$ the distances between the same points projected into the X and Y subspaces. Obviously, $\varepsilon(i) = \max(\varepsilon_x(i), \varepsilon_y(i))$ for the maximum norm. $\varepsilon(i)$ is a random fluctuating variable, and therefore $\varepsilon_x(i)$ and $\varepsilon_y(i)$ also fluctuate. If $n_x(i)$ and $n_y(i)$ are the number of points with $\|x_i - x_j\| \leq \varepsilon_x(i)/2$ and $\|y_i - y_j\| \leq \varepsilon_y(i)/2$, respectively, then MI estimate is given by the formula

$$I(X, Y) = \psi(k) - \frac{1}{k} + \langle \psi(n_x) + \psi(n_y) \rangle + \psi(n) \tag{4.13}$$

where $\langle \cdot \rangle$ denotes the averages both over all $i \in \{1, \ldots, n\}$. $\psi(x)$ is the digamma function defined as $\psi(x) = \Gamma(x)^{-1} d\Gamma(x)/dx$. It is calculated using the recursion $\psi(x + 1) = \psi(x) + 1/x$ and $\psi(1) = -E$, with $E = 0.5772156...$ the Euler–Mascheroni constant. For large x, $\psi(x) \simeq \log x - 1/2x$ holds.

According to Kraskov et al. (2004), the KNN estimator is data efficient, e.g. for $k = 1$, structures down to the smallest possible scale are resolved. It is adaptive, i.e. the resolution is adjusted according to the local data density. Moreover, it is argued to have minimal bias and is recommended for high-dimensional data sets. The free parameter of the estimator is the number of neighbors k. A large k regards a small b of the fixed binning estimators. However, the estimator does not use a fixed neighborhood size and therefore there is not a clear association of k and b. The computational cost of the algorithm is on the search of the k neighbors. It is noted that any norm can be used for the estimation of MI.

4.2.3 Kernel estimator

The kernel density MI estimator (KE) uses a smooth estimate of the unknown probability density by centering kernel functions at the data samples; kernels are used to obtain the weighted distances (Silverman, 1986; Moon et al., 1995). The kernels essentially weigh the distance of each point in the sample to the reference point depending on the form of the kernel function and according to a given bandwidth h, so that a small h produces details in the density estimate but may loose in accuracy

depending on the data size. The estimator is effective also on multi-variable cases. Let $\mathbf{x}_i$, $i = 1, .., n$ be a set of d-dimensional vectors. A fixed-width kernel density estimator with kernel function K and fixed kernel width parameter h for a point $\mathbf{x} \in \Re^d$ is

$$f(\mathbf{x}) = \frac{1}{n'h^d} \sum_{i=1}^{n'} K(\frac{\mathbf{x} - \mathbf{x}_i}{h}), \tag{4.14}$$

where n' is the number of the d-dimensional vectors. The kernel function K should satisfy two condition: $K(\mathbf{x}) \geq 0$ and $\int_{\Re^d} K(\mathbf{x})d\mathbf{x} = 1$. For the estimation of MI with kernels, the most frequent choice for K is the standard normal density (Gaussian Kernel)

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}} \exp -\frac{1}{2}\mathbf{x}^T\mathbf{x}. \tag{4.15}$$

For the estimation of the kernel density, it is suggested to linearly transform the data in order to obtain a sample with zero mean and identity covariance matrix (K., 1972; Silverman, 1986). The kernel density estimator with Gaussian kernel function and a fixed bandwidth h at a point $\mathbf{x} \in R^d$ is

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}h^d \det(\mathbf{S})^{1/2}} \sum_{i=1}^{n'} \exp\left(-\frac{(\mathbf{x} - \mathbf{x}_i)^\tau \mathbf{S}^{-1}(\mathbf{x} - \mathbf{x}_i)}{2h^2}\right), \tag{4.16}$$

where $\mathbf{S}$ is the data covariance matrix (Moon et al., 1995).

KE estimator is estimated as

$$I(X, Y) = \frac{1}{n} \sum_{i=1}^{n} log \frac{f_{X,Y}(x_i, y_i)}{f_X(x_i) f_Y(y_i)} \tag{4.17}$$

where $f_{X,Y}(x_i, y_i)$,$f_X(x_i)$, $f_Y(y_i)$ are the joint and marginal probability densities estimated as in Eq.(4.16). KE estimator has two free parameters, the bandwidth h_1 for the marginal densities of X and Y, and the bandwidth h_2 for the joint density of (X, Y). The bandwidth h_1 is related to b of the fixed binning estimators by an inverse relation, e.g. a rectangular kernel assigns a bin centered at the reference point. The advantage of KE over binning estimators is that the location of the bins is not fixed.

The KE estimator is generally defined for two variables in an d-dimensional space. However, in case of univariate time series the variable $X = X_t$ and $Y = X_{t-\tau}$ and for the dimension of the points for the estimation of the joint probability density is 2 and for the marginal ones is 1.

MI estimator based on Correlation Sum

An estimator of mutual information, which is considered as a special case of the kernel estimator, is based on the correlation sums (Manzan and Diks, 2002). The

correlation sum of a variable X reflects the mean probability that the states at two different times are close and is defined as

$$C(\mathbf{x}, m, r) = \frac{2}{n_{pairs}} \sum_{j=m}^{n} \sum_{i<j-W} \Theta(r - \|\mathbf{x}_j - \mathbf{x}_i\|). \tag{4.18}$$

where n is the time series length, m the embedding dimension, W is the Theiler window, $n_{pairs} = (n - m + 1)(n - m - W + 1)$, r is a threshold value and $\|\cdot\|$ is a norm (e.g. Euclidean) and $\Theta(\cdot)$ is the Heaviside step function ($\Theta(a) = 0$ for $a \leq 0$ and $\Theta(a) = 1$ for $a > 0$). W is used to avoid spurious correlations; two points should not be too close in time (Theiler, 1986). The mutual information of two variables X and Y is estimated as

$$I(\tau) = \log \frac{C([\mathbf{x}\,\mathbf{y}], m, r) C(\mathbf{x}, m_x, r)}{C(\mathbf{y}, m_y, r)} \tag{4.19}$$

where $C([\mathbf{x}\,\mathbf{y}], m, r)$ is the correlation sum of the two dimensional variable (X, Y), with $m = 2$, and $C(\mathbf{x}, m_x, r)$ and $C(\mathbf{y}, m_y, r)$ are the correlation sums of the variables X and Y, respectively, with $m_x = m_y = 1$. In case of one time series, $I(\tau)$ is estimated for the variables $X = X_t$ and $Y = X_{t-\tau}$.

4.3 Applications of Mutual Information

Information theoretic concepts, such as MI, are being used to extend more conventional methods in various contexts. MI is widely utilized in diverse disciplines, e.g. physics (Fraser and Swinney, 1986), image recognition (Thenevaz and Unser, 2000), speech recognition (Ellis and Bilmes, 2000), bioinformatics (Hanus et al., 2007) and clustering (Kraskov et al., 2005). MI is also used in many aspects of time series analysis, best known as a criterion to select the appropriate delay for state space reconstruction (see 3.1.2). It is also used to discriminate different regimes of nonlinear systems (Hively et al., 2000; Naa et al., 2002; Wicks et al., 2007) and to detect phase synchronization (Schmid et al., 2004; Kreuz et al., 2007). Besides nonlinear dynamics, it is used in various statistical settings, mainly as a distance or correlation measure in data mining, e.g. in independent component analysis (Roberts and Everson, 2001) and feature-based clustering (Tourassi et al., 2001; Priness et al., 2007). Mutual information has also been to EEG recordings from different channels for the prediction of epileptic seizures (Palus et al., 2001a; Chillemi et al., 2003).

In extension to other similarity measures, mutual information provides a general measure of statistical dependence between variables. It is thereby able to detect any type of functional relationship, extending its potential over linear measures. Here, MI was tested in its usefulness as a test statistic for the test of independence and the test for nonlinearity and in detecting dynamical changes of nonlinear systems. For these studies, the ED estimator was used with fixed $b = 16$ which is also the default value in the mutual information program of the TISEAN package.

4.3.1 Independence test

In time series analysis, the existence of correlations is of primary interest and there have been developed statistical tests in order to examine whether a time series is independent and whether correlations are of linear or nonlinear type. The object of this study was to find a reliable test statistic for the test of independence (P1). For this purpose, a new measure of nonlinear correlations was defined, as the sum of mutual information values for a range of lags. This measure, called cumulative mutual information $M(\tau_{max})$, can be considered as the equivalent nonlinear statistic that can be used instead of the Portmanteau test (Box and Pierce, 1970). In statistics, a portmanteau test examines whether any of a group of autocorrelations of a time series are different from zero. The most commonly used Portmanteau test has as a test statistic the weighted sum of the squares of the autocorrelations for a range of lags. The proposed statistic, $M(\tau_{max})$, was proposed for the null hypothesis of independence and is compared to a Portmanteau test. The significance and power of the test was investigated using Monte Carlo simulations on different linear and nonlinear systems. The simulation study verified the effectiveness and usefulness of $M(\tau_{max})$ as a test statistic for the test of independence.

Background

A stationary time series may be a sequence of independent variables with the same distribution (iid), may only have linear correlations, i.e. comes from a linear stochastic process, or it may also have nonlinear correlations, i.e. the stochastic process may include higher order interactions of its components. The first step in the analysis of a stationary time series $\{x_t\}, t = 1, \ldots, n$ is to test whether its terms are independent. Thus, the examined H_0 is that the time series is iid.

The most commonly used independence test is the Portmanteau test which is based on the sample autocorrelation $r(\tau)$, defined in Eq.(3.7). If a time series is iid then $r(\tau)$ follows a typical normal distribution with zero mean and standard deviation $1/\sqrt{n}$, and 95% of $r(\tau)$ values should be in $\pm 1.96/\sqrt{n}$ (Box et al., 1994). Using as a test statistic $r(\tau)$ for a fixed lag τ, H_0 is rejected at a significance level $\alpha = 0.05$ if $|r(\tau)| < 1.96/\sqrt{n}$. The autocorrelation, however, may not lead to consistent results for the different τ values. Therefore, Portmanteau test uses as test statistic the sum of $r^2(\tau)$ for a number of lags, defined as

$$Q(\tau_{max}) = n \sum_{\tau=1}^{\tau_{max}} r^2(\tau), \tag{4.20}$$

in order to eliminate the dependence on τ. $Q(\tau_{max})$ under H_0 follows a X^2 distribution with τ_{max} degrees of freedom and the rejection area is $R = \{Q(\tau_{max} > X^2_{\tau_{max},1-\alpha}\}$ (Box and Pierce, 1970).

Here, it is proposed to use the cumulative autocorrelation function defined as

$$Q(\tau_{max}) = \sum_{\tau=1}^{\tau_{max}} |r(\tau)|, \tag{4.21}$$

instead of using the test statistic of Eq.(4.20). The absolute values of $r(\tau)$ are considered in the sum, as $r(\tau)$ might also be negative. This statistic was chosen in order to be directly compared with the test statistic of the cumulative mutual information function defined as

$$M(\tau_{max}) = \sum_{\tau=1}^{\tau_{max}} I(\tau), \tag{4.22}$$

where $I(\tau)$ is the mutual information for a lag τ. The distributions of $Q(\tau_{max})$ and $M(\tau_{max})$ are not known and therefore the evaluation of the two statistic was assessed using the surrogate data test.

Set Up

The evaluation of the tests with the statistics $Q(\tau_{max})$ and $M(\tau_{max})$ was assessed by Monte Carlo simulation on four systems:

- Normal white noise (with zero mean and variance one)
- Autoregressive AR(1), with equation

$$X_t = \phi X_{t-1} + w_t \tag{4.23}$$

  where the correlation coefficient is $\phi = 0.25$ and w_t is normal white noise. In Fig.4.1, time series from one realization of white noise and AR(1) model for $n = 256$ are presented.

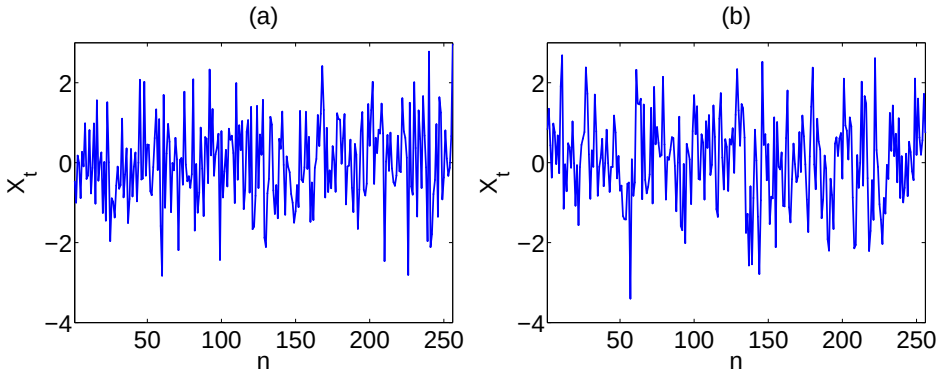

Figure 4.1: (a) One realization of normal white noise with $n = 256$. (b) One realization of AR(1) with $\phi = 0.25$ and $n = 256$.

- Henon map (Henon, 1976), which is a discrete-time dynamical system with equation

$$x_{t+1} = 1 - \alpha x_t^2 + \beta x_{t-1}. \tag{4.24}$$

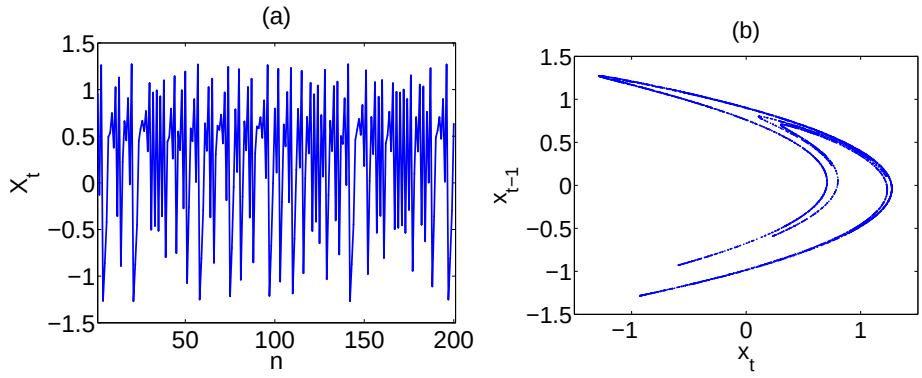

Figure 4.2: One realization of the Henon map with $n = 200$. (b) The strange attractor of the Henon map.

The Henon map depends on the coefficients α and β, which were selected here so that the map is chaotic ($\alpha = 1.4$ and $\beta = 0.3$). Fig.4.2 shows a time series with length $n = 200$ from the Henon map and its attractor.

- Ikeda map, which is a discrete-time dynamical system with equations

$$\begin{aligned} x_{n+1} &= 1 + u(x_n \cos t_n - y_n \sin t_n) \\ y_{n+1} &= u(x_n \cos t_n + y_n \sin t_n) \end{aligned} \tag{4.25}$$

where $t_n = 0.4 - 6/(1 + x_n^2 + y_n^2)$. The parameter u is set to be 0.9, so that the system is chaotic. For the simulation, the X variable is used. The Ikeda map is more complex system than the Henon map. In Fig.4.3, time series from one realization of the Ikeda map with $n = 256$ and the strange attractor of the Ikeda map are shown.

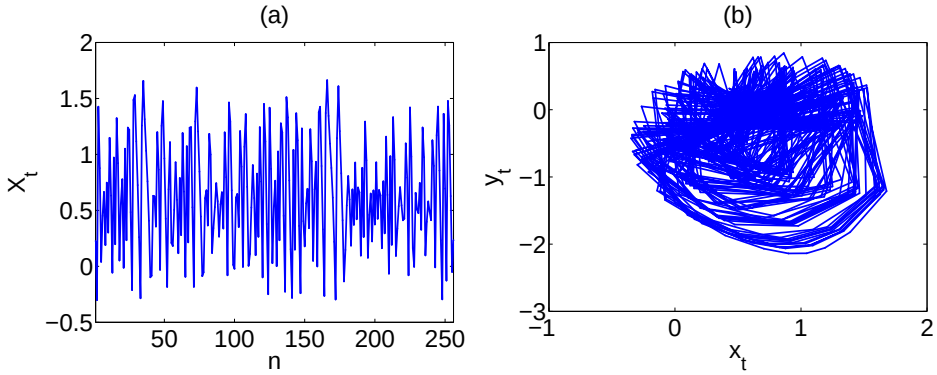

Figure 4.3: (a) One realization of the Ikeda map with $n = 256$, for the X variable. (b) The strange attractor of the Ikeda map.

Table 4.1: p-values from surrogate data test from 1000 realizations of normal white noise with test statistics $M(\tau)$ and $Q(\tau)$, for $\tau_{max} = 1, \ldots, 10$, and time series lengths $n = 128$ and $n = 2048$.

	$n = 128$		$n = 2048$	
τ_{max}	$M(\tau_{max})$	$Q(\tau_{max})$	$M(\tau_{max})$	$Q(\tau_{max})$
1	5.4	4.4	5.9	6
2	5.6	6.5	5.1	6.5
3	5.3	6.5	5.5	6.2
4	5.9	6.9	6.1	4.8
5	5.7	5.2	6	5.5
6	6.3	5	5.4	6.7
7	6.6	6.2	6	6
8	7.2	5.7	6.8	5.3
9	6.2	5.7	6.5	5.9
10	6.2	6.3	5.8	6.1

Time series of white noise are consistent to H_0, while the linear and the chaotic systems are not. 1000 realizations of each system were generated, with lengths $n = 128, 256, 512, 1024, 2048$. $Q(\tau_{max})$ and $M(\tau_{max})$ were estimated for $\tau_{max} = 1, \ldots, 10$. For the nonlinear systems, the influence of noise was also examined and time series were contaminated with additive Gaussian noise of levels $4, 8, 16, 32, 64\%$. 40 surrogates also were generated for each realization of each system in order to estimate the significance s and the p-values of the test. Surrogates were generated by randomly shuffling the time series. The test has the correct significance level if the proportion of times H_0 is rejected does not exceeds the significance level, which was set to $\alpha = 0.05$. The other three systems (AR(1), Henon map and Ikeda) that were not consistent to H_0 were used for the evaluation of the power of the test.

Results

Correct significance of the test was observed using both test statistics, from the simulation study on the time series of white noise. The p-values of the tests were slightly higher than 5%. Time series length and selection of τ_{max} does not affect the performance of the statistics. Analytical results are given in Table 4.1 for $\tau_{max} = 1, \ldots, 10$, and time series lengths $n = 128$ and $n = 2048$.

In case of the AR(1) system, $Q(\tau_{max})$ performed better than $M(\tau_{max})$. The p-values from $Q(\tau_{max})$ indicated the rejection of H_0 with high confidence, even for small time series lengths ($n \leq 256$), whereas $M(\tau_{max})$ rejected H_0 only for large n and small τ_{max}. $M(\tau_{max})$ strongly depended on the selection of τ_{max}, as p-values decreased with τ_{max}. The dependence of $Q(\tau_{max})$ on τ_{max} was also strong, but only for small n. For $n \geq 1024$, p-values gave a definite rejection of

Table 4.2: p-values from surrogate data test from 1000 realizations of AR(1) with $\phi = 0.25$ with test statistics $M(\tau)$ and $Q(\tau)$, for $\tau_{max} = 1, \ldots, 10$, and time series lengths $n = 128, 512$ and 2048.

	$n = 128$		$n = 512$		$n = 2048$	
τ_{max}	$M(\tau_{max})$	$Q(\tau_{max})$	$M(\tau_{max})$	$Q(\tau_{max})$	$M(\tau_{max})$	$Q(\tau_{max})$
1	7	77.6	37.8	100	99.8	100
2	7.6	68.7	26.8	100	99	100
3	8.5	61.8	20.7	100	94.8	100
4	7.6	55.8	17.2	99.8	91	100
5	6.9	52.5	15.5	99.8	85.5	100
6	7.5	46.9	14.4	99.7	79.5	100
7	7.2	44.5	12.1	99.2	72.9	100
8	6.6	43	11.7	98.7	69.7	100
9	6.7	40.9	11.2	98.5	66.1	100
10	7.9	39.3	11.4	98.1	60.5	100

H_0 (p-values were equal to 100%) for all τ_{max} and both statistics. Some indicative results, for $n = 128, n = 512, 2048$ are given in Table 4.2.

Finally, in case of the the chaotic systems, both statistics indicated a high power. For the Henon map, both statistics gave high percentages of rejection of H_0 (almost always 100%). Only for $n = 128$ and 64% noise level, $Q(\tau_{max})$ gave varying p-values ($75 - -97\%$) depending on the lag. For the Ikeda map, $M(\tau_{max})$ rejected H_0 with very high percentages (almost 100%) in all cases. $Q(\tau_{max})$ failed to reject H_0 for $\tau_{max} = 1$, for all n. For $n = 128$, H_0 was also not rejected with $Q(\tau_{max})$ for almost all τ_{max} and all noise levels. The same holds for $n = 256$ and 64% noise level. Some indicative results from the simulations on the Ikeda system, for $n = 128$ and noise levels $0, 16, 64\%$ are given in Table 4.3.

Conclusions

From the simulation study, it is obvious that both test statistics, $Q(\tau_{max})$ and $M(\tau_{max})$, can be used in an independence test. Depending on the type of the examined system (linear or nonlinear), one statistic might outbalance the other, however both statistics have a high significance. $Q(\tau_{max})$ presented higher power than $M(\tau_{max})$ for linear systems, while $M(\tau_{max})$ had a slightly higher power in case of nonlinear systems and mainly for small time series lengths. Time series length seemed to effect the performance of the measures only in case of the AR(1) system; $Q(\tau_{max})$ estimation for the Ikeda map was also affected but only for $n = 128$. $Q(\tau_{max})$ measures only linear correlations and therefore it was expected that it would perform better in case of linear systems, while $M(\tau_{max})$ which is a nonlinear measure was expected to be more useful in the case of the chaotic systems. These first results, indicate the usefulness of $M(\tau_{max})$ in problems, where

Table 4.3: p-values from surrogate data test from 1000 realizations of the Ikeda system with test statistics $M(\tau)$ and $Q(\tau)$, for $\tau_{max} = 1, \ldots, 10$, time series length $n = 128$ and noise levels $0, 16, 64\%$.

	0% noise		16% noise		64% noise	
τ_{max}	$M(\tau_{max})$	$Q(\tau_{max})$	$M(\tau_{max})$	$Q(\tau_{max})$	$M(\tau_{max})$	$Q(\tau_{max})$
1	98.3	14.7	98.6	14.1	98.5	8.3
2	100	93.7	100	92.9	100	55.4
3	100	89.7	100	89.1	100	49
4	100	88.2	100	85.2	100	43
5	100	84.6	100	80.4	100	43.1
6	100	78.1	100	74.9	100	39
7	100	74.1	100	71.4	100	36.5
8	100	71.5	100	67	100	34.2
9	100	66.1	100	63.4	100	31.5
10	100	64.1	100	60.2	100	31.4

possible weak correlations in a time series might not be linear.

4.3.2 Nonlinearity test

In the previous study, $M(\tau_{max})$ was proved to be an effective test statistic for test of independence. In order to extend the previous work, $M(\tau_{max})$ was also used as a statistic for nonlinearity test (P1). The examined null hypotheses in this case is that a time series is derived from a linear stochastic process. Again, the significance and the power of the test was investigated by Monte Carlo simulations on well known systems and the usefulness of $M(\tau_{max})$ was again verified also as test statistic for a nonlinearity test. $I(\tau)$ was also used as a test statistic, for comparison purposes.

Set Up

The set up of the nonlinearity test was similar to that of the independence test. For the estimation of $I(\tau)$ and $M(\tau_{max})$, τ and τ_{max} were set to be integers in $[1, 10]$. As the distribution of $I(\tau)$ and $M(\tau_{max})$ are not known, again the surrogate data test method was used. For the generation of the surrogates, the statistically transformed autoregressive process (STAP) algorithm was used (Kugiumtzis, 2002). The significance and the power of the test was examined for 100 realizations of lengths $n = 128, 256, 512, 1024, 2048$ from the following systems

- AR(1), with correlation coefficient $\phi = 0.25$.
- AR(1), with correlation coefficient $\phi = 0.5$.
- AR(9), with equation

$$X_t = 6.9627 + 1.2064X_{t-1} - 0.4507X_{t-2} - 0.1747X_{t-3}$$

$$+0.1974X_{t-4} - 0.1366X_{t-5} + 0.0268X_{t-6}$$
$$+0.0128X_{t-7} - 0.0312X_{t-8} + 0.2123X_{t-9} + w_t. \quad (4.26)$$

- Nonlinear Autoregressive model NAR(3) with equation

$$X_t = 0.25X_{t-2} - 0.4e_{t-1}X_{t-1} + 0.2e_{t-2}X_{t-2} + 0.3e_{t-3}X_{t-3} + w_t, \quad (4.27)$$

  where w_t is normal white noise.

- Henon map.

- Ikeda map.

- Lorenz system (Lorenz, 1963), given by the three differential equations

$$\begin{aligned} dx/dt &= \sigma(y_t - x_t) \\ dy/dt &= x_t(\rho - z_t) - y_t \\ dz/dt &= x_t y_t - \beta z_t \end{aligned} \quad (4.28)$$

  The parameters of the system were selected so that the system would be chaotic ($\sigma = 10, \rho = 28, \beta = 8/3$) and the sampling time was 0.02 sec. Fig.4.4 shows a time series with length $n = 200$ from the Lorenz system for variable Z and its attractor.

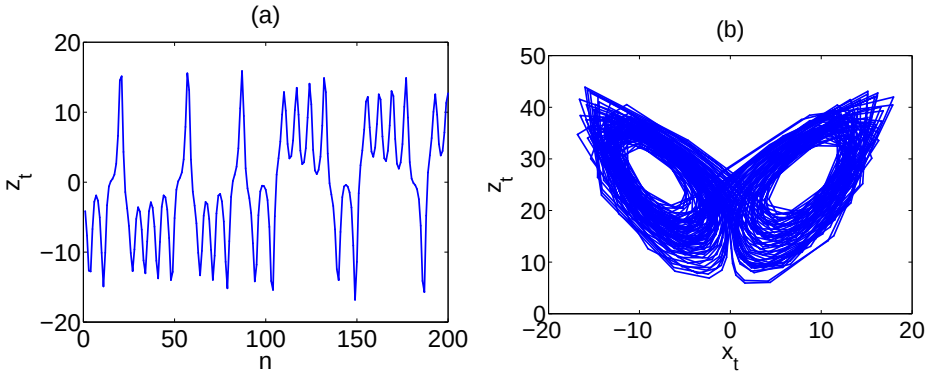

Figure 4.4: One realization of the Lorenz system with $n = 200$, for the Z variable. (b) Projection of the strange attractor of the Lorenz system on the XZ plane.

As the chaotic systems are deterministic, they were adapted to real conditions with addition of noise with levels $4, 8, 16, 32, 64\%$. The three first systems are linear and therefore consistent to H_0. The linear systems were used for the evaluation of the significance of the test. For the evaluation of the power of the test, the nonlinear systems were used.

Results

For the three linear systems, values of $I(\tau)$ and $M(\tau_{max})$ on the original time series were in the distribution of the respective estimated values from the surrogates for all lags, and thus H_0 was not rejected. An example is displayed in Fig.4.5, for the AR(1) system with $\phi = 0.5$ and $n = 128$. Even for small time series lengths,

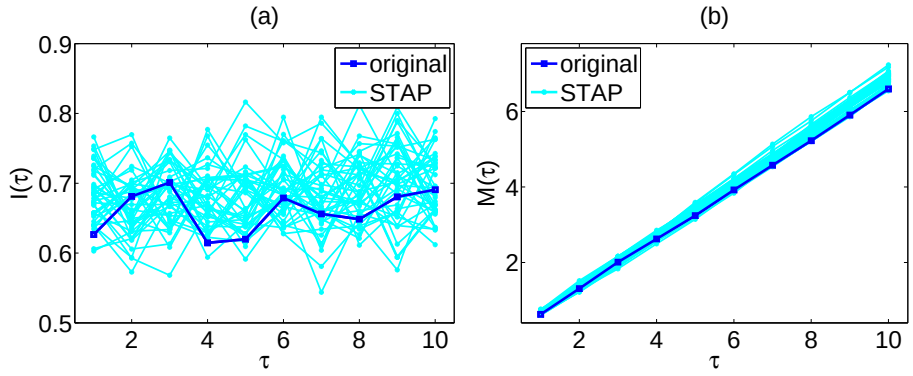

Figure 4.5: (a) Plot of $I(\tau)$ vs τ for one realization of AR(1) with $\phi = 0.5$ of length $n = 128$ and 40 STAP surrogates. (b) As in (a) but for $M(\tau_{max})$ vs τ_{max}.

the percentage of the rejection of H_0 was close to 5% for all lags.

The simulation study on the nonlinear systems showed that $M(\tau_{max})$ had higher power than $I(\tau)$, especially for small time series and high noise levels. The discrimination ability of $I(\tau)$ was affected by the selection of the lag, while $M(\tau_{max})$ discriminated the original time series from the surrogates independently of the selection of τ_{max}. An example for both statistics is displayed in Fig.4.6, from the simulations on the Henon map. The significance for $M(\tau_{max})$ was $s > 5$

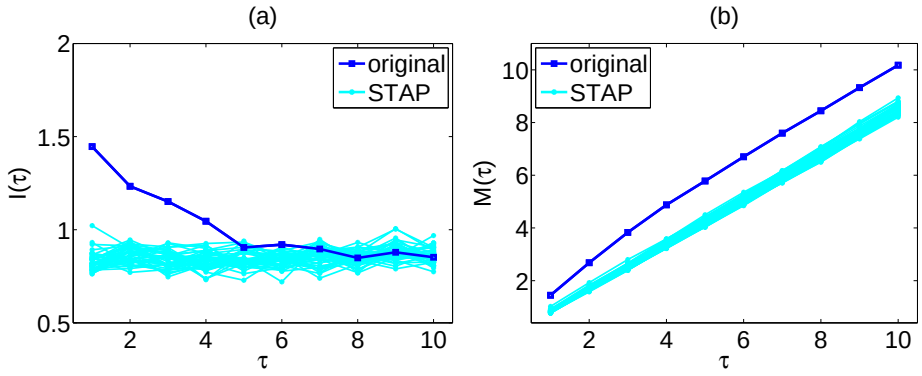

Figure 4.6: (a) Plot of $I(\tau)$ vs τ for one realization of the Henon map with $n = 128$ and 40 STAP surrogates. (b) As in (a) but for $M(\tau_{max})$ vs τ_{max}.

for all lags, while for $I(\tau)$, s decreased with τ and felt under the rejection limit of

$\alpha = 0.05$ for $\tau > 5$. $M(\tau_{max})$ outclassed $I(\tau)$, also with the addition of noise.

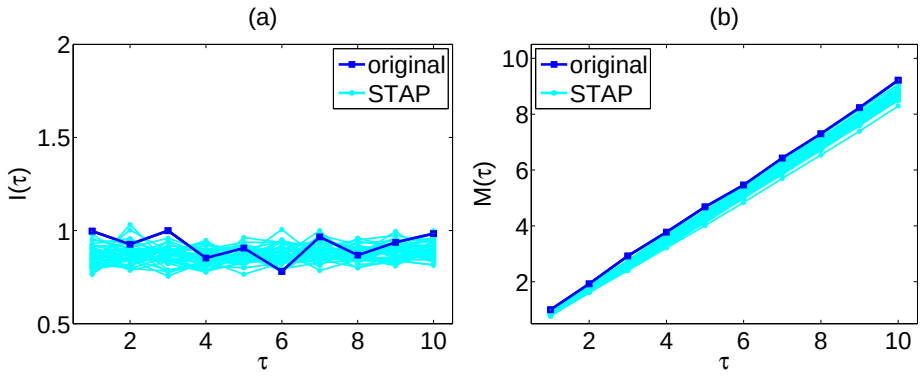

Figure 4.7: (a) Plot of $I(\tau)$ vs τ for one realization of the Henon map with $n = 128$ and noise level of 32% and 40 STAP surrogates. (b) As in (a) but for $M(\tau_{max})$ vs τ_{max}.

H_0 was rejected using $M(\tau_{max})$, even for high noise levels up to 32% for all τ_{max}, while H_0 was rejected using $I(\tau)$ only for certain lags (see Fig.4.7 for an example). However, for very high noise level, i.e. 64%, both statistics failed to discriminate the original time series from the surrogates.

The same conclusions were drawn from the simulation on the other two nonlinear systems, Lorenz and Ikeda. For the Lorenz system, $M(\tau_{max})$ had a higher power than $I(\tau)$ for all time series lengths and noise levels (see 4.8a). Addition of noise affected the power of both statistics; however $M(\tau)$ gave still a higher percentage of rejection of H_0 than $I(\tau)$. The difference in the power of the two test statistics was stronger in the case of the Ikeda map. While for large n, the percentage of rejection of H_0 is high for both statistics, for small n, the percentage of rejection of H_0 is much higher for $M(\tau_{max})$ (see Fig.4.8b). For high noise levels and small n ($n = 128, 256$), H_0 is rejected using $M(\tau_{max})$ with a percentage ranging from 20% to 40%, while for $I(\tau)$, the percentage decreased with τ and was about 5% for lags > 4. The percentage of rejection of H_0 decreased with τ for both statistics. As Ikeda map is more complex system than all the previously examined chaotic systems, it was expected that rejection of H_0 would be accomplished more difficult. Finally, results from NAR(3) also suggested using $M(\tau_{max})$ as it had higher power than $I(\tau)$. However, the rejection of H_0 required larger n; n should be ≥ 256 (see Fig.4.8c). The percentage of rejection of H_0 decreased with τ and only for $\tau \leq 3$ was H_0 rejected. As this model is of order three, it was expected that larger lags would not give any further information.

Conclusions

The simulation study showed a steady performance of $M(\tau_{max})$ as a test statistic for the nonlinearity test. The discriminating ability of this measure was generally

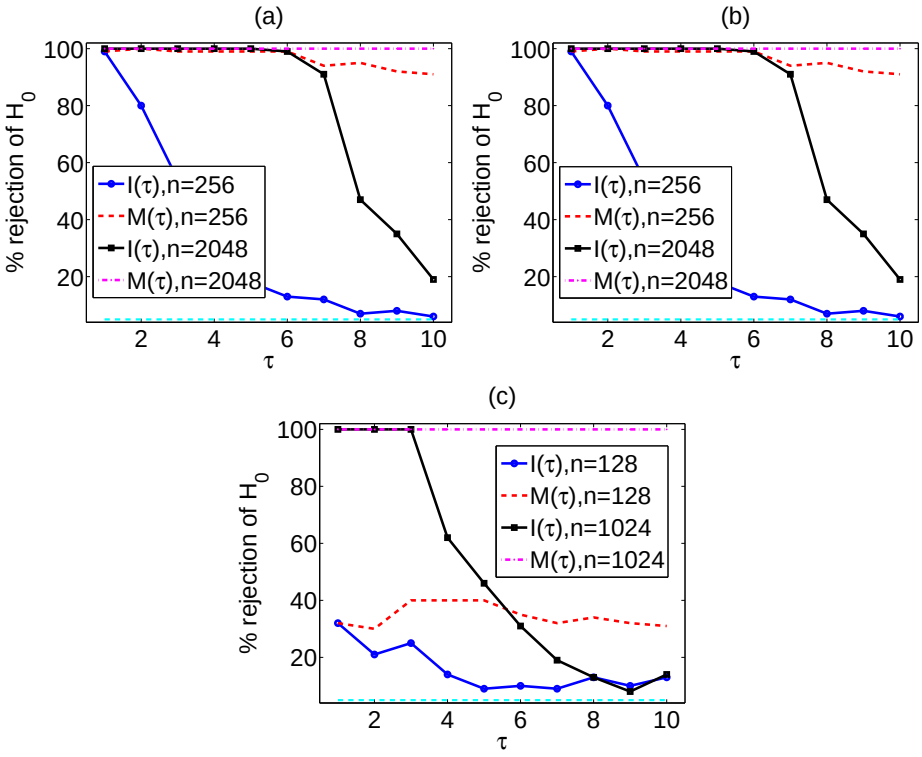

Figure 4.8: (a) Percentage of rejection of H_0 from the surrogate data test with both statistics $I(\tau)$ and $M(\tau_{max})$, for the Henon map and for time series lengths as given in the legends. (b) As in (a) but for the Ikeda map. (c) As in (a) but the NAR(3) model and for $n = 128$ and $n = 1024$.

not affected by the selection of the lag, as $I(\tau)$ did. The significance of the test using both statistics was generally correct, however for large n, $M(\tau_{max})$ gave lower percentage of rejection of H_0 than the significance level $\alpha = 0.05$. $M(\tau_{max})$ also displayed higher power than $I(\tau)$, especially for small n and high noise levels. These results accented the applicability of the statistic $M(\tau_{max})$ for nonlinearity tests. The need for an automatic and data adaptive method for the selection of τ_{max} arose from this study and was developed; however, it is going to be discussed further below.

4.3.3 Detection of dynamical changes

The last application of MI (and cumulative MI) concerned its ability to detect dynamical changes of systems (PP1). For this purpose, MI was used as a statistic for the detection of changes in the dynamical state of different nonlinear systems. Specifically, it was examined whether the changes in the characteristics of a dynamical system during its evolution could be detected using two statistical mea-

sures; MI and local linear fit. The discrimination strength of the two statistics was assessed by Monte Carlo simulations on nonlinear systems. MI was also used in a real application. The goal was to detect changes in the dynamical regime of the brain potential activity of an epileptic patient before a seizure using EEG measurements.

Set Up

$I(\tau)$ was estimated from Eq.(4.9) for time series from the same systems that have been used for the nonlinearity test. For comparative reasons, apart from MI, the normalized root mean square error (NRMSE) based on the Local Linear Fit (LLF) was also estimated. Specifically, a local linear prediction model for a time series $\{x_t\}$, $t = 1, \ldots, n$, for each time t is given as

$$\hat{x}_{t+1} = F(\mathbf{x}_t) = a_0 + a_1 x_t + a_2 x_{t_\tau} + \ldots + a_m x_{t-(m-1)\tau}, \tag{4.29}$$

where $\mathbf{x}_t$ is a reconstructed point from the time series and m is the embedding dimension. The model is supposed to be valid for the neighboring points of each point $\mathbf{x}_t$. The parameters $a_0, a_1, \ldots, a_m$ are estimated on the neighboring points. The statistic used for fitting error is

$$\text{NRMSE}(m) = \sqrt{\frac{\frac{1}{n-m-\tau}\sum_{i=m}^{n-\tau}(x_{i+\tau} - \hat{x}_{i+\tau})^2}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}}, \tag{4.30}$$

where $\bar{x}$ is the mean of the time series. The NRMSE can be estimated for any prediction model. The notation of NRMSE(m) here, estimated based on the prediction model in Eq.(4.29), will be LLF(m).

The evaluation of the two measures was assessed by Monte Carlo simulation on two nonlinear systems; the Henon map and the Lorenz system. In order to detect the changes of system characteristics, the observed data records were divided in overlapping segments using a sliding window of length $1/4$ of the segment length. Both measures were estimated from the overlapping segments. In each overlapping segment the surrogate data test was applied using as test statistics the two measures; $I(\tau)$ and $LLF(m)$. The examined null hypothesis H_0 is that the data at each segment are generated by a stationary Gaussian linear process that undergoes a nonlinear static transform. Rejection of H_0 would suggest the existence of nonlinearity and therefore the surrogate data test is applied on the data using a moving window technique in order to examine whether changes in the linearity of the system can be observed as noise levels increase. For the generation of the surrogate data consistent to this H_0 the algorithm STAP was used. To decide for the rejection of H_0, the significance s (see Eq.(3.10)) was estimated for the two discriminating statistics. Significance $s > 1.96$ suggests the rejection of H_0 at $\alpha = 0.05$.

The performance of the two statistics is investigated under different conditions of noise level, different lengths of the overlapping segment and reconstruction parameters. The variation of $I(\tau)$ with tau for a fixed embedding dimension and

the variation of LLF(m) with m for a fixed lag was also investigated. Time series lengths were chosen to be $N = 5000$ and 10000, and the lengths of the overlapping segments were $n = 500, 2500$ and 2000 for the two systems, respectively. The performance of the surrogate data test was examined for increasing noise levels in the data. Therefore, the change of the system characteristics during their evolution was simulated by monitoring the level of noise added to the observed time series. The noisy time series of length N were generated by gradually adding noise of increasing amplitudes. The levels of the Gaussian noise were $10, 20, 40, 60\%$. Thus, the original time series was comprised of 5 segments of equal length with noise levels $0, 10, 20, 40, 60\%$, respectively. For the Henon map, for estimation of $I(\tau)$, τ was set to be an integer in $[1, 10]$. For the estimation of LLF(m), m was set to be an integer in $[1, 10]$ and $\tau = 1$. For the Lorenz system, for the estimation of $I(\tau)$, tau was set to be an integer in $[1, 20]$. Finally, for the estimation of LLF(m), m is an integer in $[1, 10]$ and $\tau = 1, 5$ and 10.

Results from the simulation study

The simulation study on the Henon map showed that LLF(m) was a more effective statistical measure than MI, as the estimated significance of the test for LLF(m) was higher. The discrimination of the original time series from its surrogates was accomplished with LLF(m) as a test statistic. As expected, the significance of both tests decreased with noise level. Specifically, for overlapping segments of $n = 500$, lag $\tau = 1$ and embedding dimension $m = 1$, the two statistics gave similar results, even for high noise levels (see Fig.4.9a). For larger lags, $I(\tau)$ gave smaller values

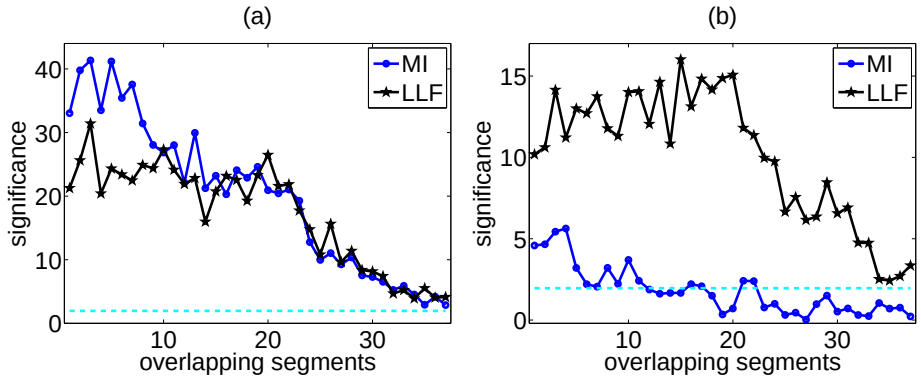

Figure 4.9: (a) Significance for the Henon map, for $N = 5000$ and $n = 500$, with test statistics MI and LLF for $m = 1$ and $\tau = 1$, respectively. (b) As in (a) but for $\tau = 10$ for $I(\tau)$ and $m = 10$ for $LLF(m)$.

of s (see Fig.4.9b). For large lags, $I(\tau)$ did not had any discriminating power ($s < 2$) for noisy time series. Besides the fact that for high noise amplitude it was hard to obtain significant discrimination, LLF discriminated also for unsuitably large embedding dimensions and even for sort time series lengths.

The effect of the noise on the significance power of the test was the first factor that was examined. The second factor was the length of the overlapping segments. Therefore, the previous results were compared with the results for the same systems but for different lengths n, e.g. for the Henon map with $N = 5000$ and $n = 2500$. In parallel, the dependence of each test statistic on its parameter is examined, i.e. the dependence of $I(\tau)$ on τ and of LLF(m) on m. Estimating $I(\tau)$ for lags 1 to 10, it is observed that the significance s is decreasing with τ. However, s is much higher than in the first case ($n = 500$), which means that the distinction of the original data and the surrogates is clearer. For LLF(m) the significance is much higher by setting $n = 2500$ than for $n = 500$ and distinction is clear for all embedding dimensions. However, no trend was observed for s, as m increased. In Fig.4.10a, the significance of both test statistics is displayed for varying τ for $I(\tau)$, and in Fig.4.10b s is given for varying m, for LLF(m).

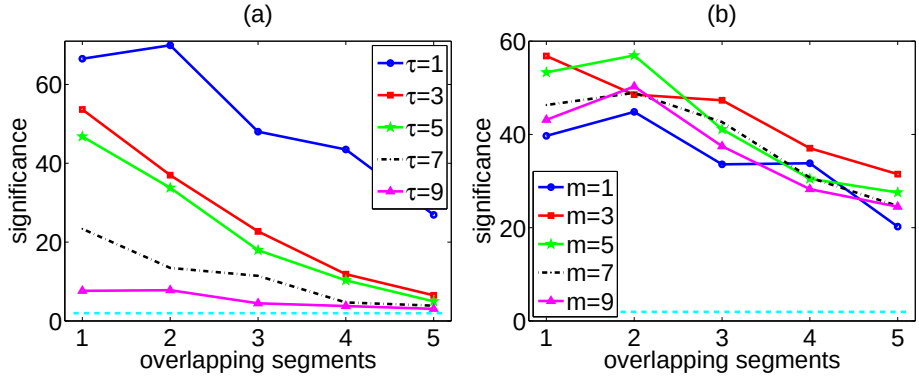

Figure 4.10: (a) Significance for varying τ (as in the legend) for $I(\tau)$ for the Henon map, for $N = 5000$ and $n = 2500$. (b) Significance for varying m (as in the legend) for $LLF(m)$ for the Henon map, for $N = 5000$ and $n = 2500$.

For the Lorenz system, with $N = 10000$ and $n = 2000$, both measures were again estimated from overlapping segments with time step $n/4$. For the computation of $I(\tau)$, the lag was set to from $\tau = 1$ up to 20. Simulations showed that from lags $\tau = 1$ to 7, s for $I(\tau)$ was gradually increasing, from $\tau = 8$ to 14 was gradually decreasing and from $\tau = 15$ to 20 was again increasing. H_0 was rejected only for $\tau = 1$ and 2. Thus, $I(\tau)$ had a higher discriminating power for the Lorentz system. The significance from LLF(m) for $\tau = 1$ increased with m and only for $m > 4$ was > 1.96, for all noise levels. For $m = 4$ to 10, s was almost the same. Therefore, large values of m seem to be more suitable for the distinction of the original data set from the surrogates. For $\tau = 5$ and $\tau = 10$, the results did not significantly differed from those for $\tau = 1$. The null hypothesis was rejected only for $m = 1$, and for $m = 2$ was rejected only for the last overlapping segments which had high noise levels. In this case, LLF(m) seems to be more effective than $I(\tau)$ for the discrimination of the original data from the surrogates, and thus for the detection of nonlinearity in the data.

Application on epileptic EEG

Finally, the two measures were tested in a real application; MI and LLF were estimated from overlapping segments of preictal EEG measurements. The two statistics were evaluated for their ability to detect changes in the dynamical evolution of EEG that are precursors of a forthcoming epileptic seizure. The surrogate data test was applied in order to quantify the power of the test statistics.

Extracranial EEG measurements from one electrode from the left frontal lobe was used. The length of the record was $n = 12000$ with sampling time 0.01s and the seizure onset was at about 100s. Overlapping segments of length $n = 3000$ (corresponding to 30 sec) were generated and the sliding step was $n/4$; measures were estimated from these segments. The last 3 segments contained data from the ictal period (during the epileptic seizure).

The significance s for $I(\tau)$ was small for the first 10 overlapping segments for all lags and there was a burst of s value at segment 11 (75s - 105s) and then decreased again for the last overlapping segments, i.e. the dramatic change of s value appeared at the time of the seizure onset (see Fig.4.11a). The estimated

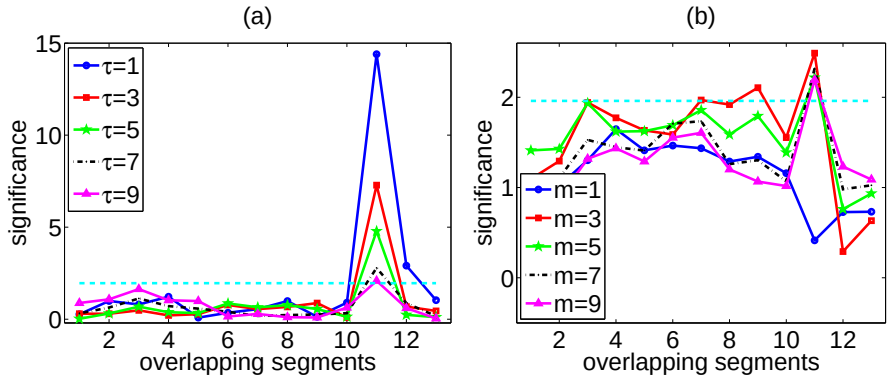

Figure 4.11: (a) Significance for different τ (as in the legend) for $I(\tau)$ for an extracranial EEG record with $N = 12000$ and $n = 3000$. (b) As in (a) but for LLF(m) (m as in the legend) and $\tau = 1$.

significance was very low ($s < 1.96$), which means that H_0 could not be rejected for all lags for the first overlapping segments.

For increasing τ (checked up to 15), s varied from non significant values ($s < 1.96$) towards significant ones (up to $s = 5$), indicating a change in its values along time up to the seizure onset. This change on s values was more obvious for MI than LLF (see Fig.4.11b). For example, for $\tau = 1$ only for large m is $s > 1.96$ in the 11th overlapping segment, indicating a change in the dynamics on the seizure onset. The rise of s at the seizure onset was not that high as for $I(\tau)$. Results were equivalent for LLF(m) estimated with $\tau = 5$ and $\tau = 10$.

Conclusions

The goal of this work was to examine whether the changes in the characteristics of a dynamical system during its evolution can be detected using the statistical measures of MI and LLF, monitoring the noise level and the length of overlapping segments. The strength of the discrimination of these statistics was estimated and compared. From the simulations on the Henon map, $I(\tau)$ had no discriminating strength for high noise levels (noise levels higher than 40%), whereas LLF(m) was able to detect the dynamical changes also for high noise levels. However, $I(\tau)$ seemed to be more effective than LLF(m) in case of the Lorenz system, for the detection of dynamical changes. As for the EEG record, values of the significance of $I(\tau)$ were very small for the preictal period and increase in the period just before the seizure indicating an increase of the stochastic component in the system for a small time period prior to the seizure onset, in agreement with other findings (Kugiumtzis and Larsson, 2000). The test did not show a clear trend of significance when LLF(m) was used as test statistic. Overall, the simulations showed that the surrogate data test equipped with suitable test statistics can detect changes in system characteristics (due to the stochastic component) when it is applied sequentially on subsequent segments of the observed signal.

4.4 Evaluation of Mutual Information Estimators

MI was proved to be useful in the previous applications; therefore the evaluation of MI estimators and the dependence of each estimator on its corresponding parameters was the next issue of interest. Mutual information estimators, based on histograms of fixed or adaptive bin size, k-nearest neighbors and kernels were evaluated and optimization their free parameters were also optimized (P6, P7, P10). The consistency of the estimators was also examined (convergence to a stable value with the increase of time series length) and the degree of the deviation among the estimators were also examined. The optimization of the parameters was assessed by quantifying the deviation of the estimated MI from its true or asymptotic value as a function of the free parameter. Moreover, some commonly used criteria for parameter selection were evaluated for each estimator. The comparative study was based on Monte Carlo simulations on time series from several linear and nonlinear systems of different lengths and noise levels. The results showed that the k-nearest neighbor is the most stable and less affected by the method-specific parameter. A data adaptive criterion for optimal binning is also suggested for the linear systems but it was found to be rather conservative for nonlinear systems. It turned out that the binning and kernel estimators give the least deviation in identifying the lag of the first minimum of mutual information from nonlinear systems, and are stable in the presence of noise.

There are some comparative studies on the MI estimators and the selection of their parameters, as well as on their performance on both linear and nonlinear dynamical systems (Wand and Jones, 1993; Steuer et al., 2002; Nicolaou and Nasuto,

2005; Khan et al., 2007). As MI estimation depends on the underlying time series, Monte-Carlo simulations are used. Comparisons of MI estimators involve the correct identification the lag of the first minimum of MI (Moon et al., 1995; Cellucci et al., 2005). Another performance criterion is the bias of estimators in the case of Gaussian processes, where the true MI is known (Cellucci et al., 2005; Trappenberg et al., 2006). MI estimators have also been compared to other linear or nonlinear correlation measures (Palus, 1995; Steuer et al., 2002; Daub et al., 2004).

Here, the performance of three types of estimators are examined, i.e. estimators based on histograms (with fixed or adaptive bin size), k-nearest neighbors and kernels. For the evaluation of MI estimators, the systems that have been used are white noise of different distributions, stochastic linear systems and dynamical non-linear systems (maps and flows of varying complexity). The performance of each estimator is examined with respect to the time series length, the distribution of noise and the noise level in the systems. Comparisons with other correlation measures are not pursued here as direct comparison is not possible due to the different scaling of the measures, even after normalization. All estimators vary in the estimation of the densities at local regions; the optimal parameter for the determination of the two-dimensional partitioning is investigated. Based on the simulation results, optimal parameters for each MI estimator are proposed with regard to the complexity of the system, the observational noise level and the time series length. It was found (Palus, 1993; Cellucci et al., 2005) that the complex algorithm of the adaptive binning estimator of Fraser and Swinney (1986) does not substantially improve the binning estimator and requires large data sets to gain accuracy; therefore it is not included in the current evaluation.

Set Up

The evaluation of the estimators is assessed by Monte-Carlo simulations on white noise, linear systems and chaotic systems of different complexity, listed in Table 4.4. A Gaussian and a skewed Gamma distribution are used to generate white noise time series, whereas for linear systems, the autoregressive model AR(1) and autoregressive moving average ARMA(1,1) are used with coefficients as given in Table 4.4, assuming both Gaussian and Gamma input white noise. The non-linear chaotic systems are the Henon map (Eq.(4.24)), the Ikeda map (Eq.(4.25)), and the Mackey-Glass system (Mackey and Glass, 1977) with differential equations

$$\frac{\mathrm{d}x}{\mathrm{d}t} = \frac{0.2x_{t-\Delta}}{1 + x_{t-\Delta}^{10}} - 0.1x_t, \tag{4.31}$$

where the parameter Δ, called delay, accounts for the system complexity. Δ is set to be 17 and 30 for low-dimensional chaos of fractal dimension about 2 and 3, respectively, and $\Delta = 100$ for high-dimensional chaos of fractal dimension about 7. Observational white noise at different levels is also assumed for the chaotic systems, given as a percentage of the standard deviation of the noise-free data.

Table 4.4: The simulation systems and their parameters. The input white noise for the linear systems (rows 4 to 7) and the observational white noise for the nonlinear systems (rows 8 to 10) have zero mean and standard deviation one. Gamma noise is skewed with $\gamma = 0.5$.The parameter notations are μ for the mean, σ for the standard deviation and γ for the skewness coefficient, φ for the coefficient of the autoregressive part for AR(1) and ARMA(1,1) and ϑ for the coefficient of the moving average part of ARMA(1,1), τ_s for the sampling time and δ for the discretization time for the Mackey-Glass system. The noise levels considered for the nonlinear systems are 20% and 40%.

Systems	*Parameters*	*Noise*
Gaussian white noise	$\mu = 0, \sigma = 1$	
Gamma white noise	$\mu = 0, \sigma = 1, \gamma = 0.5$	
AR(1)	$\varphi = 0.5, 0.9, -0.5, -0.9$	Gaussian
AR(1)	$\varphi = 0.5, 0.9, -0.5, -0.9$	Gamma
ARMA(1,1)	$\varphi = 0.9, \vartheta = 0.6$ & $\varphi = 0.7, \vartheta = 0.3$	Gaussian
ARMA(1,1)	$\varphi = 0.7, \vartheta = 0.3$ & $\varphi = 0.3, \vartheta = 0.1$	Gamma
Henon	$a = 1.4, b = 0.3$	Gaussian
Ikeda	$a = 1.0, b = 0.9, \kappa = 0.4, \eta = 6.0$	Gaussian
Mackey-Glass	$\Delta = 17, 30, 100, \tau_s = 17, \delta = 0.1$	Gaussian

Different lengths n for the generated time series from each system are considered as follows. For white noise and linear systems, n is given in powers of 2 from 5 to 13 and for nonlinear systems from 8 to 13. $I(\tau)$ is computed using all methods on 1000 realizations for each system, noise type or level, and time series length. As all linear systems are of order 1, MI is computed only for lag 1. For the nonlinear systems, $I(\tau)$ is computed up to the lag τ for which $I(\tau)$ levels-off. For the Mackey-Glass systems with $\Delta = 17$ and 30, the objective is to estimate the lag of the first minimum of $I(\tau)$. For $\Delta = 100$ MI does not exhibit a distinct minimum and therefore the lag that MI levels-off is estimated. For each estimator, $I(\tau)$ is computed for a wide range of values of the free parameter and for specific values determined by standard criteria, which are specified below.

For the binning estimators ED and EP, the number of tested bins are $b = 2, 4, 8, 16, 32, 64$, whereas b is also estimated from 10 commonly used criteria given in Table 4.5. For the choice of k of the k-nearest neighbor estimator KNN, Kraskov et al. (2004) propose to use $k = 2$ to 4 (these are also used in (Kreuz et al., 2007; Khan et al., 2007)). However for real world data one should investigate also larger values of k. Therefore, a wide range of k values is used for the simulations; k is set to be $2, 4, 8, 16, 32, 64$.

Among different kernel functions used in the literature for density estimation, and for MI estimation in particular, the common practice is to use the Gaussian

Table 4.5: Criteria for the selection of the number of bins. The parameters in the expressions are the time series length n, the standard deviation s, the interquartile range IQR, the range of the data R and the standardized skewness γ_2 as defined in (Doane, 1976). The exact expressions for criteria H8 and H9 can be found in the corresponding references, given in the third column.

Criteria	Number of bins	Reference
H1	$1 + \log_2 n$	(Sturge, 1926)
H2	$1.87(n-1)^{0.4}$	(Bendat and Piersol, 1966)
H3	$1 + \log_2 n + \log_2 \gamma_2$	(Doane, 1976)
H4	$\sqrt{n}$	(Tukey and Mosteller, 1977)
H5	$\frac{Rn^{1/3}}{3.49s}$	(Scott, 1979)
H6	$\frac{Rn^{1/3}}{2(IQR)}$	(Freedman and Diaconis, 1981)
H7	$\sqrt[3]{2n}$	(Terrell and Scott, 1985)
H8	min. of stochastic complexity	(Rissanen, 1992)
H9	mode of log of marginal posterior pdf	(Knuth, 2006)
H10	$\sqrt{n/5}$	(Cochran, 1954)

kernel in conjunction with the "Gaussian" bandwidth of Silverman (1986)

$$h = \left(\frac{4}{(d+2)n'} \right)^{1/(d+4)} \tag{4.32}$$

(n' is the number of the d-dimensional vectors) or multiples of it (Harrold et al., 2001; Steuer et al., 2002; Khan et al., 2007). In the estimation of mutual information with kernels, the range of bandwidths is usually not searched and a bandwidth is selected according to a criterion such as the "Gaussian" bandwidth (Moon et al., 1995; Steuer et al., 2002). A multiple bandwidth selection scheme for the test for independence is proposed in Diks and Panchenko (2008). Analytic and simulation studies have shown that the choice of the bandwidth is crucial and depends on the data size (Bonnlander and Weigend, 1994; Jones et al., 1996). Therefore, a wide range for the bandwidth h_1 for one dimension and h_2 for two dimensions is considered, as for b and k. Specifically, for h_1, 15 values in $[0.01, 2]$ are taken at a fixed base-2 logarithmic step and it is also set $h_2 = h_1$ and $h_2 = \sqrt{2}h_1$. The second form for h_2 accounts for the scaling of the Euclidean metric in $\Re^2$, which is used in the simulations. As for b, some well-known criteria for the choice of bandwidths are considered, given in Table 4.6. The three first criteria define bandwidth for both one and two-dimensional space. For the other criteria, h_2 is set equal either to h_1 or $\sqrt{2}h_1$.

The true (theoretical) MI $\mathcal{I}(\tau)$ is not known in general. However, for Gaussian processes $\mathcal{I}(\tau)$ is given in terms of the autocorrelation function $\rho(\tau)$ as

$$\mathcal{I}_G(\tau) = -0.5 \log(1 - \rho^2(\tau)). \tag{4.33}$$

Table 4.6: Criteria for the selection of bandwidths for one (h_1) and two (h_2) dimensions. The parameters in the expressions are $a = 1.8 - r(1)$ if $n < 200$ and $a = 1.5$ if $n \geq 200$, where $r(1)$ is the autocorrelation at lag 1, $R = 1/2\sqrt{\pi}$, s is the standard deviation and IQR is the interquartile range of the data. The exact expressions for the last four criteria can be found in the corresponding references, given in the third column.

Criteria	h_1	h_2	Reference
C1	$(4/3n)^{1/5}$	$(1/n)^{1/6}$	(Silverman, 1986)
C2	$(4/3n)^{1/5}$	$(4/5n)^{1/6}$	(Silverman, 1986)
C3	$1.06an^{-1/5}$	$an^{-1/6}$	(Harrold et al., 2001)
C4	$(\frac{8\sqrt{\pi}R}{3n})^{1/5} \min(s, \frac{\text{IQR}}{1.349})$	h_1	(Silverman, 1986)
C5		$\sqrt{2}h_1$	(Wand and Jones, 1995)
C6	L-stage direct plug-in	h_1	(Wand and Jones, 1995)
C7		$\sqrt{2}h_1$	
C8	Solve-the-equation plug-in	h_1	(Sheather and Jones, 1991)
C9		$\sqrt{2}h_1$	

(Kullback, 1959). In the lack of the true MI for the other systems, consistency of the estimators is assumed and the asymptotic value of $I(\tau)$ computed on a realization of size $n = 10^7$ is used. In this computation, for the ED and EP estimators b is set to be 64 (MI was computed for b up to 256, however MI did not substantially differed), for KNN estimator $k = 2$, and for the KE estimator $h_1 = 0.01$ and $h_2 = h_1$. Similar approach for approximating the true MI is used in (Harrold et al., 2001; Cellucci et al., 2005). The true or asymptotic MI is denoted I_∞ and is used as reference to compute the accuracy of the different estimators.

The evaluation of the estimators and their parameters is assessed separately for white noise and linear systems, for the nonlinear maps Henon and Ikeda, and for the Mackey-Glass system. First, the dependence of the estimators on their free parameters is investigated, for white noise and linear systems and for different time series lengths giving a total of $L = 126$ cases (14 systems and 9 time series lengths). For each estimator, the mean estimated MI $\bar{I}_c(l)$ from the 1000 realizations is computed for each tested value of the free parameter denoted by c, where l denotes the system case and $l = 1, ..., L$. Further, for each l the deviation $dI_c(l) = |\bar{I}_c(l) - I_\infty(l)|$ is computed, where $I_\infty(l)$ is either the true MI (for Gaussian processes) or the asymptotic value computed from each estimator. All estimators converged to the same MI under proper parameters as obtained for n increasing up to 10^7. Given the asymptotic MI $I_\infty(l)$, the optimal parameter values for each case l (system and time series length) is obtained from the minimum deviation $dI_c(l)$ with respect to c. The estimators are then compared for their optimal parameters by computing the divergence of the mean $I(\tau)$ of each estimator from $I_\infty(l)$ for all cases.

For the discrete nonlinear systems, there are $L = 48$ cases (8 maps including different noise levels and 6 time series lengths). Here, a single I_∞ for each system

cannot be obtained as the MI estimate for n increasing up to 10^7 does not converge and the MI for $n = 10^7$ is still dependent on parameter selection and varies also across estimators. Therefore, for each system the asymptotic value I_∞ of the estimator is set to be the MI computed for $n = 10^7$ and for a very fine partition. So here, the interest is in the dependence of each estimator on the free parameter and the rate of convergence towards the asymptotic value.

For the nonlinear flows derived from the Mackey-Glass system for $\Delta = 17, 30$ (noise-free and with noise), the focus is on the estimation of the first minimum of MI in order to compute the lag τ_0 that corresponds to it for each of the 1000 realizations of each system. For $\Delta = 100$ there is no clear minimum of MI and it follows a rather exponential decay. Therefore, the lag τ_0 for which MI levels off is estimated according to a criterion for levelling. In order to compare the estimators, the consistency of each estimator with n is examined, as well as the dependence of the estimation of τ_0 on the parameter selection and the variance of the estimated lags τ_0 from all cases.

Results

Results on white noise and linear systems

The MI for lag one $I(1)$ from the binning estimators ED and EP increases always with the number of bins b. Thus for white noise where $I_\infty(1) = 0$ the best choice for b is 2 that gives the smallest positive $I(1)$. For the linear processes, $I(1)$ decreases with the time series length n for each b, as shown in Fig. 4.12 for the ED and EP estimates of $I(1)$ from an AR(1) process. It is noted that for sufficiently

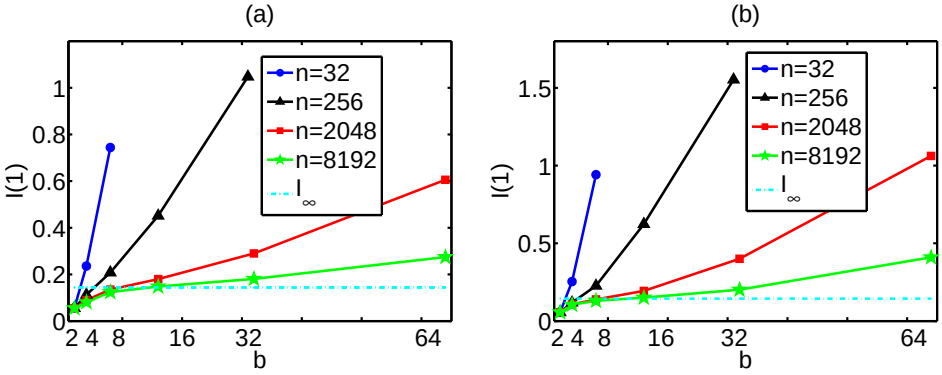

Figure 4.12: Mean $I(1)$ as a function of b from 1000 realizations of AR(1) with coefficient $\varphi = r(1) = 0.5$ and additive Gaussian noise for (a) the ED and (b) the EP estimator and for data sizes as given in the legend.

large b, $I(1)$ converges with n to the true value $\mathcal{I}_G(1) = I_\infty(1)$ given in (4.33). On the other hand, for small b, $I(1)$ underestimates $I_\infty(1)$ depending again on n. Thus the optimal b that gives the smallest $dI_c(l)$ and estimates best $I_\infty(1)$ depends

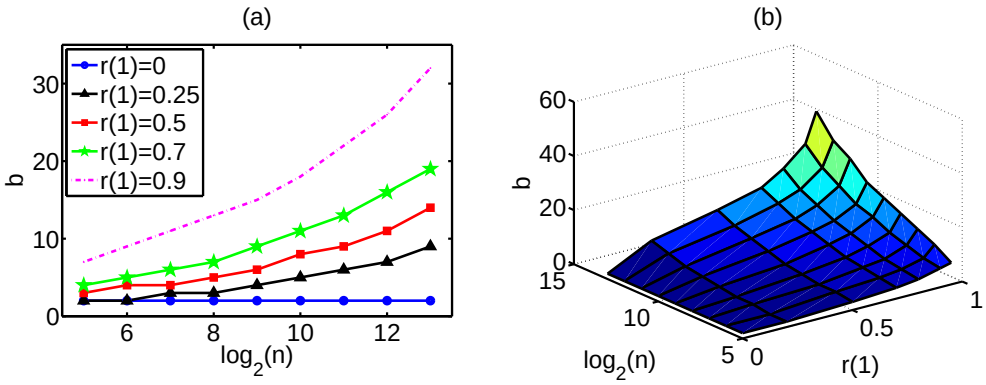

Figure 4.13: (a) Optimal number of bins for different n for AR(1) systems with lag one autocorrelation $r(1)$ as in the legend. (b) Graph of the optimal b for a range of $\log_2(n)$ and $r(1)$. The results in both panels regard the ED estimator.

on n.

It is observed that b depends also on the autocorrelation function $r(\tau)$ of the linear system. To investigate further this dependence, ED and EP estimates of $I(1)$ were computed for a wide range of $r(1)$ values of an AR(1) process. The optimal b was found (by the smallest $dI_c(l)$) to increase smoothly with n and $r(1)$, as shown in Fig. 4.13. A search for a parametric fit of optimal b regarding the graph of Fig. 4.13b resulted in the form

$$b = \alpha n^{\beta} e^{\gamma \rho^2}, \tag{4.34}$$

where the coefficients α, β, γ take similar values for the ED and EP estimators (0.65,0.25,2.11 and 0.76,0.19,1.91 respectively). Other forms have also been tested for b, polynomial models both linear and nonlinear, exponential and logarithmic ones, e.g. $b = \alpha n + \beta \rho^2$ and $b = \alpha e^{\beta \rho + \gamma n}$.

Most of the criteria in Table 4.5 tend to overestimate b. To evaluate the performance of the 10 criteria in Table 4.5 and the proposed criterion in Eq.(4.34), the total score S_c was computed for each criterion c, where $c = 1, \ldots, 11$, for all $L = 126$ tested systems and time series lengths, as

$$S_c = \frac{\sum_{l=1}^{L} (\bar{I}_c(l) - I_\infty(l))^2}{\sum_{l=1}^{L} (\bar{\bar{I}}_c(l) - I_\infty(l))^2} \tag{4.35}$$

where for each case l, $\bar{\bar{I}}_c(l) = \frac{1}{11} \sum_{c=1}^{11} \bar{I}_c(l)$ is the grand mean of the means from all criteria. A similar score was defined by Hamilton (1964), and has been used again for the evaluation of mutual information estimators (Cellucci et al., 2005). However the denominator of the formula suggested by Hamilton (1964) was the sum of the squares of I_∞ of the examined systems; as this can be zero in case of independent systems, therefore the formula was modified to be more general.

Table 4.7: Ranking and score S of the 5 criteria for b scoring lowest for the *ED* and *EP* estimators for varying lengths of time series from white noise and linear systems.

Criteria	S for *ED*	Criteria	S for *EP*
H11	0.25	H9	0.61
H7	0.94	H7	0.62
H8	1.16	H8	0.65
H5	1.20	H10	0.67
H9	1.41	H11	0.72

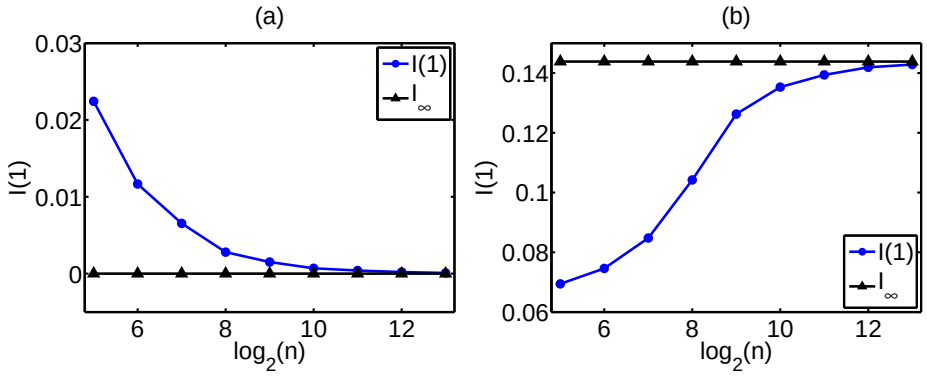

Figure 4.14: Mean estimated MI with AD estimator as a function of n from 1000 realizations of (a) normal white noise and (b) AR(1), with $r = 0.5$ and normal input white noise.

According to the score S_c, the proposed criterion in Eq.(4.34) for the optimal b, denoted H11, outperforms the other criteria when ED estimator is used, as shown in Table 4.7. For the EP estimator, criterion H9 scores lowest and H11 is ranked fifth but the differences in the scores of the best five criteria are comparatively small.

For certain bivariate distributions and Gaussian processes, it was found that the AD estimator was precise in estimating MI and converged fast to the true MI (Kraskov et al., 2004; Trappenberg et al., 2006). This result was confirmed by the simulations on the white noise and linear systems with the remark that the convergence to I_∞ is rather slow and is succeeded at large n, as shown in Fig. 4.14.

The number of nearest neighbors k in the KNN estimator determines the roughness of approximation of the density functions in Eq.(3.9), which corresponds to the roughness of the partitioning in Eq.(4.7). The simulations showed that for white noise the MI estimated by KNN is close to zero for a long range of k and the deviation from zero decreases as k approaches $n/2$ (as reported also in (Kraskov et al., 2004)). For the linear systems the optimal k is rather small. The dependence of the

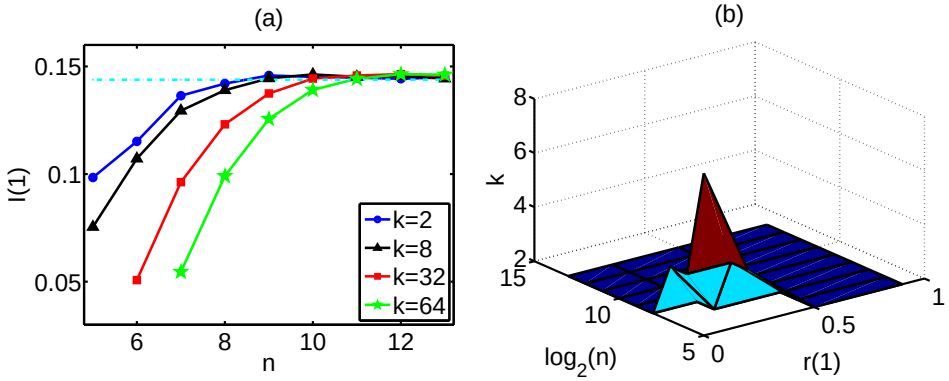

Figure 4.15: (a) Mean estimated MI with the KNN estimator as a function of n from 1000 realizations of AR(1) with $r(1) = 0.5$ and normal input white noise for k as in the legend. The dotted line stands for I_∞. (b) Graph of the optimal k for a range of $\log_2(n)$ and $r(1)$.

KNN estimator on k holds mainly for small time series as for large n the estimated MI converges to I_∞ for any k, as shown in Fig. 4.15a. Still, the convergence is slower for larger k. In any case, a highly accurate MI is attained with small k for all but very small time series. For example, for an accuracy threshold of 10^{-4} in estimating I_∞, i.e. $|I(1) - I_\infty| < 10^{-4}$, the optimal choice for k is 2 in almost all cases except for very small n and $r(1)$, as shown in Fig. 4.15b. Note that even for white noise time series of small length, $k \leq 8$ reaches this accuracy threshold (the peak in the graph of Fig. 4.15b is for $n = 2^5$ and $r(1) = 0$).

For the two dimensional bandwidth h_2, it is set $h_2 = h_1$ and $h_2 = \sqrt{2}h_1$ and it was examined the dependence of the estimated MI on h_1 and h_2 across a large range of bandwidths for white noise and linear systems. As for k of the KNN estimator, MI converges with n and faster for smaller h_1. For $h_2 = h_1$ the convergence with n is correctly towards I_∞ (see Fig. 4.16a), but for $h_2 = \sqrt{2}h_1$ MI decreases with h_1 and becomes negative (see Fig. 4.16b).

This result advocates the use of the same bandwidth for the kernel estimates of the marginal and joint distributions. It was also investigated whether there is dependence of h_1 on $r(1)$ and n. As for k of the KNN estimator, there does not seem to be any systematic dependence. Using the same threshold accuracy in estimating I_∞, the smallest optimal h_1 is always at a low level for all $r(1)$ and n and there is no apparent pattern that would suggest a particular form of dependence of h_1 on $r(1)$ or n, as shown in Fig. 4.16c. The sudden jumps in the graph of Fig. 4.16c is due to numerical discrepancies around the chosen threshold for different h_1 values.

In order to evaluate the criteria for selecting h_1 and h_2 in Table 4.6, the score defined in (Eq.(4.35)) was computed for each criterion and for all cases. The five optimal criteria and their scores for varying lengths of time series from white noise and linear systems are C1 (0.85), C3 (0.94), C2 (0.96), C4 (1.57) and C9 (1.67). The simplest criteria turned out to score lowest with best being the "Gaussian" rule

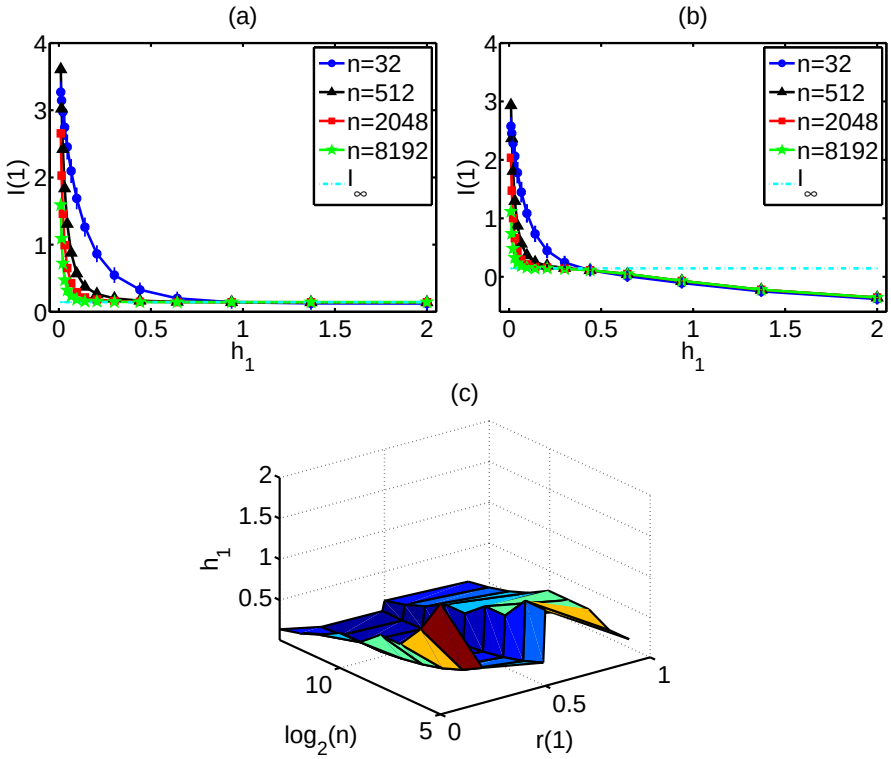

Figure 4.16: Mean estimated MI with the KE estimator as a function of h_1 from 1000 realizations of $AR(1)$ with $r(1) = 0.5$ and normal input white noise for (a) $h_1 = h_2$, and (b) $h_1 = \sqrt{2}h_2$, and n as in the legend. (c) Graph of the optimal h_1 for a range of $\log_2(n)$ and $r(1)$.

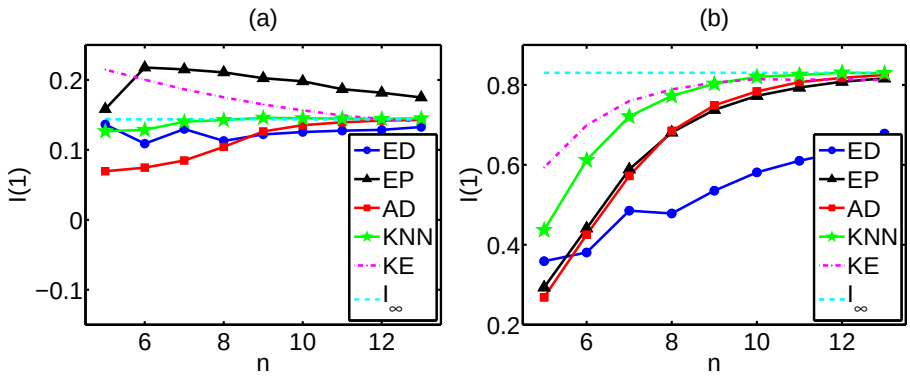

Figure 4.17: (a) Mean estimated MI vs n for the estimators given in the legend from simulations on AR(1) with $r(1) = 0.5$ and normal input white noise. For each estimator the optimal free parameter is considered, i.e. H11 for ED, H9 for EP, $k = 2$ for KNN and C1 for KE. (b) As (a) but for $r(1) = 0.9$.

of Silverman C1 (see also Eq.(4.32)).

Summarizing the results on white noise and linear systems, it turns out that fixed binning estimators are the most dependent on the free parameter, the number of bins b, whereas for KNN and KE estimators a small number of neighbors k and bandwidth h_1, respectively, turns out to be sufficient for all but very small time series length n and weak autocorrelation $r(\tau)$. In such cases, binning estimators can approximate I_∞ better with a relatively small b. An expression for this involving n and $r(1)$ is provided. All estimators are consistent but converge at different rates to the true or asymptotic MI I_∞, as shown in Fig. 4.17 for the AR(1) system with weak and strong autocorrelation.

In general, the KNN estimator converges fastest. To this respect, the parameter-free AD estimator would be the second best choice after the KNN estimator because it showed a slower convergence rate. KE estimator has about the same convergence rate as AD and is not significantly affected by parameter selection (for $h_2 = h_1$), however it would not be preferred due to its computational cost. The estimation accuracy of each estimator is quantified by the index $dI_{c_{opt}} = \sum_{i=1}^{L}(\bar{I}_{c_{opt}}(l) - I_\infty(l))^2$, where $L = 126$ and c_{opt} is the optimal free parameter found in the simulations above, i.e. H11 for b in ED and H9 for b in EP, $k = 2$ for KNN, and C1 for the bandwidths in KE. The smallest index $dI_{c_{opt}} = 0.254$ was obtained by ED, followed by KE (0.302) and KNN (0.496), whereas AD and EP scored worse (1.865 and 2.231, respectively). The numerical analysis on the linear systems and noise showed that for the three aspects of estimation considered, i.e. parameter dependence, rate of convergence, and accuracy of estimation, no estimator ranks first but KNN and KE turn out to perform overall best.

Results on nonlinear maps

In terms of chaotic systems, let us first note that MI can be viewed as a measure on the reconstructed attractor projected on $\Re^2$, i.e. on points $[x_t, x_{t-\tau}]'$. Due to the fractal structure in all scales of the chaotic attractor, $\mathcal{I}(\tau)$ defined in terms of a partition (see Eq.(4.7)) increases with finer partition towards the limit of $\mathcal{I}(\tau)$ given in Eq.(3.9) for the continuous space. On the other hand, for the estimation of entropy, and particularly the Kolmogorov-Sinai or metric entropy, it is postulated that there exists a so-called generation partition that gives the expected entropy value and further refinement to this partition does not increase further the computed entropy (Walter, 1975; Cohen and Procaccia, 1985). However, with regard to the Shannon entropy, it was observed that only an upper limit of MI from the KNN estimator with $k = 1$ could be considered. The estimation algorithm of KNN does not allow for a finer partition, whereas increasing b for the binning estimators or decreasing h_1 for the KE estimator within the tested range does not seem to lead to convergence of MI.

The true $\mathcal{I}(\tau)$ in Eq.(3.9) is not known since the joint distribution of $[x_t, x_{t-\tau}]'$ is also not known. This prevents the direct comparison of the estimators and the search for optimal free parameters. The presence of noise in chaotic time series sets a limit to the scale where fractal details can be observed and consequently to the finiteness of the partition when estimating $\mathcal{I}(\tau)$. In that case, an asymptotic I_∞ does exist and the performance of the estimators in terms of the free parameter and time series length can be compared, also for different noise levels. In the following, it is aimed to delineate the differences among estimators in estimating I_∞ and particularly in converging to I_∞ with respect to the time series length and their free parameter.

The discussion above would suggest that the estimated MI should always increase as the partition gets finer, but in practice this requires a sufficient time series length n. For the binning estimators the optimal number of bins b, i.e. the b giving largest MI and minimum $|I_\infty(\tau) - \bar{I}_c(\tau)|$, is not always the largest (limited to $b = 64$ in this study) but increases with n, as shown in Fig. 4.18a for the ED estimator and the noise-free Henon map. In the same figure the limits for optimal b from the suggested criterion H11 in Eq.(4.34) (lower for $\mathrm{r}(1) = 0$ and upper for $\mathrm{r}(1) = 1$) are shown with dotted lines and are well beyond the optimal bins found for small lags. For this system, $I(\tau)$ decreases smoothly with τ and therefore the optimal b decreases as well. For $\tau = 10$, $I(\tau)$ levels off for small n and then $b \simeq 2$ is optimal, but as n increases more bins give indeed larger values of MI. Thus as n increases weak MI for large τ becomes significant and can be distinguished from the plateau of independence only when a larger b is used for the binning estimator. However, for a fixed b MI converges to I_∞ with n, even for noise-free data (Fig. 4.18c).

With the addition of noise, $I(\tau)$ decreases and the optimal number of bins drops, as shown in Fig. 4.18b and d respectively for 20% additive noise on the Henon time series. The stronger the noise component is, the more the deterministic

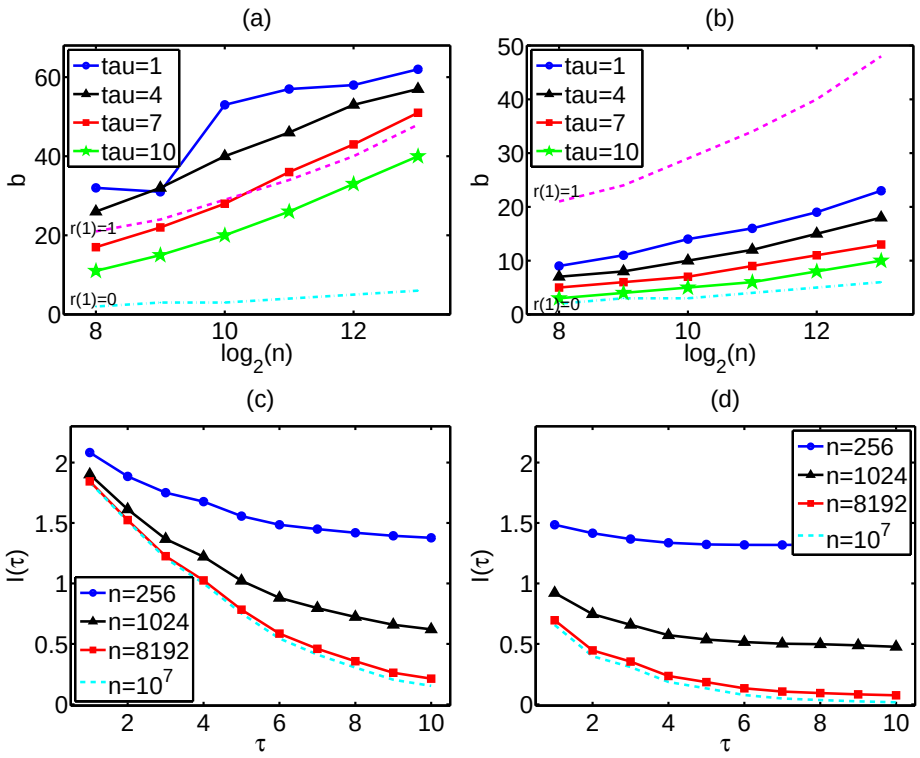

Figure 4.18: (a) Optimal number of bins b as a function of the time series length n for the ED estimator from 1000 realizations of the Henon map for different lags, as given in the legend. (b) Same graph as in (a) but for the Henon map with 20% additive noise. In both plots the dotted lines give the optimal number of bins b from the suggested criterion in (4.34) assuming r(1)=0 and r(1)=1, as given in the plot. (c) Mean estimated MI with ED estimator as a function of τ from 1000 realizations of the Henon map for $b = 32$, and n as in the legend. (d) As (c) but for Henon map with 20% additive noise.

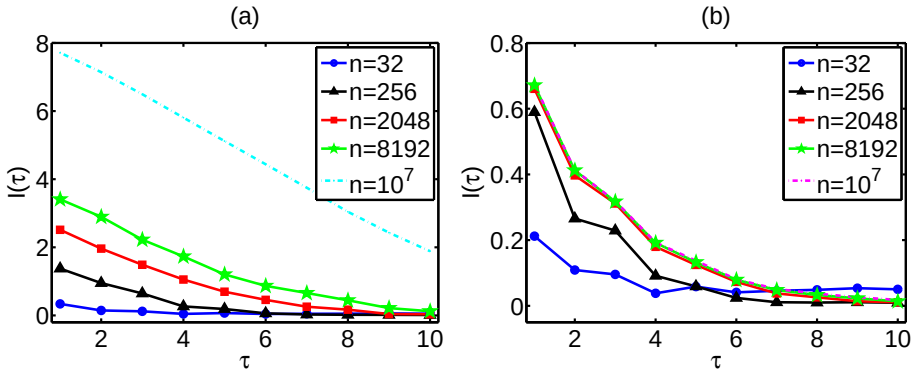

Figure 4.19: Mean estimated MI with AD estimator as a function of τ from 1000 realizations of the Henon map with no noise in (a) and with 20% noise in (b).

structure is masked and the faster the estimated MI levels towards zero with the lag. For the noisy chaotic data, the pattern of the dependence of the binning estimates of MI to n and b is closer to the one observed for the linear systems. For example, the range of optimal b in Fig. 4.18b is at the level of b given by the suggested criterion H11 for $r(1)$ ranging from 0 to 1. The results on EP estimator are similar.

In line with the ED and EP estimators, the AD estimator does not converge with n to I_∞ for the nonlinear systems unless the fine partition is limited by the presence of noise, as shown in Fig. 4.19 for the Henon map. The increase of n directs the algorithm of AD to make a finer partition which results in a larger $I(\tau)$. The effect of n on the adaptive estimator decreases with the increase of the noise level.

As pointed earlier, there is a loose relationship between the number of nearest neighbors k in the KNN estimator and the number of bins b in the binning estimators, i.e. small k corresponds to large b. The lower limit $k = 1$ corresponds to the finest partition for the given data, and the analogue b could be formidably large and is not reached in the study as b goes up to 64 (the same stands for b up to 256). Thus direct comparison to binning estimators when k is very small cannot be drawn. For noise-free chaotic time series, very fine partitions are sought and this agrees with the suggestion in Kraskov et al. (2004) to use small k at the order of 3, which was also used in other simulation studies (Kreuz et al., 2007; Khan et al., 2007). In Fig. 4.20a, it is shown for the noise-free Henon map that $I(\tau)$ increases with decreasing k. For small n, a large value of k gives a poor estimation of the densities and consequently of $I(\tau)$. For a fixed k, MI increases with n (see Fig. 4.20b). Assuming a fixed k the effect of n on the KNN estimator is large similarly to the effect of n on the AD estimator as there in no convergence of MI with n, contrary to the fixed-bin estimators. In agreement to the binning estimators, the MI from the KNN estimator decreases with the noise level. Therefore the dependence of KNN estimator on n is smaller and $I(\tau)$ converges faster to I_∞ with n (Fig. 4.20c). Further, for larger n the estimation is the same regardless of the value

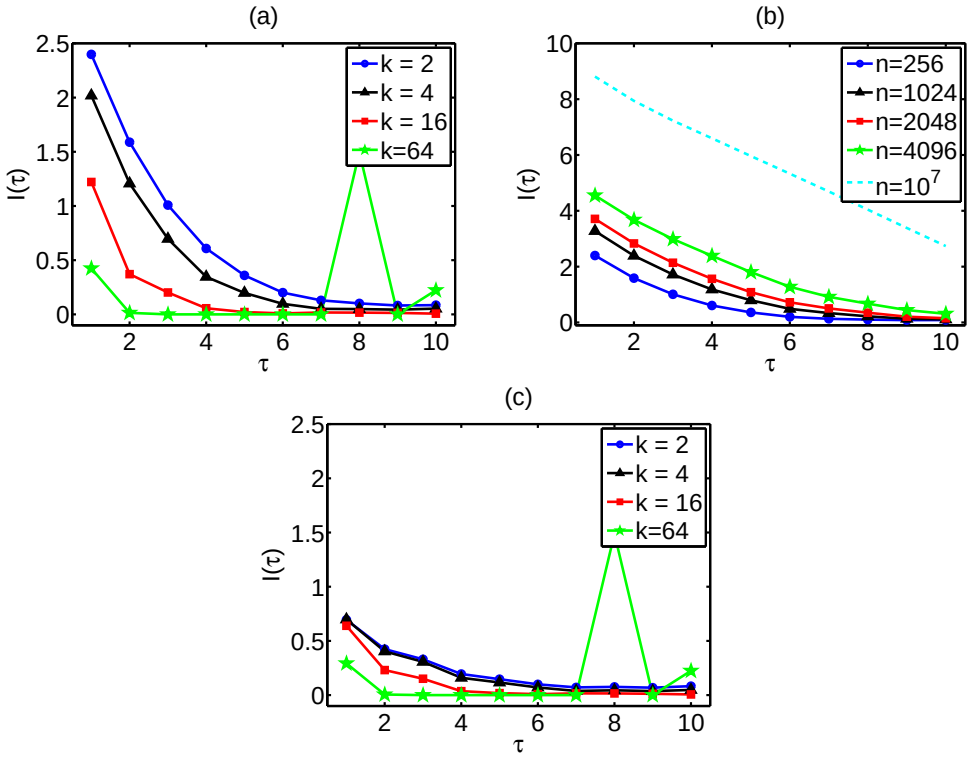

Figure 4.20: (a) Mean estimated MI with KNN estimator as a function of τ from 1000 realizations of the Henon map, for $n = 256$ and k as in the legend. (b) As in (a) but for $k = 2$ and n as in the legend. (c) As in (a) but for Henon map with 20% additive noise.

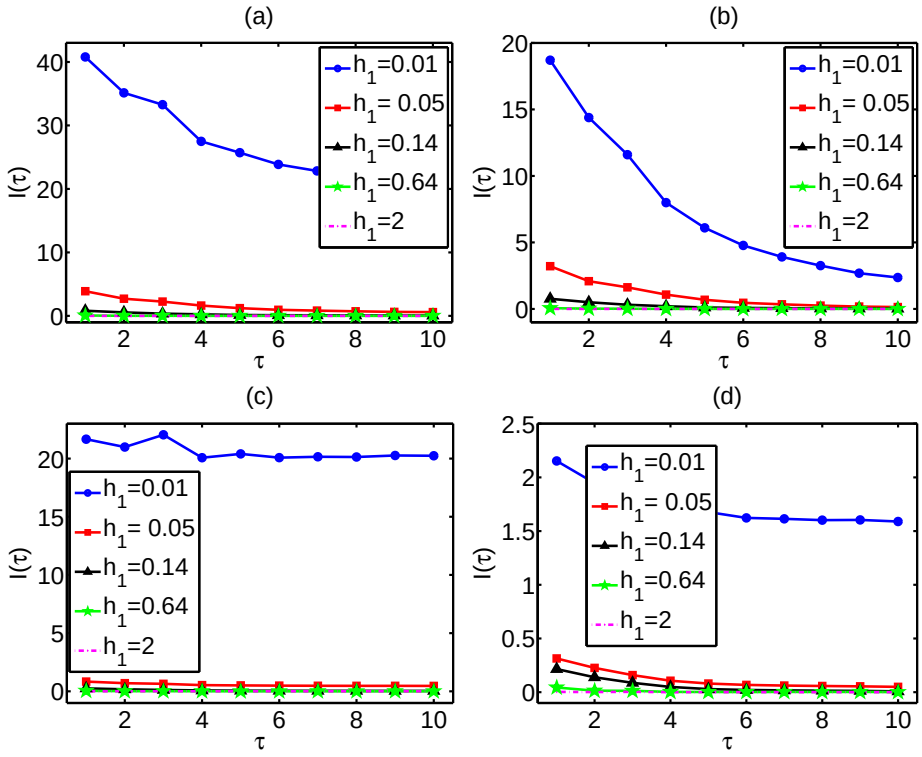

Figure 4.21: (a) Mean estimated MI with KE estimator as a function of τ from 1000 realizations of the Henon map, for $n = 512$ and bandwidths as in the legend. (b) As in (a) but for $n = 4096$ and bandwidths as in the legend. (c) and (d) are the same as (a) and (b) respectively but for the Henon map with 20% additive noise.

of k.

The dependence of the KE estimator on the bandwidth h_1 is similar to the dependence of the KNN estimator on k. As shown in Fig. 4.21a and b for the noise-free Henon map, $I(\tau)$ increases with the decrease of h_1 ranging from 0.01 to 2. It is noted that such extremely large values of $I(\tau)$ for very small bandwidth h_1 do not occur by any other estimator. Given that the KNN estimator for $k = 1$ sets an upper limit for the estimated $I(\tau)$ on the given time series, larger $I(\tau)$ obtained by the KE estimator are superficially overflown estimates due to the use of an unsuitably small h_1 for the given time series. This systematic bias for very small h_1 is more pronounced with the addition of noise as it persists at the same level for larger τ (see Fig. 4.21c). For the noisy data, $I(\tau)$ decreases and differences with respect to the partitioning parameters are smaller, a feature observed also with the other estimators (see Fig. 4.21c and d). Also, the estimated $I(\tau)$ is rather stable to the change of n.

Regarding the 9 criteria for selecting h_1 (and at cases h_2, see Table 4.6), the estimated bandwidths vary with the criterion but within a small range, e.g. for $n =$

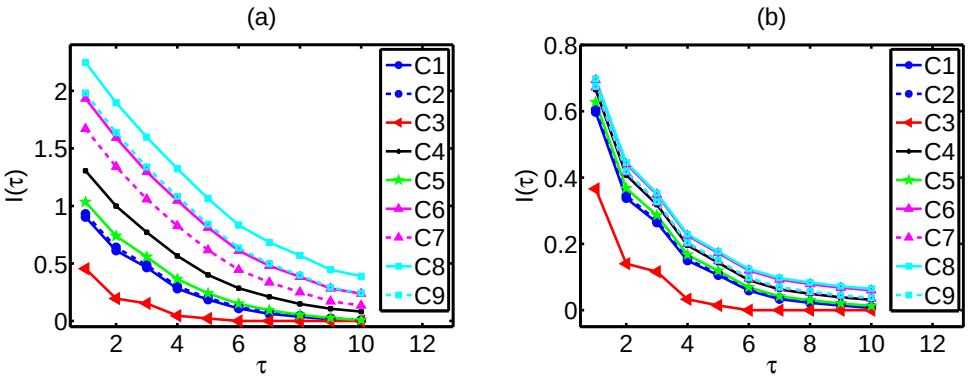

Figure 4.22: (a) Mean estimated MI with KE estimator as a function of τ from 1000 realizations of the noise-free Henon map with $n = 8192$ and nine bandwidth selection criteria as given in the legend. (b) As (a) but for 20% additive noise.

256 they are bounded in $[0.13, 0.35]$ except C3 that always gives larger bandwidths and in this case $h_1 \simeq 0.7$). Deviations of the estimated bandwidths hold for larger n but at smaller magnitudes, e.g. for $n = 8192$, they are bounded in $[0.03, 0.18]$ and for C3 $h_1 \simeq 0.37$). All criteria depend on n in a similar way and estimate smaller bandwidths as n increases giving larger $I(\tau)$ (see Fig. 4.22a for $n = 8192$). When noise is added to the time series, the estimated $I(\tau)$ using different bandwidth selection criteria converge and are rather stable to the change of n (see Fig. 4.22b).

Contrary to linear systems, for noise-free nonlinear maps, the estimated MI does not converge to an asymptotic I_∞ and even for very large time series the MI values computed by different estimators vary, as tested for $n = 10^7$. For increasing n, a finer partition gives larger MI regardless of the selected estimator. The closest approximation to the finest partition for a large n is succeeded by the KNN estimator using a very small k, say $k = 2$ for $n = 10^7$. This turned out to be indeed an upper bound of the estimated MI for large n. For the other estimators, restrictions to the partition resolution, i.e. smallest h_1 for KE and largest b for the binning estimators, bound the estimated MI to smaller values. For example, bins up to $b = 256$ for ED and EP estimators underestimate MI for $n = 10^7$, meaning that b has to increase towards computationally prohibitive large magnitudes to succeed an adequately fine partition for this data size. In the same way, the bandwidth h_1 has to decrease accordingly with n and for large n the KE estimator turns out to be computationally ineffective. The presence of noise sets a lower limit to the partition resolution and allows for an asymptotic MI value I_∞ to which all estimators converge with n for suitably fine partition.

The results on the different estimators were only given for the Henon map in order to facilitate comparisons, but qualitatively similar results are obtained from the same simulations on the Ikeda map. Indicative, some figures indicating the effect of the time series length and the free-parameter of the estimators on Ikeda map are displayed in Fig.4.23.

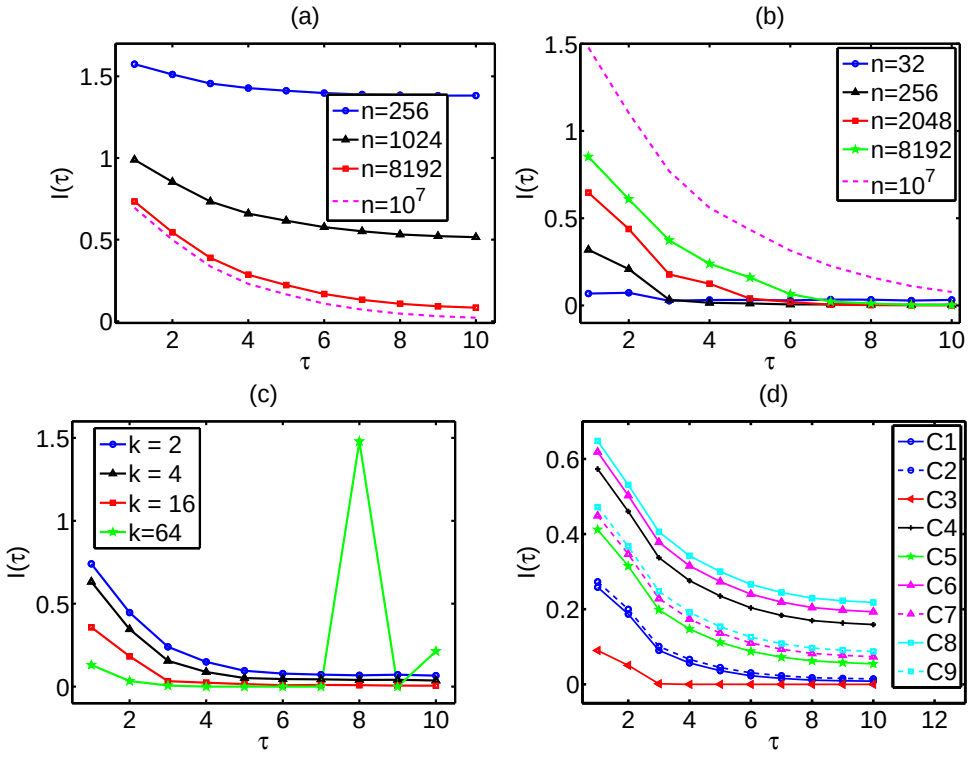

Figure 4.23: (a) As Fig.4.18c, (b) as in Fig.4.19a, (c) as in Fig.4.20a and (d) as in Fig.4.22a, but for the Ikeda map.

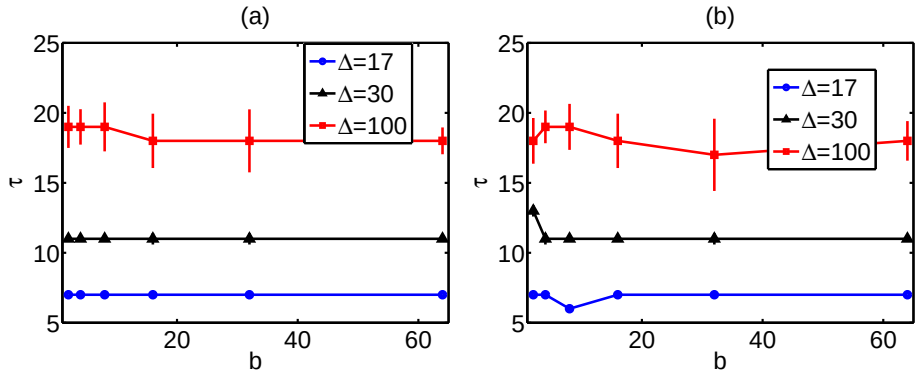

Figure 4.24: (a) Mean estimated τ_0 and standard deviation as error bar as a function of b from all time series lengths using the ED estimator on the Mackey-Glass system with $\Delta = 17, 30, 100$, as given in the legend. (b) As in (a) but for the EP estimator.

Results on nonlinear flows

When using MI on nonlinear flows the interest is often in extracting the lag τ_0 of the first minimum of MI. The estimate of τ_0 with the different MI estimators is examined for the Mackey-Glass system with delays $\Delta = 17, 30$ and 100, that regards increasing complexity of correlation dimension being roughly 2,3 and 7, respectively (Grassberger and Procaccia, 1983).

The simulations using the ED and EP estimators showed that the same τ_0 is estimated for all b, all n and noise levels, and for $\Delta = 17$ and $\Delta = 30$. For $\Delta = 100$, the lag for which MI levels off was estimated and there was some variation in the estimation of τ_0 (see Fig. 4.24). Although $I(\tau)$ increases with b, τ_0 does not vary with b. Moreover, the estimate of τ_0 is stable with n and the addition of noise.

AD estimator is also not affected by n, when computing the lag of the minimum MI τ_0 in the Mackey-Glass system (see Fig. 4.25a). Addition of noise does not affect the mean τ_0, as shown in Fig. 4.25b. From simulations on the Mackey-Glass system with $\Delta = 100$, it was observed that the mean estimated lag that MI levels off, holds for increasing n (see Fig. 4.25c) and addition of noise does not affect it.

The estimation of τ_0 using the KNN estimator on the Mackey Glass systems varies more with n and k than for the binning estimators, as shown in Fig. 4.26a and b. With the addition of noise, the variance of the estimated τ_0 decreases with respect to k, and the mean is rounded to the same integer for all k. For the Mackey Glass system with $\Delta = 100$, it is observed that there is consistency with k and n, with MI for all k having the same shape and therefore giving the same lag for levelling off (see Fig. 4.26c).

Finally, the mean estimated τ_0 using the KE estimator on the realizations of each of the three Mackey-Glass systems is stable against changes in the time series

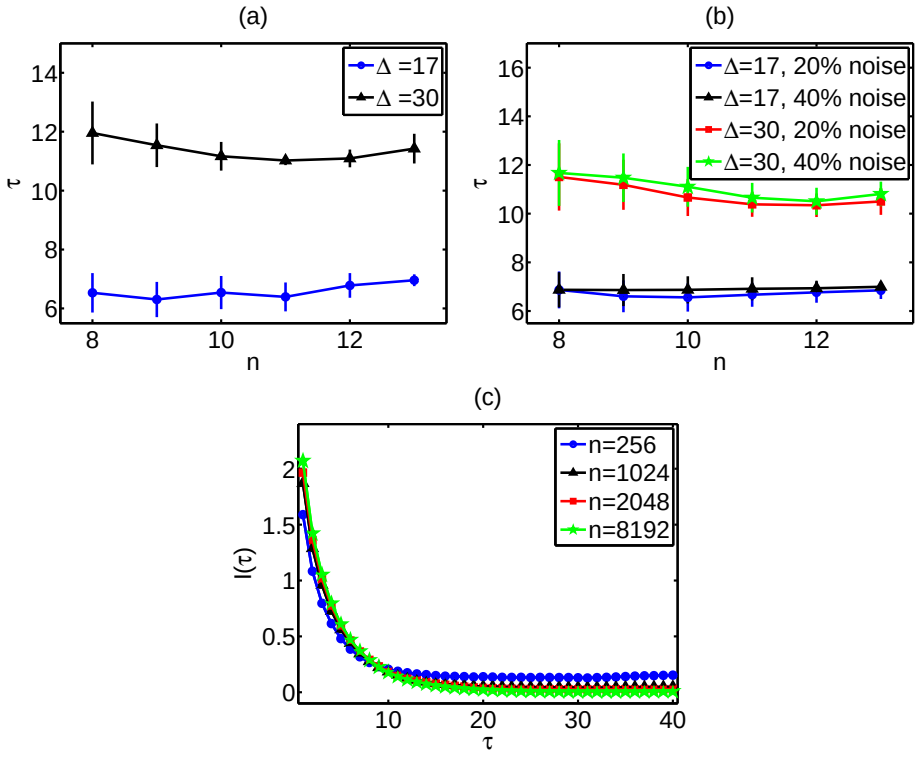

Figure 4.25: (a) Mean estimated τ_0 and standard deviation as error bar as a function of n using the AD estimator on 1000 realizations of the Mackey-Glass system with $\Delta = 17, 30$, as given in the legend. (b) The same as in (a) but for additive noise with levels $20, 40\%$, as given in the legend. (c) Mean estimated MI with the AD estimator as a function of τ from 1000 realizations of the Mackey-Glass with $\Delta = 100$, for n as in the legend.

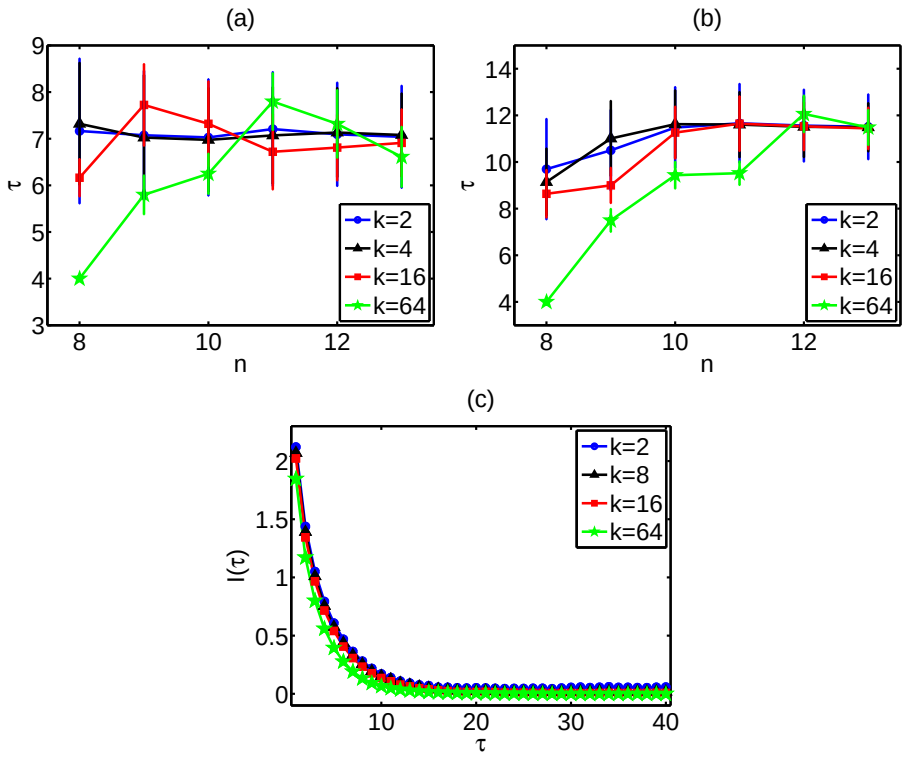

Figure 4.26: Mean estimated τ_0 and standard deviation as error bar as a function of n using the KNN estimator on 1000 realizations of the Mackey-Glass system with (a) $\Delta = 17$, and (b) $\Delta = 30$. (c) Mean estimated MI with the KNN estimator as a function of τ from 1000 realizations of the Mackey-Glass with $\Delta = 100$, for k as in the legend and $n = 2048$.

length and bandwidth as for the binning estimators.

The simulations showed that all estimators identify sufficiently τ_0 as the shape of the MI function is not affected significantly by n or by the addition of noise. ED and KE estimators are the estimators of choice for this task, as they give smaller variation in the estimation of τ_0 compared to the other estimators.

Conclusions

MI estimators are sensitive to their free parameter, with binning estimators (ED and EP) being the most affected. There is a loose correspondence among the different free parameters b, k and h_1 depending also on the time series length. Thus the differences in the performance of the estimators can be explained to some degree by the coarseness of the partition as determined by the free parameter. The choice of b for the binning estimators determines the bin size of the partition. The analogue of the bin size for the KNN estimator is the size of neighborhoods given by the number of neighbors k and for the KE estimator is the size of the efficient support of the kernel approximation given by the bandwidth h_1 (and h_2). The simulation results have quantified the correspondence of the different free parameters and showed that for large time series, a suitable refined partition can be easily accommodated by a very small k or h_1, whereas for the binning estimators the requirement for a very large b renders the binning estimator computationally ineffective. To this respect, the KNN estimator adapts easily to a refined partition by setting, say, $k = 2$, as does the adaptive binning estimator (AD) that has no free parameter, whereas h_1 has to be further investigated at ranges of small values.

The optimization of the parameters of the estimators is very crucial, even more than the choice of the estimator. Therefore, the focus was on estimating the optimal free parameter of each estimator in order to fairly evaluate the estimators. For linear systems, different selection criteria for the optimal free parameter were evaluated and based on the simulation study a criterion for the fixed-binning estimators for the optimal b is proposed. This criterion is given as a function of the autocorrelation and the time series length n. The parameter-free AD estimator tends to overestimate MI compared to the other estimators, indicating that the in-build partition algorithm of AD terminates at a very fine partition. The KNN estimator turns out to be the least sensitive to its free parameter. For example, $k = 2$ that gives a very fine partition does not deviate much for smaller time series where larger k are more appropriate. The simulation results on the linear systems have shown that the KE estimator depends less than the binning estimators on the free parameter for the selected ranges of h_1 and b, respectively.

For noise-free nonlinear systems, all estimators lack consistency, i.e. the estimated MI does not converge with n to an asymptotic value. Therefore, optimal parameter cannot be derived for these systems. The optimal parameter values found for the linear systems tend to give conservative estimates of MI for the nonlinear systems, for which a finer partition is required. This is accommodated by a small k in the KNN estimator. Indeed the simulation study on the different chaotic systems

has shown that the KNN estimator has the least variance with the free parameter k than all other estimators. For noisy nonlinear systems, the MI from all estimators converge with n to an upper limit set by noise and KNN estimator for $k = 2$ turned out to reach this limit faster.

For the computation of the lag τ_0 of the first minimum of MI, the binning estimators ED and EP as well as the KE estimator seem to perform best. For the Mackey-Glass system, although $I(\tau)$ may vary with the free parameter of the estimator and n, τ_0 is rather stable. The addition of noise does not seem to effect the estimation of τ_0.

The KE estimator has the highest computational cost and the fixed-binning estimators become computationally intractable when b has to be very large, as for long chaotic time series. On the other hand, the KNN estimator is rather fast for long time series that require small k for which neighbor search is faster. The computation efficiency of the AD estimator is comparable to that of KNN and these two estimators seem to be the most appropriate for all practical purposes in terms of computational efficiency, parameter selection (small k for KNN and no free parameter for AD) and accuracy of estimation (with KNN scoring better than AD).

It is noted that the consistency of estimators of MI on linear systems is not indicative of the behavior of the estimators on nonlinear systems. Although consistency of estimators is claimed in some recent works, this might be due to the use of only linear systems or noisy real data, such as EEG.

Chapter 5

Evaluation of Univariate Correlation and Information Measures in Detecting Dynamical Changes

Correlation reflects the concept of mathematical association, and is usually used to quantify the linear association between variables. Information measures, e.g. mutual information, are measures that quantify the correlation or dependence of variables and can be considered as generalized correlation measures. Here, a variety of existing linear and nonlinear measures are presented, but also modifications of them are proposed. All the presented measures along with those discussed in Ch.4 are evaluated in their ability to discriminate between different dynamical systems or states of a system and detect the dynamical changes of a system.

5.1 Correlation and Information measures

5.1.1 Linear decorrelation time or decay time

The decorrelation time, ts, is a measure of the memory of a system. Different classes of autocorrelation structures can be distinguished with respect to the form of their decay for large time lags, e.g. for exponentially decay of $r(\tau)$ systems exhibit short-range correlations. The decorrelation time is equal to the lag at which the autocorrelation function for the first time attains the value of zero. It can also be defined as the lag such that $r(\tau)$ first drops below $1/e$ (this definition is used here).

5.1.2 Nonlinear decorrelation time

A measure of the memory of a system is defined here, that encounters also nonlinear correlations. The nonlinear decorrelation time, dct, measures the general decorrelation time of a time series. It is defined as the first lag for which the mutual information $I(\tau)$ levels off, i.e. the lag for which $I(\tau)$ converges to the limit of zero mutual information. Theoretically, for very large values of τ where no correlations between the terms x_t and $x_{t-\tau}$ exist, MI tends to be zero. In practice, MI is always positively biased and therefore converges to a small positive value. Computationally, the levelling of $I(\tau)$ is defined here and for its estimation it is examined when MI first enters and stays for three consecutive lags at the 'zero-correlation area'. This area is estimated from $I(\tau)$ values for a range of very large lags, depending on the system. If m_I and sd_I are the mean and standard deviation of the $I(\tau)$ values for the range of the very large lags, then the 'zero-correlation area' is $m_I \pm 2sd_I$. dct is data dependent and can be used as an additional correlation measure. The estimation scheme of dct is shown in Fig.5.1. Estimation problems arise

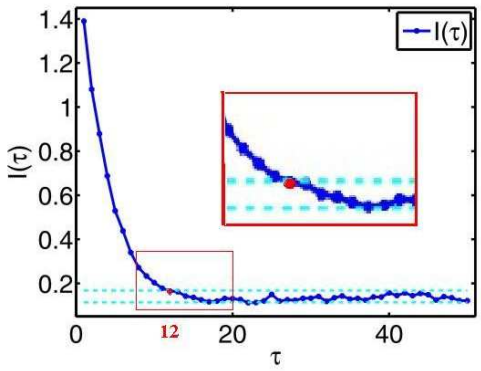

Figure 5.1: Plot of $I(\tau)$ vs τ and the zero-correlation area (within the dotted lines). The limits of the zero-correlation area are estimated using the last ten MI values ($I(\tau)$ for $\tau = 41, \ldots, 50$). Here, $dct = 12$.

especially for small time series and the selection of the range of very large lags for the determination of the 'zero correlation area' should be carefully determined. The suggestion of the three at least consecutive values of mutual information in the 'zero correlation area' in order to identify the nonlinear decorrelation time was rather empirical.

5.1.3 Declination from normality

In an attempt to provide a measure of solely nonlinear correlations of a time series, a new measure has been proposed (P2), called declination from normality. This measure is estimated as the difference of $I(\tau)$ from the expected mutual informa-

tion, $I_G(\tau)$, under the assumption of normality

$$dI(\tau) = I(\tau) - I_G(\tau), \tag{5.1}$$

where $I_G(\tau)$ has been defined in Eq.(4.33) in terms of the autocorrelation function. The cumulative function of declination from normality, $cdI(\tau_{max})$, is defined equivalently to the cumulative autocorrelation $Q(\tau_{max})$ and cumulative mutual information function $M(\tau_{max})$, as the sum of $dI(\tau)$ for lags $\tau = 1, \ldots, \tau_{max}$:

$$cdI(\tau_{max}) = \sum_{\tau=1}^{\tau_{max}} dI(\tau). \tag{5.2}$$

Again, the definition of the cumulative function of declination from normality aims to eliminate the influence of the selection of the lag on the measure.

Transforms of time series have been used in many applications and for a variety of reasons. The simpler one is the standardization in order to have mean zero and standard deviation one. The transform of time series to have uniform marginals dates back to the idea of Spearman (Spearman, 1904) to use ranks. This transform has already been suggested for the computation of mutual information (Pompe, 1993; Kraskov et al., 2004). The transform is static and is defined as

$$u_i = F_X(x_i), \tag{5.3}$$

where F is the empirical cumulative distribution of the time series.

A positive value of $dI(\tau)$ may imply nonlinearity in the examined time series or may be due to deviations of the distribution of the time series from normality. Therefore, in order to dump the effect of non-normality of the data, a transform of the time series is considered. The static transform in order to have a normal marginal distribution is defined as

$$y_t = \Phi^{-1}(F_X(x_t)), \tag{5.4}$$

where Φ is the cumulative density function of the standard normal distribution. The measures $dI(\tau)$ and $cdI(\tau)$ can thus be estimated on time series $\{y_t\}$, instead of being estimated on $\{x_t\}$. In Fig.5.2, an example of an EEG segment and its distribution before and after the transform to normal marginal distribution is displayed. This transform renders the amplitude normal (Gaussian) without alerting the dynamics of the time series and eliminates the effect of heavy tails and outliers in the signal amplitude distribution. The transform of the data to have gaussian empirical distribution has been proposed (Palus, 1995), in order to perform the surrogate data test for nonlinearity. The effects of the static nonlinearities or non-Gaussian distributions of the data under study on the nonlinearity tests are prevented and only the true dynamical nonlinearity is detected. It is noted that this transform does not make the time series Gaussian but only their marginals. Also, a positive value of the measure could be obtained not just due to nonlinearity but also from time series that that are non-monotonic transforms of a Gaussian process.

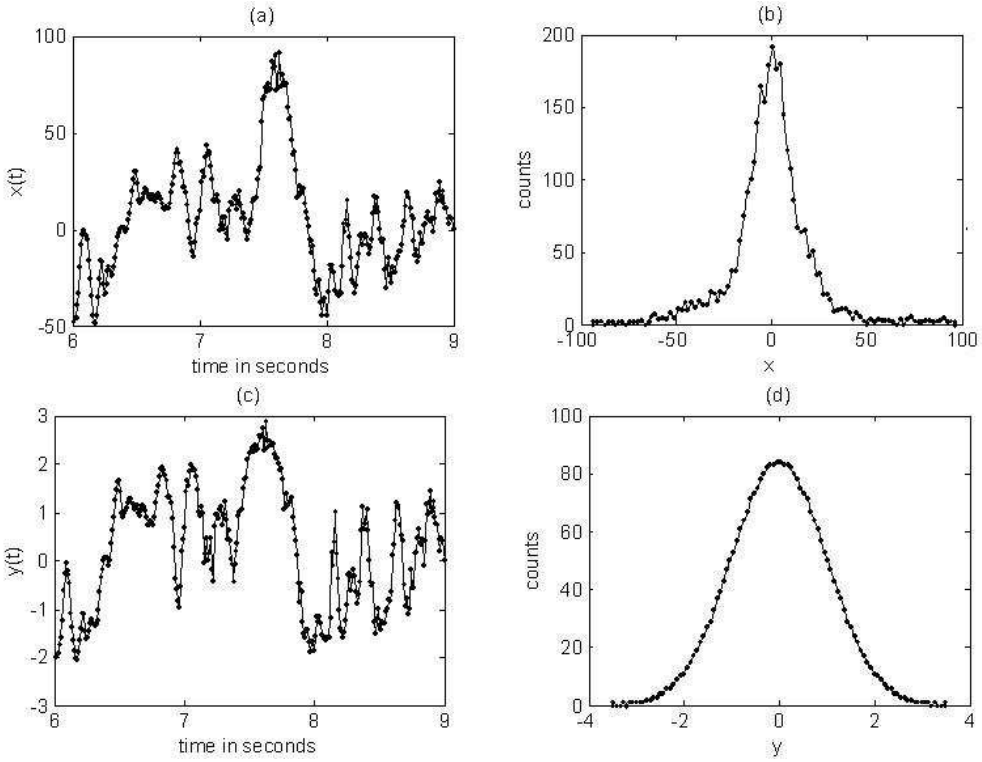

Figure 5.2: (a) Graph of an EEG segment vs time and its histogram in (b). The same for the "Gaussianized" segment in (c) (with normal marginal distribution) and its histogram in (d).

5.2 Entropy measures

The Shannon entropy has already been introduced in Ch.4. The Shannon entropy has been defined on different variables under different names, e.g. spectral entropy when the variable is the power spectrum at different frequencies. Here, two forms of Shannon entropy, Tsallis entropy and Sample entropy are presented.

5.2.1 Shannon entropy on two variables

The first form of the Shannon entropy is defined for the two-dimensional variable $(X_t, X_{t-\tau})$ instead of the univariate variable of a time series $\{x_t\}$. In applications, very often the Shannon entropy is estimated on the variable X of the univariate time series. The suggestion of estimating Shannon entropy on the two dimensional variable $(X_i, X_{i-\tau})$ has the advantage that incorporates also characteristics of the dynamics of the examined system. The Shannon entropy of a univariate variable X is referred to the marginal distribution of a time series, and no information about the dynamics of the examined systems is involved. For the estimation of the Shannon entropy, denoted by $ShEnt(\tau)$, the equidistant binning scheme (see Ch.4) has been

used here in the applications.

5.2.2 Shannon entropy on variables from recurrence quantification analysis

The second form of the Shannon entropy is based on the Recurrence Quantification Analysis (RQA) (Eckmann et al., 1987). RQA is a method of nonlinear data analysis for the investigation of dynamical systems, which quantifies the number and duration of recurrences of the system trajectory. RQA uses the recurrence plots, tools which visualize the recurrence behavior of the trajectory of the examined dynamical system reconstructed from the time series. For a time series $\{x_t\}$, the reconstructed vectors $\mathbf{x}_t$ of the state space are formed typically with the method of delays using an embedding dimension m and delay τ, so that $\mathbf{x}_t = (x_t, x_{t-\tau}, \ldots, x_{t-(m-1)\tau})'$. The recurrence matrix $\mathbf{R}$, is a square matrix with $(m-1)\tau$ columns and lines. $\mathbf{R}$ is filled with ones and zeros indicating which points are neighbors. If $\mathbf{R}_{ij}$ is the (i,j) cell of this matrix, then $\mathbf{R}_{ij} = 1$ if the reconstructed points $\mathbf{x}_i$, $\mathbf{x}_j$ are neighbors, otherwise $\mathbf{R}_{ij} = 0$. The points $\mathbf{x}_i$, $\mathbf{x}_j$ are neighboring points if their distance (e.g. Euclidean) is less than a fixed radius ϵ_x. The recurrence plot has black dots for each pair (i,j) if $\mathbf{x}_i$, $\mathbf{x}_j$ are neighboring points, otherwise has white dots. The main diagonal of $\mathbf{R}$ is for $i = j$ and therefore has elements equal to one. A schematic diagram of the construction of the recurrence plot from a time series is presented in Fig.5.3. A line parallel to the diagonal

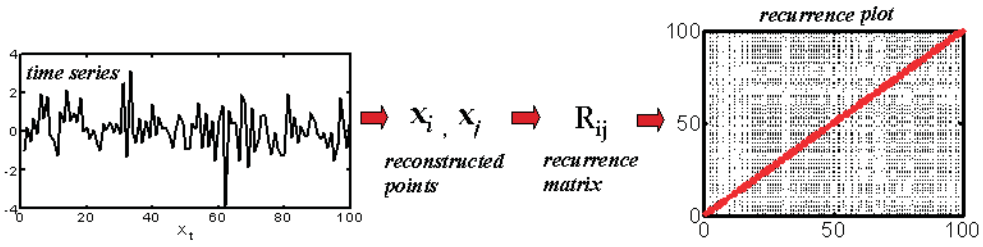

Figure 5.3: Schematic diagram of construction of an RQA plot from a time series. The thick (red) line of the RQA plot represents the main diagonal of the recurrence matrix $\mathbf{R}$.

line in the recurrence plot occurs when a segment of the trajectory "runs" parallel to another segment, i.e. the trajectory visits the same region of the phase space at different times, and is estimated from the recurrence matrix as $R_{i+k,j+k} = 1$, for $k = 1, \ldots, l$, where l is the length of the so-called line. The direction of these diagonal structures can differ. Diagonal lines parallel to the main diagonal represent the similar evolution of trajectories for a certain time period. If l is the variable of the length of the diagonal lines which are parallel to its main diagonal, the Shannon

entropy using the probability distribution of the diagonal line lengths is defined as

$$ShEnt_{RQA}(m,\tau) = -\sum_{l=1}^{l_{max}} p(l)\ln p(l), \tag{5.5}$$

where l_{max} is the maximum length of the diagonal lines parallel to the main diagonal of $\mathbf{R}$. $ShEnt_{RQA}$ is given as a function of the embedding parameters, although m and τ are not included in the right hand side of Eq.(5.5). However, the embedding parameters define the points and their neighborhoods and therefore the lengths of the diagonal lines in $\mathbf{R}$. $ShEnt_{RQA}(m,\tau)$ reflects the complexity of the deterministic structure in the system. Similarly to this measure, one can define Shannon entropy also on other variables, e.g. the length of the vertical lines of the recurrence plot.

5.2.3 Tsallis entropy

A system with macroscopic memory is called non-extensive. Tsallis entropy (Tsallis, 1988) is the non-extensive entropy that reflects the macroscopic memory of the system. For a discrete variable X, Tsallis entropy is defined as

$$S_q(X) = \frac{1}{q-1}(1 - \sum_x p^q(x)), \tag{5.6}$$

where $p(x)$ denotes the probability distribution of the variable X and q is a real parameter which is also reflecting the non-extensivity of the system. When $q \to 1$, the Tsallis entropy recovers the Shannon entropy for any probability distribution. Similarly to the first form of Shannon entropy, Tsallis entropy, denoted by $TsEnt(\tau, q)$, is estimated here on the two-dimensional variable $(X_t, X_{t-\tau})$ using the equidistant binning scheme.

5.2.4 Sample entropy

Sample entropy (Richman and Moorman, 2000) is a measure of the complexity of a system, which examines whether consecutive values of a time series recur with a declination (as in RQA). Sample entropy, $SampEn(m)$ is the negative natural logarithm of the conditional probability that two data segments of length m, being similar at tolerance r, are also similar when their length is augmented to $m+1$. In the computations, identical segments are not considered. Sample entropy is defined as

$$SamEnt(m) = -\ln(A/B) = -\ln\frac{\sum_{i=1}^{n-m\tau} A_i}{\sum_{i=1}^{n-m\tau} B_i}, \tag{5.7}$$

where A and B are the number of recurrences of $(m+1)$ and m consecutive points of the time series, respectively. B_i is the number of $\mathbf{x}_j^m$ such that the distance of $\mathbf{x}_i^m$ and $\mathbf{x}_j^m$ is less than r, where $\mathbf{x}_i^m$ and $\mathbf{x}_j^m$ are m-dimensional vectors and A_i

is the number of $\mathbf{x}_i^{m+1}$ within tolerance r of $\mathbf{x}_j^{m+1}$ for the $(m+1)$-dimensional vectors $\mathbf{x}_i^{m+1}$ and $\mathbf{x}_j^{m+1}$. Self-counting of points is excluded by requiring that $i \neq j$. Although m and r are critical in determining $SampEn(m)$, there exists no optimization rule for their values. It is usually set $r \in [0.1, 0.25]$ and $m = 1$ or 2 for data records of lengths $n \in [100, 5000]$.

Sample entropy is an extension of approximate entropy, a measure of regularity closely related to the Kolmogorov entropy (Pincus, 1991), developed to reduce its bias. The difference of the two measures is that approximate entropy counts each sequence as matching itself (allows $i = j$), a practice carried over at sample entropy to avoid the occurrence of $\ln 0$ in the calculations. In addition to eliminating self-matches, the sample entropy algorithm is simpler than the approximate entropy algorithm, requiring approximately one-half as much time to calculate and is largely independent of the record length.

5.2.5 Permutation entropy

Permutation entropy (Bandt and Pompe, 2002) is a complexity measure for time series analysis based on the concept that a dynamical system can be suitably represented and analyzed using a symbolic sequence. For the estimation of permutation entropy, the values $1, 2, \ldots, m$ are considered. There are $m!$ permutations of them and π denotes a specific permutation of them. For a time series $\{x_t\}, t = 1, \ldots, n$, all segments with length m are considered (or reconstructed vectors with embedding dimension m), and $N(\pi)$ is the number of them with the same order as π. The relative frequency of each permutation π is given as $p(\pi) = \frac{N(\pi)}{n-m+1}$ and (normalized) permutation entropy is defined as

$$PermEnt(m) = -\frac{\sum p(\pi) log p(\pi)}{(m-1)\log(m!)} \tag{5.8}$$

$PermEnt(m)$ is a measure of the departure of the time series under study from a complete random one; the smaller it is, the more regular the time series is.

5.3 Applications of Information Measures

5.3.1 Evaluation of three types of correlation measures in discriminating regimes of dynamical systems

One of the most important problems in nonlinear time-series analysis is the detection of dynamical changes in complex systems and the designation of the dynamical changes, e.g. through external stimulus. Dynamical change detection of a system is a developing area of time series analysis with many applications, e.g. for prediction of epileptic seizures, earthquakes or weather. A number of methods have been proposed to detect dynamical changes based on characteristic dynamical system invariants from state space based methods (Trulla et al., 1996; Hively

et al., 2000; Tykierko, 2008), synchronization measures (Mormann et al., 2000) and information theory (Hively et al., 2000).

Here, linear and nonlinear measures of correlation are compared in distinguishing different regimes of a dynamical system on the frame of time series analysis (PP2). The linear measures are the autocorrelation function $r(\tau)$ for certain delays and the sum of autocorrelations (Portmanteau) until a maximum delay, $Q(\tau_{max})$. Similarly, the nonlinear measures that have been considered are the mutual information $I(\tau)$ on the same delays and the sum of $I(\tau)$ for the same maximum delay, $M(\tau_{max})$. It is also examined whether the discrimination of the dynamical regimes can be achieved equally well, or even better, using as measures the p-values of the surrogate data test for nonlinearity. In this way, the p-value can be considered as an indirect measure of departure from linear correlations. To assess the three correlation measure types (linear measures, nonlinear measures and p-values of the surrogate data test), Monte Carlo simulations on well-known linear and nonlinear systems are used, varying their complexity by monitoring control parameters of the system (AR model, Mackey Glass).

Measures

Three types of measures were used in this study; measures of linear correlation, measures of nonlinear correlation and measures of nonlinear deviation from linearity. The autocorrelation function $r(\tau)$ (defined in Eq.(3.7)) is estimated for lags $\tau = 1, 5, 10, 20, 30$, and the cumulative autocorrelation function $Q(\tau_{max})$ (defined in Eq.(4.21)) is estimated for $\tau_{max} = 40$. For nonlinear correlations, mutual information $I(\tau)$ (defined in Eq.3.9) is estimated for the same lags $\tau = 1, 5, 10, 20, 30$, and the cumulative mutual information function $M(\tau_{max})$ (defined in Eq.(4.22)) is estimated for $\tau_{max} = 40$. As a measure of nonlinear deviation from linearity, the p-values from surrogate data test for nonlinearity are used. The p-values are defined parametrically as from Eg.(3.10). As test statistics, the previous nonlinear measures are used ($I(1), I(5), I(10), I(20), I(30), M(40)$). The surrogate data capture only the linearity of each time series. Thus, a p-value < 0.05 indicates the presence of nonlinearity in the time series.

Set Up

For the simulations, 1000 realizations from well-known systems with time series lengths $n = 1000, 2000, 4000$ were generated. The linear systems considered were the autoregressive AR(9) model defined in Eq.(4.26) under a quadratic transform ($y_t = x_t^2$) and under a cubic transform ($z_t = x_t^3$). Representative plots from one realization of each system are displayed in Fig.5.4. The Mackey-Glass system (Eq.(4.31)) was also used (discretization step $\tau_s = 17$, sampling time $\delta = 0.1$) with $\Delta = 30$ for low-dimensional chaos of fractal dimension about 3, and $\Delta = 70$ and 100 for high-dimensional chaos of fractal dimension about 6 and 7, respectively. Representative plots from one realization of each system are displayed in

Fig.5.5. Observational white noise at levels 20% and 50% is also added on the chaotic systems, given as a percentage of the standard deviation of the noise-free data. The aim was to discriminate between the different dynamical systems. The discrimination of each pair of systems was assessed with the ROC curves.

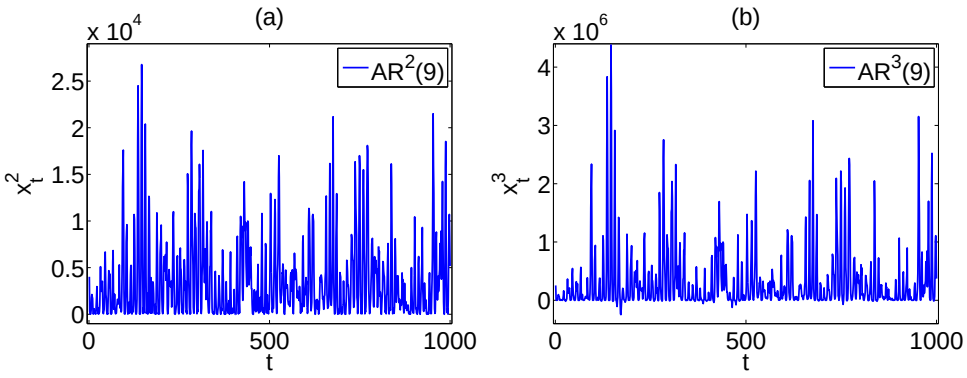

Figure 5.4: One realizations of AR(9) models with $n = 1000$ under a quadratic transform in (a) and under a cubic transform in (b).

Results

The discrimination of the measure profiles from the different systems is presented here. The first discrimination is between the quadratic and the cubic transforms of the AR(9) model. In order to discriminate two systems, the AUC values from the ROC curve was computed for the two corresponding measure profiles. The sample distributions of some measures for the square and cubic transform of AR(9) model for $n = 1000$ are shown in Fig.5.6. For $n = 1000$, $p_{I(1)}$ with $AUC = 0.95$ has the highest discriminating power, follows $I(1)$ with $AUC = 0.91$ and $r(1)$ with $AUC = 0.86$. The ranking of the measures as far as their discriminating power is concerned is invariant to time series length, however the AUC values increase with n. For $n = 2000$, the AUC values of $r(1)$, $I(1)$ and $p_{I(1)}$ are 0.97, 0.90 and 0.98, respectively and for $n = 4000$ they increase to 0.98, 0.99 and 0.99, respectively. The ROC curves for these measures are shown in Fig.5.7.

The discrimination of Mackey-Glass with $\Delta = 30$ and $\Delta = 70$ is succeeded with many of these measures and with a high discriminating power, even for $n = 1000$. The five measures with the highest discriminating power and the corresponding AUC values for all time series lengths are given in Table 5.1. Again, the time series length does not significantly change the ranking of the measures concerning their discriminating power. Addition of noise also does not affect the discriminating power of the measures or the ranking; even for noise level of 50%, the same measures give AUC values equal to one. The discrimination of the Mackey-Glass system with $\Delta = 70$ and $\Delta = 100$ was also succeeded with some measures even for $n = 1000$ and 50% noise level. The five measures with the highest discriminat-

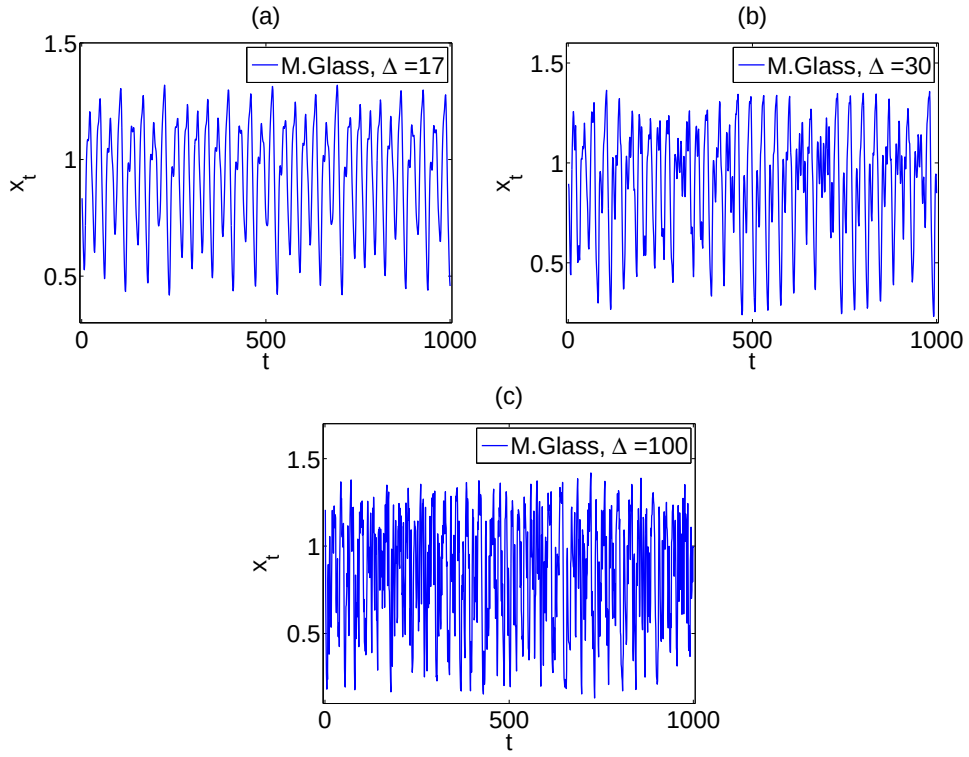

Figure 5.5: One realizations of the Mackey-Glass system with $n = 1000$ and $\Delta = 17$ in (a), with $\Delta = 30$ in (b) and with $\Delta = 100$ in (c).

Table 5.1: Five optimal measures and their AUC values from the discrimination of the Mackey-Glass system with $\Delta = 30$ and $\Delta = 70$, for $n = 1000, 2000, 4000$.

	n=1000		n=2000		n=4000	
A.A	measure	AUC	measure	AUC	measure	AUC
1.	$r(1)$	1	$r(1)$	1	$r(5)$	1
2.	$r(5)$	1	$r(5)$	1	$I(1)$	1
3.	$I(1)$	1	$I(1)$	1	$I(10)$	1
4.	$M(40)$	0.99	$M(40)$	1	$I(30)$	1
5.	$I(5)$	0.98	$I(5)$	0.99	$M(40)$	1

ing power and the corresponding AUC values for all time series lengths are given in Table 5.2. The optimal measures from discrimination among the Mackey-Glass systems including all n and noise levels are $r(1)$, $I(1)$, $r(5)$, $p_{I(10)}$ and $M(40)$ with AUC values (mean from all discriminations) $0.99, 0.99, 0.96, 0.95, 0.94$, respectively.

Finally, measures were tested on their ability to discriminate the quadratic transform of AR(9) and the Mackey-Glass system with $\Delta = 30$, which has a low complexity. The five optimal measures (out of the 18 measures) with the highest

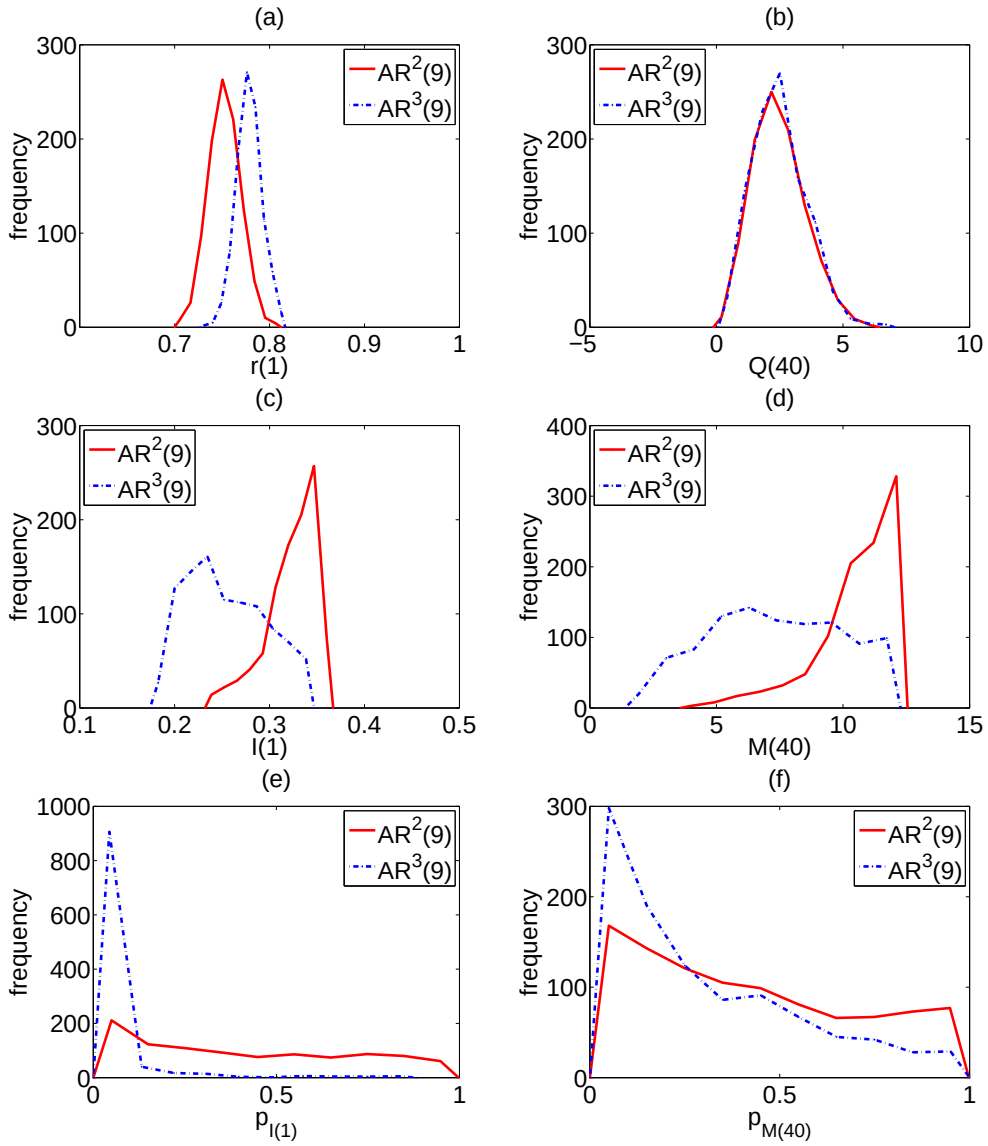

Figure 5.6: (a) Plot of the distributions of the $r(1)$ measure from 1000 realizations of the AR(9) model of length $n = 1000$ with quadratic and cubic transform. (b), (c), (d), (e), (f) as in (a) but for $Q(40)$, $I(1)$, $M(40)$, $p_{I(1)}$ and $p_{M(40)}$.

discriminating power for all lengths and noise levels are $r(5)$,$r(10)$,$r(20)$,$p_{I(10)}$,$I(1)$ with AUC values $1, 1, 1, 0.99, 0.99$ respectively. Both linear and nonlinear measures seem to have a significant discrimination performance.

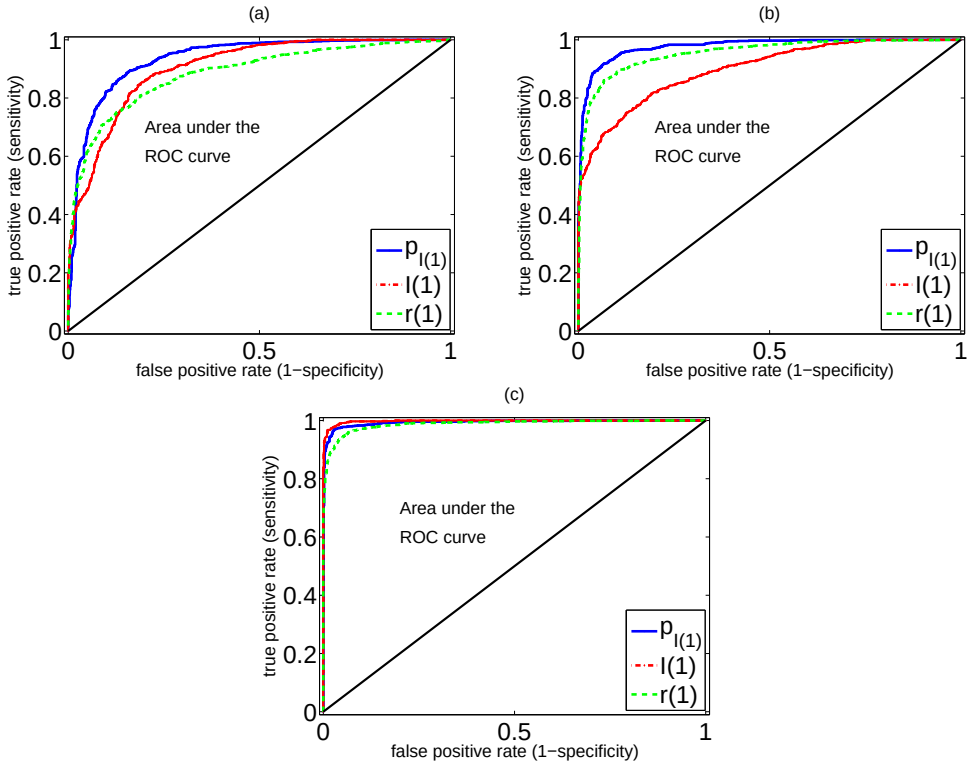

Figure 5.7: (a) ROC curves of $r(1)$, $I(1)$ and $p_{I(1)}$ from the discrimination between the quadratic and the cubic transform of AR(9), for $n = 1000$. (b) and (c) as in (a) but for $n = 2000$ and $n = 4000$ respectively.

Table 5.2: Five optimal measures and their AUC values from the discrimination of the Mackey-Glass system with $\Delta = 70$ and $\Delta = 100$, for $n = 1000, 2000, 4000$.

	n=1000		n=2000		n=4000	
A.A	measure	AUC	measure	AUC	measure	AUC
1.	$r(1)$	1	$p_I(10)$	1	$p_M(40)$	1
2.	$I(1)$	0.99	$I(1)$	1	$p_I(10)$	1
3.	$p_I(10)$	0.98	$r(1)$	1	$I(1)$	1
4.	$r(5)$	0.82	$p_M(40)$	0.96	$r(1)$	1
5.	$M(40)$	0.79	$p_I(5)$	0.95	$p_I(20)$	0.99

Conclusions

From this simulation study, it is observed that linear and nonlinear measures perform equally well. The cumulative measures do not outstand in performance, however they are included in the list of the five optimal measures in many discrimination settings. The p-values from the surrogate data test turned out also to have a satisfactory discriminative power, however there is a high computational cost for

the generation of the surrogates.

5.3.2 Evaluation of information measures in detecting dynamical changes

The previous study was extended in order to include a wide variety of information measures. The aim was again to investigate the discriminating power of the information measures in order to detect dynamical changes in time series from linear and chaotic systems. The evaluation included existing and developed linear and nonlinear correlation measures and measures of entropy and was assessed by Monte Carlo simulation. This simulation study was more extensive than the previous one, including stochastic linear and nonlinear systems and chaotic systems. Moreover, the influence of the marginal distributions of the time series on the measures was investigated. Thus, the measures were estimated also from two transforms of the time series in order to possess marginal uniform and marginal Gaussian distribution. The statistical analysis of the results was based on ROC curves. The study verified the effectiveness of the information measures in discriminating between different dynamical systems and in detecting smooth changes of dynamical systems (P9).

Simulation systems

The first type of systems are the following linear and nonlinear stochastic systems:

- Normal white noise, $w_t \sim N(0,1)$
- Autoregressive process AR(1), with $\phi = 0.2$
- Threshold Autoregressive process (TAR), with equation

$$X_t = \begin{cases} -0.5X_{t-1} + w_t & , \quad if\ X_{t-1} < 1 \\ 0.4X_{t-1} + w_t & , \quad otherwise \end{cases} \tag{5.9}$$

- Autoregressive process with conditional heteroskedasticity ARCH(1), with equation

$$X_t = \sqrt{1 + 0.4X_{t-1}^2} + w_t \tag{5.10}$$

- Generalized Conditional Autoregressive heteroskedasticity GARCH(1,1)

$$X_t = \sqrt{h_t} w_t, \tag{5.11}$$

  where $h_t = 0.01 + 0.8h_{t-1} + 0.15h_{t-1}^2$

- Bilinear process (BL), with equation

$$x_t = 0.6w_{t-1}x_{t-2} + w_t \tag{5.12}$$

The parameters of the models were chosen so that the time series have zero autocorrelations (except AR(1)) and were standarized to have zero mean and standard deviation one in order to avoid possible discriminations due to different autocorrelations or scaling of values. In Fig.5.8a, time series from one realization of each (standarized) stochastic system is shown.

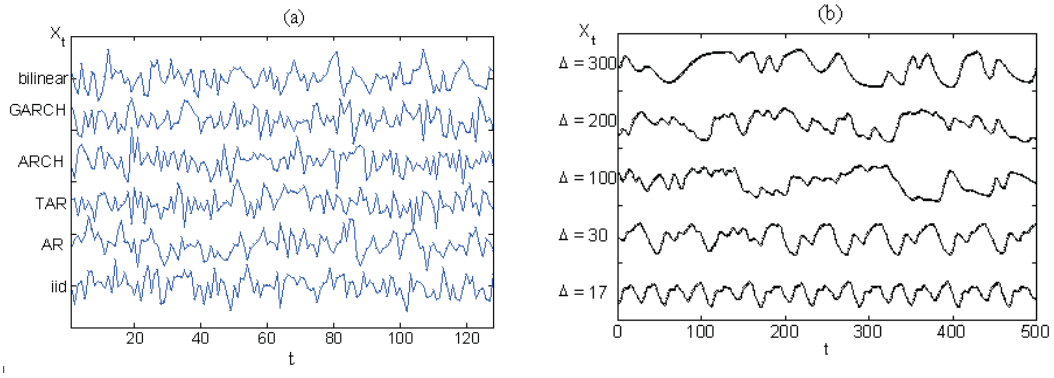

Figure 5.8: (a) One realization of each stochastic system $n = 128$. (b) One realization of the Mackey Glass system with $\Delta = 17, 30, 100, 200, 300$ as shown in the legend, with $n = 500$.

The chaotic system of Mackey-Glass was also considered (the time series were generated for discretization time 0.1 and sampling time 17). For the Mackey-Glass systems, Δ is set $17, 30, 100, 200, 300$ and $n = 2000, 4000$. Realizations of the Mackey-Glass system with different Δ values can be considered as systems with different complexity where the complexity increases with Δ. In Fig.5.8b, time series from one realization of each the Mackey-Glass for the different Δ is shown.

Measures

A variety of information measures is used in this study; linear measures, nonlinear measures, and entropy measures. The linear measures that have been used are the linear decorrelation time, ts, and the cumulative autocorrelation function, $Q(\tau_{max})$. For the stochastic systems, $\tau_{max} = 3$ and 6 and for the Mackey-Glass systems $\tau_{max} = 5, 10, 20, 50$ and 100.

The nonlinear measures that have been used are the nonlinear decorrelation time dct and the cumulative mutual information $M(\tau_{max})$. MI was estimated using ED, EP, AD and KNN estimators and thus dct and $M(\tau_{max})$ were estimated using the four estimators. The estimated measures are denoted by dct_{ED}, dct_{EP}, dct_{AD}, dct_{KNN} and $M_{ED}(\tau_{max})$, $M_{EP}(\tau_{max})$, $M_{AD}(\tau_{max})$, $M_{KNN}(\tau_{max})$, respectively. The cumulative declination from linearity, $cdI(\tau_{max})$, is estimated using all the previously mentioned estimators (notations are as for $dI(\tau)$, in order to indicate also the estimation scheme that has been used). Again, for the

stochastic systems τ_{max} is set to be 3 and 6, and for the Mackey-Glass systems $\tau_{max} = 5, 10, 20, 50$ and 100.

Shannon entropy, $ShEnt(\tau)$, was estimated on $(X_t, X_{t-\tau})$ using the binning scheme with equidistant cells. Tsallis entropy, $TsEnt(\tau, q)$, was also estimated in the same way and q was set to be 1.5 and 3. For the stochastic systems, the Shannon and Tsallis entropy were estimated for $\tau = 1$, while for the Mackey-Glass system for $\tau = 1, 5, 10, 20, 50, 100$. Shannon entropy based on RQA, $ShEnt_{RQA}(m, \tau)$, was estimated for $m = 2$ and $\tau = 1$. Sample entropy, $SamEnt(m)$, was estimated for $m = 2$ for the stochastic systems and for $m = 2, 5, 10, 15, 20$ for the Mackey-Glass systems. The permutation entropy, $PermEnt(m)$, was estimated for $m = 5, 6, 7$. For the stochastic process, τ is set to be 1 as the systems are of small order, while for the Mackey-Glass system which is a flow, there is a need to set also larger values. Selection of m for sample entropy for the stochastic processes is the same as for the Shannon entropy in order to be comparable. For the Mackey-Glass systems, m takes also larger values and not only the value 2, as the complexity of the systems may be high (for $\Delta = 100$). Selection of m for permutation entropy is based on the literature (Bandt and Pompe, 2002; Cao et al., 2004).

Discrimination of dynamical systems with different complexity

The first goal is to investigate whether information measures can discriminate between classes of the stochastic systems. Specifically, the discrimination of the white noise and the AR(1) model from the rest of the stochastic systems is investigated. The measures are estimated for 100 realizations of each system, for lengths $n = 128, 256, 2048$. All the measures were also estimated for two transformations of the time series in order to have uniform (Eq.5.3) and normal marginal distribution (Eq.5.4).

In order to examine if the measures can discriminate between systems with different complexity, the chaotic system of Mackey-Glass was considered. Again, measures were estimated from 100 realization of the system Mackey-Glass for $\Delta = 17, 30, 100, 200, 300$ and time series lengths $n = 2000$ and 4000. In Fig.5.8b, time series from one realization of the Mackey-Glass for each parameter values is shown.

Statistical analysis For the statistical evaluation of the measures, the ROC curves were used. The samples of the values of the measures from the 100 realizations of each time series will be called here measure profiles, although they are not timely depended. The measure profiles that correspond to the different systems or dynamical states are compared in order to assess whether they come from the same distribution. AUC values are estimated in order to quantify the discrimination ability of each measure. An example is given in Fig.5.9, where the profile of $Q(3)$ from the white noise and the AR(1) process seems to significantly differ and therefore the classification gives a high AUC value.

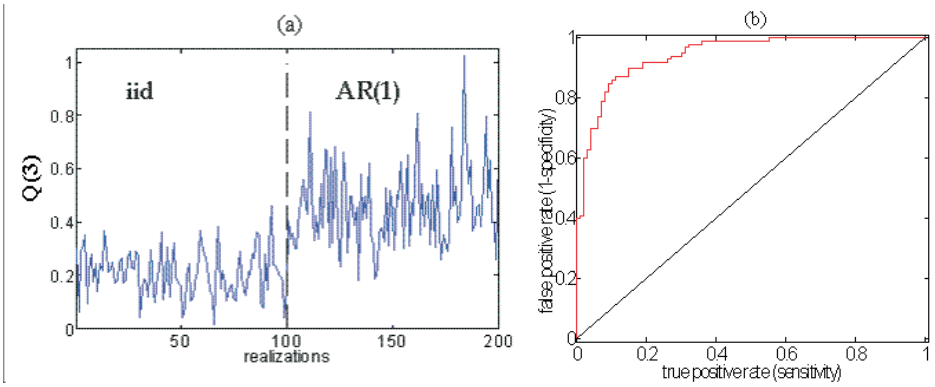

Figure 5.9: (a) Samples of cumulative autocorrelation $Q(3)$ from the white noise and AR(1) for 100 realizations of length $n = 128$. (b) ROC curve for the discrimination of the two systems (AUC=0.944).

Detection of transition of system dynamics

Mackey-Glass system was used as an example in order to detect the transition of system dynamics. The transition was realized by a gradual increase of Δ. It is investigated whether the information measures can detect the dynamical changes during the smooth transition of the system from one state (for a certain Δ) to another one (for a different Δ). This transition is simulated with step by step variation (increase) of Δ (see Fig.5.10). Specifically, a realization of Mackey-Glass system

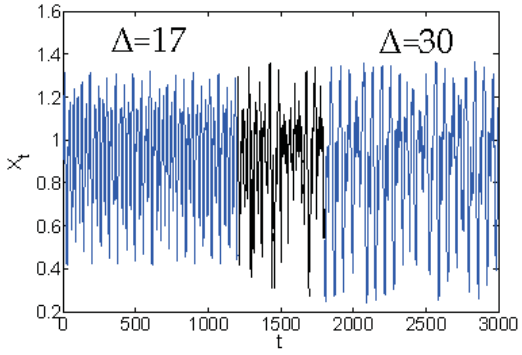

Figure 5.10: One realization of the Mackey-Glass system. The first part of the time series has $\Delta = 17$ (first state), the last part has $\Delta = 30$ (second state) and the middle part is generated by increasing values of the parameter from $\Delta = 17$ up to 30 (transition period).

with length 2.800.000 points was formed, for Δ increasing from 17 up to 300. For $\Delta = 17, 30, 100, 200, 300$ (five states), the segments had length 400.000 (100 segments of length $n = 4000$ or 200 segments of length $n = 2000$) and the in-between

segments (four transition periods with gradual increase of Δ) had length 200.000 (50 segments of length $n = 4000$ or 100 segments of length $n = 2000$). The measures were computed from non-overlapping segments of lengths $n = 2000$ and 4000, from this realization that included five different states of the Mackey-Glass system and four transition periods.

Statistical analysis For the detection of the transition and the dynamical changes in the Mackey-Glass system, each case is determined by the measure profile comprised of the values of two states and the transition period between them. Then, the measure profile is divided into two samples, let them be A and B, of varying lengths (minimum length 20) and each time increasing the size of A by one and thereby decreasing B by one (step 1). AUC values are computed for all possible samples, in order to find when the best discrimination is detected, i.e which is the maximum AUC value and when this is obtained. A schematic representation of this process is shown in Fig.5.11.

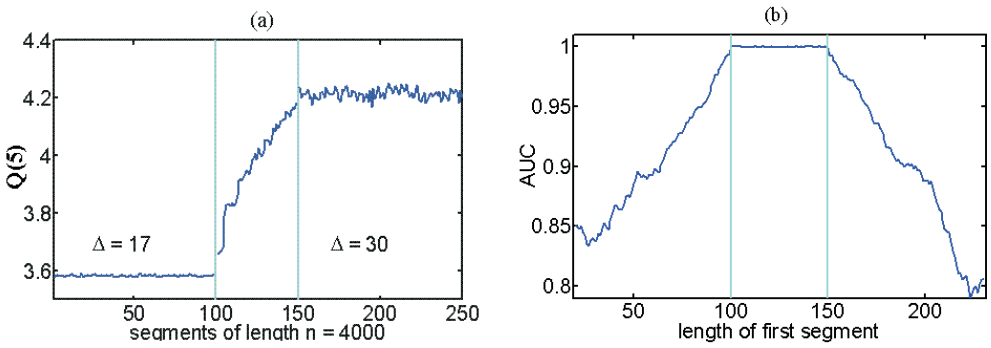

Figure 5.11: (a) Time series of values of cumulative autocorrelations $Q(5)$ for the Mackey-Glass system (segments of length $n = 4000$) for $\Delta = 17$ and 30 (two states) and $\Delta \in [18, 29]$ (transition period). (b) AUC values from all possible samples from the split of the measure profile of $Q(5)$ in two samples A and B.

Results

Results for the discrimination of the stochastic systems The discrimination of white noise from the rest of the stochastic systems is accomplished with some of the information measures, with optimal measure being the Shannon entropy estimated using the RQA method. The time series length is a determinant factor, as AUC values increase with n, however it does not substantially affect the ranking of the measures based on their discriminative ability. Table 5.3 presents the five optimal measures of this discrimination for $n = 128, 256, 2048$.

The transform of the time series to normal marginal distribution improves the discriminating ability of the measures, while the transform to uniform marginal

Table 5.3: The five optimal measures and the AUC values from the discrimination of the white noise from the other stochastic systems, for all $n = 128, 256$ and 2048.

$n = 128$		$n = 256$		$n = 2048$	
Measure	AUC	Measure	AUC	Measure	AUC
$ShEnt_{RQA}(2,1)$	0.84	$ShEnt_{RQA}(2,1)$	0.92	$SamEnt(2)$	0.99
$SamEnt(2)$	0.73	$SamEnt(2)$	0.84	$ShEnt_{RQA}(2,1)$	0.99
$M_{ED}(3)$	0.69	$cdI_{AD}(6)$	0.75	$M_{AD}(6)$	0.99
$cdI_{AD}(3)$	0.69	$cdI_{AD}(3)$	0.74	$M_{EP}(3)$	0.99
$Q(6)$	0.69	$TsEnt(1,3)$	0.74	$M_{AD}(3)$	0.98

Table 5.4: The five optimal measures for all n and the mean AUC (mAUC) values for the discrimination of the white noise from the other stochastic systems, for the original time series x_t, for the transformed y_t with normal marginal distribution and u_t with uniform marginal distribution.

x_t		y_t		u_t	
Measure	mAUC	Measure	mAUC	Measure	mAUC
$ShEnt_{RQA}(2,1)$	0.92	$ShEnt_{RQA}(2,1)$	0.94	$ShEnt_{RQA}(2,1)$	0.83
$SamEnt(2)$	0.86	$TsEnt(1,1.5)$	0.88	$TsEnt(1,1.5)$	0.81
$cdI_{AD}(3)$	0.80	$ShEnt(1)$	0.86	$ShEnt(1)$	0.81
$cdI_{AD}(6)$	0.80	$TsEnt(1,3)$	0.85	$Q(3)$	0.81
$M_{EP}(3)$	0.79	$SamEnt(2)$	0.84	$TsEnt(1,3)$	0.80

distribution has the opposite effect. The overall optimal measures for the discrimination of white noise for original time series and the transformed time series are presented in Table 5.4.

For the discrimination of AR(1) system from the rest of the stochastic systems, it is observed again a good discrimination using the information measures. The optimal measures, as expected, were the cumulative autocorrelations, as AR(1) was the only system with significant autocorrelations. Table 5.5 presents the five optimal measures of this discrimination for all n, for the original and the transformed time series.

The overall optimal measures and the respective mean AUC values for all the discriminations between the stochastic systems and for all n are presented in Table 5.6. $ShEnt_{RQA}(2,1)$ is the optimal measure for all time series (original and transforms). The transform to normal marginal distribution does not seem to substantially improve the discriminating ability of the measures, however the transform to uniform marginal distribution gives lower AUC values.

Results for the discrimination of dynamical systems with different complexity A high discrimination of the different states of the Mackey-Glass system for $\Delta = 17 - 30$ and $\Delta = 30 - 100$ was accomplished with almost all informa-

Table 5.5: As in Table 5.4 but for the discrimination of AR(1) from the other stochastic processes.

x_t		y_t		u_t	
Measure	mAUC	Measure	mAUC	Measure	mAUC
$Q(3)$	0.96	$Q(3)$	0.96	$Q(3)$	0.96
$Q(6)$	0.93	$Q(6)$	0.93	$Q(6)$	0.92
$M_{AD}(3)$	0.81	$cdI_{ED}(3)$	0.85	$M_{AD}(3)$	0.82
$ShEnt_{RQA}(2,1)$	0.81	$cdI_{ED}(6)$	0.83	$ShEnt(1)$	0.81
$cdI_{AD}(3)$	0.79	$M_{AD}(3)$	0.81	$TsEnt(1,1.5)$	0.81

Table 5.6: The overall optimal measures and the mean AUC values for all the discriminations and all n.

x_t		y_t		u_t	
Measure	mAUC	Measure	mAUC	Measure	mAUC
$ShEnt_{RQA}(2,1)$	0.87	$ShEnt_{RQA}(2,1)$	0.87	$ShEnt_{RQA}(2,1)$	0.83
$Q(3)$	0.83	$cdI_{ED}(3)$	0.82	$TsEnt(1,1.5)$	0.81
$Q(6)$	0.82	$Q(3)$	0.82	$ShEnt(1)$	0.81
$SamEnt(2)$	0.81	$Q(6)$	0.81	$Q(3)$	0.81
$cdI_{AD}(3)$	0.80	$ShEnt(1)$	0.81	$TsEnt(1,3)$	0.80

tion measures (AUC=1). The discrimination for $\Delta = 100 - 200$ was accomplished only for the cumulative measures for $\tau_{max} = 100$, dct_{EP}, dct_{AD} and $ShEnt_{RQA}(m,\tau)$, for $m = 10, 15, 20$. As system becomes more complex (for larger Δ values), the effect of the selection of the free parameters of each measure is substantial. A satisfactory discrimination for $\Delta = 200 - 300$ is achieved only with $ShEnt_{RQA}(m,\tau)$, for $m = 15, 20$. Indicatively, the measure profile of $Q(5)$ for all states of the Mackey-Glass system is presented in Fig.5.12. The discrimina-

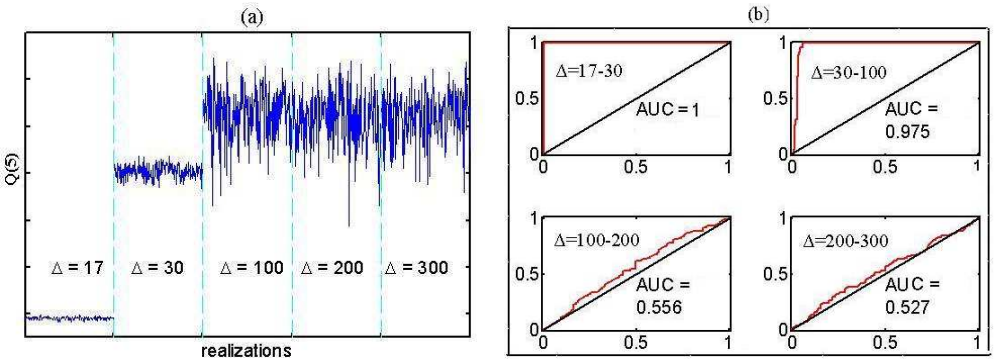

Figure 5.12: (a) The measure profile of $Q(5)$, for the five states of the Mackey-Glass systems as given in the graph. (b) The respective ROC curves from the discriminations of the Mackey-Glass systems.

Table 5.7: The five optimal measures and the corresponding AUC values from the discrimination of the Mackey-Glass system with $\Delta = 100$ and $\Delta = 200$ and the discrimination of the Mackey-Glass system with $\Delta = 200$ and $\Delta = 300$.

	$\Delta = 100$-$\Delta = 200$		$\Delta = 200$-$\Delta = 300$	
A.A.	Measures	AUC	Measures	AUC
1.	$cdl_{AD}(100)$	1	$ShEnt_{RQA}(15, 1)$	0.71
2.	$cdl_{EP}(100)$	1	$ShEnt_{RQA}(20, 1)$	0.71
3.	$cdl_{ED}(100)$	1	dct_{AD}	0.68
4.	$cI_{AD}(100)$	1	dct_{EP}	0.64
5.	$cdl_{KNN}(100)$	0.99	$cdl_{KNN}(100)$	0.88

tion of the first two states from the others is obvious just from the measure profile, however the discrimination of the three last states is not. From the discriminations of all states of Mackey-Glass system for all n, the optimal measures are found to be $M_{ED}(100)$, $M_{KNN(100)}$, $cdI_{AD}(100)$, $cdI_{ED}(100)$, $cdI_{KNN}(100)$ with mean AUC values equal to 0.88, for all the measures.

Results for the transition of systems dynamics Information measures detected also smooth dynamical changes in the Mackey-Glas system, while increasing Δ. The transition from $\Delta = 17$ to 30 and from $\Delta = 30$ to 100, was accomplished with almost all information measures (with AUC=1) and correctly within the transition period. The transition from $\Delta = 100$ to 200 was also accomplished with AUC=1, but only with the cumulative measures for $\tau_{max} = 100$. The transition from $\Delta = 200$ to 300 was accomplished also satisfactory only with the Shannon entropy (RQA) with $m = 15, 20$. In Table 5.7, the five five optimal measures and the corresponding AUC values from the discrimination of the Mackey-Glass systems with $\Delta = 100$ and $\Delta = 200$ and the discrimination of the Mackey-Glass systems with $\Delta = 200$ and $\Delta = 300$ are presented.

Conclusions

Information measures turned out to be useful in discriminating between different dynamical systems and are able to detect smooth changes of dynamical systems. The time series length does not affect the ranking of the measures as far as their discriminating ability is concerned. As expected, for systems with high complexity, the selection of a suitable measure is quite difficult and the same stands for the selection of the free parameters of each measure. The transform of the original time series to have normal marginal distribution is particularly useful for the discrimination of white noise from other systems. However, the transform to uniform marginal distribution seem to deteriorate the discriminating performance of the measures.

Chapter 6

Applications of Univariate Information Measures on EEG

Babloyantz and Destexhe (1986) were the first to report changes in the scalp EEG of epileptic patients using nonlinear methods (fractal analysis and Lyapunov exponents). A decrease in the fractal dimension during the seizure was detected. Their study and the general trend of applying nonlinear methods to time series has triggered the onset of works based on EEG analysis. Although the first studies on EEG had shown diverging results, however studies on epilepsy around 90's applying nonlinear methods have shown promising results.

Time series analysis has been applied in EEG data for the detection of epileptic seizures (Lehnertz et al., 2000; Hirsch et al., 2006) and for the detection of the epileptic focus area (Gersch and Goddard, 1970; Andrzejak et al., 1999). Both linear and nonlinear measures have been used for this reason, e.g. autoregressive modelling (Rogowski et al., 1981), Lyapunov exponents (Pardalos et al., 2003), entropy (Quiroga et al., 2000a) and synchronization measures (Le Van Quyen et al., 2005). Recently, nonlinear methods of time series analysis, some of them with high computational complexity, have been proved to be more useful than linear methods in the problem of epileptic seizure prediction (Lehnertz et al., 2000; Iasemidis et al., 2001). However, the advantage of nonlinear measures has been questioned (McSharry et al., 2003).

Many studies estimated the changes in the dynamics of the brain during the normal functioning (interictal state) and the preictal state. However, there are no consistent statistical indications that allow the use of these methods in clinical practice yet; there is still a long way to go before a statistical significance for the predictability of any measure in terms of its sensitivity and specificity can be established (Mormann et al., 2006; Gudmundsson et al., 2007).

Here, univariate information measures are evaluated for their ability to predict the short-term onset of a seizure by examination of the preictal EEG. Measures are also used in order to investigate their usefulness in discriminating different states of the epileptic brain, aiming to detect the onset of an epileptic seizure.

6.1 Short-term Prediction of Epileptic Seizures from Preictal EEG Records and Discrimination of Different Brain States using Statistical Tests

The first aim of this study is to investigate the existence of any trends in the values of the measures (measure profiles) estimated from consecutive segments of the preictal EEG. The detection of such a trend can be considered as an indication of changes in the brain dynamics of the patient before the seizure onset. For this problem, two straightforward comparable information measures are used here; the cumulative autocorrelation function $Q(\tau_{max})$ and the cumulative mutual information function $M(\tau_{max})$, both estimated for the same lags. For the detection of the nonlinear correlations, the p-values from the surrogate data test for nonlinearity were also used with test statistics the mutual information $I(\tau)$ and $M(\tau_{max})$.

The second aim of this study is to discriminate between the EEG records from the early and late preictal state, using the afore mentioned measures. Here, a late preictal state about two minutes before the seizure onset and an early preictal state from one or more hours before the seizure onset have been used. For the detection of any changes in the correlation structure of the EEG time series at different brain states, statistical tests for the means and the medians have been applied on the measure profiles from the different states, i.e. the discrimination between the early and late preictal state was assessed by comparison tests of the measure profiles (t-test for means and the Wilcoxon rank sum test for medians). ROC curves could not be used as the distributions of the measure profiles could not be reliably estimated due to small sample sizes.

The study showed that a short-term prediction of epileptic seizures from the late preictal EEG of a patient can be accomplished using the information measures, although results varied with the channel and the patient. The information measures were also able to discriminate between the two states (early and late preictal), especially when the early preictal records were from many hours before the seizure onset. Again, the performance of each measure varied with the patient and the channel (P3).

EEG data

Multichannel extracranial EEG from 8 patients were used for the evaluation of the information measures. For all the patients, late preictal records from about one or two minutes before the seizure onset are examined (for two patients, two preictal records exist), and earlier preictal records from one to many hours before the seizure were examined for six patients. For the early preictal states larger than two minutes records could be considered, however the duration of the records was restrained to two minutes to coincide with those from the late preictal state. A 25 channel system was used for the six patients and a 63 channel system for the other two patients. The notations of the records for each patient are given in Table 6.1.

Table 6.1: Notations of EEG records from late preictal and earlier preictal states for the patients. The indexes s and l denote records from 1 hour before the seizure onset and from many hours before seizure, respectively.

Patient	late preictal	early preictal
25 channels		
1.	A	-
2.	B	B_s, B_l
3.	C	C_s
4.	D	D_l
5.	E_1, E_2	E_l
6.	F	F_s, F_l
63 channels		
7.	G	G_l
8.	H	-

Measures

The measures of linear correlation that were estimated are the autocorrelation function $r(\tau)$ for lags $\tau = 1, 5, 10, 20, 30$, and the cumulative autocorrelation function $Q(\tau_{max})$ for $\tau_{max} = 40$. The measures of nonlinear correlation that were estimated are the mutual information $I(\tau)$ for the same lags ($\tau = 1, 5, 10, 20, 30$), and the cumulative mutual information $M(\tau_{max})$ for $\tau_{max} = 40$. As measures of nonlinear deviation from linearity, the p-values from the surrogate data test for nonlinearity are used, with test statistics the previous nonlinear measures, i.e. $I(1)$, $I(5)$, $I(10)$, $I(20)$, $I(30)$ and $M(40)$. The maximum lag τ_{max} for the estimation of the cumulative measures is set to be 40, as it was observed that $I(\tau)$ of the EEG records tend to reach the 'zero-correlation area' around this lag.

Set Up

For each record and channel, time series of 30sec are generated from successively overlapping segments of the preictal states, with a time step of 15sec. A similar method for the analysis of the EEG can be found in Kugiumtzis and Larsson (2000). The measures are estimated from the overlapping time series and new time series are generated with the values of the measures from these overlapping segments (measure profiles).

The measure profiles from the late preictal states are examined in order to investigate the existence of a trend for the short-term prediction of the seizure. Therefore, for the examination of any changes in the correlation structure of the patients EEG, measures profiles are examined for possible increase or decrease of their values as seizure comes. As an example, Fig.6.1a and b show the EEG time series from patient A from two channels, for the first and the sixth overlapping segment of the record. No conclusion can be drawn for the changes of the dynamics of the

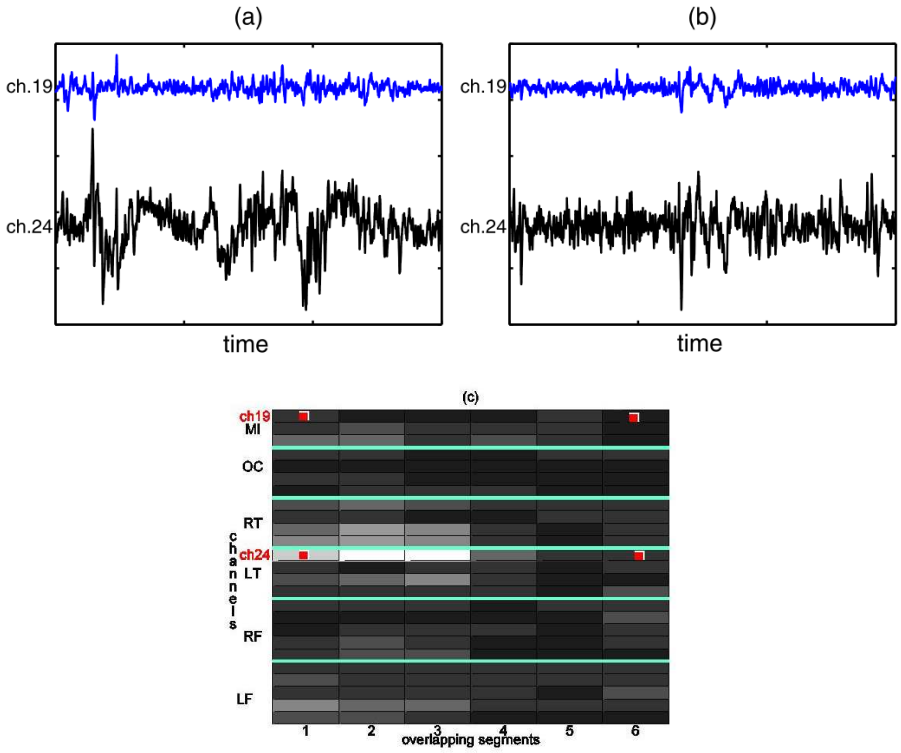

Figure 6.1: (a) EEG time series from patient A from channel 19 (MI) and channel 24 (LT) from the first overlapping segment. (b) As in (a) but for the last overlapping segment. (c) $I(10)$ values from measure profiles of the preictal state of patient A in a gray scale (black for the lowest and white for the highest value).

brain just from the EEG signal. However, by examination of the measure profiles, changes in the values of the measures may be observed. In Fig.6.1c, the change of $I(10)$ values are presented in a gray scale (black for the lowest and white for the highest value). A decrease in $I(10)$ indicates an increase of the stochasticity or complexity of the brain.

Based on this pilot study of the effectiveness of the measures, a more thorough examination is applied in order to ascertain the existence of a trend in the measure profiles extracted from the late preictal state. A linear regression is used for the values of each measure against the overlapping segments in order to investigate the null hypotheses H_0 that there is no trend. The test statistic was defined in Eq.(3.13). The corresponding p-values of the test statistic T indicate the rejection or not of H_0. If $p < 0.05$ then H_0 is rejected and that means that there is a trend and thus there is an indication for the onset of the seizure. An example of the detection of positive and negative trends or no trend detection from the measure profiles of patient A are displayed in Fig.6.2.

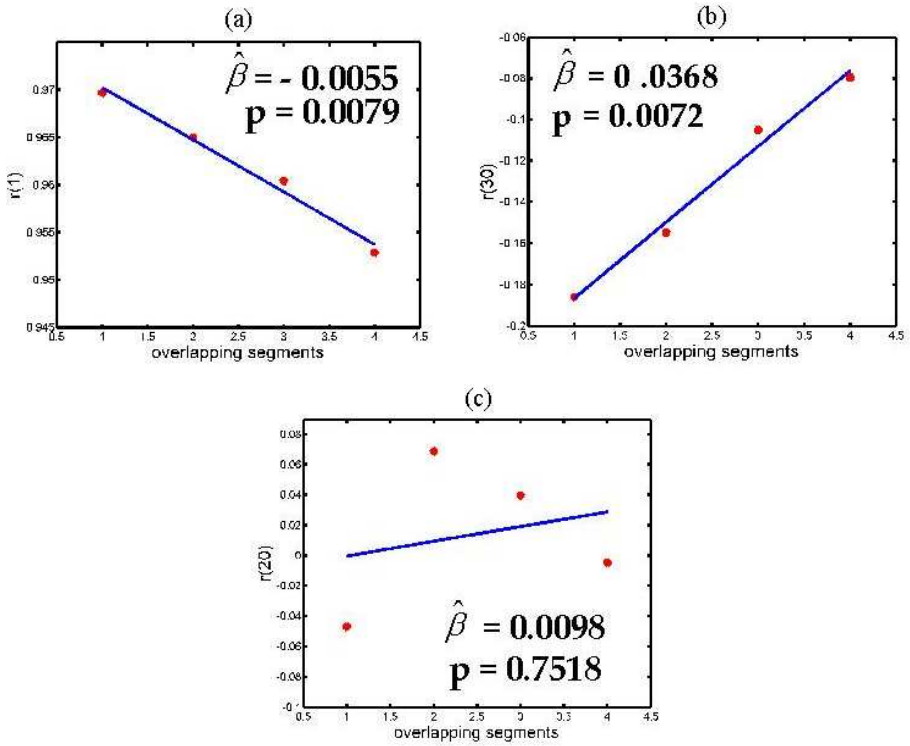

Figure 6.2: (a) Negative trend of $r(1)$ values from the first 4 overlapping segments of the preictal record of patient A, from channel 1. (b) As in (a) but for the positive trend of $r(30)$ values, from channel 2. (c) As in (a), but for $r(20)$; no trend is detected.

6.1.1 Results for the short-term trend detection on late preictal EEG

The examination of the late preictal measure profiles for the trend detection is assessed considering two factors; the channels (for the different brain areas) and the measures. The results of the trend detection are represented qualitatively in graphs where x-axis shows the channels and y-axis the measures. The trend result is denoted at each cell of the graph with the gray color indicating negative trend, black indicating positive trend and white indicating that there is no statistically significant trend ($p > 0.05$). If a segment has many artifacts and cannot be used in the analysis, a cross (X) is displayed in the graph. In Fig.6.3, results from trend detection from preictal records of all patients with 25 channels are shown. The existence of a trend is not obvious in general, but only for certain channels and measures for each patient. Only for patient A there seems to exist a significant negative trend, i.e. decrease of values of the measures in many channels. The same results are observed from the examination of the preictal records from the patients with 63 channels.

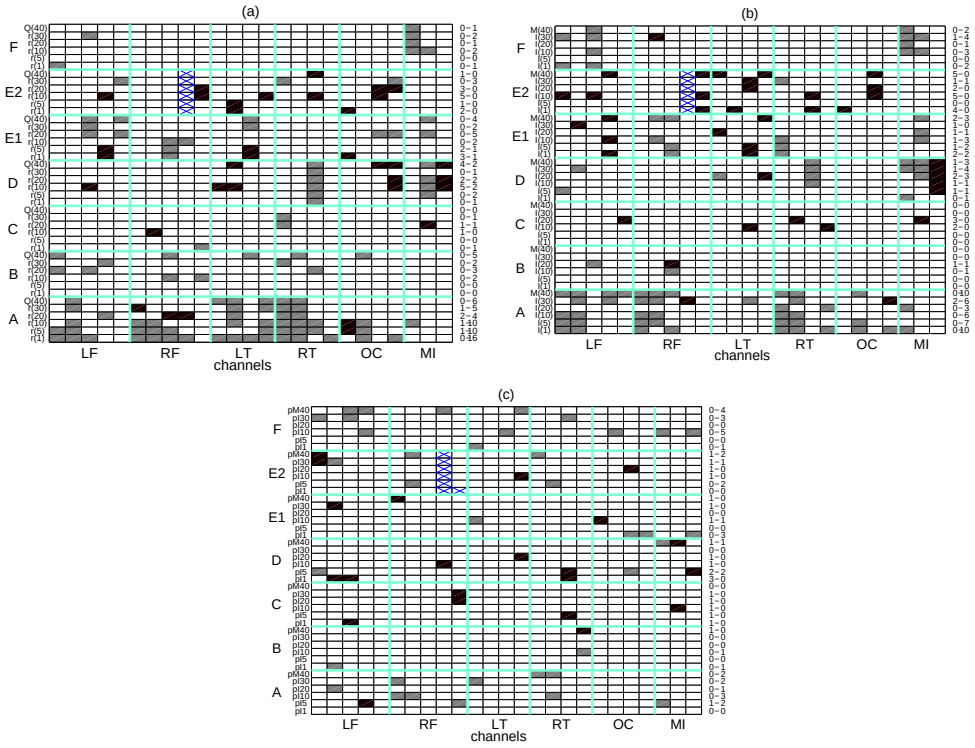

Figure 6.3: Results of the existence of trend from preictal records from the six patients with 25 channels. In (a) are the results for the linear measures $r(\tau)$ and $Q(40)$, in (b) for the nonlinear measures $I(\tau)$ and $M(40)$ and in (c) for the p-values $p_I(\tau)$ and $p_M(40)$. In the right side of the graph, the sums of positive-negative trends are given for each patient.

Conclusions

From the above results, it seems that the detection of trends in the measure profiles of the information measures is accomplished only for certain measures and channels for each patient. In most cases a trend was detected with a correlation measure, it was found to be negative, indicating the loss of correlation just before the seizure onset.

6.1.2 Results for the discrimination between early and late preictal states

The factors that are examined are again the channels, the measures and their discriminating ability, i.e. their ability to discriminate between the early and late preictal states. For the patients with 63 channels, there are two records from the late preictal state and one record from an earlier preictal state, many hours before the seizure (patient G). The discrimination between the states with the t-test is clear for

most of the channels and measures, but for the p-values. In most cases, the measure profiles from the last 2 min before the seizure onset are greater than from 2 min data segments before the seizure. The results from this discriminations are displayed in Fig.6.4. In the x-axis of the plot are the channels and in the y-axis are the measures, grouped for each comparison. Most measures (excluding p-values) in-

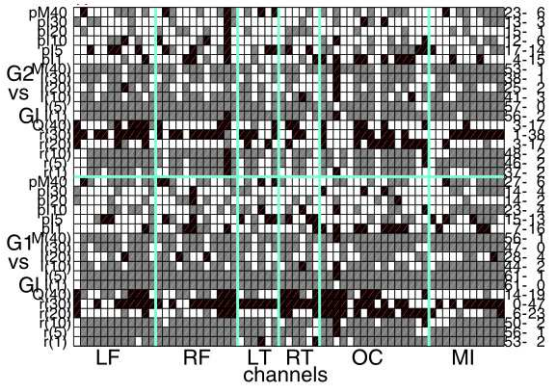

Figure 6.4: Discrimination of late preictal and early preictal states for patient G. The gray color is set when the values of a measure are greater for the late preictal state than the the values from the early preictal state, the black is used for the opposite case and white indicates that there is no discrimination between the two states. In the right side of the plots, the sums of gray - black cells are given.

dicate a significant discriminating ability. However, there is no systematic increase or decrease of the values, although in most cases a decrease in the correlations of the brain is observed.

For the patients with 25 channels, the discrimination between the two states is clear. For these records, there is also a clear discrimination even with the p-values at two out of the five cases (see Fig.6.5). For most patients (but for F) the values of the correlation measures from the late preictal states are higher, which indicates that brain functioning away from the seizure is more complicated (or stochastic) than close to the seizure.

The discrimination between the two states for the early preictal records from only 1 hour before the seizure onset, as expected, is not as clear as for records of about 4 hours prior to seizure onset (see Fig.6.6). Again, most measures (excluding p-values) indicate a significant discriminating ability, however there is no systematic increase or decrease of the values. In most cases, correlation measures decrease. In some patients, it seems that the discrimination is succeeded only at certain brain areas, e.g. in patient F, for brain areas LF and RF almost all nonlinear measures discriminate the two states. At other cases the selection of the measure is more important, e.g. measure $I(5)$ in patient F discriminates the two states for most brain areas.

For the same records, the Wilcoxon rank sum test was applied in order to com-

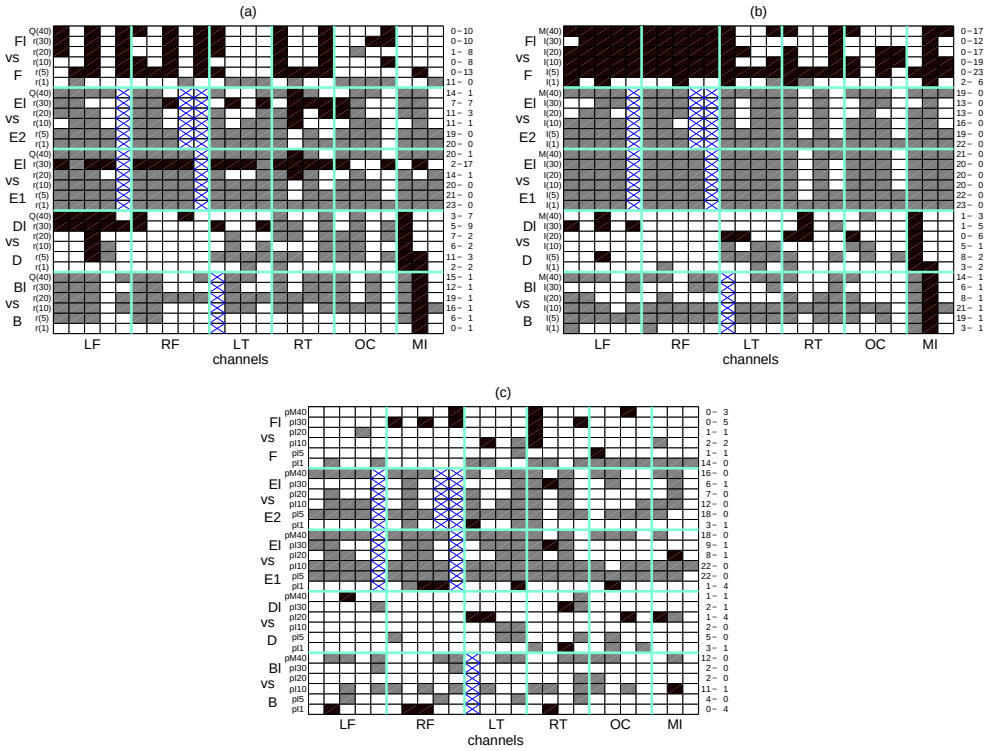

Figure 6.5: As in Fig.6.4 but for the patients with 25 channels. (a) Discrimination of the late preictal and the early preictal states with $r(\tau)$ and $Q(\tau_{max})$ and comparisons of late preictal to about four hours prior to seizure onset. (b) and (c) as in (a) but for test statistics $I(\tau)$, $M(\tau_{max})$ and the p-values, respectively.

pare the results from those from the t-test. It was observed that there were no significant differences in the results. As an example, results from the discrimination of the two states for patient G are shown in Fig.6.7.

Conclusions

The information measures were able to discriminate between the two states, late preictal and early preictal, especially when the early preictal records were from many hours before the seizure. The discrimination of the two states was statistically significant for most patients and at large proportion of channels and measures. However, results differentiate with the patient, channel and measure. The tests for the mean or median of the measure profiles gave equivalent results. The study has not found any systematic superiority of the nonlinear measures concerning their discriminating power over the linear ones.

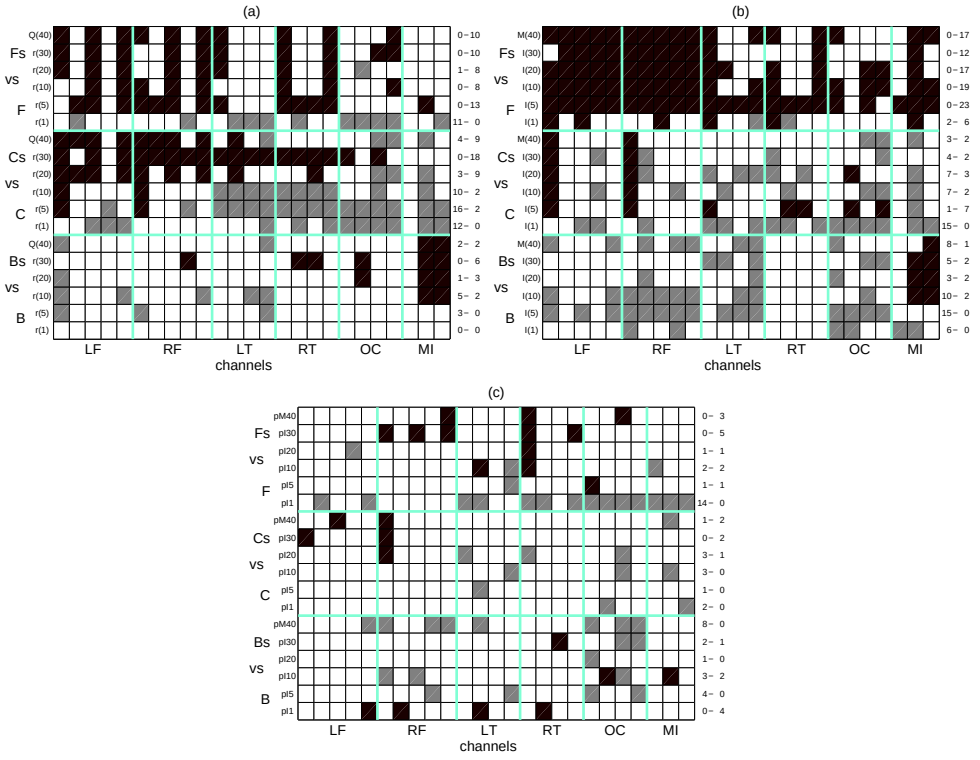

Figure 6.6: As in Fig.6.5 but for comparisons of late preictal to about one hour prior to seizure onset.

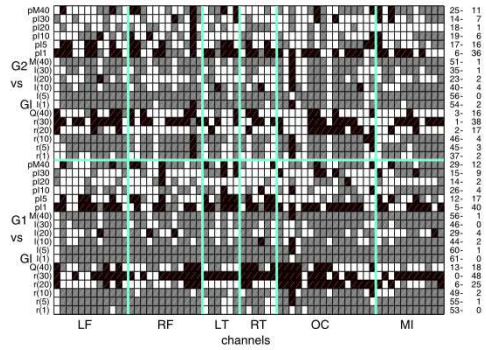

Figure 6.7: As in Fig.6.4 but the discrimination is assessed with the Wilcoxon rank sum test.

6.2 Evaluation of a large set of Correlation and Entropy Measures in Discriminating Preictal States using Statistical Tests

This study evaluates the discriminating power of a large set of correlation measures measures. The detection of dynamical changes in the preictal states is investigated,

at the conditions of typical clinical practice, where intracranial and scalp EEG recordings are delivered without preprocessing. The study concentrates again on univariate correlation measures that have been used in the prediction of seizures but also the correlation measures that have been defined in the previous chapter. EEG time series are transformed here to have normal marginal distribution and some measures are estimated from the modified time series instead of the original EEG time series. The hypothesis tests applied for each channel and epoch showed good discrimination of the preictal states and allowed for the selection of optimal measures. The results that are presented here are a part of a more general work (P2); in cooperation with the other co-writers, a majority of measures was included in the evaluation, e.g. measures based on the features of the time series (minima and maxima of the time series).

Set Up

Four EEG epochs were used for this study; three scalp EEG recordings with 25 channels based on a 10–20 system, referred to as A,B and C, and one intracranial EEG recording with 28 channels, referred to as D. All EEG data were band-passed at 0.5–70 Hz and sub-sampled at 100 Hz.

Data windows of 10 min duration from early, intermediate and late preictal states were used, corresponding to periods of approximately 4h, 1h and 10min before the seizure onset and denoted as e, i, and l, respectively. Each 10min long data window was split to 20 successive segments of 30 s and the measures were estimated for each of them. For epoch A, larger data windows of 50 min of state e and the last 60 min of the preictal state (merging states i and l) were considered, segmented in the same way.

Measures

Let $\{x_t\}$ be a segment of the original EEG time series and $\{y_t\}$ the corresponding transformed time series with normal marginal distribution (as defined in Eq.(5.4)). The nonlinear decorrelation time, dct, was computed from the original time series. The autocorrelation function was estimated for both original and transformed time series, for lags $\tau = 1, 5, 10, 20, 30$, and let us denote them as $r_x(\tau)$ and $r_y(\tau)$. Similarly, the mutual information, $I_x(\tau)$ and $I_y(\tau)$, was also estimated for both time series at lags $\tau = 1, 5, 10, 20, 30$. The cumulative autocorrelation functions, $Q_x(dct)$ and $Q_y(dct)$, and the cumulative mutual information functions, $I_x(\tau_{max})$ and $I_y(\tau_{max})$, were estimated for $\tau_{max} = dct$ (nonlinear decorrelation time). The measure of declination from normality (defined in Section 5.1.3), $dI_y(\tau)$, was estimated only from the transformed time series $\{y_t\}$ at lags $\tau = 1, 5, 10, 20, 30$, and the cumulative declination from normality, $cdI_y(\tau)$, was estimated for $\tau_{max} = dct$. Moreover, as dct was estimated from all the overlapping segments from each epoch, the mean of dct over all EEG segments of each epoch, denoted by $\langle dct \rangle$, was also estimated. All the cumulative measures were estimated also at $\tau_{max} = \langle dct \rangle$.

Statistical evaluation

The Student test for means (equal variances not assumed) as well as the Wilcoxon rank sum test for medians were applied to the three pairs of preictal states. In order to derive the discriminating power of each measure over all channels, a score s_q was assigned to each measure q for each preictal state pair comparison and epoch as follows. Let p_{qj}, $q = 1, \ldots, n$ and $j = 1, \ldots, m$ be the p-values of the test for two preictal states of one epoch for all n measures and m channels. It is set $p_{qj} = 0$ if $p_{qj} > \alpha$ for a predefined significance level α. For each channel j the non-zero p-values were sorted at decreasing order and a score s_{qj} was set for each q equal to its rank r_{qj}, i.e. if there are n_j non-zero entries for the channel j the measure q with the lowest p-value gets the largest score for this channel, equal to n_j. The score of each measure q at each channel j is thus defined as

$$s_{qj} = \begin{cases} r_{qj} & : \quad p_{qj} < \alpha \\ s_{qj} = 0 & : \quad else \end{cases} \tag{6.1}$$

Then, scores of each measure q were averaged over all channels

$$s_q = \frac{1}{m} \sum_{j=1}^{m} s_{qj}. \tag{6.2}$$

In this way, the score of the performance of each measure in each channel is balanced by the discriminating power of the other measures in the channel. So, a channel that appears to be less relevant to the preictal activity will have less effect on the score of the measures. Finally, the average performance score S_q for each measure is defined as the average of the scores s_q over all three preictal state comparisons and all four epochs.

Results

The correlation measures were estimated on the multi-channel EEG segments at each of the three preictal states (e, i, l) and for each of the four epochs (A, B, C, D). Autocorrelations of the measure profiles were also estimated in order to investiagate the correlations of their terms, however no evidence of serial dependence was found. The variance and the shape of the distribution for each measure changed a lot across the sample sets, also due to the existence of the artifacts that were not removed. The hypothesis tests for the means and medians on the measure profiles from each group were assessed in order to discriminate the three preictal states.

The Student test and the Wilcoxon rank sum test gave similar results, but there was large variation of the test results across preictal state comparisons, channels, measures, and epochs, as expected. In Fig.6.8, the Student test results are shown for all three pair comparisons with all measures, for the epoch B. For this epoch it seems that the late preictal state is distinguished well and with many measures whereas the early and intermediate preictal states do not bear a clear difference.

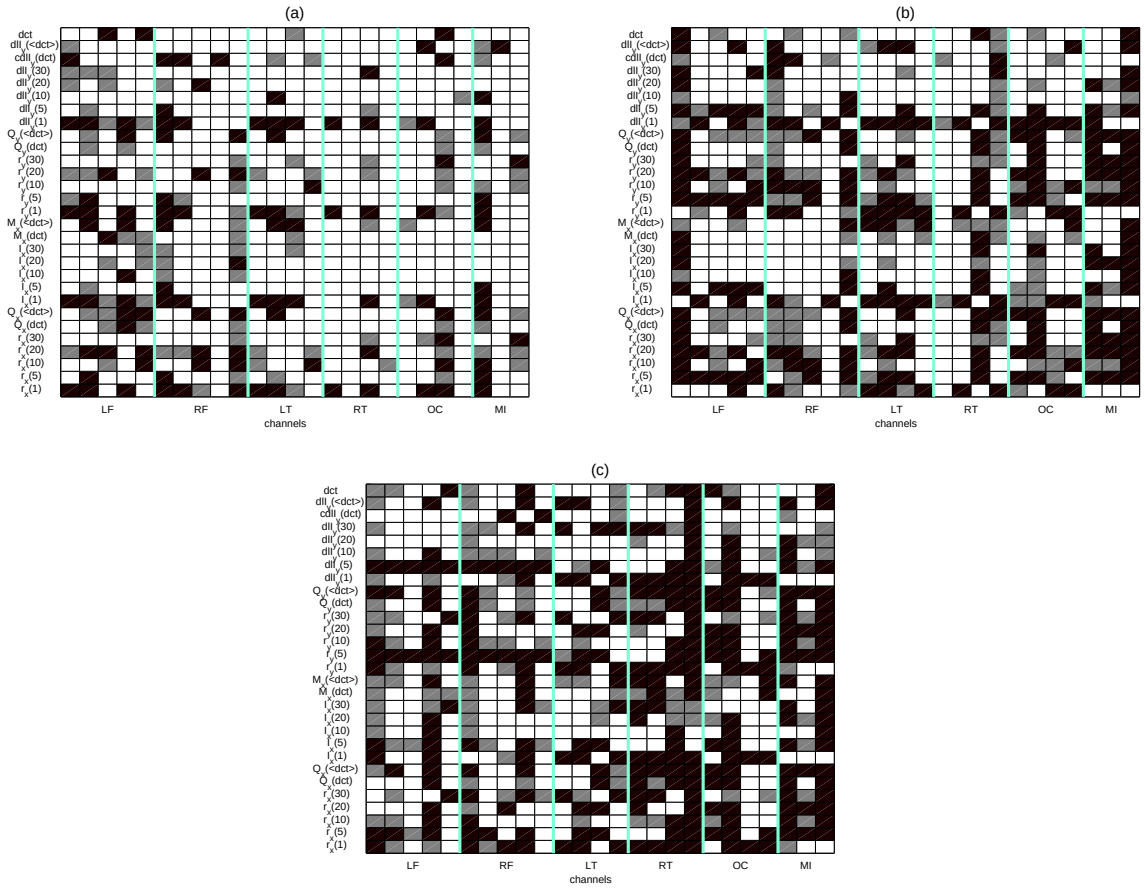

Figure 6.8: The Student test results for epoch B based on the p-values: white cells when $p \geq 0.5$, grey cells when $0.01 < p < 0.05$, and black cells when $p < 0.01$. The channels are organized in brain areas as left and right frontal (LF and RF), left and right temporal (LT and RT), occipital (OC) and middle (MI). The correlation measures are shown for the e–i comparison in (a), for the e–l comparison in (b), and for the i–l comparison in (c).

This can be observed mostly in the right temporal, occipital and middle brain areas. Moreover, some measures seem to detect better and more consistently the differences across channels.

The performance of the measures varied in the other three epochs, so that conclusive results could not be obtained by simple eye-ball judgement. Therefore, the summary results of the average performance score S_q are reported. The 10 best scores S_q from the p-values of the Student test in the set of the 29 correlation measures, are shown in Table 6.2. The autocorrelation measures outperformed the mutual information measures, with the 'normal' autocorrelation measures (estimated from the transformed time series $\{y_t\}$) scoring always better than the respective

Table 6.2: The 10 best measures and their scores S_q in the set of correlation measures taken over all pairs of states and epochs.

A.A	Measure	S_q
1.	$r_y(5)$	76.8
2.	$Q_y(\langle dct \rangle)$	72.0
3.	$r_x(5)$	69.6
4.	$r_y(10)$	66.1
5.	$Q_x(\langle dct \rangle)$	65.6
6.	$r_x(10)$	59.9
7.	$r_y(20)$	57.5
8.	$r_x(20)$	50.2
9.	$r_y(30)$	47.6
10.	$dlI_y(5)$	47.0

autocorrelation measures on the original time series.

The performance of the measures appeared to be dependent on the choice of the time window in each preictal state. To investigate this effect, the same analysis was made on 5 data windows of 10min each for e ($\pm$ 20 min to the initial selected 10 min data window), and 5 data windows of 10 min each for i (from 60 min to 10 min prior to seizure onset) from epoch A. The measures were estimated on successive segments of 30 s duration and the tests were applied on the measure profiles as before. The measure profiles varied across channels as shown in Fig.6.9 for two adjacent channels in the middle brain area for the measure $r_y(5)$.

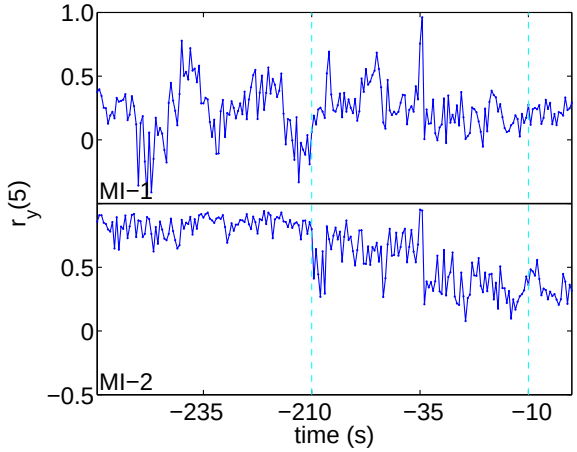

Figure 6.9: Measure profiles for two channels in the middle area of the brain (upper and lower panel) over an early preictal state ([-260,-210] min with respect to seizure onset) followed by an intermediate and late preictal state ([-60,-10] min and [-10,0] min), separated by vertical lines. Results are displayed for measure $r_y(5)$.

Table 6.3: The 10 best measures and their scores S_q in the set of correlation measures taken over all pairs of states and epochs.

A.A	Measure	S_q
1.	$r_y(10)$	76.8
2.	$r_y(5)$	72.0
3.	$r_x(10)$	69.6
4.	$Q_y(\langle dct \rangle)$	66.1
5.	$r_x(5)$	65.6
6.	$I_x(1)$	59.9
7.	$r_y(1)$	57.5
8.	$r_y(20)$	50.2
9.	$dlI_y(10)$	47.6
10.	$Q_x(\langle dct \rangle)$	47.0

The Student test was applied with all measures on all 25 combinations of the 5 data windows in e and i for the A epoch. The average performance scores S_q were computed on basis of the 25 cases and the best 10 correlation measures are shown in Table 6.3. The results are in agreement to those in Table 6.2. The 'normal' autocorrelation measures are still the best among the correlation measures (with $r_y(10)$ now being the optimal measure). Here, also the mutual information for lag 1 is among the best correlation measures. The analysis on the selected data windows from the e,i and l states from the 4 epochs and on several data windows from the e and i states from one epoch suggests that the measures with the best discriminating power are the 'normal' autocorrelation at small lags as well as the cumulative 'normal' autocorrelation.

Conclusions

The evaluation of the correlation measures in detecting statistically significant differences between early, intermediate and late preictal states, showed that simple and computationally effective measures, such as the 'normal' autocorrelation function, perform better than other well-known and computationally intensive measures, such as the mutual information. Certainly, the evaluation of the measures is not complete and comparison to other measures that are reported to discriminate preictal states is missing here. The analysis was done on small data windows prone to artifacts that could affect the results on the measure evaluation. However, when the same analysis was done on larger data windows for one epoch the results were about the same. An automatic scoring approach was developed for the evaluation of the measures that downweighs the effect of channels when they do not exhibit changes in the measure values across the preictal states. Alternative scoring schemes were also tested and they all tended to give the same results. The scores from the Student tests showed a rather consistent tendency as to the performance

of the measures ranking the 'normal' autocorrelation for lag 5 first and then the cumulative 'normal' autocorrelation.

6.3 Evaluation of Information and Complexity Measures in Discriminating States using ROC Curves

6.3.1 A pilot study

The previous study was generalized to include all the previously introduced information measures and a few complexity measures. Moreover, different estimators of mutual information were also evaluated in this application with real data. The aim of the study was again the detection of different dynamics in the brain activity of an epileptic patient in order to discriminate different states (late preictal and early preictal). For the statistical evaluation of the measures, the ROC curves were used. Most of the information measures turned out to effectively discriminate the different dynamical states of the brain (P9).

EEG data

The extracranial EEG records from patient G of Table 6.1 were used and the records from each state had one hour duration; the late preictal record was measured from one hour before the seizure up to the seizure onset and the early preictal record was from four up to three hours before the seizure onset. For the discrimination of the two states, measurements from three channels are used; from the middle (MI), left (LF) and right frontal (RF) part of the brain. In Fig.6.10, EEG time series from the three brain areas from both states are presented.

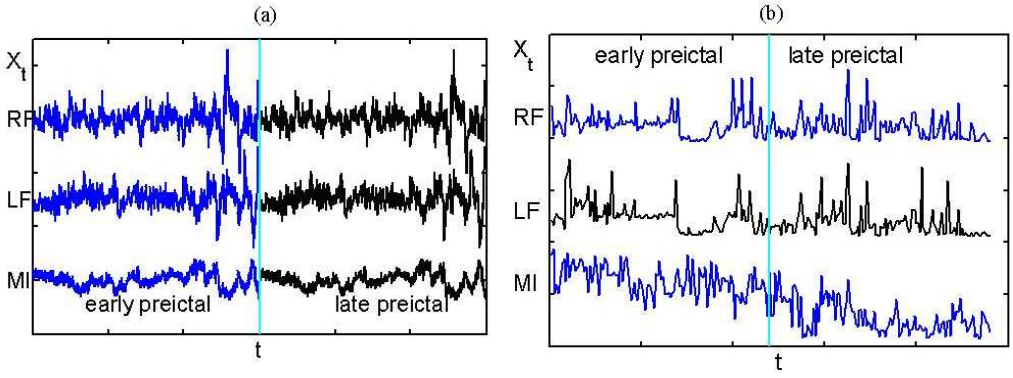

Figure 6.10: (a) EEG time series of 30sec from the two states from the three channels. (b) Time series of cumulative autocorrelation $cr(100)$ for both states and the three channels.

Set Up

Here, EEG records from each state and channel are used in order to generate consecutive non-overlapping time series of 30sec. Measures are estimated from the consecutive non-overlapping segments from both states for the three channels. The reason that the segments are not overlapping allows the statistical evaluation of the measures by the ROC curves, as the measure profiles can be considered as samples with independent terms.

Measures

The correlation measures that were used in this study are ts (linear decorrelation time), dct (nonlinear decorrelation time), $r(\tau)$ and $I(\tau)$ for lags $\tau = 1, \ldots, \tau_{max}$, and the cumulative measures $Q(\tau_{max})$, $M(\tau_{max})$ and $cdI(\tau_{max})$. MI was estimated with ED, EP, AD and KNN estimators. Shannon entropy was estimated using the binning scheme with equidistant cells for the two dimensional variable $(X_i, X_{i-\tau})$ for lags $\tau = 1, \ldots, \tau_{max}$ and with the method of RQA for $m = 2, 5, 10, 15, 20$. Tsallis entropy is estimated for $q = 1.5$ and 3 from the two dimensional variable $(X_i, X_{i-\tau})$ for lags $\tau = 1, \ldots, \tau_{max}$ with the equidistant binning scheme, sample entropy is estimated for $m = 2$, and the permutation entropy is estimated for $m = 5, 6, 7$.

Results

First of all, the selection of the channel seems to be crucial. For the channel 18 from the middle part of the brain, it is observed that almost all measures discriminate between the two states. The ten optimal measures based on the AUC values are given in Table 6.4. For the other two channels though, measures do not seem to have a high discrimination performance as AUC values are very low (< 0.75).

Conclusions

Information measures can be used for the discrimination of different states in epilepsy. However, the selection of the channels is crucial and also the selection of the parameters of the measures, something that remains open for further investigation. Linear and nonlinear measures seem to perform equally well, independently of the selection of the channel.

6.3.2 A large scale study

Over the last years the measures that have been used for seizure prediction have increased but only few works focused on evaluating and comparing a variety of different measures, as in Mormann et al. (2005). Therefore, there is an obvious need to review and validate a large number of proposed measures. Along these lines, a first work of our group tested a number of different measures of scalar

Table 6.4: Optimal measures for the discrimination between late and early preictal states, based on AUC values.

	Channel 18		Channel 20		Channel 21	
A.A	Measure	AUC	Measure	AUC	Measure	AUC
1	$Q(100)$	0.947	$Q(5)$	0.711	$SamEn(2)$	0.676
2	$Q(50)$	0.940	$Q(10)$	0.708	$M_{ED}(5)$	0.661
3	dct_{EP}	0.935	$SamEn(2)$	0.707	$Q(5)$	0.655
4	dct_{KNN}	0.926	$Q(20)$	0.701	$M_{ED}(10)$	0.651
5	$M_{ED}(100)$	0.925	$M_{AD}(20)$	0.700	$M_{ED}(20)$	0.646
6	$M_{ED}(50)$	0.924	$M_{EP}(20)$	0.711	$M_{EP}(5)$	0.645
7	$Q(20)$	0.921	ts	0.708	$Q(10)$	0.643
8	ts	0.919	$M_{KNN}(20)$	0.707	ts	0.642
9	$M_{AD}(100)$	0.919	$M_{ED}(20)$	0.701	$Q(20)$	0.641
10	$M_{EP}(100)$	0.918	$M_{AD}(10)$	0.700	$M_{AD}(5)$	0.639

time series analysis in discriminating early and late preictal stages (P2). In order to have a more complete and exhaustive evaluation of the measures, this study was extended to include also complexity measures that have been used in the univariate EEG analysis (P5). The scalp multi-channel EEG recordings from 7 patients were used; the EEG measurements were of at least 3 h duration and up to the seizure onset. The interest was again in the power of the measures in discriminating the late preictal state (minutes before seizure onset) to earlier preictal states (hours before seizure onset). For this, the ROC curves were used and the measures were ranked according to their mean AUC values from all the discriminations.

EEG data

The EEG data that were used, comprised of 7 multichannel scalp EEG records, each regarding a single episode of a generalized tonic clonic seizure from each patient, covering at least 3 h prior to the seizure onset. The EEG recording system of the two first patients (denoted A and B) comprised of 63 channels, while of the other 5 patients (denoted from C to G) comprised of 25 channel. The frontopolar channels were excluded, bringing down the number of channels to 19 and 53, respectively.

Measures

The evaluation included both information and complexity measures. It was observed in (P2), that correlation measures would perform discrimination tasks on preictal EEG better when the data were 'gaussianized' (transformed to have normal marginal distribution). Therefore, many of the measures were computed on the transformed time series $\{y_t\}$, instead of the original EEG time series $\{x_t\}$. Specifically, only the nonlinear decorrelation time, dct, was estimated from $\{x_t\}$.

From $\{y_t\}$ were estimated the linear decorrelation time t_e, the autocorrelation function $r_y(\tau)$ and mutual information $I_y(\tau)$ and their respective cumulative measures $Q_y(\tau_{max})$ and $M_y(\tau_{max})$. MI and cumulative MI were used in order to quantify both linear and nonlinear dependencies for the same lags as above, where the estimation was based on histogram of 16 equidistant bins (Chillemi et al., 2003; Nicolaou and Nasuto, 2005). In an attempt to include measures of solely nonlinear correlations, declination from normality measures, $dI(\tau)$ and $cdI(\tau_{max})$, were also computed from $\{y_t\}$.

The additional correlation measures, estimated from $\{y_t\}$, that were used in this study are the following:

- Autocorrelation computed for each lag τ by the Spearman's rank correlation coefficient, denoted bt r_y^S (Spearman, 1904). The Spearman's rank correlation coefficient is defined as

$$r_y^S = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}, \tag{6.3}$$

  where d_i is the difference between the ranks of the corresponding values of X and Y, and n is the number of values of the data set.

- Cumulative Spearman's autocorrelation, Cr_y^S, defined as the sum of r_y^S for lags $1, \ldots, \tau_{max}$.

- Autocorrelation computed for each lag τ by the Kendal's τ rank correlation coefficient of $(X_t, X_{t-\tau})$, denoted by r_y^K, which is a measure used to quantify the degree of correspondence between two rankings and assess the significance of this correspondence (Kendall, 1938). It can take values between -1 and 1. For perfect agreement between the two rankings the coefficient has value 1 and for perfect disagreement the coefficient has value -1. If the rankings are completely independent, the coefficient has value 0. Kendall's tau rank coefficient is defined as

$$r_y^K = \frac{4P}{n(n-1)} - 1 \tag{6.4}$$

  where n is the number of items, and P is the sum, over all the items, of items ranked after the given item by both rankings.

- Cumulative Kendal's autocorrelation, Cr_y^K, defined as the sum of r_y^K for lags $1, \ldots, \tau_{max}$.

- Bicorrelation function, r_y^b, which measures linear serial dependencies within a window, is defined as

$$r_y^b = \frac{1}{n-s} \sum_{t=1}^{n-s} z(t) z(t+r) z(t+r) \tag{6.5}$$

where $z_t = (x_t - \overline{x})/s_x$ is the standarized time series of $\{x_t\}$ ($\overline{x}, s_x$ are the mean and standard deviation of $\{x_t\}$). The bicorrelation is based on the third-order joint moments, and is also referred as three-point autocorrelation.

- Hinich bicorrelation function,Cr_y^b, was developed by Hinich and Patterson (1995) as a test statistic designed to detect episodes of transient serial dependencies within a data series. It is defined as

$$Cr_y^b = \sum_{s=2}^{L} \sum_{r=1}^{s-1} (n-s)(r_y^b)^2 \tag{6.6}$$

  where $L = n^{0.4}$.

The autocorrelation functions (Pearson, Spearman, Kendal) were estimated at lags $\tau = 5, 10, 20, 30$ and r_y^b was estimated at $\tau = 5, 10, 20$. For the respective cumulative measures, two choices of τ_{max} were considered; the fixed value $\tau_{max} = 40$ and the nonlinear decorrelation time dct.

The additional complexity measures, estimated from $\{y_t\}$, that were used in this study are the following:

- Correlation sum or correlation integral, $C(\mathbf{y}, m, r)$ (see Eq.(4.18)). The correlation sum is used instead of the popular measure of fractal dimension, correlation dimension, as reliable estimation of the fractal dimension requires scaling of the correlation integral with the distance r that hardly can be maintained on preictal EEG data. Therefore, $C(\mathbf{y}, m, r)$ was defined as a measure of point density at a fixed distance ($r = 0.2$), in line with other studies on EEG (Lerner, 1996; McSharry et al., 2003; Andrzejak et al., 2006). Our modification in using $\{y_t\}$ instead of the original EEG time series in the computation of the correlation sum has the advantage of suppressing the dependence of r on the amplitude changes in the original EEG. The reconstructed parameters that were used was for both low and high dimension ($m = 5, 10$) for a fixed delay ($\tau = 10$).

- In an inverse manner, the distance r_0 that regards $C(\mathbf{y}, m, r) = 0.1$ was considered, which has been used as a measure of determinism (Ehlers et al., 1998). The radius for a particular value of the correlation integral and embedding dimension is the size of a m-dimensional cube containing the fraction of points given by the correlation integral. It is the inverse measure of the density of points on the embedded trajectory. The denser a trajectory is, the more deterministic the system is.

- Approximate entropy, $ApEnt(m, r)$, is a measure of the complexity of both deterministic and stochastic signals closely related to sample entropy (see Section 5.2). Pincus (1991) defined approximate entropy as

$$ApEnt(m, r) = \ln \frac{C(\mathbf{x}, m, r)}{C(\mathbf{x}^{m+1}, m+1, r)} \tag{6.7}$$

where $C(\mathbf{x}, m, r)$ and $C(\mathbf{x}^{m+1}, m+1, r)$ are the correlation sum for the vectors $\mathbf{x}$ and $\mathbf{x}^{m+1}$, with dimensions m and $m+1$, respectively. $ApEnt(m, r)$ quantifies the unpredictability of fluctuations in a time series; it is a scale and model independent measure that assigns a single non-negative number to a time series to quantify the orderliness or regularity of the measurements. Larger values of it denote greater apparent process randomness. Approximate entropy, $ApEnt_y(m, r)$, was estimated here for the 'gaussianized' time series $\{y_t\}$ and the parameters for its estimation were $r = 0.2$ and $m = 5$ and 10.

Set Up

Each EEG record was split into consecutive non-overlapping segments $\{x_t\}$ of 30sec duration giving segments of 3000 length. The 'gaussianized' segments $\{y_t\}$ are derived from the non-overlapping segments $\{x_t\}$ from Eq.5.4. The measures were estimated from all the consecutive segments (of $\{x_t\}$ or $\{y_t\}$ type) over a period of at least 3 h for each patient; the measure profiles comprised of at least 360 values for each measure.

In (P2) and Section 6.2, the discriminating power of some of the listed measures was assessed by means of statistical comparison of two groups (t-test, non-parametric tests) regarding smaller preictal periods. Here, larger periods were considered spanning 30 min and resulting in 60 samples in each group giving a sufficient estimate of the distribution of the measure for this period (preictal state). Thus, the comparison of the measures can be extented to the whole distribution rather than the mean or median, and thus ROC curves were used for the statistical analysis (Armitage et al., 2002). It is noted that ROC analysis requires that the samples in each group are independent. However, there may be the case that significant correlations within a preictal state are found, probably due to trends within the period, such as an upward trend in the late preictal state. Since such a trend is itself evidence in favor of discrimination, the independence assumption in the computation of ROC and AUC was overlooked.

Results

While it was somehow expected that no measure would single out, it appears that some measures occurred more often among the highest ranked measures for a range of seizure episodes, such as the autocorrelation at lag 10 (Spearman, Kendall and Pearson). Interestingly, more advanced methods were not included among the best measures. The AUC values of a number of measures suggested the discrimination between the two preictal stages in many of the 7 seizure episodes. However, among all the measures it was plausible that some measures would score high in AUC. To investigate this further, the same AUC calculations were repeated using the same late preictal state and other earlier preictal states and found that there was some variation in the AUC scores of the measures but the same simple measures

performed similarly.

Groups of measure values were consider at different periods of 30 min long (60 values) that regarded different preictal stages. As the interest was in discriminating the late preictal state from earlier preictal states, the group of the late preictal state in the period [-30,0] (with reference to the seizure onset at time 0 as identified by a clinician) was compared by means of AUC, with other groups at earlier periods of the same length. For each group comparison, the AUC of a measure was computed for every channel and the average AUC over all channels was the score of the measure. The five measures scoring highest in terms of the average AUC for the comparison of groups in the periods [-30,0] and [-190,-160] are given for 3 patients in Table 6.5.

Table 6.5: Five measures that ranked highest in terms of their AUC values, and the respective AUC values, for 3 patients. Results are from the periods [-30,0] and [-190,-160], where seizure onset is considered at time 0.

Patient	A		E		G	
A.A	Measure	AUC	Measure	AUC	Measure	AUC
1.	$r_y^K(10)$	0.735	$r_y^S(10)$	0.847	ts	0.972
2.	$r_y^S(10)$	0.734	$r_y^K(10)$	0.846	$r_y^K(30)$	0.920
3.	$r_y(10)$	0.720	$r_y(10)$	0.843	$r_y^S(30)$	0.913
4.	$r_y^S(5)$	0.704	$r_y^S(20)$	0.802	$r_y^S(5)$	0.913
5.	$r_y^K(5)$	0.703	$r_y^K(20)$	0.802	$r_y^K(5)$	0.913

The total average of AUC was computed from comparisons of the late preictal stage to 5 consecutive stages of 30 min long, starting at 190 min prior to seizure onset and ending at 40 min prior to seizure onset. The best measure from all the comparisons for the patients A, D, E, F, G are $r_y^K(10)$, ts, $r_y^K(20)$, ts, $r_y^K(30)$, respectively and the total AUC values are 0.661, 0.822, 0.864, 0.786, 0.914, respectively. In (P2), a total number of 53 measures were included in the evaluation and results for patients B and C showed that the correlation measures were not among the 10 optimal measures. Though the ranking of the measures was not the same as for a single comparison, the same simple measures occurred consistently among the 10 best. Again the highest AUC scores suggested a significant discrimination of the late preictal state from the earlier preictal one. It is noted that Spearman, Kendall and Pearson autocorrelation gave similar results, with Pearson somehow worse.

Conclusions

The study in (P2) was comprehensive with respect to the measures of scalar time series analysis applied to EEG, however here only the results on correlation and complexity measures have been discussed. Early and late preictal records from scalp EEG were used without artifact rejection. The performance of the measures

was evaluated in terms of their ability to discriminate different periods of the preictal record; a late preictal stage (30 min before seizure onset) and an earlier preictal stage of the same duration. Spearman and Kendall autocorrelation that seemed to outclass other more complex measures, have not been used till now in the EEG analysis. Some of the studied measures, such as the cumulative autocorrelation measures and declination from normality, that have been introduced in order to capture different characteristics of the EEG signal, were included among the ten best measures but only for certain patients. The performance of each measure varied with the epileptic episode and the channel, suggesting that the level of the discrimination of the preictal stages depends on the physiological conditions. The correlation measures, such as the Spearman autocorrelation at lag 10 and the decorrelation time gave high scores in many cases, but lack consistent high score in all seizure episodes. Measures detecting subtle characteristics in the signal, such as the bicorrelation, performed relatively poorly; the same holds for some measures based on nonlinear dynamics theory, such as the complexity measures.

Chapter 7

Evaluation of Causality Measures in Detecting the Direction of Information Flow

The interaction or coupling between variables or sub-systems of a complex dynamical system is a developing area of nonlinear dynamics and time series analysis (Pereda et al., 2005; Hlavackova-Schindler et al., 2007; Palus and Vejmelka, 2007). The detection and characterization of interdependence among interacting components of complex systems can give information about their functioning and a better understanding of the system dynamics. Information flow is a ubiquitous and essential feature of many complex physical phenomena, such as climatic processes (Jevrejeva et al., 2003), electric circuits (Bezruchko et al., 2003) and the brain potential system (Arnhold et al., 1999; Pikovsky et al.; Palus and Stefanovska, 2003; Pereda et al., 2005; Dhamala et al., 2008). Here, the most commonly used causality measures are reviewed, evaluated and finally improved to gain statistical significance. The effectiveness of the measures is then tested for the identification of the strength and direction of the information flow in the brain dynamics of epileptic patients.

Given a set of time series observations, it is essential to assess whether they originate from coupled or decoupled systems, detect the hidden causal dependencies between them and understand which system is the driver and which is the responder. Granger causality (Granger, 1969) was the leading approach for a long time inferring the direction of interactions. Granger investigated the dependencies between time series and whether one time series is useful in forecasting another. The concept of Granger causality is based on the predictability of time series and uses linear models. If the prior knowledge of a time series improves the prediction of another, the former Granger-causes the latter. Many measures have been developed based on the concept of Granger causality (Baccala and Sameshima, 2001; Winterhalder et al., 2005; Schelter et al., 2006) and extended it in order to incorporate also nonlinear relationships between time series (Pijn and Lopes

Da Silva, 1993; Chen et al., 2004). Many model-free techniques have also been developed incorporating the concept of Granger causality. The recently developed measures of interaction go beyond the standard cross-correlation and exploit nonlinear properties of dynamical systems. The causal measures are divided in three main categories; those based on phase and event synchronization (Rosenblum and Pikovsky, 2001; Quiroga et al., 2002; Smirnov and Bezruchko, 2003), on reconstruction of the state spaces (Cenys et al., 1992; Schiff et al., 2000; Arnhold et al., 1999; Quiroga et al., 2000a; Andrzejak et al., 2003; Bhattacharya et al., 2003; Feldmann and Bhattacharya, 2004; Romano et al., 2007; Chicharro et al., 2008), and on information theory (Schreiber, 2000; Palus et al., 2001a; Marschinski and Kantz, 2002; Staniek and Lehnertz, 2008; Vejmelka and Palus, 2008). The causal information measures make no assumptions on the system dynamics as opposed to phase or event synchronization measures that assume strong oscillatory behavior or distinct event occurrences, respectively, and the state space methods that require local dynamics being preserved in neighborhoods of reconstructed points. In Sec. 7.1, causality measures of state space dynamics, synchronization and information theory are reviewed.

Some works evaluating different causality measures exist, but most studies cannot be considered to be universal as only certain types of measures were used and the conclusions do not point to the same measures as different selection of measures are included in these studies (Smirnov and Andrzejak, 2005; Lungarella et al., 2007; Kreuz et al., 2007; Palus and Vejmelka, 2007). Therefore, state space, synchronization and information causality measures are evaluated in Sec.7.2 (P8). The study focuses on information causal measures; the study on the probability density estimators (P6, P7, P10) for the mutual information on scalar time series is extended and is used for the estimation of measures detecting the information flow that involve joint and conditional probability density estimation from bivariate time series. Other information measures that are evaluated are the mean conditional mutual information (Vejmelka and Palus, 2008) and the coarse-grained transinformation rate (Palus et al., 2001a).

The evaluation of the measures is assessed by Monte Carlo simulations on known dynamical systems. Unidirectionally and bidirectionally coupled nonlinear systems are considered and the strength of the coupling is varied in order to examine whether measures can correctly detect the direction of the causal effects and whether they give spurious causal effects in case of uncoupled systems. Therefore the power of the measures is examined, i.e. how sensitive the measure are in detecting interaction and identifying its direction. Measures are also evaluated on EEG for the detection of the information flow in the brain of an epileptic patient. The results from the simulation study are used in order to interpret the performance of the measures on EEG.

The simulation study underlined the need to render the statistical significance of the causality measures as most measures are biased and indicated causal effects when they were not present. Aspects influencing the robust estimation of the causality measures are the finite time series length, the presence of noise and the

reconstruction parameters for the estimation of the measures. The estimation of the measures gets even more complicated when the possibly coupled systems are non-identical or when they have different complexity. Thus, in Sec.7.3 a scheme for improving the statistical significance of the causality measures is improved (P11). First of all, a correction applicable to a group of similarly defined state space measures is presented, where the main idea is to restrict the neighborhoods of the reconstructed points of the systems and slightly modify the estimation procedure of the measures. A second modification of the measures discussed here is under the frame of effective entropy (Marschinski and Kantz, 2002), i.e. the subtraction of the value of a measure estimated by randomly shuffling the driving time series from its original value. Here, instead of using only one 'surrogate' value, it is suggested to subtract the mean estimated value of a number of surrogates in order to decrease the bias of the estimate of the surrogate value. Finally, a new surrogate approach is introduced which is again applicable to all causality measures. As thorough investigation for the validity and usefulness of the causality measures should start with a test of significance, i.e. a measure should not identify coupling (or interaction) in any direction when it is not present, therefore the use of surrogates is inevitable. For the generation of the surrogates, the reconstructed vectors of the driving time series are randomly shuffled, instead of randomly shuffling the driving time series. A number of modifications of the causality measures are developed, based on the frame of the introduced surrogates and on each measure's estimation scheme.

In Sec.7.4, the improved causality measures are evaluated on known systems. The effectiveness of the measures in detecting the causal effects is examined with respect to the systems complexity, time series length, noise level and the embedding parameters and specifically the embedding dimension for the reconstruction of the state space of the examined systems. The evaluation of the measures is completed in Sec.7.5, where measures are evaluated on EEG recordings; the direction of interactions among the different brain areas is investigated.

7.1 Causality Measures

The main interest is the interaction of nonlinear dynamical systems and therefore measures that have the power to detect this type of interaction are reviewed, skipping model-based, correlation and coherence measures used in linear multivariate analysis (Kaminski, 2005; Winterhalder et al., 2005) and methods of mutual nonlinear prediction, e.g. see Faes et al. (2008).

Let $\{x_t\}$ and $\{y_t\}$, $t = 1, \ldots, n$, denote two simultaneously observed univariate time series derived from the dynamical systems X and Y, respectively. Let also consider that the two systems are unidirectionally coupled and X is the driving system and Y is the response system. The notation $X \rightarrow Y$ is used in order to indicate the effect of X system on Y, while $Y \rightarrow X$ is indicating the opposite effect. The causality measures will be introduced only for the direction $X \rightarrow Y$ here, as they are equivalently defined for the opposite direction. More-

over, all the measures are defined in order to allow for different embedding parameters for the corresponding variables X and Y of the time series. Let m_x and m_y be the embedding dimensions and τ_x and τ_y the delays for the two systems, respectively. The reconstructed vectors of the two state spaces are defined as $\mathbf{x}_t = (x_t, x_{t-\tau_x}, \ldots, x_{t-(m_x-1)\tau_x})'$ and $\mathbf{y}_t = (y_t, y_{t-\tau_y}, \ldots, y_{t-(m_y-1)\tau_y})'$, where $t = 1, \ldots, N'$ and $N' = n - \max\{(m_x-1)\tau_x, (m_y-1)\tau_y\}$. The steps ahead or time horizon to address the interaction is denoted by h.

7.1.1 State space measures

Nonlinear interdependence measures

Five measures based on state space reconstruction and on neighboring distances, called nonlinear interdependence measures (NI), are defined for detecting the strength and direction of causal effects. This class of interaction measures involves the state space reconstruction from each time series using standard delay embedding. The measures operate on the neighboring reconstructed points.

Let $r_{t,j}$ and $s_{t,j}$, $j = 1, \ldots, k$, denote the time indices of the k-nearest neighbors of $\mathbf{x}_t$ and $\mathbf{y}_t$, respectively. For each point $\mathbf{x}_t$, the mean square Euclidean distance to its k neighbors is defined as

$$R_t^{(k)}(X) = \frac{1}{k} \sum_{j=1}^{k} ||\mathbf{x}_t - \mathbf{x}_{r_{t,j}}||^2. \tag{7.1}$$

The Y-conditioned mean square Euclidean distance is defined by replacing the true nearest neighbors of a point $\mathbf{x}_t$ by the equal time partners of the closest neighbors of $\mathbf{y}_t$ as

$$R_t^{(k)}(X|Y) = \frac{1}{k} \sum_{j=1}^{k} ||\mathbf{x}_t - \mathbf{x}_{s_{t,j}}||^2. \tag{7.2}$$

The mean square distance of $\mathbf{x}_t$ to all points in $\{\mathbf{x}_t\}$, $t = 1, \ldots, N'$ (except $\mathbf{x}_t$) is given as

$$R_t(X) = \frac{1}{N-1} \sum_{t \neq j} ||\mathbf{x}_t - \mathbf{x}_j||^2. \tag{7.3}$$

Different interdependence measures are derived from these mean distances. The first one is

$$\mathrm{S}_{X \to Y} = \frac{1}{N'} \sum_{t=1}^{N'} \frac{R_t^{(k)}(X)}{R_t^{(k)}(X|Y)} \tag{7.4}$$

(Arnhold et al., 1999). The original notation of $\mathrm{S}_{X \to Y}$ is $\mathrm{S}(X|Y)$, however this notation was misunderstood in many papers and therefore it should be specially stressed which is the direction of the causal effects. By construction $0 < \mathrm{S}_{X \to Y} \leq 1$ holds. Values of $\mathrm{S}_{Y \to X}$ close to zero suggest independence of X and Y, while significant positive values of $\mathrm{S}_{X \to Y}$ suggest dependence of Y on X. The measure

is non-symmetric, so if $\mathrm{S}_{X\to Y} > \mathrm{S}_{Y\to X}$ then Y depends more on X than vice versa. It has been reported that this measure is robust against noise and capable of detecting weak inter-dependence (Arnhold et al., 1999; Quiroga et al., 2002; Krug et al., 2007).

The second measure is defined as

$$\mathrm{H}_{X\to Y} = \frac{1}{N'}\sum_{t=1}^{N'} \log \frac{R_t(X)}{R_t^{(k)}(X|Y)} \tag{7.5}$$

(Arnhold et al., 1999). $\mathrm{H}_{X\to Y}$ has no upper bound. It is zero if X and Y are completely independent, while it is positive if nearness in X implies also nearness in Y for equal time partners. This measure was found to be more robust against noise and easier to interpret than $\mathrm{S}_{X\to Y}$ (Quiroga et al., 2000b), as the quantity $R_t(Y)$ is much less dependent on the structure and dimensionality in X than $R_t^{(k)}(X)$.

The third measure is defined as

$$\mathrm{N}_{X\to Y} = \frac{1}{N'}\sum_{t=1}^{N'} \frac{R_t(X) - R_t^{(k)}(X|Y)}{R_t(X)} \tag{7.6}$$

(Quiroga et al., 2000a). Note that $\mathrm{N}_{X\to Y} = 1$ holds only if $R_t^{(k)}(X|Y) = 0$ for all t. Even in the case of identical synchronization $\mathrm{N}_{X\to Y} < 1$, and this is determined by $R_t^{(k)}(X|Y)$, which is strongly influenced by autocorrelations and finite dimensionality of X.

Andrzejak et al. (2003) in order to minimize the unwanted influence of autocorrelations and finite dimensionality of the variables, suggested a modification of measure N, as

$$M^*_{X\to Y} = \frac{1}{N'}\sum_{t=1}^{N'} \frac{R_t(X) - R_t^{(k)}(X|Y)}{R_t(X) - R_t^{(k)}(X)}. \tag{7.7}$$

In order to avoid the negative values that $M^*_{X\to Y}$ can take, the fourth NI measure is defined as

$$\mathrm{M}_{X\to Y} = \max\{M^*(X|Y), 0\}. \tag{7.8}$$

$\mathrm{M}_{X\to Y}$ is defined in complete analogy to $\mathrm{S}_{X\to Y}$, in order to test whether Y is dependent on X. For independent dynamics, $\mathrm{M}_{X\to Y}$ should tend to zero while in the case of identical synchronization, it will reach their maximal value of 1.

Chicharro et al. (2008) built a similar form to $M^*_{X\to Y}$ in Eq.(7.7) using rank statistics instead of distance statistics. For each point $\mathbf{x}_t$, let $g_{t,j}$ denote the rank of the distance $\|\mathbf{x}_t - \mathbf{x}_j\|$ among all distances for $j = 1, \ldots, N'$ and $j \neq t$. Note that the K first ranks are $g_{t,s_{t,j}}$ for $K = 1, \ldots, k$, and the mean rank of k-closest points as $G_t^{(k)}(X) = k/2$. The Y-conditioned mean rank is defined as

$$G_t^{(k)}(X|Y) = \frac{1}{k}\sum_{j=1}^{k} g_{t,r_{t,j}}, \tag{7.9}$$

and the fifth NI rank-based measure is defined as

$$\mathrm{L}_{X\to Y} = \frac{1}{N'}\sum_{t=1}^{N'} \frac{G_t(X) - G_t^{(k)}(X|Y)}{G_t(X) - G_t^{(k)}(X)}. \tag{7.10}$$

This measure is normalized to one for identical synchronization, and it is symmetrically distributed around zero for independent linear stochastic processes. Chicharro et al. (2008) argue that bias and false indication of directional coupling due to the autocorrelation of the involved processes are minimized, whereas the NI measures based on distance statistics are more affected by the different degree of complexity of the respective dynamics of the systems.

Mean conditional probability of recurrence

Recalling that neighboring points correspond to nearby trajectories or equivalently recurrences of a single (chaotic) trajectory at the neighborhood of reference, NI measures bear a great deal of similarity to the measure of *mean conditional probability of recurrence* (MCR) (Romano et al., 2007). According to the visualizing method of recurrence plots (Eckmann et al., 1987), the recurrence matrixes of X and Y are respectively $R_{i,j}^X = \Theta(\varepsilon_x - \|\mathbf{x}_i - \mathbf{x}_j\|)$ and $R_{i,j}^Y = \Theta(\varepsilon_y - \|\mathbf{y}_i - \mathbf{y}_j\|)$, where Θ is the Heaviside function and ε_x, ε_y are distances (radius when the distance metric is Euclidean). In addition, the joint recurrence matrix of (X, Y) is defined as $J_{i,j}^{X,Y} = \Theta(\varepsilon_x - \|\mathbf{x}_i - \mathbf{x}_j\|)\Theta(\varepsilon_y - \|\mathbf{y}_i - \mathbf{y}_j\|)$. The concept of recurrence has been used to quantify a weaker form of synchronization, whereas MCR is an extension of it that detects the direction of the coupling. MCR, indicating the influence of X on Y, is defined as

$$\mathrm{MCR}_{X\to Y} = \frac{1}{N'}\sum_{i=1}^{N'} \frac{\sum_{j=1}^{N'} J_{i,j}^{X,Y}}{\sum_{j=1}^{N'} R_{i,j}^{Y}}. \tag{7.11}$$

In loose terms, MCR estimates the probability of recurrence on Y when there is synchronous recurrence on X. The independence or the direction of the dependence of X and Y is indicated by comparing the estimated values of the measure from the two directions, i.e. if $\mathrm{MCR}_{X\to Y} > \mathrm{MCR}_{Y\to X}$ then X drives Y and if $\mathrm{MCR}_{X\to Y} \simeq \mathrm{MCR}_{Y\to X}$ then the coupling is symmetric (same strength of coupling in both directions). The main difference between the NI measures and MCR is that MCR uses counts of neighboring points according to a distance threshold, whereas NI uses the distances for a fixed number of neighboring points. Romano et al. (2007) suggested to estimate the values of the thresholds ϵ_x and ϵ_y so that in case of no interactions both mean probabilities of recurrences $\langle p(\mathbf{x}_i)\rangle$ and $\langle p(\mathbf{y}_i)\rangle$ are equal to 0.01, but obviously in real applications no such optimization of the thresholds is possible.

7.1.2 Synchronization measures

Directionality index

Reverting from state space dynamics to phase dynamics, the measure of *directionality index* (DI) quantifies the degree the phase dynamics of one oscillator is influenced by the phase dynamics of another oscillator (Rosenblum and Pikovsky, 2001). To form the DI measure it is assumed that the two time series indeed exhibit oscillating behavior, so that phases $\phi_X(t)$, $t = 1, \ldots, N$, can be extracted from $\{X_t\}_{t=1}^N$, typically using the Hilbert transform (Hilbert, 1953) (the same for Y, also for the following expressions). It is also assumed that the phase increments $\Delta_Y(t) = \phi_Y(t+\tau) - \phi_Y(t)$ are generated by an unknown two-dimensional map $\Delta_Y(t) = F_Y[\phi_X(t) - \phi_Y(t)]$ fitted by a finite Fourier series $\sum_{m,l} A_{m,l} e^{im\phi_X + il\phi_Y}$. The function F_X is defined in full analogy to F_Y. The fitted functions for X and Y are used to quantify the cross dependencies of phase dynamics of the two systems given as

$$c_X^2 = \int_0^{2\pi} \int_0^{2\pi} \frac{\partial F_X}{\partial \phi_X} \mathrm{d}\phi_X \mathrm{d}\phi_Y, \qquad c_Y^2 = \int_0^{2\pi} \int_0^{2\pi} \frac{\partial F_Y}{\partial \phi_Y} \mathrm{d}\phi_X \mathrm{d}\phi_Y. \tag{7.12}$$

The directionality index is defined as:

$$d_{X,Y} = \frac{c_Y - c_X}{c_X + c_Y}, \tag{7.13}$$

where $d_{X,Y}$ close to -1 or 1 suggests unidirectional coupling driven by Y or X, respectively, while $d_{X,Y} \simeq 0$ suggests symmetric bidirectional coupling or no coupling. The main disadvantage of this method is that it is not always possible to extract phases from scalar time series.

Event delay and event synchronization

Event synchronization is the relative timing of certain 'events' in the time series (like spikes, local minima or maxima), i.e. quasi-simultaneous appearances of these events in the two time series (Quiroga et al., 2002). Let $c^\tau(X|Y)$ denote the number of times an event appears in X shortly after it appears in Y (for details see Quiroga et al. (2002)), allowing a time lag τ between two synchronous events. The strength of the coupling, termed as *event synchronization* (ES), is expressed by the normalized total of synchronized events

$$Q_\tau = \frac{c^\tau(Y|X) + c^\tau(X|Y)}{\sqrt{n_x n_y}}, \tag{7.14}$$

and the coupling direction, termed as *event delay* (EvD), by the normalized difference

$$q_\tau = \frac{c^\tau(Y|X) - c^\tau(X|Y)}{\sqrt{n_x n_y}}, \tag{7.15}$$

where n_x and n_y are the total number of occurrences of events in X and Y. The measures are normalized so that $0 \leq Q_\tau \leq 1$ and $-1 \leq q_\tau \leq 1$. For $Q_\tau = 1$ the events of the signals are fully synchronized while $Q_\tau = 0$ suggests no synchronization. When q_τ is close to 1 an event in X is likely to precede an event in Y and thus X drives Y, and respectively when q_τ is close to -1 suggests that Y drives X. As also mentioned in Quiroga et al. (2002), τ is not fixed but adapted to the time interval between events at each step (it is half of the minimum of times from the current event to the preceding and to the succeeding event for both X and Y).

7.1.3 Information measures

Mean conditional mutual information

For a scalar time series, mutual information was already defined in terms of a delay, $I(X,Y) = I(x_t, x_{t-\tau}) = I(\tau)$. For bivariate time series a natural extension is the cross mutual information $I(x_t, y_{t-\tau})$. This measure is considered insufficient to measure interaction correctly because it is influenced by the self-dynamics of each system (e.g. see Schreiber (2000)). Instead of $y_{t-\tau}$, the difference $\Delta_h y_t = y_{t+h} - y_t$ is considered, and then the (cross) mutual information conditioned on y_t is $I(x_t, \Delta_h y_t | y_t)$. The average of $I(x_t, \Delta_h y_t | y_t)$ for delays up to a maximum delay $\tau_{\max}$ is the measure of *mean conditional mutual information* (MCMI) (Vejmelka and Palus, 2008)

$$i_{X \to Y} = \frac{1}{\tau_{\max}} \sum_{\tau=1}^{\tau_{\max}} I(x_t, \Delta_h y_t | y_t) \tag{7.16}$$

that quantifies the information transferred from X to Y at a later time conditioning on the current state of Y, i.e. the level of driving of X to Y. $i_{Y \to X}$ is defined similarly.

Coarse-grained transinformation rate

A different expression for the conditional mutual information is used to define the measure of *coarse-grained transinformation rate* (CGTR) (Palus, 1996; Palus et al., 2001b). Averaging on time increments τ as for MCMI, CGTR measures the average rate of the net amount of information transferred from X to Y and is defined as:

$$Ci_{X \to Y} = \frac{1}{\tau_{\max}} \sum_{\tau=1}^{\tau_{\max}} I(x_t, y_{t+\tau} | y_\tau) - \frac{1}{2\tau_{\max}} \sum_{\tau=-\tau_{\max}, \tau \neq 0}^{\tau_{\max}} I(x_t, y_{t+\tau}). \tag{7.17}$$

If $Ci_{X \to Y} \simeq 0$ then there is no information flow, while if $Ci_{X \to Y} > 0$ then X influences Y. If $Ci_{X \to Y} > Ci_{Y \to X}$ then X influences Y more than vice versa.

Transfer entropy

Transfer entropy (TE) is an information theoretic measure which takes into account the dynamics of information transport and detects the directed exchange of information between two systems. As defined by Schreiber (2000), TE quantifies the information flow from X to Y by the amount of information explained in Y at one step ahead from the state of X, accounting for the concurrent state of Y. The concept of transfer entropy extends the Shannon entropy for transition probabilities and quantifies how the conditioning on X change the transition probabilities of Y. It has been shown that with proper conditioning, transfer entropy $\mathrm{TE}_{X\to Y}$ is the exact equivalent to the conditional mutual information $I(y_{t+\tau}|x_t, y_t)$ (Hlavackova-Schindler et al., 2007; Palus and Vejmelka, 2007).

TE is defined here in accordance with the state space measures as

$$\mathrm{TE}_{X\to Y} = \sum p(y_{t+h}, \mathbf{x}_t, \mathbf{y}_t) \log \frac{p(y_{t+1h}|\mathbf{x}_t, \mathbf{y}_t)}{p(y_{t+h}|\mathbf{y}_t)}, \tag{7.18}$$

where $p(y_{t+h}, \mathbf{x}_t, \mathbf{y}_t)$, $p(y_{t+h}|\mathbf{x}_t, \mathbf{y}_t)$, and $p(y_{t+h}|\mathbf{y}_t)$ are the joint and conditional probability mass functions. The time horizon h is also here introduced instead of the time step one that was originally used in the definition of TE. TE can also be defined in terms of entropies as

$$\mathrm{TE}_{X\to Y} = H(\mathbf{x}_t, \mathbf{y}_t) - H(y_{t+h}, \mathbf{x}_t, \mathbf{y}_t) + H(y_{t+h}, \mathbf{y}_t) - H(\mathbf{y}_t) \tag{7.19}$$

Let X be a continuous, possibly vector-valued, random variable. For a fixed small r, the entropy of a variable X can be estimated in term of the correlation sum as

$$H(X) \simeq \ln C(\mathbf{x}, m, r) + m \ln r \tag{7.20}$$

(Manzan and Diks, 2002), where m is the dimension of the vectors $\mathbf{x}$. Let us denote the correlation sums $C([y_{t+h}\ \mathbf{x}_t\ \mathbf{y}_t], 1 + m_x + m_y, (1 + m_x + m_y)^{1/2}r)$, $C(\mathbf{y}_t, m_y, m_y^{1/2}r)$, $C([\mathbf{x}_t\ \mathbf{y}_t], m_x + m_y, (m_x + m_y)^{1/2}r)$ and $C([y_{t+h}\ \mathbf{y}_t], 1 + m_y, (1 + m_y)^{1/2}r)$ as $C(y_{t+h}, \mathbf{x}_t, \mathbf{y}_t)$, $C(\mathbf{y}_t)$, $C(\mathbf{x}_t, \mathbf{y}_t)$ and $C(y_{t+h}, \mathbf{y}_t)$, respectively. Then, TE is defined as

$$\mathrm{TE}_{X\to Y} = \log \frac{C(y_{t+h}, \mathbf{x}_t, \mathbf{y}_t)C(\mathbf{y}_t)}{C(\mathbf{x}_t, \mathbf{y}_t)C(y_{t+h}, \mathbf{y}_t)} \tag{7.21}$$

(Grassberger and Procaccia, 1983; Pawelzik and Schuster, 1987). Here, the Euclidian norm for the estimation of the correlation sums is used and the radius at each correlation sum is normalized by the corresponding dimension of the respective vectors. In all applications, time series were normalized to have zero mean and standard deviation one and r was set to 0.2, which is a recommended value for the estimation of the correlation sum and for many state space measures based on neighborhoods, e.g. see Pincus (1991); Govindan et al. (2007).

7.1.4 Relationships of causality measures

MCR defines in a rather direct way the conditional probability of close points in Y given they are close in X, whereas NI measures attempt to approximate the conditional probability indirectly through distances. To this respect the MCR method is closer to information measures such as TE. However, TE involves also transition probabilities, which constitute an advantage of incorporating the dynamical structure of the systems. Romano et al. (2007) claim that MCR needs smaller number of data points than TE. They argue that TE gives values larger than zero from both directions in case of purely unidirectional coupling and this is due to the finite length of the time series, whereas MCR seems not to have this problem. However, all measures have bias.

7.2 Evaluation of Causality Measures on known systems

The causality measures that have been presented, are here evaluated on their effectiveness in identifying correctly the direction of the coupling between two interacting systems. The evaluation of the measures is assessed by Monte Carlo simulations on systems of different types and for varying coupling strengths. From the five NI measures, only S was used here. For the information measures, two estimators of the probability functions have been used; based on the equidistant binning scheme and on the correlation sums.

7.2.1 Simulation Set-Up

The evaluation of the measures was assessed by means of Monte Carlo simulation on different types of systems. Causality measures were computed on coupled systems for increasing coupling strengths in order to evaluate the ability of the measures to detect the degree and direction of the coupling. 100 realizations were generated from the different simulation systems (linear and nonlinear, chaotic systems, unidirectional and bidirectional), for time series lengths $n = 1024, 2048, 4096$. Analytically, the simulation systems were:

- Two unidirectionally coupled AR(1) models

$$\begin{aligned} x_{t+1} &= 0.5x_t + e_t^x \\ y_{t+1} &= 0.5y_t + cx_t + e_t^y \end{aligned} \tag{7.22}$$

  where e_t^x and e_t^y are normal iid with mean zero and variance one, and the coupling strength c is set to 0, 0.1, 0.2, 0.3, 0.4, 0.5.

- Two unidirectionally coupled Henon maps

$$\begin{aligned} x_{t+1} &= 1.4 - x_t^2 + 0.3x_{t-1} \\ y_{t+1} &= 1.4 - cx_t y_t + (1-c)y_t^2 + 0.3y_{t-1} \end{aligned} \tag{7.23}$$

with coupling strengths $c = 0, 0.1, 0.2, 0.3, 0.4, 0.5$.

- A Rössler system (x_1, y_1, z_1)

$$\begin{aligned} \dot{x_1} &= -6(y_1 + z_1) \\ \dot{y_1} &= -6(x_1 + 0.2y_1) \\ \dot{z_1} &= -6(0.2 + z_1(y_1 - 5.7)) \end{aligned} \tag{7.24}$$

driving a Lorenz system (x_2, y_2, z_2)

$$\begin{aligned} \dot{x_2} &= 10(x_2 + y_2) \\ \dot{y_2} &= 28x_2 - y_2 - x_2 z_2 + cy_1^2 \\ \dot{z_2} &= x_2 y_2 - \frac{8}{3} z_2 \end{aligned} \tag{7.25}$$

with coupling strengths $c = 0, 0.5, 1, 1.5, 2$.

- Two bidirectionally coupled Henon maps

$$\begin{aligned} x_{t+1} &= 1.4 - x_t^2 + 0.3x_{t-1} + c_2(x_t^2 - y_t^2) \\ y_{t+1} &= 1.4 - y_t^2 + 0.3y_{t-1} + c_1(y_t^2 - x_t^2) \end{aligned} \tag{7.26}$$

with coupling strengths (c_1, c_2) = (0.05,0.05), (0.1,0.05), (0.1,0.1), (0.15,0.05) and (0.2,0.05).

For all systems, the range of the coupling strength was bounded from zero (independent systems) up to the level of almost complete synchronization. All the time series were first normalized to have zero mean and standard deviation one. The parameters for the estimation of the measures were set as follows: the number of the bins for the information measures when estimated with the equidistant (ED) scheme is $b = 8$, the radius for finding neighbors is $r = 0.15$ (but for transfer entropy where $r = 0.2$), the number of neighbors for the state space measures is $k = 10$ and the Theiler window is $W = 10$ (to exclude time correlated points). For all systems apart from the unidirectionally coupled Rössler–Lorenz system, the embedding parameters were set to be $m_x = m_y = 2$ and $\tau_x = \tau_y = 2$, the time horizon $h = 1$, the maximum lag for the average of time increments in information measures $\tau_{\text{max}} = 10$, and for transfer entropy also the case for $h = 1$ and $\tau_x = \tau_y = 1$ was considered. For the unidirectionally coupled Rössler–Lorenz system it was set $m_x = 5, m_y = 5, h = 5, \tau_x = \tau_y = 5, \tau_{\text{max}} = 20$, and for the transfer entropy it was set $m_x = m_y = 2, h = 5$ and $tau_x = \tau_y = 5$.

7.2.2 Results

For the unidirectionally coupled systems, the directionality measures should increase with coupling strength at the direction of the coupling (at the direction $X \rightarrow Y$) and be stable (and close to zero) at the opposite direction ($Y \rightarrow X$)

and in case of no causal effects ($c = 0$). In order to evaluate how well the measures follow this expected behavior, the mean values and standard deviation of the measures from the simulations were computed and displayed for the different settings of systems, time series length, noise level and state space reconstruction.

Results on the unidirectionally coupled AR(1) system

The results from the simulations on the unidirectionally coupled AR(1) system indicated the superiority of the information causality measures. The state space measures and DI could not discriminate between the linear stochastic systems, even for large time series lengths, due to the lack of nonlinear dynamics (in local state space) and phase dynamics, respectively. On the other hand, EvD regarding the local maxima of the time series, could discriminate the directionality at a level that increases with the coupling strength, as shown in Fig. 7.1a. Both estimators

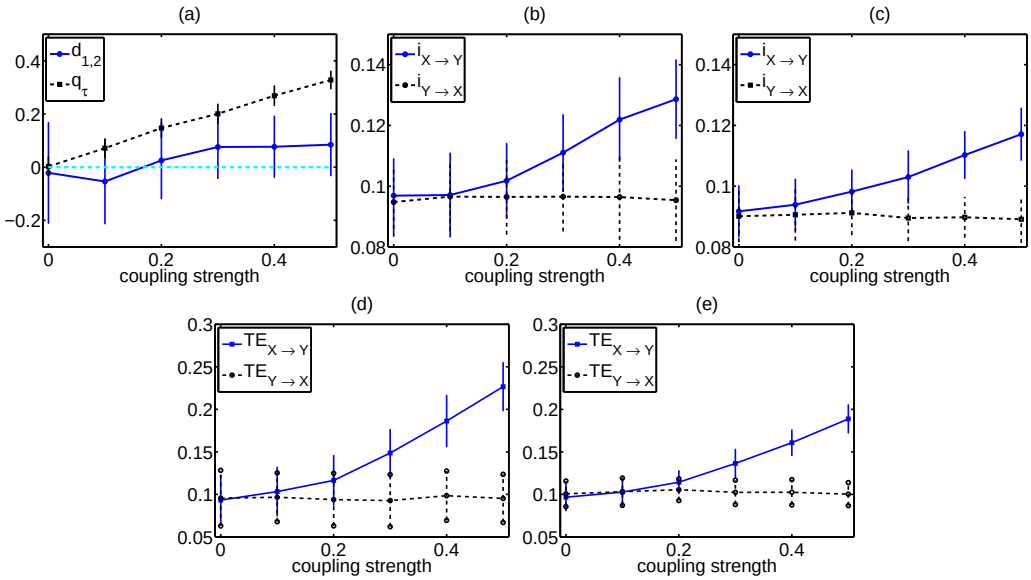

Figure 7.1: Mean and standard deviation (shown as error bars) of the directionality measures for 100 realizations of unidirectionally coupled $AR(1)$ system. (a) DI and EvD as shown in the legend, for $n = 4096$. The light grey dotted line indicates the zero level. (b) MCMI computed by correlation sums, $n = 1024$. (c) As in (b) but for equidistant partition. (d) TE computed by correlation sum, $n = 1024$. (e) As in (d) but for equidistant partition.

of MCMI (with correlation sum and equidistant binning) detected the directionality of the coupling correctly (see Fig. 7.1b and c), and even for small time series lengths, while CGTR did not seem to detect the directionality of the coupling. Both estimators of TE (from binning and correlation sum) detected the direction of information flow correctly (see Fig. 7.1d and e). However, all measures but EvD

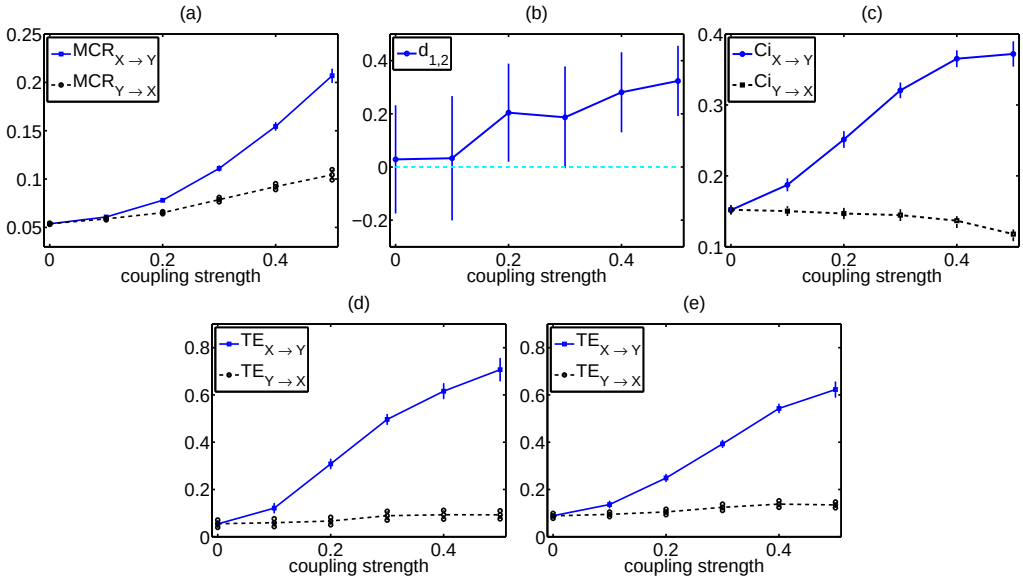

Figure 7.2: Average and standard deviation (shown as error bars) of the directionality measures for 100 realizations of unidirectionally coupled Henon system. (a) MCR and (b) DI for $n = 4096$. (c) CGTR, for $n = 1024$. (d) TE computed by correlation sums, for $n = 1024$. (e) As in (d) but for equidistant partition.

were positively biased for $c = 0$ and for the direction $Y \rightarrow X$ the mean estimated values of the measures are in the limits of 0.1 instead of zero.

Results on the unidirectionally coupled Henon map

For the unidirectionally coupled Henon map, MCR detected the increase of the coupling in the correct direction, but it wrongly indicated the same but at a lesser degree in the other direction (see Fig. 7.2a). On the other hand, S could not detect the direction of information flow correctly. DI detected correctly the strength and the directionality of the coupling but with a large variance, even for large time series lengths, and did not allow for a clear indication of directionality (see Fig. 7.2b). EvD could not detect any information flow. On the other hand, all information measures detected correctly the strength and the directionality of the coupling even for small time series lengths (results for CGTR and TE are shown in Fig. 7.2c-e).

Results on the unidirectionally coupled Rossler-Lorenz system

For the unidirectionally coupled Rössler–Lorenz system, the detection of the directionality of the coupling was expected to be more difficult as the two flows are of different type. The simulations showed that EvD, NI, DI and MCMI could not detect the direction and strength of the coupling, whereas MCR and CGTR could both detect it, even for small time series lengths. Using the TE measure

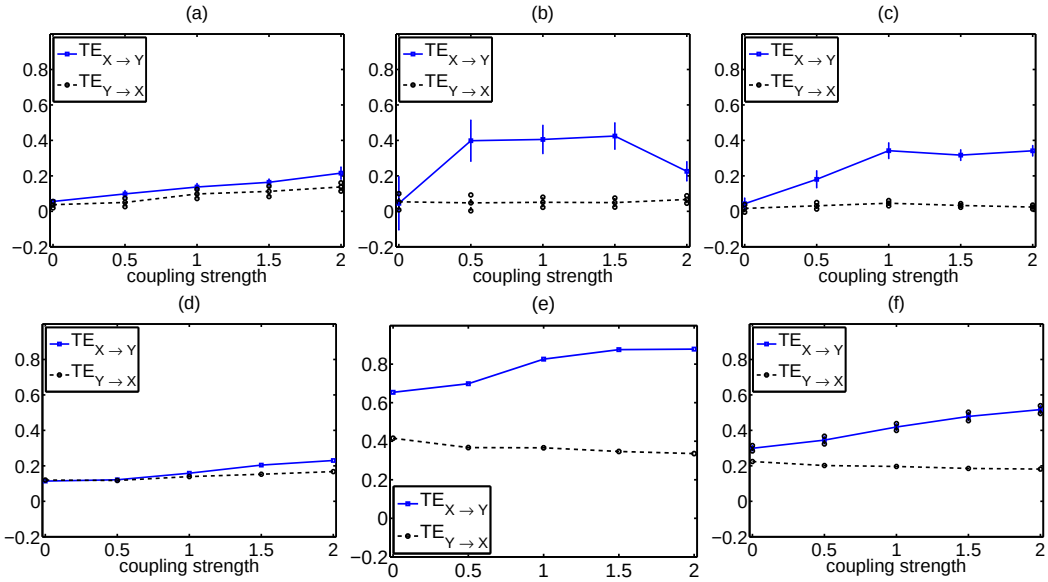

Figure 7.3: Mean and standard deviation (shown as error bars) of the directionality measures for 100 realizations of unidirectionally coupled Rössler–Lorenz system. (a) TE computed by correlation sums, for $h = 1, m_x = 2, m_y = 1$ and $n = 1024$. (b) As in (a) but for $h = 5, m_x = m_y = 2, \tau_x = \tau_y = 5$ and $n = 1024$. (c) As in (b) for $n = 2048$. (d), (e), (f) As in (a), (b) and (c), respectively, but for equidistant partition.

with the parameters $h = 1, m_x = 2$, $m_y = 1$ and $\tau_x = \tau_y = 1$, only slight difference of the mean estimated values of TE from the two directions was observed (see Fig. 7.3a, d). However, estimation of TE using proper reconstruction of the state spaces, significantly improves the performance of the measure. For $h = 5, m_x = m_y = 2, \tau_x = \tau_y = 5$, TE correctly detected the direction of the coupling, and the estimator based on the correlation sums seems to be more effective than the equidistant one (see Fig. 7.3b and e). The detection of the direction of the causal effects is improved with the increase of the time series length as shown in Fig. 7.3c and f. The strong dependence of TE on the data size is due to the estimation of the probability densities regarding the reconstructed points rather than the samples.

Results on the bidirectionally coupled Henon map

For the bidirectionally coupled Henon maps, the synchronization measures (for phase and event) failed to detect the direction of stronger couplings, as there was no distinct phases or events in the time series. Regarding the state space measures, MCR correctly estimated the driver-response interaction only when it was present as shown in Table 7.1, whereas S measure gave very small values independently

Table 7.1: Mean estimated values of the causality measures from 100 realizations of the bidirectionally coupled Henon map for $n = 1024$.

	Coupling strengths				
Measures	(0.05,0.05)	(0.1,0.05)	(0.1,0.1)	(0.15,0.1)	(0.2,0.05)
$\text{MCR}_{X\to Y}$	0.0874	0.1018	0.1655	0.1363	0.1873
$\text{MCR}_{Y\to X}$	0.0875	0.0940	0.1649	0.1133	0.1453
$\text{S}_{X\to Y}$	0.0022	0.0032	0.0075	0.0051	0.0101
$\text{S}_{Y\to X}$	0.0022	0.0032	0.0075	0.0045	0.0074
$i_{X\to Y}$(Cor. Sum)	0.1519	0.2273	0.3876	0.3159	0.3609
$i_{Y\to X}$(Cor. Sum)	0.1473	0.1555	0.3870	0.1820	0.1847
$i_{X\to Y}$(hist. based)	0.1398	0.1769	0.2267	0.2172	0.2551
$i_{Y\to X}$(hist. based)	0.1386	0.1387	0.2269	0.1503	0.1519
$Ci_{X\to Y}$	0.1749	0.2036	0.1119	0.2336	0.2629
$Ci_{Y\to X}$	0.1734	0.1663	0.1124	0.1609	0.1391
$\text{TE}_{X\to Y}$(Cor. Sum)	0.1007	0.1818	0.2787	0.3121	0.3984
$\text{TE}_{Y\to X}$(Cor. Sum)	0.0990	0.1044	0.2804	0.1323	0.1318
$\text{TE}_{X\to Y}$(hist. based)	0.1021	0.1415	0.1660	0.2075	0.3021
$\text{TE}_{Y\to X}$(hist. based)	0.1024	0.0999	0.1658	0.1224	0.1300

of the coupling setting. All information measures proved to be effective (see Table 7.1), giving about the same values for both directions in case of equal coupling strengths c_1 and c_2. In cases of stronger coupling strength in one direction, TE gave the larger difference among the values from the two directions. For example, for $c_1 = 0.15$ and $c_2 = 0.1$, $\text{TE}_{X\to Y}$ increased by about 136% from $\text{TE}_{Y\to X}$ for the estimator using correlation sums and 70% for the estimator using histograms, whereas the increase of MCMI was about 74% (correlation sums estimator) and about 45% (estimator using histograms) and for CGTR about 45%. Note that these results are for $n = 1024$, and for larger time series the difference in the presence of stronger coupling in one direction was clearer.

7.2.3 A pilot study: effectiveness of the causality measures on EEG

In order to examine the effectiveness of the measures on real applications, a pilot study was conducted using EEG data from an epileptic patient and the measures were tested on their ability to detect the propagation of the epileptic activity. Specifically, it was investigated whether changes in the directionality of the information flow occur in the different brain areas of an epileptic patient at two different time periods (states) prior to seizure onset. Extra-cranial multichannel EEG recordings from one patient with back left temporal (LT) lobe epilepsy were used. The measures were estimated and compared on recordings from 60 to 50 min before the seizure onset (early preictal state) and on recordings from 10 min before and up to seizure onset (late preictal state). Two electrodes from the epileptic focus area (LT1

and LT2) were selected in order to investigate the information flow to other areas close or farther from it, i.e. from channels from the occipital area (OC1, OC2), one middle channel (MI) and one right temporal (RT) channel. Therefore, for each of the two periods and for each channel, 20 consecutive non-overlapping time series of 30 sec were derived and all directionality measures were computed for pairs of LT1 and LT2 to all other selected channels. The parameters of the directionality measures were set to be $k = 10$, $W = 10$, $r = 0.15$, $m_x = m_y = 2$, $h = 5$, $\tau_x = \tau_y = 5$, $b = 8$ and $\tau_{\max} = 20$.

Results

Only few causality measures detected a slight change in the information flow from the epileptic focus area to the other brain areas when comparing the period about an hour prior to seizure to the period just before the seizure onset. A slight decrease was observed in the information flow from the focus to the other areas (mostly from LT to OC and MI) moving from early to late preictal state. However, this decrease on information flow could be observed only with some measures, as shown in Fig. 7.4.

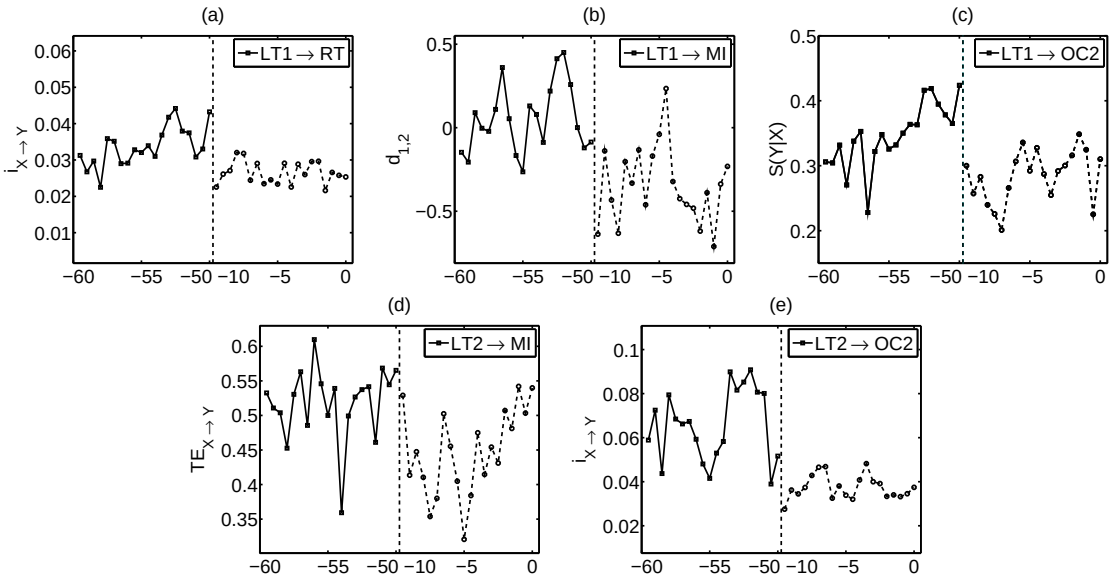

Figure 7.4: Estimated values from the two preictal states of (a) MCMI (histogram based) from channel LT1 to RT, (b) DI from LT1 to MI, (c) NI from LT1 to OC2, (d) TE (histogram based) from LT2 to MI, and (e) MCMI (histogram based) from LT2 to OC2. The states are indicated by the time in sec, with reference to time 0 at seizure onset and they are separated by a vertical dashed line.

7.2.4 Conclusions

From the simulation study, it was educed that information measures and specifically transfer entropy were able to detect the directionality in coupled systems independently of the type of the system (linear or non-linear, map or flow). Moreover, information measures seemed to perform better for short time series lengths than state space and synchronization measures did. The suggested generalized form of TE using reconstructed points and larger time increment ($h > 1$) that is particularly useful for flows, seemed to perform well, however the selection of the embedding parameters seems to be crucial. In the simulations, the standard values as given in the literature for the measure parameters that were used also here, should be further investigated in order to be optimized as significantly effect the performance of the measures. Although transfer entropy seemed to perform better than the other causality measures in the simulation study, in the real application it did not perform any better.

7.3 Improving Statistical Significance of Causality Measures

The evaluation of the causality measures emerged the need for a further investigation of their effectiveness in terms of their free parameters and the state space reconstruction parameters, the time series length and the noise level of the data. Further investigation of the causality measures concluded in a number of modifications for the improvement of their statistical significance.

A main drawback of almost all the causality measures is that for weak causal structures and noisy time series, the detection of the causal effects becomes difficult. Especially for non-identical dynamical systems, the detection of the causal interdependencies should be carefully investigated. When estimating the values of a measure from both directions, it is generally difficult to interpret them without having any additional information about the underlying system. Bias due to finite time series length may be exhibited and this bias may be different in each direction, especially in case of nonidentical interacting systems. Moreover, most causality measures do not give values in the limit of zero in case of no causal effects, and therefore this difference from zero can be misinterpreted as weak coupling.

A possible solution to these problems is provided by testing the significance of the results using surrogate data. Surrogate time series help to rule out spurious conclusions about the existence and the direction of the couplings. When testing for the directionality of the information transfer, the surrogates should replicate the dynamics of each system but destroy any existing coupling among them. After calculating a measure, significance levels are determined using the surrogate data method. The null hypothesis examined determines the type of surrogates one should use (Andrzejak et al., 2003). The null hypothesis in case of the investigation of the causal effects is that two time series are uncoupled. A common technique

to generate surrogates is by altering some of the temporal structures of the time series, so that if the systems are coupled, cause and effect are separated. However, by randomly shuffling each time series, its self-dynamical structures are destroyed, as well.

In order to overcome these problems and establish proper statistical significance, modifications of the causality measures based on the concept of surrogate data test and their estimation scheme are proposed here. The first modification is based on the definition of effective transfer entropy (ETE) (Marschinski and Kantz, 2002), which uses surrogates extracted by randomly shuffling the driving time series (the detailed definition of ETE is given further down). A simple but not insignificant modification of ETE is also suggested here in order to establish proper statistical significance, i.e. the original estimation of ETE uses only one surrogate value, however here a number of surrogates are used.

The second and most important modification of the causality measures is based on the estimation procedure of each measure and the frame of surrogates. The surrogates are extracted by randomly shuffling the reconstructed vectors of the driving time series in order to keep the significant dynamical properties of each system, as the random shuffling of a time series does not preserve the dynamical properties of the underlying system. Therefore, an approach such as of ETE would not work properly unless the driving system is very low-dimensional. The 'corrected' measures are estimated by generating surrogates based on the estimation procedure of each measure. Normalized version of the measures (by division with the standard deviation of the randomly shuffled samples and point surrogates) were also formed.

The effectiveness of the measures in detecting correctly the direction of the information flow will be discussed in the context of unidirectionally coupled chaotic systems with known patterns of interaction. The behavior of each estimator with respect to its free parameter is assessed by Monte Carlo simulations on coupled nonlinear systems (maps and flows) for a range of increasing coupling strengths. In the numerical experiments, the main interest is in the detection of the coupling directionality and strength under different settings of dynamics complexity, time series length, noise level and embedding dimensions for the reconstruction of the two state spaces. All these factors can be sources of bias in the estimation of the directionality measures. Subsequently, the estimators may at cases lack correct statistical significance, which is attempted to be established here.

Synchronization measures were not included in this evaluation. The information measures of mean conditional mutual information and coarse-grained transinformation rate, were rather similar to transfer entropy and therefore were also not included in the study.

7.3.1 Modifications of Causality Measures

Modifications of NI measures

First of all, let us consider the reconstructed points $\mathbf{x}_t$ and $\mathbf{y}_t$ at a time t of two

systems X (driving) and Y (response), with $r_{t,j}$ and $s_{t,j}$, $j = 1, \ldots, k$ being the indexes of their k nearest neighbors.

Bhattacharya's modification

Bhattacharya et al. (2003) proposed a modification that yields all NI measures using inter-point distances. This correction was proposed in order to emphasize the contribution of one system explaining the dynamics of the other system and diminishing its own influence. Specifically, $2k$ neighbors are considered for each point $\mathbf{x}_t$, ordered from smallest to largest distance with time indices $r_{t,j}$, where now $j = 1, \ldots, 2k$. In the computation of the squared mean Euclidean distances involving the k nearest neighbors, only the farthest k neighbors are included, e.g. the term $R_t^{(k)}(X)$ will be replaced by

$$Rb_t^{(k)}(X) = \frac{1}{k} \sum_{j=k+1}^{2k} ||\mathbf{x}_t - \mathbf{x}_{r_{t,j}}||^2. \tag{7.27}$$

The k nearest vectors are discarded in an attempt to attenuate the influence of self-dynamics and increase the influence of the dynamics of the driver. Similarly are also the other terms defined. Indicatively, the modified measure H is defined as

$$\text{HB}_{X \to Y} = \frac{1}{N'} \sum_{t=1}^{N'} \log \frac{R_t(X)}{Rb_t^{(k)}(X|Y)} \tag{7.28}$$

where

$$Rb_t^{(k)}(X|Y) = \frac{1}{k} \sum_{j=k+1}^{2k} ||\mathbf{x}_t - \mathbf{x}_{s_{t,j}}||^2. \tag{7.29}$$

NI measures estimated based on Bhattacharya's modification will be denoted by adding a letter B after the notation of each measure, similarly to HB.

New modifications of NI measures

The 'modified' NI measures A more advanced modification of the estimation algorithm of NI measures is proposed here that aims to eliminate the influence of each time series to itself and also the spurious detection of interdependencies. Specifically, it is suggested to extract the common indexes of the k neighboring points of X and Y and use only the remaining non-common points when computing the conditioned mean squared Euclidean distances. Let us again denote by $ns_{t,j}$ the non common indexes of $r_{t,j}$ and $s_{t,j}$ of the points $\mathbf{x}_t$ and $\mathbf{y}_t$, respectively, whereas j can vary from 0 to k and let us again denote as k^* the number of the non common indexes. For the estimation of the NI measures, $R_t^{(k)}(X|Y)$ is replaced by

$$R_t^{(k^*)}(X|Y) = \frac{1}{k^*} \sum_{j=1}^{k^*} ||\mathbf{x}_t - \mathbf{x}_{ns_{t,j}}||^2. \tag{7.30}$$

Thus, the modified measure H will be defined as

$$\mathrm{HM}_{X\to Y} = \frac{1}{N'}\sum_{t=1}^{N'} \log \frac{R_t(X)}{R_t^{(k^*)}(X|Y)} \tag{7.31}$$

Similarly are all NI measures defined and notated; the letter M will be added after the original notation of each measure, e.g. HM is the corresponding modified measure of H.

The 'corrected' NI measures The second modification of the NI measures defined here is based on the frame of surrogates. For the generation of the surrogates, the indexes of the k-nearest neighbors of any point of the driving system X are replaced with the indexes of the k-nearest neighbors of any other randomly selected point of the state space. This proposed estimation method of surrogates is used instead of randomly shuffling the reconstructed vectors of the driving process and estimating each measure; it is computationally very fast, as the distances between all the points are already known. Specifically, the conditioned terms of NI measures and $R_t^{(k)}(X)$ will be changed, however the terms $R_t(X)$, $R_t(Y)$ and $R_t^{(k)}(Y)$ will not be changed. As an indicative example, the estimation of the term $R_t^{(k)}(X|Y)$ will be given here in detail. Therefore, let us consider again the reconstructed points of the two systems at a time t and let t_i be a randomly selected time for which the respective reconstructed point of the driving time series X is $\mathbf{x}_{t_i}$ and the indexes of its k nearest neighbors are $s_{t_i,j}, j = 1,\ldots,k$. Then, $R_t^{(k)}(X|Y)$ is replaced by

$$R_t^{(k)}(X^*|Y) = \frac{1}{k}\sum_{j=1}^{k} ||\mathbf{x}_t - \mathbf{x}_{s_{t_i,j}}||^2. \tag{7.32}$$

The 'surrogate' value of the H will be then defined as

$$\mathrm{Hs}_{X\to Y} = \frac{1}{N'}\sum_{t=1}^{N'} \log \frac{R_t(X)}{R_t^{(k)}(X^*|Y)} \tag{7.33}$$

The 'corrected' measures are estimated as the difference of the original value of each measure from the mean value of the measure estimated from the surrogates, e.g. the 'corrected' H measure will be defined as

$$\mathrm{CH}_{X\to Y} = \mathrm{H}_{X\to Y} - \frac{1}{M}\sum_{l=1}^{M} \mathrm{Hs}_{l,X\to Y} \tag{7.34}$$

where $\mathrm{Hs}_{l,X\to Y}$ is a surrogate value of H and M is the number of surrogates. Similarly are all NI measures defined. This correction can be also applied on the Bhattacharya's modified measures and the 'modified' measures defined here.

The 'normalized' NI measures Let us denote $SD_{\mathrm{Hs}_{X\to Y}}$ the standard deviation of the M surrogate values of measure H. The respective normalized H measure is defined as the fraction of the 'corrected' one to the standard deviation of the surrogate values

$$\mathrm{nH}_{X\to Y} = \mathrm{CH}_{X\to Y}/SD_{\mathrm{Hs}_{X\to Y}}. \tag{7.35}$$

Similarly are all the normalized NI measures defined and denoted.

Modifications of MCR

The 'effective' MCR The first modification of MCR is based on the frame of surrogates which are extracted by randomly shuffling the reconstructed vectors of the driving time series X. MCR is estimated as in Eq. (7.11), however the joint matrix $J_{i,j}^{X,Y}$ is replaced by the matrix which is extracted after randomly shuffling its lines. Let us denote $Js_{i,j}^{X,Y}$ the new matrix that is extracted after randomly shuffling the lines of $J_{i,j}^{X,Y}$. A 'surrogate' value of MCR is estimated as

$$\mathrm{MCRs}_{X\to Y} = \frac{1}{N'}\sum_{i=1}^{N'}\frac{\sum_{j=1}^{N'} Js_{i,j}^{X,Y}}{\sum_{j=1}^{N'} R_{i,j}^{Y}}. \tag{7.36}$$

The 'effective' MCR is defined as

$$\mathrm{EMCR}_{X\to Y} = \mathrm{MCR}_{X\to Y} - \frac{1}{M}\sum_{l=1}^{M}\mathrm{MCRs}_{l,X\to Y} \tag{7.37}$$

where $\mathrm{MCRs}_{l,X\to Y}$, $l = 1,\ldots,M$ is each surrogate value of the measure.

The 'normalized' MCR The normalized MCR is the extension of the 'effective' MCR and is defined as

$$\mathrm{nMCR}_{X\to Y} = \mathrm{EMCR}_{X\to Y}/SD_{\mathrm{MCRs}_{X\to Y}} \tag{7.38}$$

where $SD_{\mathrm{MCRs}_{X\to Y}}$ is the standard deviation of the M surrogates values of MCR.

The 'corrected' MCR A more sophisticated modification of MCR is defined here. 'Corrected' MCR is estimated again using the frame of surrogates which are extracted by randomly shuffling the reconstructed vectors of the driving time series. The lines of $J_{i,j}^{X,Y}$ are randomly shuffled and M new matrices are extracted denoted by $Js_{l,i,j}^{X,Y}$, where $l = 1,\ldots,M$. Fot the estimation of the corrected MCR (CMCR), the $\sum_{j=1}^{N'} J_{i,j}^{X,Y}$ in Eq.(7.11) is replaced by $\frac{1}{M}\sum_{l=1}^{M}\sum_{j=1}^{N'} Js_{l,i,j}^{X,Y}$. A 'corrected' surrogate value is estimated as

$$\mathrm{MCRs}_{X\to Y} = \frac{1}{N'}\sum_{i=1}^{N'}\frac{\frac{1}{M}\sum_{l=1}^{M}\sum_{j=1}^{N'} Js_{l,i,j}^{X,Y}}{\sum_{j=1}^{N'} R_{i,j}^{Y}}. \tag{7.39}$$

and CMCR is defined as

$$\text{MCR}_{X\to Y} = \text{MCR}_{X\to Y} - \text{MCRs}_{X\to Y}. \tag{7.40}$$

Modifications of TE measures

Effective transfer entropy

Marschinski and Kantz (2002) proposed a modification of transfer entropy (TE), called effective transfer entropy (ETE), defined as the difference of TE computed on the original bivariate time series from the TE computed on a surrogate bivariate time series, where the driving time series X is randomly shuffled

$$\text{ETE}_{X\to Y} = \text{TE}_{X\to Y} - \text{TE}_{X_{\text{shuffled}}\to Y}. \tag{7.41}$$

The use of a randomly shuffled surrogate aims at setting a significant threshold in the estimation of TE. The approach of ETE can be used for the estimation of any other causal measure. Here, for the estimation of ETE, instead of one random permutation of the time series, a number of M random permutations of the driving time series X are considered and therefore $\text{TE}_{X_{\text{shuffled}}\to Y}$ is the mean of the corresponding values from all the permutations:

$$\text{ETE}_{X\to Y} = \text{TE}_{X\to Y} - \frac{1}{M}\sum_{l=1}^{M}\text{TE}_{X_{l,\text{shuffled}}\to Y}, \tag{7.42}$$

where $l = 1, \ldots, M$.

New modifications of TE measures

The 'corrected' TE The corrected TE (CTE), is again defined on the frame of surrogates, but instead of randomly shuffling the driving time series, the reconstructed vectors of the driving process are randomly shuffled. For the estimation of CTE, Eq.(7.21) is used, where the correlation sums $C(y_{t+1}, \mathbf{x}_t, \mathbf{y}_t)$ and $C(\mathbf{x}_t, \mathbf{y}_t)$ are replaced by a mean value estimated after the random shuffling of the reconstructed vectors of the driving time series, which are defined as

$$C_s(y_{t+1}, \mathbf{x}_t, \mathbf{y}_t) = \frac{1}{M}\sum_{l=1}^{M} C(y_{t+1}, \mathbf{x}_{t_l}, \mathbf{y}_t) \tag{7.43}$$

and

$$C_s(\mathbf{x}_t, \mathbf{y}_t) = \frac{1}{M}\sum_{l=1}^{M} C(\mathbf{x}_{t_l}, \mathbf{y}_t) \tag{7.44}$$

where t_l are random time indices. Then, a 'surrogate' TE value is estimated as

$$\text{TEs}_{X\to Y} = \log\frac{C_s(y_{t+1}, \mathbf{x}_t, \mathbf{y}_t)C(\mathbf{y}_t)}{C_s(\mathbf{x}_t, \mathbf{y}_t)C(y_{t+1}, \mathbf{y}_t)}. \tag{7.45}$$

and CTE is defined as

$$\text{CTE}_{X\to Y} = \text{TE}_{X\to Y} - \text{TEs}_{X\to Y} \tag{7.46}$$

Therefore, for the definition of CTE, two different suggestions are combined. Firstly, instead of using surrogates by randomly shuffling the driving time series, surrogates are extracted by randomly shuffling the reconstructed vectors of the driving time series. Secondly, instead of taking the mean of a number of 'surrogate' TE values as in Eq.7.42, a single 'surrogate' TE value is extracted by taking a mean at each component at the estimation formula of TE. In order to investigate the significance of the type of surrogates that are used and of the proposed estimation scheme (by taking mean at each component for the estimation of a 'surrogate' TE value), another two modified measures are considered, CETE and ECTE.

The corrected effective TE (CETE) Let us denote X_l, $l = 1, \ldots, M$, the time series that are generated by randomly shuffling the reconstructed vectors of the time series of the driving system X and $\text{TE}_{X_l\to Y}$ the respective estimated TE values considering X_l instead of the original time series X. Corrected effective transfer entropy (CETE) is defined as the difference of the original TE value from the mean of the corresponding 'surrogate' values of TE, estimated by considering surrogates by randomly shuffling the reconstructed vectors of the driving time series X, as

$$\text{CETE}_{X\to Y} = \text{TE}_{X\to Y} - \frac{1}{M}\sum_{l=1}^{M}\text{TE}_{X_l\to Y}. \tag{7.47}$$

The effective corrected transfer entropy (ECTE) Another modification of TE, effective corrected effective transfer (ECTE), is derived as a CTE, i.e. by considering the mean at each component of TE formula, but the surrogates that are used here are extracted by randomly shuffling the reconstructed vectors of the driving time series. The investigation of the effectiveness of CETE and ECTE will provide information on the significance of the type of surrogate that are used and the estimation procedure that is suggested here by considering the mean at each component of TE in order to estimate a surrogate value of TE.

The 'normalized' TE measures Two more measures are defined, nETE and nCETE, which are the corresponding normalized versions of ETE and CETE, respectively. nETE is defined as

$$nETE = ETE/SD_{\text{TE}_{X_{\text{shuffled}}\to Y}} \tag{7.48}$$

where $SD_{\text{TE}_{X_{\text{shuffled}}\to Y}}$ is the standard deviation of M surrogate values of TE extracted by randomly shuffling the driving time series X. Similarly, nCETE is defined as the fraction of CETE to the standard deviation of the surrogate values of TE extracted by randomly shuffling the reconstructed vectors of the time series X.

Symbolic transfer entropy

Staniek and Lehnertz (2008) derived the symbolic transfer entropy (STE) as the transfer entropy defined on rank-points formed by the reconstructed vectors of X and Y. Thus, for each vector $\mathbf{y}_t$, the ranks of its components assign a rank-point $\hat{\mathbf{y}}_t = [r_1, r_2, \ldots, r_{m_y}]$, where $r_j \in \{1, 2, \ldots, m_y\}$ for $j = 1, \ldots, m_y$. Following this sample-point to rank-point conversion, the sample y_{t+h} in Eq.(7.18) is taken as a rank point at time $t + h$, $\hat{\mathbf{y}}_{t+h}$, and STE is defined as

$$\text{STE}_{X\to Y} = \sum p(\hat{\mathbf{y}}_{t+1}, \hat{\mathbf{x}}_t, \hat{\mathbf{y}}_t) \log \frac{p(\hat{\mathbf{y}}_{t+1}|\hat{\mathbf{x}}_t, \hat{\mathbf{y}}_t)}{p(\hat{\mathbf{y}}_{t+1}|\hat{\mathbf{y}}_t)}. \tag{7.49}$$

STE is estimated here based on Eq.(7.19), defined on the rank-points.

Mew modifications of STE

The 'effective' STE Effective STE (ESTE) is defined similarly to ETE, i.e. as the difference of STE from the mean estimated 'surrogate' STE values extracted from Eq.(7.49) for M random permutations $X_{l,\text{shuffled}}$ of the time series X:

$$\text{ESTE}_{X\to Y} = \text{STE}_{X\to Y} - \frac{1}{M}\sum_{l=1}^{M} \text{STE}_{X_{l,\text{shuffled}}\to Y} \tag{7.50}$$

The 'corrected' STE Corrected STE (CSTE) is defined based on surrogates extracted by randomly shuffling the reconstructed points of the driving time series. The means of $H(\mathbf{x}_t, \mathbf{y}_t)$ and $H(y_{t+1}, \mathbf{x}_t, \mathbf{y}_t)$ are respectively estimated after randomly shuffling the reconstructed points of the driving time series X, as

$$H_s(\mathbf{x}_t, \mathbf{y}_t) = \frac{1}{M}\sum_{l=1}^{M} H(\mathbf{x}_{t_l}, \mathbf{y}_t), \tag{7.51}$$

and

$$H_s(y_{t+1}, \mathbf{x}_t, \mathbf{y}_t) = \frac{1}{M}\sum_{l=1}^{M} H(y_{t+1}, \mathbf{x}_{t_l}, \mathbf{y}_t) \tag{7.52}$$

where $t_l, l = 1, \ldots, M$ are random time indices. Therefore, a mean surrogate value is estimated as

$$\text{STEs}_{X\to Y} = H_s(\mathbf{x}_t, \mathbf{y}_t) - H_s(y_{t+1}, \mathbf{x}_t, \mathbf{y}_t) + H(y_{t+1}, \mathbf{y}_t) - H(\mathbf{y}_t) \tag{7.53}$$

and CSTE is defined as

$$\text{CSTE}_{X\to Y} = \text{STE}_{X\to Y} - \text{STEs}_{X\to Y} \tag{7.54}$$

More modifications of STE Corrected effective symbolic transfer entropy (CESTE) and effective corrected symbolic transfer entropy (ECSTE) are estimated on the rank-points formed by the reconstructed vectors of the driving system, in complete analogy to CETE and ECTE, respectively. For CESTE surrogates are extracted by randomly shuffling the reconstructed vectors of the driving time series, while for ECSTE surrogates are extracted by randomly shuffling the driving time series. Moreover, the two normalized measures nESTE and nCESTE are again the corresponding normalized versions of ESTE and CESTE, respectively, estimated only by diving ESTE and CESTE by the respective standard deviation of the surrogate values of the measures.

7.4 Evaluation of Improved Causality measures on known systems

The evaluation of the measures is assessed by Monte Carlo simulations on nonlinear systems. Directionality coupling measures are computed respectively from 100 realization of unidirectionally coupled systems for increasing coupling strengths, from both directions $X \rightarrow Y$ (the correct dependence direction) and $Y \rightarrow X$. The ability of the measures to detect the degree and the direction of coupling is evaluated from the simulation systems. Analytically, the simulation systems are

- Two unidirectionally coupled Henon maps, as defined in Eq.(7.23) with coupling strengths $c = 0$, 0.15, 0.1, 0.2, 0.3, 0.4 and 0.5.
- Two unidirectionally coupled Mackey-Glass systems

$$\begin{aligned} \frac{\mathrm{d}x}{\mathrm{d}t} &= \frac{0.2x_{t-\Delta_1}}{1+x_{t-\Delta_1}^{10}} - 0.1x_t \\ \frac{\mathrm{d}y}{\mathrm{d}t} &= \frac{0.2y_{t-\Delta_2}}{1+y_{t-\Delta_2}^{10}} + c\frac{0.2x_{t-\Delta_1}}{1+x_{t-\Delta_1}^{10}} - 0.1y_t \end{aligned} \tag{7.55}$$

  for Δ_1 and Δ_2 taking the values 17, 30, 100 (all combinations for the driving and response system) and with coupling strengths $c = 0$, 0.05, 0.1, 0.15, 0.2, 0.3, 0.4 and 0.5.

7.4.1 Set Up

The time series lengths for the coupled Henon maps are $n = 512, 1024, 2048$ and for the Mackey-Glass $n = 2048$. The noise levels are 5% and 20%, given as percentage of the standard deviation of the time series (additive Gaussian noise with mean zero). Apart from the influence of time series length and noise level, we are also interested in examining the influence of the state space reconstruction parameters, and specifically of the embedding dimension selection, on the performance of the measures. Therefore, we investigate this dependence for a range of rational

choices of m_x and m_y. Specifically, for the unidirectionally coupled Henon map system, the embedding dimensions of the two systems vary as $m_x = 1, \ldots, 5$ and $m_y = 1, \ldots, 5$ (25 cases). For the symbolic information measures, the embedding dimensions cannot be set equal to one as there will be no different symbolic patterns and therefore m_x and m_y vary from 2 to 5 (16 cases). For the unidirectionally coupled Mackey-Glass system, with $\Delta_1 = 30$ and $\Delta_2 = 100$, m_x and m_y varied from 3 up to 10 (64 cases). For the other combinations of Δ_1 and Δ_2 taking the values 17, 30, 100 (8 cases), measures were estimated for equal embedding dimensions, m_x and m_y ranging from 3 to 10. This was indicated from the results of the simulation study on both the unidirectionally coupled Henon map and the coupled Mackey-Glass system, with $\Delta_1 = 30$ and $\Delta_2 = 100$. Finally, the number of neighboring points, k, for the state space measures was set to 10, 20 and 40 for the time series lengths $n = 512, 1024$ and 2048, respectively.

7.4.2 Results

In the sequel, we quantify the effect of embedding dimension, time series length and noise, on each of the studied measures. In all the results, X represents the true driving time series and Y the response time series.

Results on NI measures

The possibility of predicting driver-response relationships from asymmetries in NI measures has already been studied, however results are contradictory as far as the effectiveness of these measures is concerned (Arnhold et al., 1999; Quiroga et al., 2000a; Smirnov and Andrzejak, 2005). The modification introduced by Bhattacharya et al. (2003) that aims to emphasize the contribution of the other system and suppress the influence of the original time series to itself does not seem to substantially improve the performance of the measures. It is also claimed that asymmetries in NI measures reflect mainly the different degrees of complexity of the two systems at the level of resolution at which these measures are most sensitive Quiroga et al. (2000a), and that NI measures are somewhat heuristic in their definition, and only provide a rather ad hoc way of evaluating the statistical relationship between time series (Lungarella et al., 2007).

Results on the unidirectionally coupled Henon maps

The simulation study on the unidirectionally coupled Henon maps showed the effectiveness of the NI measures in detecting the direction of the causal effects. The measure H (see Eq.(7.5)) detected correctly the direction of the information flow for almost all combinations of the embedding dimensions but for $m_y = 1$, even for small time series lengths. The mean estimated values of $\mathrm{H}_{Y \to X}$ were correctly in the limit of zero for all coupling strengths up to $c = 0.5$, and increased for larger values of c. The modified measure HB (Bhattacharya's correction) almost

coincided with H at all cases and therefore wont be separately discussed. On the other hand, HM (the proposed modification by excluding the common neighbors) was also effective in detecting the direction of the interdependence giving slightly lower values than H. An advantage of HM was observed for $c = 0.6$ and 0.7, as $\mathrm{HM}_{Y\to X}$ was closer to the limit of zero than $\mathrm{H}_{Y\to X}$. The performance of the three measures H, HB and HM was not significantly affected by the time series length or the addition of noise. Indicative results of the performance of H and HM are displayed in Fig.7.5.

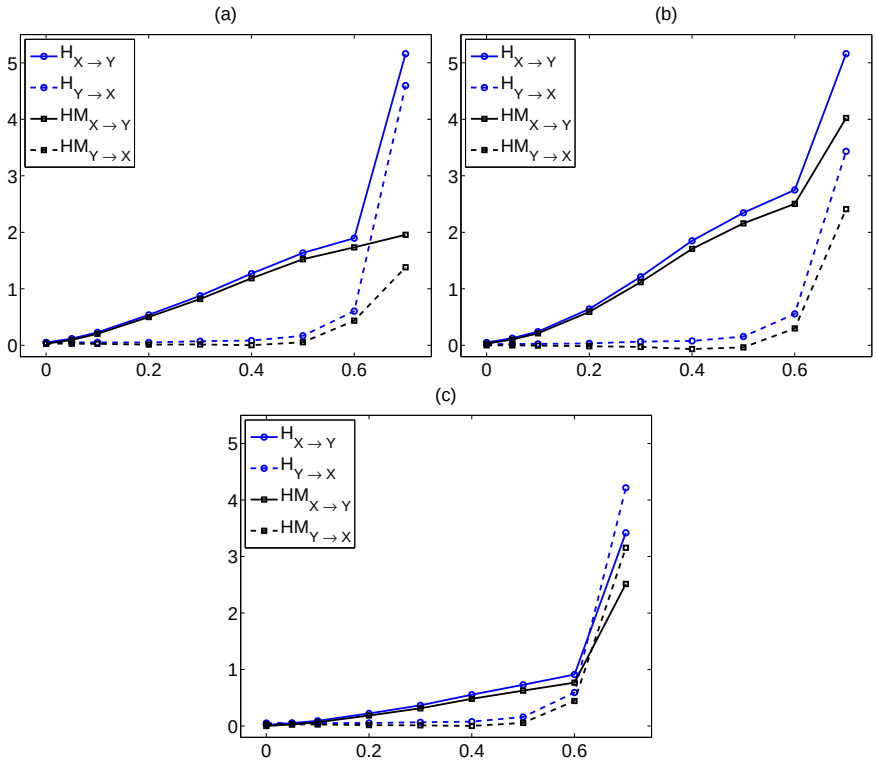

Figure 7.5: (a) Mean estimated values of H and HM for both directions from 100 realizations of the noise-free unidirectionally coupled Henon map with $n = 512$ and $m_x = 2, m_y = 2$. (b) As in (a) but for $m_x = 2$ and $m_y = 5$. (c) As (a) but for $m_x = 5$ and $m_y = 2$.

The corrected measures CH, CHB and CHM seem to worsen the performance of the measures as the range of their estimated values were not in the limit of zero for the uncoupled case and for the direction $Y \to X$. The normalized measures nH, nHB and nHM, gave similar results with H, HB and HM, respectively and only the range of the estimated values differed. Again, the estimated values of the original values and Bhattacharya's modification coincided. An indicative result of the performance of the corrected and normalized H measures is displayed in Fig.7.6.

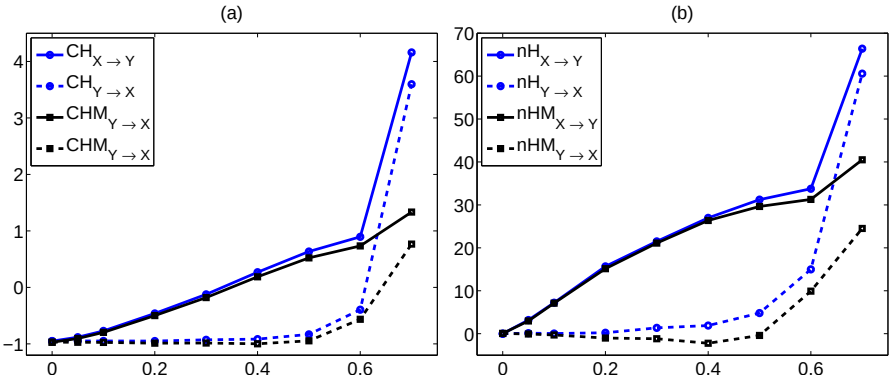

Figure 7.6: (a) Mean estimated values of CH and CHM for both directions from 100 realizations of the noise-free unidirectionally coupled Henon map with $n = 512$ and $m_x = 2$, $m_y = 2$. (b) As in (a) but for the normalized measures nH and nHM.

The measure S (see Eq.(7.4)) and the corresponding modifications of it, proved to be ineffective in detecting any causal effect. The estimated values of the measures S from both directions were very small for all coupling strengths but $c = 0.7$. However the estimated values on the direction $X \rightarrow Y$ were larger than the opposite direction. The values for both directions increased with c. Corrected measures performed similarly, giving again negative instead of zero for $c = 0$. However, the normalized measures seem to detect correctly the causal effects. The slight differences of the estimated values of the measures from the two directions are revealed as measures are extracted by dividing with standard deviations which are also very small, and therefore normalized measures attain high values. Results for $n = 2048$ for all variants of the S measure and for $m_x = 2$ and $m_y = 2$ are presented in Fig.7.7.

The measure L (see Eq.(7.10)) proved to be effective in detecting the direction of the information flow, even for small time series lengths and large noise levels. However, for larger c, although the causal effects are correctly detected, the estimated values of L from the direction $Y \rightarrow X$ also increase with c. LB coincides again with L. LM seems to significantly improve the performance of the measure but only for coupling strengths $c = 0.2 - 0.5$, as for $c \geq 0.5$ $\mathrm{LM}_{Y \rightarrow X}$ increases with c. In Fig.7.8, results for $n = 512$ are displayed for different embedding dimensions and noise levels.

The corrected measures give again values around zero for $c = 0$ and for the direction $Y \rightarrow X$, however the mean estimated values of the correct direction $X \rightarrow Y$ are higher than the opposite one. The normalized measures are effective, and specifically nLM detect correctly the direction of the interdependencies. The advantage of nLM over nL (and nLB) is the much slower increase of the estimated values of it at the direction $Y \rightarrow X$ as c increases (see Fig.7.9).

Measure M (see Eq.(7.8)) and the corresponding modifications of it perform

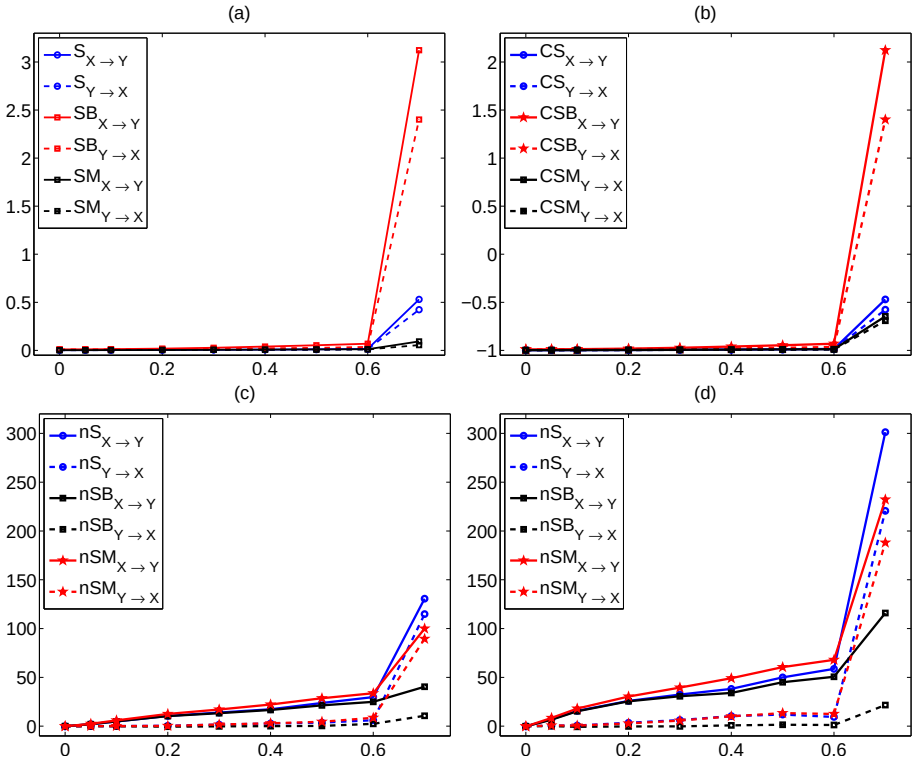

Figure 7.7: (a) Mean estimated values of S, SB and SM for both directions from 100 realizations of the noise-free unidirectionally coupled Henon map with $n = 2048$ and $m_x = 2$, $m_y = 2$. (b) As in (a) but for CS, CSB and CSM. (c) As (a) but for nS, nSB and nSM and $n = 512$. (d) As in (c) but for $n = 2048$.

similarly to L. The direction of the information flow is correctly detected for all embedding dimensions, and the estimated values of the measures from the direction $Y \rightarrow X$ also increase with c. The suggested modified measures improve the performance of the measures, as previously. The corrected and normalized measures also perform similarly to the corresponding L measures. The same conclusions hold for measure N and its modifications, whereas the increase at the direction $Y \rightarrow X$ is observed for $c > 0.5$.

Results on the unidirectionally coupled Mackey-Glass system

The results on the unidirectionally coupled Mackey-Glass system showed that the performance of NI measures (and their modifications) is not invariant to the underlying systems that are examined. There was no consistency in the performance of all measures as some performed better and some worse compared to the previously discussed results on the unidirectionally coupled Henon maps. All NI measures (original and the two modifications which are not based on surrogates) detected the causal effects for most of the combinations of the embedding dimensions for

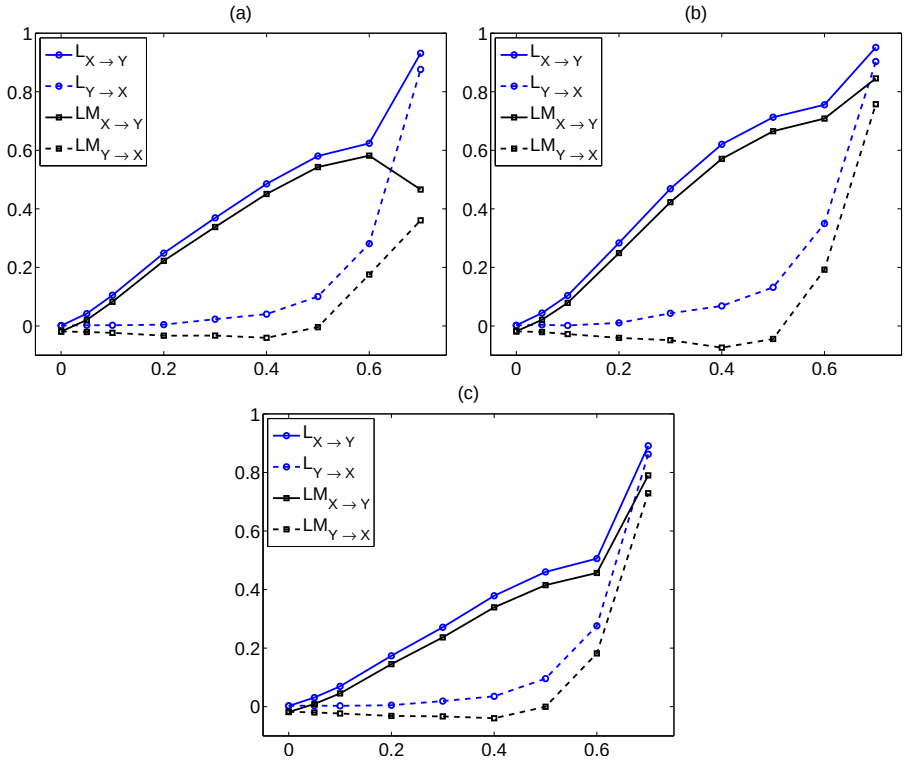

Figure 7.8: (a) Mean estimated values of L and LM for both directions from 100 realizations of the noise-free unidirectionally coupled Henon map with $n = 512$ and $m_x = 2$, $m_y = 2$. (b) As in (a) but for $m_x = 2$ and $m_y = 5$. (c) As in (b) but for $m_x = 5$ and $m_y = 2$.

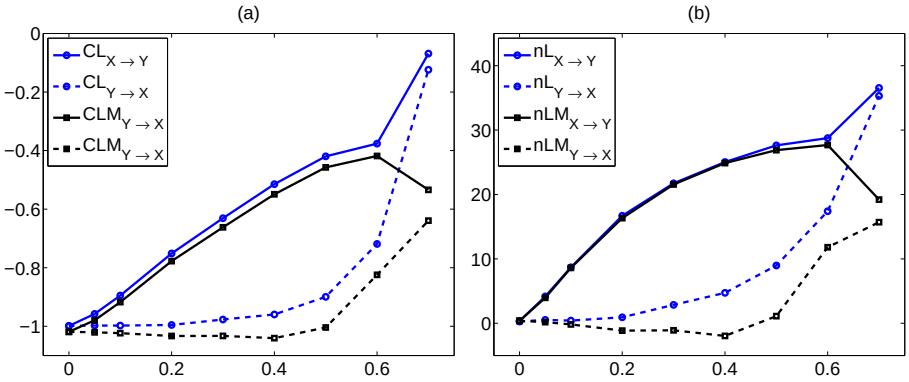

Figure 7.9: (a) Mean estimated values of CL and CLM for both directions from 100 realizations of the noise-free unidirectionally coupled Henon map with $n = 512$ and $m_x = 2$, $m_y = 2$. (b) As in (a) but for nL and nLM.

$m_x \leq m_y$, however the estimated values of the measures from both directions increased with c (see Fig.7.10). The increase at the direction $Y \rightarrow X$ is observed even for very small coupling strengths as $c = 0.01$ (see Fig.7.10d), while for the Henon map it was observed only for strong causal effects ($c > 0.4$) and for the modified measures, such an increase was observed close to synchronization. The S measures (original and modified) that were previously found to be ineffective, however in the case of the coupled Mackey-Glass systems, correctly detected the direction of the interdependencies for all m_x and m_y. The drawback of the false increase at the $Y \rightarrow X$ direction was however again observed at all cases. Indicative results on the unidirectionally coupled Mackey-Glass systems with $\Delta_1 = 30$ (driving system) and $\Delta_2 = 100$ (response system) are displayed in Fig.7.10. Addition

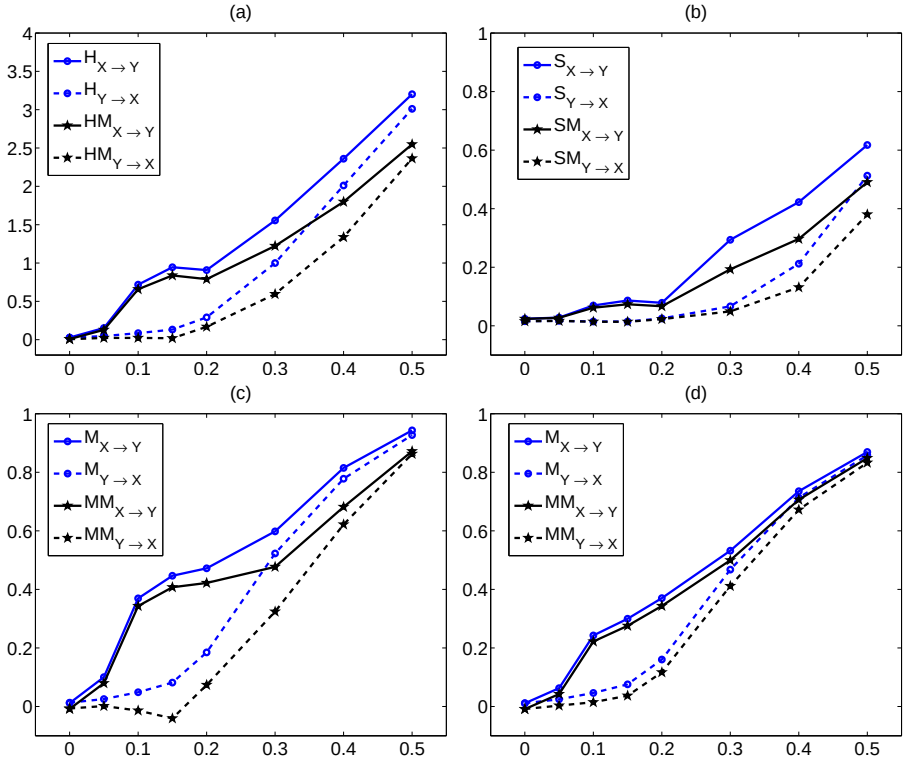

Figure 7.10: (a) Mean estimated values of H and HM for both directions from 100 realizations of the noise-free unidirectionally coupled Mackey-Glass systems ($\Delta_1 = 30$ for the driving and $\Delta_2 = 100$ for the response system) with $n = 2048$ and $m_x = m_y = 3$. (b) As in (a) but for S and SM, and $m_x = m_y = 4$. (c) As in (a) but for M and MM and $m_x = m_y = 4$. (d) As in (c) but for 20% noise level.

of high noise levels affects the performance of the measures and the problem of the false increase at the direction $Y \rightarrow X$ deteriorated. The performance of the corrected measures is the same as for the unidirectionally coupled Henon maps, giving negatively biased values (as previously around -1) in case of no causal effects. The

normalized measures performed similarly to the corresponding original measures, and differ only on the scaling of the estimated values. Results for different Δ_1 and Δ_2 values are similar, whereas the drawback of the false increase of causal effects on the $Y \rightarrow X$ direction is more obvious.

Results on MCR measures

The measure MCR, similarly to NI measures, seems to be able to detect the direction of the information flow for certain parameter selections. Selection of the embedding dimensions, time series length and noise level are factors that significantly affect the performance of the measure. A substantial problem of MCR is that it increases also in the opposite direction (with no causal effect) with the increase of the coupling strength. Positively biased estimates of MCR are not clearly indicating the existence of causal effects, i.e. equal positive values of the measure may be interpreted as bidirectional causality in cases of no causal effects.

Results on the unidirectionally coupled Henon maps

The direction of the information flow was detected with MCR only for $m_x \leq m_y$, independently of n. Although MCR detected correctly the direction of the information flow, the estimated values of the measure were not always in the limits of zero when no coupling existed, whereas MCR was positively biased particularly for small time series and large embedding dimensions. The problem of the false increase of the estimated values of the NI measures at the direction $Y \rightarrow X$ is also observed for MCR but it was not that apparent.

The proposed corrected measure, CMCR, detected correctly the direction of the information flow only for $m_x \leq m_y$, and thus it was also affected by the selection of the embedding dimensions. However, CMCR seemed to be less biased than MCR, giving values closer to the zero level than MCR when no coupling existed (see Fig.7.11). By increasing the time series length and the embedding dimensions, MCR and CMCR converged to the same values. CMCR seems to be more effective than MCR, particularly for small time series and for small values of the embedding dimensions. EMCR coincided with CMCR (differences in the order of 10^{-10}), for all time series lengths and combinations of the embedding parameters. Therefore, results are not displayed for EMCR. The estimated normalized measure, nMCR, could not effectively detect the direction of the information low. The estimated standard deviation of the surrogates were very small and therefore nMCR gave very large values, as in the case of normalized NI measures. Indicative results for the unidirectionally coupled Henon map are displayed in Fig.7.11.

Addition of small noise levels did not substantially affect MCR and CMCR, even for small time series lengths (see Fig.7.12a). However, addition of larger noise level (20%) substantially affected the performance of the measures (see Fig.7.12b), whereas the estimated values of MCR and CMCR from the two directions are closer, and specifically for small c no detection of the causal effects is posible. The

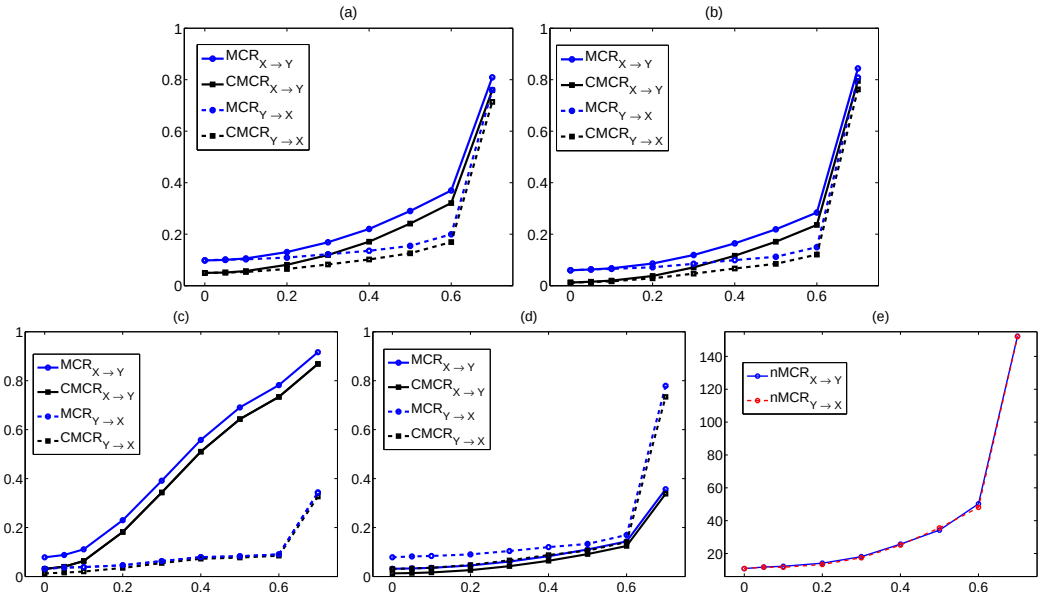

Figure 7.11: (a) Mean estimated values of MCR and CMCR for both directions from 100 realizations of the noise-free unidirectionally coupled Henon map with $n = 512$ and $m_x = m_y = 2$. (b) As in (a) but for $n = 2048$. (c) As in (b) but for $m_x = 2$ and $m_y = 5$. (d) As in (b) but for and $m_x = 5, m_y = 2$. (e) As in (b) but for nMCR.

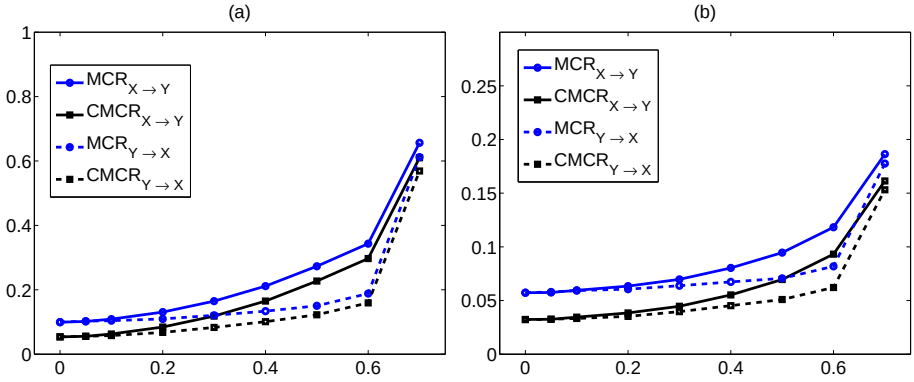

Figure 7.12: (a) Mean estimated values of MCR and CMCR for both directions from 100 realizations of the unidirectionally coupled Henon maps with $n = 512$ and noise of 5% noise level, and $m_x = m_y = 2$. (b) As in (a) but for $n = 2048$ and 20% noise level.

estimated measure profiles for the different coupling strengths increase in magnitude with the embedding dimensions. For $c = 0$, only for $m_x = m_y$ were the estimated values of the measures from the two direction equal, giving values larger than zero as was expected.

Results on the unidirectionally coupled Mackey-Glass systems

The detection of the direction of flow in the case of the unidirectionally coupled Mackey-Glass systems was not established with MCR measures. For $\Delta_1 = 30$ and $\Delta_2 = 100$, the estimated values of the two measures were almost equal for all embedding dimensions. The mean estimated values of the measures at the direction $X \rightarrow Y$ were larger than on the opposite direction only for $m_x \leq m_y$, however the estimated values of both measures increased dramatically with the coupling strength at both directions (see Fig.7.13a and b). Mean estimated MCR values were positively biased and increased with the increase of the embedding dimensions (increase of either m_x or m_y or both). The performance of the measures on the

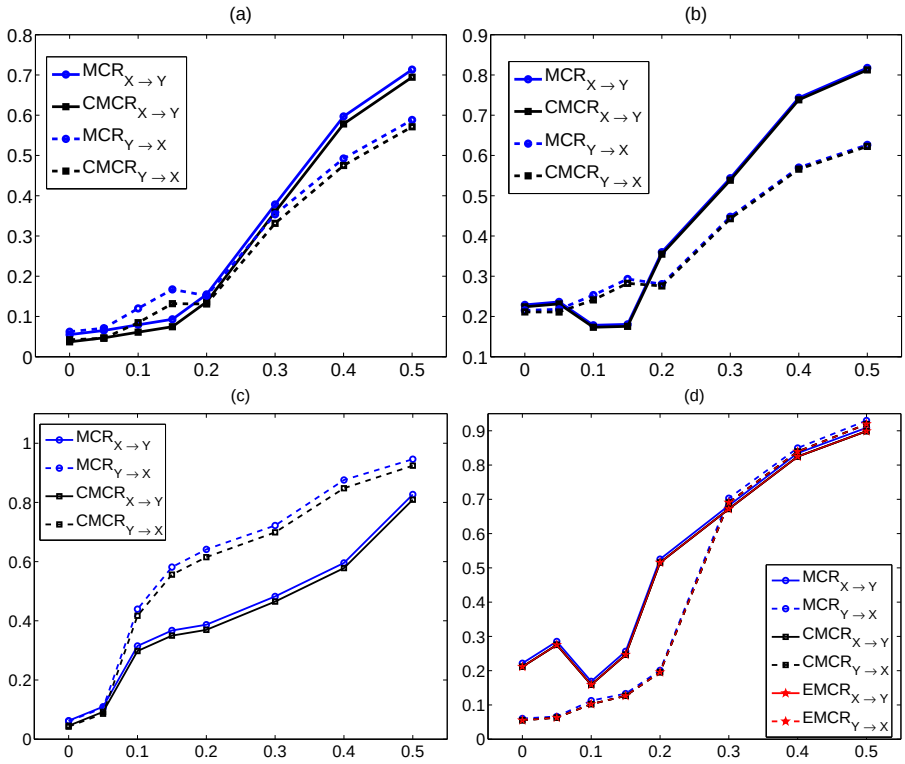

Figure 7.13: (a) Mean estimated values of MCR and CMCR for both directions from 100 realizations of the noise-free unidirectionally coupled Mackey-Glass systems ($\Delta_1 = 30$ for the driving and $\Delta_2 = 100$ for the response system) with $n = 2048$ and $m_x = m_y = 3$. (b) As in (a) but for $m_x = m_y = 6$. (c) As in (a) but for $\Delta_1 = 17$ and $\Delta_2 = 17$. (d) As in (b) but the mean estimated values of EMCR are also displayed and $\Delta_1 = 17$ and $\Delta_2 = 30$.

different unidirectionally coupled Mackey-Glass systems, and specifically when $\Delta_1 = \Delta_2$ was very poor. MCR and CMCR still were ineffective in detecting correctly the causal effects.

Results on TE measures

TE measures and mainly CTE were more effective in detecting the direction of the information flow than the state space measures. The causal relationships were correctly detected for all the simulation systems, although the estimated values of the measures were positively biased at some cases.

Results on the unidirectionally coupled Henon maps

TE and CTE correctly detected the direction of the information flow for $m_x \geq m_y$, contrary to NI and MCR. The estimated values of the measures from both directions were in the limits of zero for $c = 0$ for the most combinations of the embedding dimensions, even at small time series lengths, and specifically at the $Y \rightarrow X$ direction, values were in the limit of zero for all coupling strengths when $m_x \geq m_y$. On the other hand, ETE gave negative values for $c = 0$ for most embedding dimensions, and the mean estimated values of ETE from the two directions were not always equal for $c = 0$. ETE is affected by the selection of the embedding dimensions much more than TE and CTE. Indicative results on the unidirectionally coupled Henon maps are displayed in Fig.7.14.

The estimated values of the measures decrease with the addition of a small amount of noise only ETE seems to be significantly affected even for small noise levels contrary to TE and CTE that are not significantly influenced. However, addition of high noise levels affects the effectiveness of all measures in detecting the causal effects (see Fig.7.15).

The mean estimated values of ECTE for the direction $X \rightarrow Y$ are larger than the values for the opposite direction when $m_x \geq m_y$, indicating that the measure detects the direction of the information flow. However, ECTE is negatively biased as is ETE as well, i.e. for $c = 0$ the values of the measures are negative and $\text{ECTE}_{Y \rightarrow X}$ is negative for all c and almost all embedding dimensions. With the addition of high noise level (20%), the estimated values of the measure are increased and the measure performs better as $\text{ECTE}_{X \rightarrow Y}$ =$\text{ECTE}_{Y \rightarrow X} \simeq 0$ for $c = 0$, but only for some embedding dimensions (see Fig.7.16). On the other hand, CETE seems to perform similarly to CTE as the direction of the information flow was correctly detected and the estimated values of CETE from both directions were in the limit of zero for $c = 0$ for $m_x \geq m_y$ (see Fig.7.16). With addition of high noise (20%), CETE was still effective in detecting the causal effects for most embedding dimensions under the condition $m_x \geq m_y$. Here, smaller embedding dimensions should be considered for the effective performance of the measure compared to the embedding dimensions that seem to be effective for CTE. Therefore, the modifications of TE based on surrogates generated by randomly shuffling the reconstructed vectors of the driving system, i.e. CTE and CETE, perform better than the modifications of TE based on surrogates generated by randomly shuffling the time series of the driving system. However, the estimation procedure is not that significantly affecting the effectiveness of the measures, i.e. considering the mean

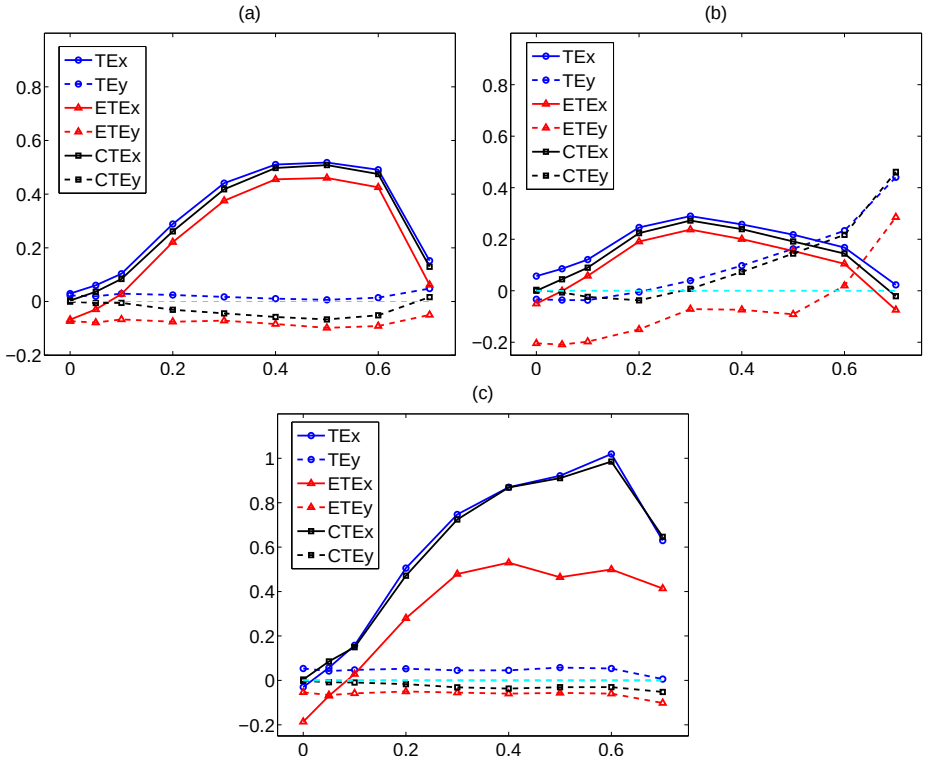

Figure 7.14: (a) Mean estimated values of TE, ETE and CTE for both directions from 100 realizations of the noise-free unidirectionally coupled Henon maps, with $n = 512$ and $m_x = m_y = 2$. (b) and (c) as in (a) but for $m_x = 2, m_y = 4$ and $m_x = 4, m_y = 2$, respectively.

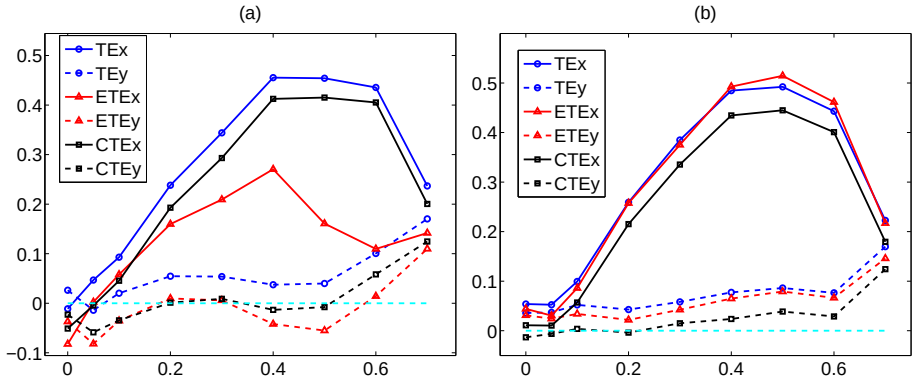

Figure 7.15: (a) Mean estimated values of TE, ETE and CTE for both directions from 100 realizations of the unidirectionally coupled Henon maps with 20% noise, $n = 512$ and $m_x = m_y = 3$. (b) As in (a) but for $n = 1024$.

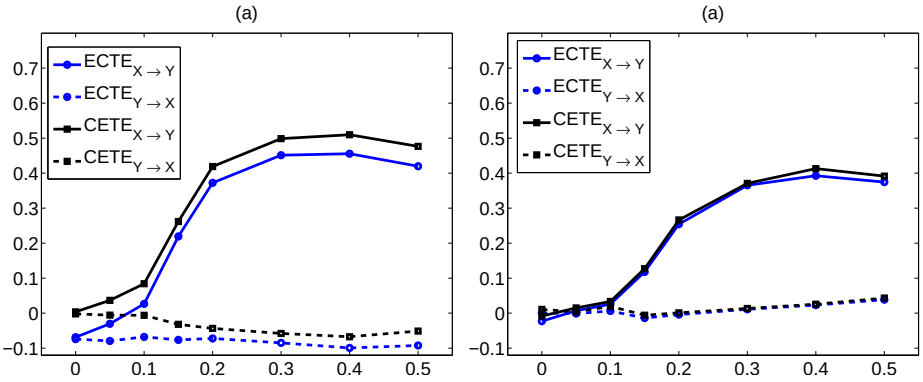

Figure 7.16: (a) Mean estimated values of ECTE and CETE for both directions from 100 realizations of the unidirectionally coupled Henon maps with $n = 512$ and $m_x = m_y = 2$. (b) As in (a) but for 20% noise.

at each component of the measure as for the estimation of CTE or the overall mean from surrogates as for the estimation of ETE.

Finally, the normalized measures nETE and nCETE detected correctly the direction of the causal effects, also for embedding dimensions that did not satisfied the condition $m_x \geq m_y$. However, the mean estimated values of the measures were not systematically equal for $c = 0$ or at the limit of zero (see Fig.7.17). Both measures were significantly affected by the selection of the embedding parameters, with nETE being more sensitive on the time series length. Addition of high level of noise improved the performance of the measures, as the estimated values of the measures from the two directions converged to the limit of zero for $c = 0$ (see Fig.7.17d).

Results on the unidirectionally coupled Mackey-Glass systems

In case of the unidirectionally coupled Mackey-Glass systems with $\Delta_1 = 30$ for the driving system and $\Delta_2 = 100$ for the response system, TE correctly detected the direction of the information flow for $m_x \geq m_y$, and $m_x = m_y$ seems as the most appropriate choice for the embedding dimensions. In case of no causal effects ($c = 0$), TE values from both directions were in the same range and in the limits of zero for $m_x \geq m_y$. ETE was found to be ineffective is case of the unidirectionally coupled Mackey-Glass systems; the direction of the information flow was detected only for very few embedding dimensions and for most embedding dimensions ETE$x \neq$ ETEy for $c = 0$. CTE outperformed again TE in detecting correctly the direction of the information flow for $m_x \geq m_y$. Addition of high level of noise, spoiled the effectiveness of the measure only in the case of small time series and large embedding dimensions ($m_x, m_y = 4$ or 5). CTE was more sensitive on the selection of the embedding dimensions in the case of noisy time series. Indicative results on the unidirectionally coupled Mackey-Glass systems

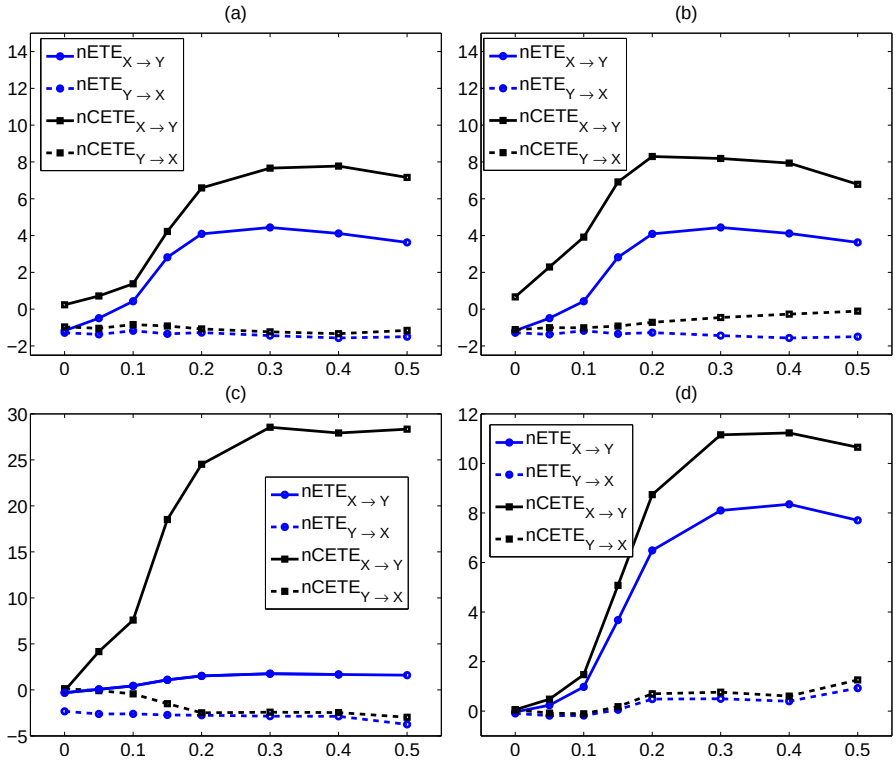

Figure 7.17: (a) Mean estimated values of nETE and nCETE for both directions from 100 realizations of the unidirectionally coupled Henon maps with $n = 512$ and $m_x = m_y = 2$. (b) As in (a) but for $n = 2048$. (c) As in (b) but for $m_x = 5, m_y = 2$. (d) As in (c) but for 20% noise level.

are displayed in Fig.7.18.

CETE was again effective for $m_x \geq m_y$, giving values around zero for $c = 0$ for almost all embedding dimensions. On the other hand, ECTE was very sensitive on the selection of the embedding dimensions and only in few cases correctly detected the direction of the causal effects or gave values around zero for $c = 0$ (see Fig.7.19). Addition of high level of noise affects the effectiveness of both measures; CETE was not that effective for $c = 0$ giving either negative values or slightly larger values at the $Y \rightarrow X$.

nETE detected correctly the causal effects only for pair ($m_x = m_y = 3$), ($m_x = m_y = 4$), ($m_x = 4, m_y = 3$) and ($m_x = 5, m_y = 3$). Addition of high level of noise, significantly affected the performance of both measures; nCETE was found to be effective only for few embedding dimensions whereas m_x and m_y were rather small (see Fig.7.20).

The results on the unidirectionally coupled Mackey-Glass systems with different Δ_1 and Δ_2, were in agreement with the previously observed results. CTE and CETE again outperformed the other measures, whereas ETE and nETE were again

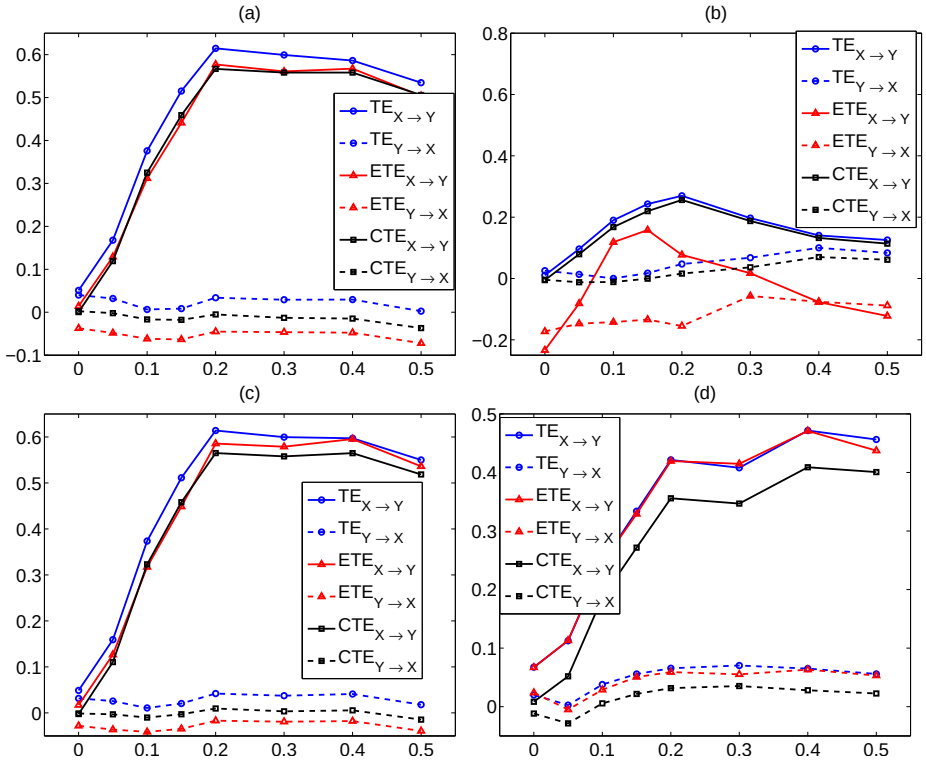

Figure 7.18: (a) Mean estimated values of TE, ETE and CTE for both directions from 100 realizations of the noise-free unidirectionally coupled Mackey-Glass systems ($\Delta = 30$ for the driving and $\Delta = 100$ for the response system) with $n = 2048$ and $m_x = 4$ and $m_y = 3$. (b) As in (a) but for $m_x = 5$ and $m_y = 5$. (c) As in (a) but for 5% noise level. (d) As in (c) but for 20% noise level.

found to be ineffective at most cases.

The performance of all the measures was worse in the case of identical unidirectionally coupled Mackey-Glass systems. It is noted that the performance of all the measures was worse in the case of identical unidirectionally coupled Mackey-Glass, i.e. in case $\Delta_1 = \Delta_2$, probably due to the fact that the systems have the same phases and get synchronized very fast, i.e. the trajectories of the driven system converges to the trajectories of the response system faster while the rate of convergence when the systems are of different complexity is slower. However, the same range of coupling strengths were considered for all systems in order to have comparable results for each coupling strength. Indicative results on the performance of TE, ETE and CTE on coupled Mackey-Glass systems with $\Delta_1 = \Delta_2 = 17$ and $\Delta_1 = \Delta_2 = 30$, and $\Delta_1 = 17$, $\Delta_2 = 100$, respectively, are displayed in Fig.7.21.

Overall, the performance of the measures were generally consistent in the examined systems, indicating the superiority of CTE and CETE compared to the other TE measures. These two modifications of TE which are based on the introduced

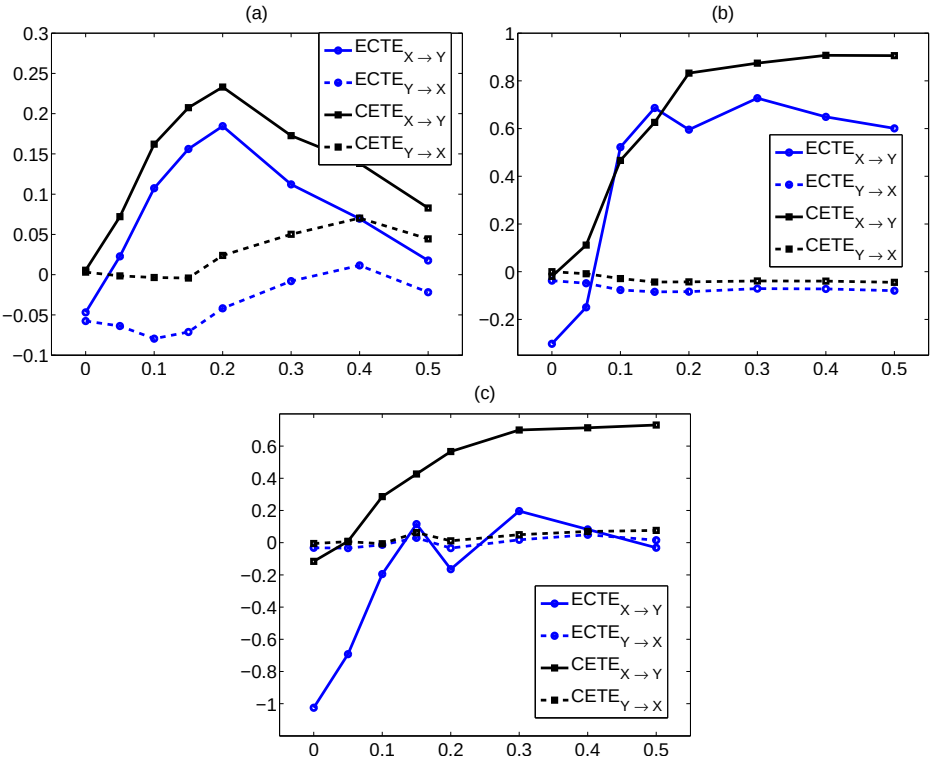

Figure 7.19: (a) Mean estimated values of CETE and ECTE for both directions from 100 realizations of the noise-free unidirectionally coupled Mackey-Glass systems ($\Delta_1 = 30$ for the driving and $\Delta_2 = 100$ for the response system) with $n = 2048$ and $m_x = m_y = 4$. (b) As in (a) but for $m_x = 7$ and $m_y = 3$. (c) As in (b) but for 20% noise level.

surrogates (extracted by randomly shuffling the reconstructed vectors of the driving time series) performed similarly. Their main advantage is in case of no causal effects ($c = 0$), whereas in most cases gave values in the limit of zero, while TE was positively biased, and ETE and ECTE were negatively biased or their mean estimated values from the two direction were not equal.

Results on STE measures

Measures based on symbolic transfer entropy seem to be more affected by the selection of the embedding dimension than the previous measures. The modified measures based on surrogates extracted by randomly shuffling the reconstructed vectors of the driving time series performed similarly; they were less sensitive in the noise and gave the most consistent results in case of no causal effects giving values in the limits of zero for $c = 0$.

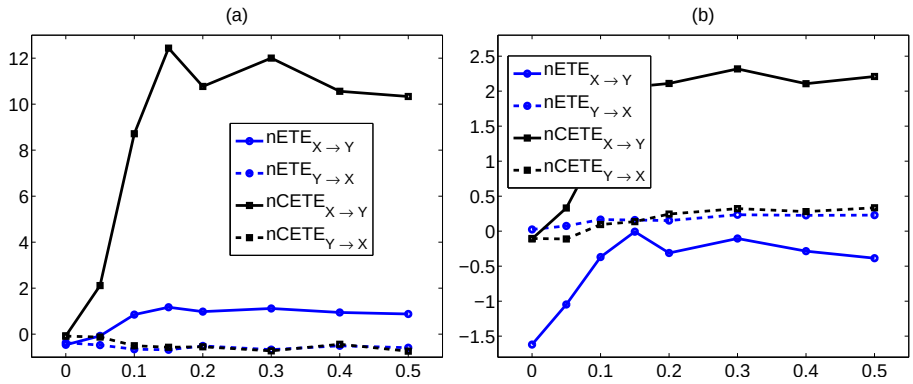

Figure 7.20: (a) Mean estimated values of nETE and nCETE for both directions from 100 realizations of the noise-free unidirectionally coupled Mackey-Glass systems ($\Delta_1 = 30$ for the driving and $\Delta_2 = 100$ for the response system) with $n = 2048$ and $m_x = 5$ and $m_y = 3$. (b) As in (a) but for 20% noise level.

Results on the unidirectionally coupled Henon maps

STE, ESTE and CSTE detected correctly the direction of information flow for $m_x \geq m_y > 2$ in the case of the unidirectionally coupled Henon maps, for all time series lengths. STE was positively biased for embedding dimensions larger than 3, whereas ESTE on the other hand was negatively biased. CSTE was more effective compared to ESTE and STE. Moreover, $\text{CSTE}_{X\to Y} = \text{CSTE}_{Y\to X} \simeq 0$ for $c = 0$ for all embedding dimensions. Indicative results are presented in Fig.7.22. CSTE was robust against noise, as addition of high level noise did not substantially influence its performance, even for small n (see Fig.7.22c).

ECSTE was not effective in detecting the direction of the information flow and in most cases its estimated values were negatively biased, ans especially for small time series lengths. However, for larger time series lengths, its performance was improved indicating correctly the direction of the information flow. On the other hand, CESTE correctly detected the causal effects and for $m_x \geq m_y$ and $m_x, m_y < 5$, gave values at the limit of zero. It was also insignificantly affected by the addition of noise (see Fig.7.23).

The normalized measures performed similarly to the corresponding normalized measures of TE. nESTE was significantly affected by the time series length, the embedding dimensions and the noise level of the time series. Only large noise-free time series and for embedding dimensions smaller than 4 and for $m_x \geq m_y$, was the direction of the information flow correctly detected. On the other hand, nCESTE was effective even for small time series lengths and noisy time series for most cases for $m_x \geq m_y$ (see Fig.7.24).

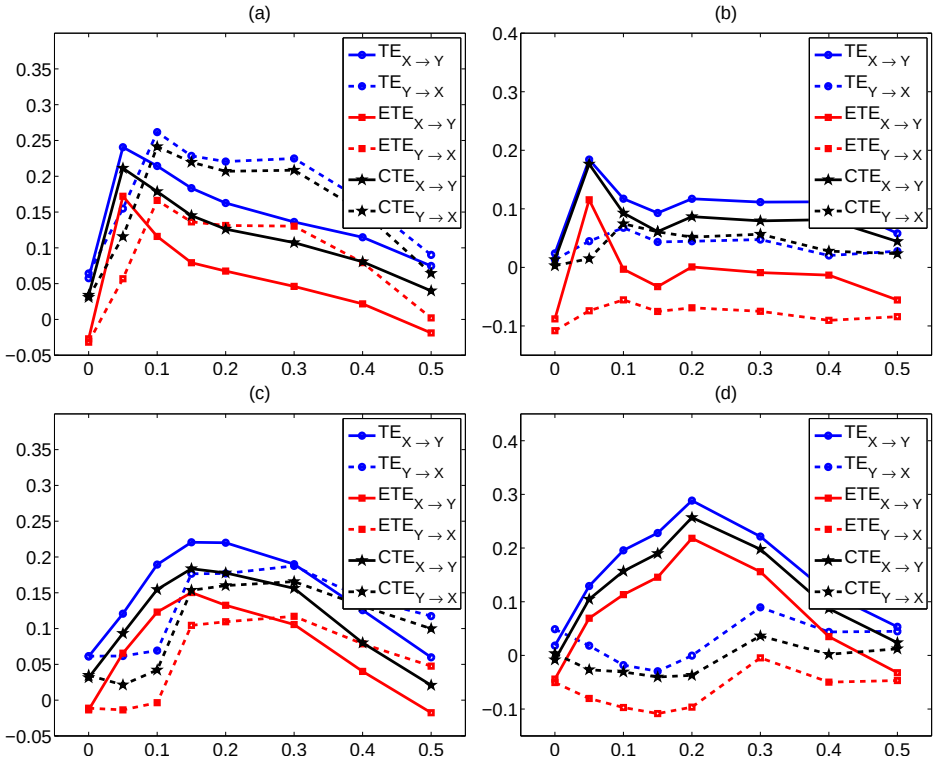

Figure 7.21: (a) Mean estimated values of TE, ETE and CTE for both directions from 100 realizations of the noise-free unidirectionally coupled Mackey-Glass systems with $\Delta_1 = \Delta_2 = 17$, $n = 2048$ and $m_x = m_y = 3$. (b) As in (a) but for $m_x = m_y = 4$. (c) As in (a) but for $\Delta_1 = \Delta_2 = 30$. (d) As in (a) but for $\Delta_1 = 17$ and $\Delta_2 = 100$.

Results on the unidirectionally coupled Mackey-Glass systems

The results on the unidirectionally coupled Mackey-Glass systems with $\Delta_1 = 30$ and $\Delta_2 = 100$, do not substantially differentiate from previous results. CSTE, CESTE and nCSTE outperformed the other measures, correctly detecting the direction of the causal effects for almost all cases for $m_x \geq m_y$. Their main advantage was in the case of no causal effects ($c = 0$), giving values around zero for all embedding dimensions, even in cases of incorrectly detecting the direction of the information flow. CESTE and ECSTE performed similarly to CESTE and ESTE, respectively. Indicative results on the performance of STE, ESTE and CSTE are displayed in Fig.7.25. nESTE has not detected correctly the direction of the information flow, while nCESTE performed similarly to CESTE, but just in a different scale of values.

The results on the unidirectionally coupled Mackey-Glass systems with different Δ_1 and Δ_2 showed the insufficiency of the measures. CSTE and CESTE were again the only measures to give values closer to zero for $c = 0$. However, even

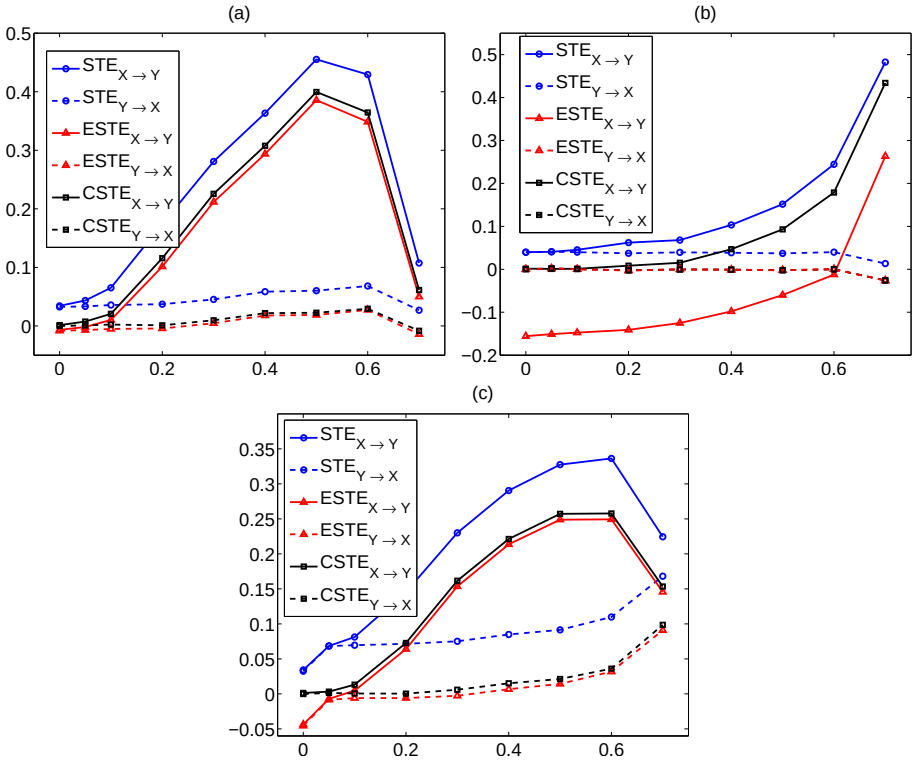

Figure 7.22: (a) Mean estimated values of STE, ESTE and CSTE for both directions from 100 realizations of the noise-free unidirectionally coupled Henon maps, with $n = 512$ and $m_x = 3, m_y = 3$. (b) As in (a) but $m_x = 5$ and $m_y = 2$. (c) As in (a) but for 20% additive noise.

these measures were unable to detect correctly the causal effects for most of the selected embedding dimensions (see Fig.7.26).

7.4.3 Quantitative results from the evaluation of the causality measures

The previous comparisons were rather qualitative and results were presented mainly by graphs. In order to obtain also quantitative summary results for the performance of the measures, t-tests for means were conducted on the 100 samples of the estimated measures from each of the two directions in order to examine whether there is a significant discrimination of the two directions. Two main settings were examined.

The first setting aims at assessing the ability of a measure to detect a causal effect only when is present and therefore it is examined whether it gives values in the limit of zero when there are no interactions among the examined systems. Therefore, we test the null hypothesis that the two systems are uncoupled and expect to

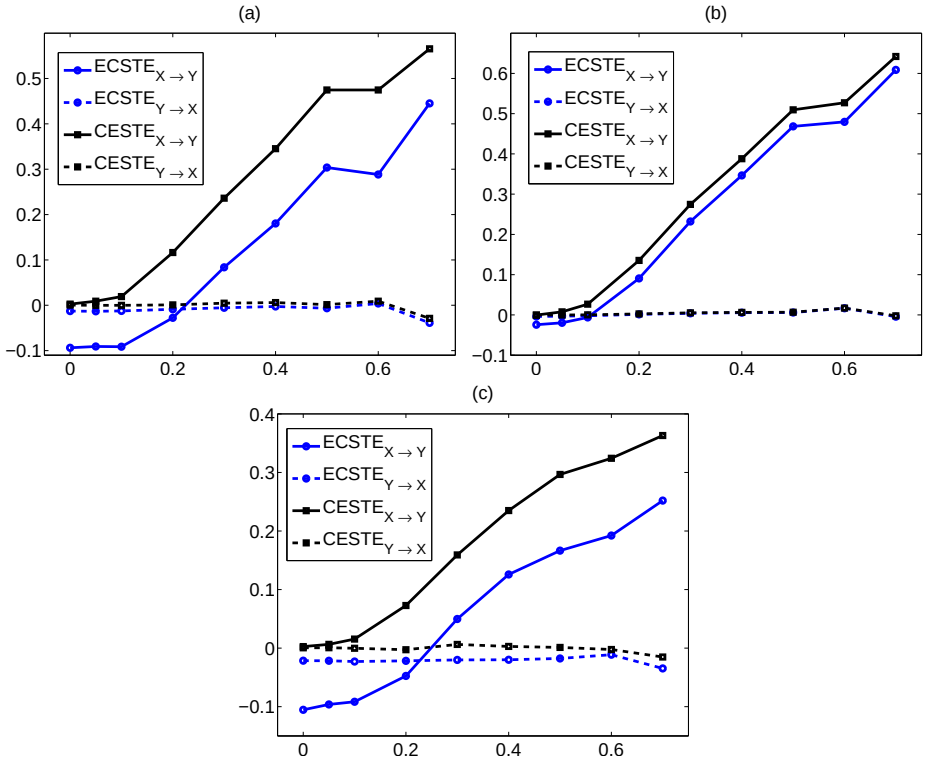

Figure 7.23: (a) Mean estimated values of ECSTE and CESTE for both directions from 100 realizations of the noise-free unidirectionally coupled Henon maps, with $n = 512$ and $m_x = 4$ and $m_y = 3$. (b) As in (a) but for $n = 2048$. (c) As in (a) but for 20% noise level.

obtain insignificant measure value for both directions. Thus, the distribution of the measures in either direction should contain zero. The examination of the proper performance of each measure for this setting is decomposed in the three following sub-tests/ hypothesis:

- H_0: The mean of the measure distribution for the direction $X \to Y$ is zero.
- H_0: The mean of the measure distribution for the direction $Y \to X$ is zero.
- H_0: The means of the measures in the two directions are equal.

The third H_0 is examined just for the cases where one or both of the first two hypotheses are not satisfied, e.g. if the values of a measure are positively biased then the first two hypotheses will be rejected, however the estimated values of this measure might be in the same range. Each accepted H_0 scores one, so a score equal to three suggests that the three conditions are fulfilled for the examined measure for the setting of no coupling (and therefore the p-values of the three tests are larger than the significant level α). A score equal to two suggests that two conditions are valid and so on.

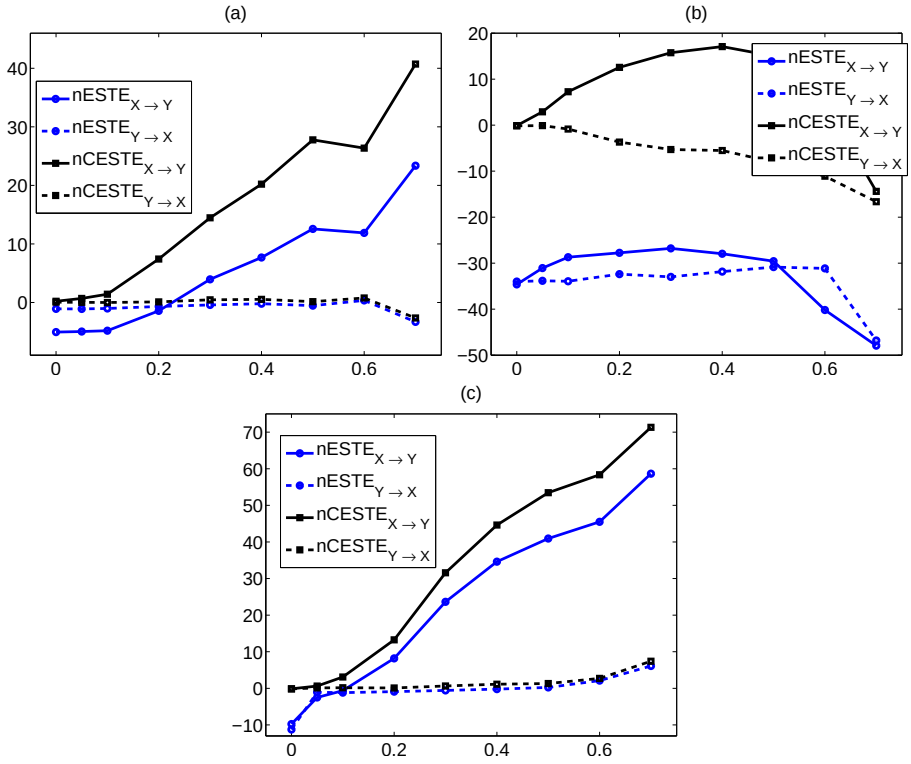

Figure 7.24: (a) Mean estimated values of nECSTE and nCESTE for both directions from 100 realizations of the noise-free unidirectionally coupled Henon maps, with $n = 512$ and $m_x = 4$ and $m_y = 3$. (b) As in (a) but for $n = 2048$. (c) As in (a) but for 20% additive noise.

Based on the estimated scores from the examination of the setting of no coupling, the modified NI measures do not seem to significantly improve the performance of the original measures, however there is an obvious superiority of the measures L, M and N. Indicative results are displayed in Table 7.2 for the noise-free unidirectionally coupled Henon maps. Results for the NI measures on the noisy cases are not displayed analytically as the are insignificant differences of the estimated scores compared to the noise free case, even for high noise levels (20%). Similar scores are obtained also in case of the Mackey-Glass systems.

The measures MCR and CMCR performed poorly, scoring very low. In the case of the uncoupled Henon maps, both measures gave scores equal to zero, and only for equal embedding dimensions the estimated scores were equal to one (only the third H_0 is satisfied), even for large noise-free time series. The same conclusions are observed from the estimated scores on the uncoupled Mackey-Glass systems; MCR and CMCR gave zero scores at almost all cases.

In case of the information causality measures, the corrected measures CTE and CSTE performed properly, even in the case of small time series lengths ($n = 512$).

Table 7.2: Scores of the NI measures from the examination of the setting of no coupling, from the 100 realizations of the noise-free uncoupled Henon maps (for $c = 0$) with $n = 512$.

		scores									
m_x	m_y	H	HM	L	LM	S	SM	M	MM	N	NM
1	1	1	1	3	1	1	1	2	1	2	1
1	2	0	0	3	1	0	0	3	1	3	1
1	3	0	0	3	1	0	0	3	1	3	1
1	4	0	1	3	1	0	0	3	1	3	1
1	5	0	1	3	1	0	0	3	1	3	1
2	1	0	0	3	1	0	0	3	1	3	1
2	2	1	1	3	1	1	1	3	1	3	1
2	3	0	0	3	1	0	0	3	1	3	1
2	4	0	0	3	1	0	0	3	1	3	1
2	5	0	1	3	1	0	0	3	1	3	1
3	1	0	0	3	1	0	0	3	1	3	1
3	2	0	0	3	1	0	0	3	1	3	1
3	3	1	1	3	1	1	1	3	1	3	1
3	4	0	0	3	1	0	0	2	1	2	1
3	5	0	1	3	1	0	0	3	1	3	1
4	1	0	0	3	1	0	0	3	1	3	1
4	2	0	0	3	1	0	0	3	1	3	1
4	3	0	0	3	1	0	0	2	1	2	1
4	4	1	1	3	1	1	1	1	1	1	1
4	5	1	2	3	1	0	0	1	1	2	1
5	1	0	1	3	1	0	0	2	1	2	1
5	2	0	1	3	1	0	0	3	1	3	1
5	3	0	1	3	1	0	0	3	1	3	1
5	4	1	2	3	1	0	0	1	1	1	1
5	5	1	3	3	1	1	1	1	1	1	1

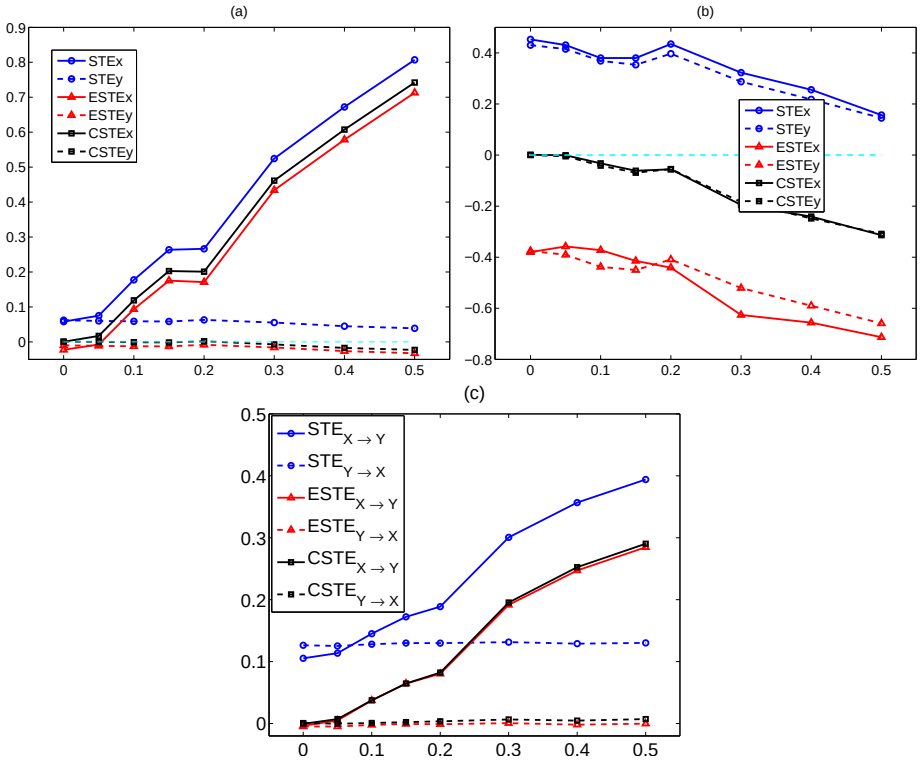

Figure 7.25: (a) Mean estimated values of STE, ESTE and CSTE for both directions from 100 realizations of the noise-free unidirectionally coupled Mackey-Glass systems with $\Delta_1 = 30$ and $\Delta_2 = 100$, $n = 2048$, $m_x = 4$ and $m_y = 3$. (b) As in (a) but for $m_x = m_y = 5$. (c) As in (a) but for 20% additive noise.

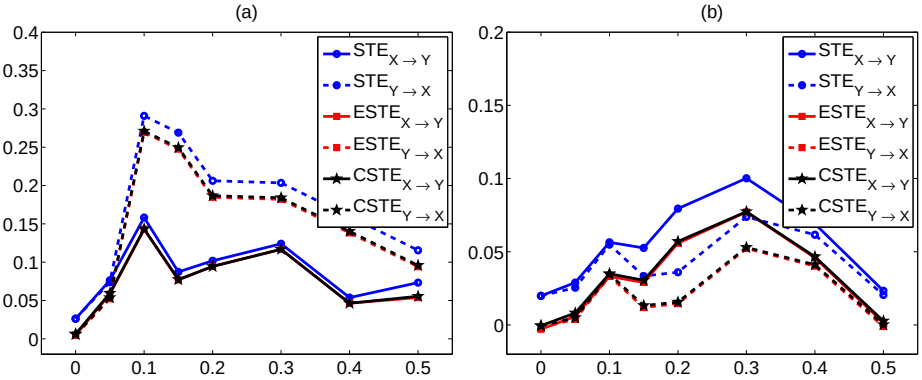

Figure 7.26: (a) Mean estimated values of STE, ESTE and CSTE for both directions from 100 realizations of the noise-free unidirectionally coupled Mackey-Glass systems with $\Delta_1 = \Delta_2 = 17$, $n = 2048$ and $m_x = m_y = 3$. (b) As in (a) but for $\Delta_1 = 17$ and $\Delta_2 = 100$.

CSTE requires larger embedding dimensions than CTE in order to be effective and therefore give score equal to 3. The performance of CTE and CSTE is also not significantly affected by the addition of noise, even in case of high noise levels (20%). Indicative results are displayed in Table 7.4 for the noise-free unidirectionally coupled Henon maps and in Table 7.4 for 20% additive noise. The scores for the other modified information measures are not shown as the qualitative study has shown that ECTE performed similarly to ETE and CETE performed similarly to CTE.

The same conclusions are drawn for the performance of the information causality measures from the estimated scores on the coupled Mackey-Glass systems, for $c = 0$. CTE and CSTE again outperform compared to the other information measures (see Table 7.5 for the noise free Mackey-Glass systems with $\Delta_1 = 17$ and $\Delta_2 = 30$ and Table 7.6 for 20% additive noise).

The second setting aims at assessing the effectiveness of a measure to detect correctly the direction of the causal effect when unidirectionally interaction exists. For a specific coupling strength ($c > 0$), the same three sub-tests are performed, but now the proper performance of a measure suggests the following:

- Rejection of the first H_0, i.e. a measure for the direction $X \rightarrow Y$ is significantly different and moreover greater than zero.
- No rejection of the second H_0, i.e. a measure for the direction $Y \rightarrow X$ is at the zero level.
- Rejection of the third H_0, i.e. mean measure should be significantly greater in the direction $X \rightarrow Y$.

The scores for the original NI measures for the second setting, were expected to be lower than three, due to the problem of the false causal effects at the $Y \rightarrow X$ direction (and therefore the second H_0 is not rejected). The modified NI measures based on the extraction of the common neighbors, seem to improve the performance of the original NI measures. Indicative results are displayed in Table 7.7 for the noise-free unidirectionally coupled Henon maps. M and N performed similarly to L and therefore their scores are not displayed. Increase of the time series length or addition of noise insignificantly affected the performance of the measures, and therefore results are not analytically displayed. In case of the unidirectionally coupled Mackey-Glass system, NI measures performed similarly giving score 2 at almost all cases.

The corrected measures (CMCR, CTE and CSTE) performed almost similarly to the original measures while the effective ones (ETE and ESTE) had no fixed performance. Indicative results for the effectiveness of the measures are given in Table 7.8 for the unidirectionally coupled Henon maps (for $c = 0.4$) and in Table 7.9 for the unidirectionally coupled Mackey-Glass systems with $\Delta_1 = 100$ and $\Delta_2 = 17$. MCR measures proved to be ineffective in case of the unidirectionally coupled Mackey-Glass systems and gave scores equal to one at all cases.

Table 7.3: Scores of the TE and STE measures from the examination of the setting of no coupling, from the 100 realizations of the noise-free uncoupled Henon maps (for $c = 0$) with $n = 512$.

		scores					
m_x	m_y	TE	ETE	CTE	STE	ESTE	CSTE
1	1	1	3	3	-	-	-
1	2	1	1	3	-	-	-
1	3	1	1	3	-	-	-
1	4	0	0	3	-	-	-
1	5	1	1	3	-	-	-
2	1	0	0	3	-	-	-
2	2	1	1	3	1	2	2
2	3	0	0	3	0	3	2
2	4	0	0	3	0	1	1
2	5	0	0	3	0	0	1
3	1	1	1	3	-	-	-
3	2	0	0	3	1	2	1
3	3	1	1	3	1	1	3
3	4	1	0	3	0	0	3
3	5	1	0	3	0	0	2
4	1	1	1	3	-	-	-
4	2	0	0	3	0	0	0
4	3	1	0	3	0	0	3
4	4	3	1	3	1	1	3
4	5	1	0	3	0	0	3
5	1	0	1	3	-	-	-
5	2	0	0	3	1	1	3
5	3	1	0	3	0	0	3
5	4	3	0	3	0	0	3
5	5	3	1	3	1	1	3

Table 7.4: Scores of the TE and STE measures from the examination of the setting of no coupling, from the 100 realizations of the uncoupled Henon maps (for $c = 0$) with $n = 512$ and 20% noise level.

		scores					
m_x	m_y	TE	ETE	CTE	STE	ESTE	CSTE
1	1	1	3	2	-	-	-
1	2	0	3	1	-	-	-
1	3	0	3	1	-	-	-
1	4	0	3	3	-	-	-
1	5	0	1	3	-	-	-
2	1	0	3	3	-	-	-
2	2	1	3	3	1	1	2
2	3	0	3	3	0	1	2
2	4	0	1	2	0	0	1
2	5	0	1	3	0	0	1
3	1	0	3	3	-	-	-
3	2	0	3	3	1	0	1
3	3	1	3	3	1	1	3
3	4	0	1	3	0	0	3
3	5	2	0	3	0	0	2
4	1	0	3	3	-	-	-
4	2	0	1	3	0	0	0
4	3	1	1	3	0	0	3
4	4	2	1	3	1	0	3
4	5	3	0	3	0	0	3
5	1	0	0	2	-	-	-
5	2	0	1	3	1	0	3
5	3	2	1	3	0	0	3
5	4	3	0	3	0	0	3
5	5	3	1	3	1	1	3

Table 7.5: Scores of the TE and STE measures from the examination of the setting of no coupling, from the 100 realizations of the uncoupled Mackey-Glass systems (for $c = 0$) with $\Delta_1 = 17$ and $\Delta_2 = 30$ and $n = 2048$.

		scores					
m_x	m_y	TE	ETE	CTE	STE	ESTE	CSTE
3	3	0	0	2	1	1	1
4	4	1	0	3	0	0	2
5	5	1	0	3	0	0	3
6	6	1	0	2	0	0	2
7	7	2	0	1	0	0	2
8	8	3	3	1	0	0	2
9	9	2	3	1	0	0	3
10	10	1	3	3	0	0	3

Table 7.6: Scores of the TE and STE measures from the examination of the setting of no coupling, from the 100 realizations of the uncoupled Mackey-Glass systems (for $c = 0$) with $\Delta_1 = 17$ and $\Delta_2 = 30$, $n = 2048$ and 20% noise level.

		scores					
m_x	m_y	TE	ETE	CTE	STE	ESTE	CSTE
3	3	0	1	2	1	1	1
4	4	2	0	3	0	0	3
5	5	3	0	3	0	0	3
6	6	2	1	2	0	0	3
7	7	1	1	1	0	1	3
8	8	1	0	3	0	0	3
9	9	3	0	3	0	1	3
10	10	0	0	0	0	1	2

7.4.4 Conclusions

NI measures correctly detected the direction of the causal effects, except for S measure and its variants that seemed to poorly perform at certain cases, e.g. on the coupled Henon maps. The variation of the embedding dimensions seems to insignificantly effect the performance of NI measures; only at few cases where m_x was very small, NI measures proved to be ineffective. The main drawback of these measures seems to be the false detection of increase of the causal effects at the direction $Y \rightarrow X$, as well. Addition of noise effects the estimated values of the measures from both directions, and stresses the problem of the false detection of causal effects at the wrong direction. The 'modified' measures based on the extraction of the common indexes of the neighboring points of X and Y, improved the performance of the NI measures, in contrast to the measures based on the modification of Bhattacharya, which performed similarly to the original measures. The

Table 7.7: Scores of the NI measures from examination of the second setting concerning the correct detection of the causal effects when coupling strength is $c = 0.4$, from 100 realizations of the noise-free uncoupled Henon maps with $n = 512$.

		scores					
m_x	m_y	H	HM	L	LM	S	SM
1	1	2	2	2	2	1	2
1	2	2	3	1	1	1	2
1	3	2	2	1	1	1	2
1	4	2	2	1	1	1	2
1	5	2	2	1	1	1	2
2	1	2	3	2	2	1	2
2	2	2	2	2	2	1	3
2	3	2	2	2	2	1	2
2	4	2	2	2	1	1	2
2	5	2	2	1	1	1	2
3	1	2	2	2	2	1	2
3	2	2	3	2	2	1	3
3	3	2	2	2	2	1	2
3	4	2	2	2	2	1	2
3	5	2	2	1	1	1	3
4	1	2	2	2	2	1	2
4	2	2	3	2	2	1	2
4	3	2	2	2	2	1	2
4	4	2	2	2	2	1	2
4	5	2	2	2	2	1	2
5	1	1	1	2	2	1	3
5	2	2	3	2	2	1	2
5	3	2	2	2	2	1	2
5	4	2	2	2	2	1	2
5	5	2	2	2	2	1	2

Table 7.8: Scores of the MCR, TE and STE measures from examination of the second setting concerning the correct detection of the causal effects when coupling strength is $c = 0.4$, from 100 realizations of the noise-free uncoupled Henon maps with $n = 2048$.

		scores							
m_x	m_y	MCR	CMCR	TE	ETE	CTE	STE	ESTE	CSTE
1	1	1	1	2	2	2	-	-	-
1	2	2	2	2	2	2	-	-	-
1	3	2	2	2	2	2	-	-	-
1	4	2	2	2	2	2	-	-	-
1	5	2	2	1	1	1	-	-	-
2	1	1	1	2	3	3	-	-	-
2	2	2	2	2	2	2	2	2	2
2	3	2	2	3	2	2	2	2	2
2	4	2	2	3	2	2	2	1	1
2	5	2	2	2	2	2	1	1	1
3	1	1	1	2	3	3	-	-	-
3	2	2	2	2	2	2	2	3	3
3	3	2	2	2	2	2	2	2	2
3	4	2	2	2	2	2	2	2	2
3	5	2	2	2	2	2	2	2	2
4	1	1	1	2	3	3	-	-	-
4	2	1	1	2	2	2	2	3	3
4	3	2	2	2	2	2	2	2	2
4	4	2	2	2	2	2	2	2	2
4	5	2	2	2	2	2	2	2	3
5	1	1	1	2	3	3	-	-	-
5	2	1	1	2	2	2	2	2	3
5	3	2	2	2	2	2	2	1	2
5	4	2	2	2	2	2	2	1	2
5	5	2	2	2	2	2	2	2	2

Table 7.9: Scores of TE and STE measures from examination of the second setting concerning the correct detection of the causal effects when coupling strength is $c = 0.1$, from 100 realizations of the noise-free uncoupled Mackey-Glass systems with $n = 2048$, $\Delta_1 = 100$ and $\Delta_2 = 17$.

		scores					
m_x	m_y	TE	ETE	CTE	STE	ESTE	CSTE
3	3	2	2	2	2	3	2
4	4	2	3	2	1	2	2
5	5	2	2	2	1	2	2
6	6	2	1	2	1	2	2
7	7	2	1	2	1	2	2
8	8	2	2	2	1	2	2
9	9	2	0	3	2	2	2
10	10	2	0	3	1	2	2

performance of the measures on the different coupled systems seemed to be consistent.

The MCR measures also correctly detected the direction of the causal effect in the case of the coupled Henon maps, but were positively biased and therefore scored very low at the performance tests. Moreover, MCR measures were also found to lack consistency across systems, as they failed to detect the correct direction of the causal effect in case of the coupled Mackey-Glass systems.

The causality information measures and specifically the proposed 'corrected' measures based on surrogates extracted by randomly shuffling the reconstructed vectors of the driving time series, performed better than the original measures. The corrected measures, consistently gave values in the limits of zero when there was no causal effects and that is their main benefit of using them. CTE was less sensitive than CSTE in the selection of the embedding dimensions, however CSTE was more robust against noise.

7.5 Evaluation of Improved Causality Measures on EEG

The simulation study is indicative of the effectiveness of each measure in detecting causal effects, however the evaluation and comparison of the causality measures on a real application is also essential. Therefore, a similar pilot study as in Sec.7.2.3 is also conducted here, including also the suggested modified measures that seemed to be effective in the simulation study. Multichannel intracranial EEG recordings from one epileptic patient are considered here and the measures are evaluated in their ability to detect changes in the information flow among brain areas corresponding to EEG. Therefore, EEG recordings from an early preictal state and from a late preictal state are used. The aim of the study is to investigate whether any measure can detect a significant change of the interactions of any pair of channels

between the two different states. An indication of such a change suggests a change of the brain dynamics of the patient before the seizure onset and is essential as it can be used as a precursor of a seizure.

7.5.1 Set Up

One intracranial EEG record of 28 channels is used in this application, as intracranial EEG are not that noisy. The causality measures were estimated from both directions for a number of pairs of channels as suggested by the clinician participating in this joint work. The channels were either from the same brain area or from the opposite brain area, symmetrically selected. Specifically, nine different combinations of pairs of channels are considered, as:

- two left frontal channels
- two left temporal channels
- two left occipital channels
- two right frontal channels
- two right temporal channels
- two right occipital channels
- a left frontal and a right frontal channel
- a left temporal and a right temporal channel
- a left occipital and a right occipital channel

Data windows of 1 hour duration from each state (early and late preictal) are considered. Specifically, the early preictal state corresponds to recordings from 4 hours up to 3 hours before the seizure onset and the late preictal state corresponds to recordings from the last one hour before the seizure onset up to the seizure onset. Each one hour long data window is split to 120 successive non-overlapping segments of 30 s and the causality measures are estimated for the pairs of segments as given earlier, for both directions.

Not all the modifications of the measures are included in this study; measures that were ineffective in the simulation study, e.g. ECTE, would probably fail in a real application, where data are derived from more complex systems and are noisy. The measures considered for the analysis are the NI measures and the corresponding 'modified' measures (e.g. HM), while the 'corrected' measures are not be included in this application. MCR, TE and STE and the corresponding 'corrected measures' CMCR, CTE and CSTE which scored best in the performance tests. CETE and CESTE performed similarly to CTE and CSTE, respectively, and therefore are also not used. For the estimation of the measures, the following parameters have been used: $m_x = m_y = 5$ (embedding dimensions), $h = 5$ (time

horizon), $\tau_x = \tau_y = 5$ (lags), $k = 40$ (number of neighbors) and $r = 0.2$ times the standard deviation of the data (radius).

7.5.2 Results

Causality measures were found to be in general not effective in this application, for the detection of changes in the direction of the interactions among the different brain areas, from the two states (early and late preictal). In most cases, the estimated values of the measures indicated a bidirectional form of causality, and this was present at both states. Only few causality measures detected a slight change in the information flow between the early and late preictal state. The state space measures seem to be more effective here than the information measures. The original and the 'corrected' measures gave rather similar results (slightly lower values than the original measures) in almost all cases, for all measure types. Although the patient had a generalized seizure, a change in the causal effects among the channels on the right part of the brain was mostly observed (see Fig. 7.27).

7.5.3 Conclusions

Although causality measures and specifically some of the suggested modified ones that have been introduced here, were pretty effective on the simulation study, in the real application performed poorly. A further investigation of the parameters of the measures and mainly of the embedding dimensions should be conducted in order to have a more clear understanding of the effect of the parameters on the performance of the measures. The fact that there is a bidirectional causal effect observed from all channels, probably reflects the fact that there are interactions among all brain areas and therefore it is hard to draw conclusions without any further investigation of the type of the interactions, e.g. whether causal effects are direct or indirect.

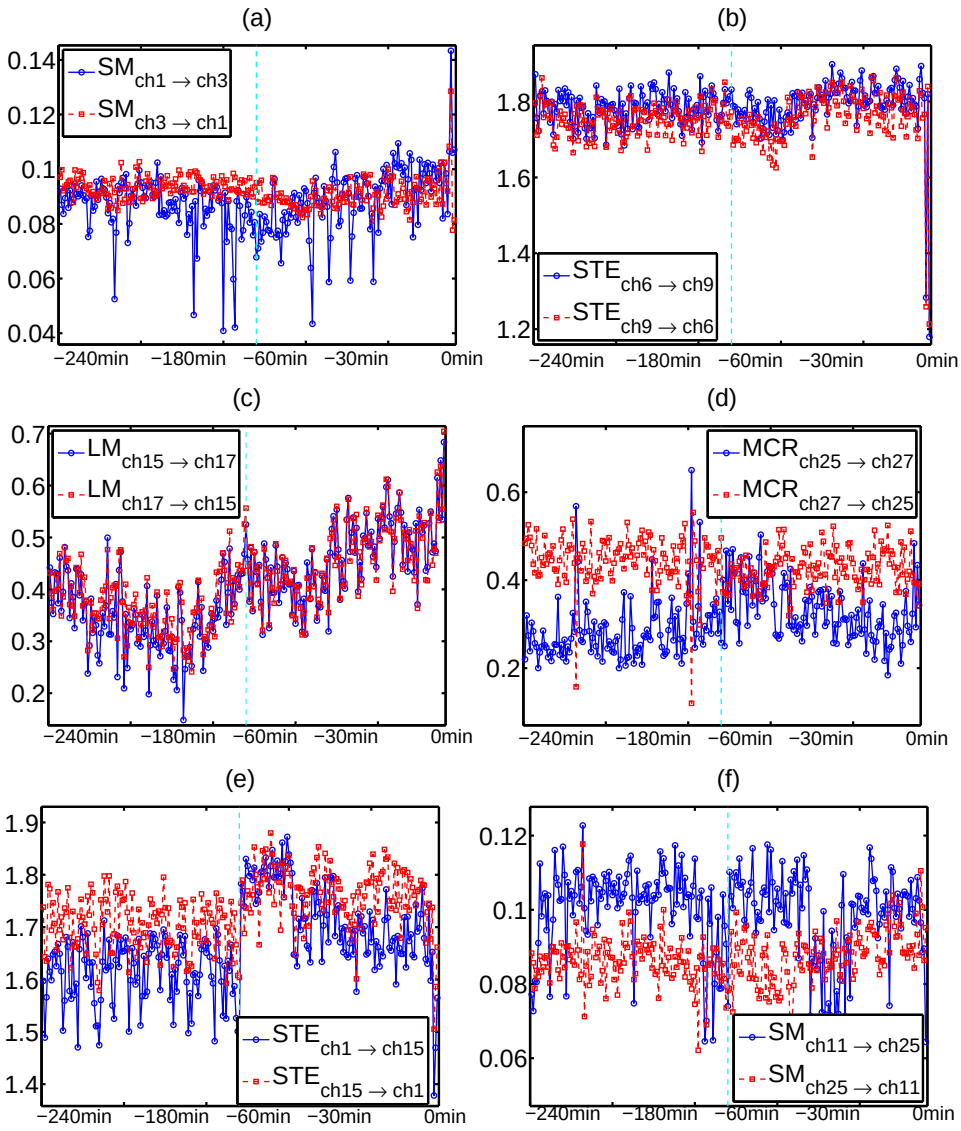

Figure 7.27: (a) Estimated values of SM, from both states (early and late preictal) and both directions, for channels as in the legend. (b) As in (a) but for STE. (c) As in (a) but for LM. (d) As in (a) but for MCR. (e) As in (a) but for STE. (f) As in (a) but for SM. The states are indicated by the time in min, with reference to time 0 at seizure onset and the two states are separated by a vertical dashed line.

Chapter 8

Conclusions

8.1 General Overview

The aim of this thesis was to find a reliable statistic with high statistical significance, for the prediction of the onset of epileptic seizures from multi-channel EEG recordings. The study concentrated on linear and nonlinear information measures, however it was also extended in order to include other types of measures (entropy and complexity measures), and resulted in the modification of some existing measures and in the extraction of some new measures.

The simple extension of the mutual information to the cumulative mutual information in order to eliminate the effect of the selection of a specific lag, was found to be an efficient statistic for independence and nonlinearity test, but also for the independence and nonlinearity test, as well as for the detection of dynamical changes in systems. In addition, an automatic and data adaptive method for the selection of the maximum lag for the estimation of the cumulative mutual information was developed. The estimated lag can be considered as the nonlinear decorrelation time and can be used as a measure indicating the general memory of a dynamical system. An interesting development of the surrogate data test for nonlinearity using the statistic of mutual information was to use the p-values of the test as a measure of the purely nonlinear correlation of the system.

Special emphasis was given in this thesis on the mutual information, as it is a useful tool in time series analysis, necessary for the reconstruction of the state space of the time series, and commonly used in EEG analysis. Specifically, the statistical properties of the most known mutual information estimators were studied. The estimators do not follow a known distribution and they are all biased due to the finite time series length and due to the finite time series length and due to constraints in their realization, e.g. finite number of cells for the binning estimators. The different mutual information estimators were evaluated based on extensive simulations and the most robust estimator was found for the given conditions. Moreover the corresponding free parameter of each estimator was optimized.

The evaluation of mutual information estimators was extended and more infor-

mation measures, entropy measures and complexity measures, were tested in their ability to discriminate among different dynamical systems, i.e. systems with different complexity or different dynamical states of the same system. Mutual information appears as a better discriminating statistic than the autocorrelation function because it estimates linear and nonlinear correlations. A new measure that quantifies only nonlinear correlations was defined as the difference of the estimated mutual information from the transformed time series from the 'normal' mutual information, which is theoretically known based on the autocorrelation function for normal processes. The respective cumulative measure was also defined to suppress the dependence of the lag. A comprehensive evaluation of all the discussed univariate measures was assessed by Monte Carlo simulations on linear and chaotic systems where their complexity was varied with control parameters. The simulation study verified the usefulness of the information measures and specifically of the cumulative ones, as far as their discriminating ability is concerned.

The applications of the information measures on epileptic EEG gave a more comprehensive understanding of the effectiveness of the measures and the problems that arise when testing real data. Some interesting points were revealed from the comparison of the different measures. First of all, the level of the discrimination of the preictal stages depends on the physiological conditions. The EEG records that were considered, are extracted mainly from patients with generalized tonic clonic seizures. It was also observed that the discrimination performance of the measures varied with the epileptic episode and channel, suggesting that the physiological activity at the different brain areas varies with the seizure. There is a common notion that the seizures start at specific locations even though denoted as primarily generalized seizures.

A short-term prediction of the onset of the epileptic seizure was accomplished in an horizon of 1–2 minutes with many measures by quantifying the changes in the correlations on the EEG time series. The discrimination of different states of the brain was also investigated, specifically the discrimination of the late preictal state (minutes before the seizure onset) and earlier preictal states (one or more hours before the seizure onset). Cumulative measures were proved to be quite effective, especially for comparing late preictal states to early preictal states many hours before the seizure onset. Although results varied with the patients, the channel and the measure, information measures seem to be able to detect changes in the brain dynamics and the discrimination of different states was statistically significant for a number of channels and measures. The results did not show though a clear discrimination of the late preictal state (last 30min) from earlier periods (from about 3h to 40 min prior to seizure onset) as no measure was effective at all seizure episodes, but different measures could discriminate the different states at different seizure episodes. Overall, many measures could succeed significant discrimination of the late preictal state from many earlier preictal stages suggesting that the seizure prediction is at cases possible. This needs a further investigation including more episodes as well as records going further backwards in time from the seizure onset (interictal state).

The study of the univariate measures was extended in order to include bivariate measures that indicate the strength and the direction of the interaction between two systems. Again, a number of different measures were reviewed and evaluated and some of the existing measures were modified in order to improve their statistical significance. The study focused again on information measures, however also state space and synchronization measures were included. Investigation of the influence of the time series length, noise level and the selection of the embedding dimension for the reconstruction of the state space of the examined systems was also assessed. The suggested modifications evidently improved the statistical significance of the causality measures and particularly the information transfer measures. The corrected measures scored highest to performance tests contacted on simulated data. Although the results from the simulation study concerning the modified measures were very promising, on real applications measures were pretty insufficient, as no consistent detection of information flow changes was observed in the different brain areas of the examined patients.

8.2 Suggestions for future work

The cumulative measures (autocorrelation and mutual information, declination from normality) were proved to be very effective in the simulation study in discriminating among different dynamical states. The selection of the maximum lag that was optimized; however considering the nonlinear decorrelation time might not be the most suitable choice for all applications. The performance of the measures on EEG recordings have not verified any advantage of the nonlinear measures over the linear ones, e.g. Spearman's and Kendal's autocorrelation coefficients seemed to be more useful than the complexity measures. The fact that simple measures can discriminate preictal stages as good as or even better than measures that are more conceptually advanced and computationally intensive has been previously reported (Kugiumtzis and Larsson, 2000; Mormann et al., 2007). This suggests a more thorough evaluation of simpler statistical measures.

The 'gaussianization' of the time series seemed to improve the statistical significance of the measures especially when used for the independence test, but also for the discrimination of the different dynamical systems. These results suggest a further investigation of the usefulness of this transform. Moreover, the investigation and generation of different transformation schemes would be an interest subject for future work. It would also be very interesting to apply the developed methods either preprocessed extracranial EEG or intracranial EEG, where the level of observational noise is moderate and the recordings contain fewer artifacts, however a large scale study is essential in order to draw any conclusions. Moreover, the limitations of the EEG studies in terms of data resources does not allow for conclusive results, and therefore all the studies should be extended to include more EEG epochs.

The necessity to describe whether different brain areas interact is of great im-

portance in the neuroscience field and many approaches have been developed for this purpose (Kaminski and Blinowska, 1991; Baccala and Sameshima, 2001). Obviously, the pairwise analysis of multivariate data may yield misleading results. The existing and developed causal measures that have been examined here do not distinguish among direct and indirect causal effects (Kus et al., 2004; Winterhalder et al., 2006). Therefore, causality measures discriminating among direct and indirect causal patterns should be also examined and possibly modified. The identification of the epileptic focus area from the EEG recordings using the causality measures could also be investigated.

Bibliography

Abarbanel, H., Brown, R., Sidorowich, J., and Tsimring, L. (1993). The analysis of observed chaotic data in physical systems. *Review of Modern Physics*, 65:1331–1392.

Abarbanel, H., Masuda, N., Rabinovich, M., and Tumer, E. (2001). Distribution of mutual information. *Physics Letters A*, 281(5-6):368–373.

Adelson, P., Nemoto, E., Scheuer, M., Painter, M., Morgan, J., and Yonas, H. (1999). Noninvasive continuous monitoring of cerebral oxygenation peri-ictally using near infrared spectroscopy: a preliminary report. *Epilepsia*, 40:1484–1489.

Andrzejak, R., Mormann, F., Kreuz, T., Rieke, C., Kraskov, A., Elger, C., and Lehnertz, K. (2003). Testing the null hypothesis of the nonexistence of a preseizure state. *Physical Reviwe E*, 67(1):010901.

Andrzejak, R., Mormann, F., Widman, G., Kreuz, T., Elger, C., and Lehnertz, K. (2006). Improved spatial characterization of the epileptic brain by focusing on nonlinearity. *Epilepsy Research*, 69(1):30–44.

Andrzejak, R., Widman, G., Lehnertz, K., David, P., and Elger, C. (1999). Nonlinear determinism in intracranial EEG recordings allows focus localization in neocortical lesional epilepsy. *Epilepsia*, 40(7):171–172.

Armitage, P., Berry, G., and Matthews, J. (2002). *Statistical Methods in Medical Research (4th edition)*. Blackwell Science, Oxford.

Arnhold, J., Grassberger, P., Lehnertz, K., and Elger, C. (1999). A robust method for detecting interdependences: Application to intracranially recorded EEG. *Physica D*, 134(4):419–430.

Babloyantz, A. and Destexhe, A. (1986). Low-dimensional chaos in an instance of epilepsy. *Proceedings of the National Academy of Sciences*, 83:3513–3517.

Baccala, L. and Sameshima, K. (2001). Partial directed coherence: a new concept in neural structure determination. *Biological Cybernetics*, 84(6):463474.

Balatoni, J. and Renyi, A. (1956). On the notion of entropy (Hungarian). 1:9–40.

Bandt, C. and Pompe, B. (2002). Permutation entropy - a complexity measure for time series. *Physical Review Letter*, 88:174102.

Barnard, G. (1963). Discussion of professor Bartlett's paper. *Journal of Royal Statistical Society B*, 25:294.

Bendat, S. and Piersol, A. (1966). *Measurements and Analysis of Random Data*. John Wiley and Sons, New York.

Bezruchko, B., Ponomarenko, V., and Rosenblum, M.G.and Pikovsky, A. (2003). Characterizing direction of coupling from experimental observations. *Chaos*, 13(1):179–184.

Bhattacharya, J., Pereda, E., and Petsche, H. (2003). Effective detection of coupling in short and noisy bivariate data. *IEEE Transactions on Systems, Man, and Cybernetics B*, 33(1):85–94.

Birkhoff, G. (1931). Proof of the ergodic theorem. In *National Academy of Sciences*, volume 17, pages 656–660.

Blinnikov, S. and Moessner, R. (1998). Expansions for nearly Gaussian distributions. *Astronomy and Astrophysics, Supplement Series*, 130(1):193–205.

Bonnlander, B. and Weigend, A. (1994). Selecting input variables using mutual information and nonparametric density estimation. In *International Symposium on Artificial Neural Networks (ISANN"94)*, pages 42–50, Taiwan.

Box, G., Jenkins, G., and Reinsel, G. (1994). *Time series Analysis: Forecasting and Control*. Prentice Hall, Englewood Clifs, NJ.

Box, G. and Pierce, D. (1970). Distribution of the autocorrelations in autoregressive moving average time series models. *Journal of American Statistical Association*, 65:1509–1526.

Bruce, H. (1980). *A Multidisciplinary Handbook of Epilepsy*. Charles C. Thomas, Springfield.

Cao, Y., Tung, W., Gao, J., Protopopescu, V., and Hively, L. (2004). Detecting dynamical changes in time series using the permutation entropy. *Physical Revie E*, 70(4):046217.

Cellucci, C., Albano, A., and Rapp, P. (2005). Statistical validation of mutual information calculations: Comparison of alternative numerical algorithms. *Physical Review E*, 71:066208.

Cenys, A., Lasiene, G., Pyragas, K., Peinke, J., and Parisi, J. (1992). Analysis of spatial correlations in chaotic systems. *Acta Physica Polonica B*, 23(4):357–365.

Chen, Y., Rangarajan, G., Feng, J., and Ding, M. (2004). Analyzing multiple nonlinear time series with extended granger causality. *Physics Letters A*, 324:26–35.

Chicharro, D., Ledberg, A., and Andrzejak, R. (2008). A new measure for the detection of directional couplings based on rank statistics. In *BMC Neuroscience, 7th Annual Computational Neuroscience Meeting*, volume 9, page 148, Portland, USA.

Chillemi, S., Balocchi, R., Garbo, A., D'Attellis, C., Gigola, S., Kochen, S., and Silva, W. (2003). Discriminating preictal from interictal states by using coherence measures. In *IEEE Engineering in Medicine and Biology Society (25th Annual International Conference)*, volume 3, pages 2319–2322.

Chui, C. (1992). *An Introduction to Wavelets*. San Diego: Academic Press, San Diego.

Cocatre-Zilgien, J. and Delcomyn, F. (1992). Identification of bursts in spike trains. *Journal of Neuroscience Methods*, 10(1):19–30.

Cochran, W. (1954). Some methods for strengthening the common X2 tests. *Biometrics*, 41(4):417–451.

Cohen, A. and Procaccia, I. (1985). Computing the kolmogorov entropy from time signals of dissipative and conservative dynamical systems. *Physical Review A*, 31(3):1872–1882.

Cohen, D. and Cuffin, B. (1983). Demonstration of useful differences between the magnetoencephalogram and electroencephalogram. *Electroencephalography and Clinical Neurophysiology*, 56:38–51.

Cover, T. and Thomas, J. (1991). *Elements of Information Theory*. John Wiley and Sons, New York.

Darbellay, G. and Vajda, I. (1999). Estimation of the information by an adaptive partitioning of the observation space. In *IEEE Transactions on Information Theory*, volume 45, pages 1315–1321.

Daub, C., Steuer, R., Selbig, J., and Kloska, S. (2004). Estimating mutual information using B-spline functions: An improved similarity measure for analysing gene expression data. *BMC Bioinformatics*, 5(1):118.

Delamont, R., Julu, P., and Jamal, G. (1999). Changes in a measure of cardiac vagal activity before and after epileptic seizures. *Epilepsy Research*, 35(87–94).

Dhamala, M., Rangarajan, G., and M., D. (2008). Analyzing information flow in brain networks with nonparametric granger causality. *NeuroImage*, 41:354362.

Diks, C. and Manzan, S. (2002). Tests for serial independence and linearity based on correlation integrals. *Journal Studies in Nonlinear Dynamics and Econometrics*, 6(2):1005.

Diks, C. and Panchenko, V. (2008). Rank-based entropy tests for serial independence. *Studies in Nonlinear Dynamics and Econometrics*, 12(1).

Doane, D. (1976). Aesthetic frequency classifications. *The American Statistician*, 30(4):181–183.

Dobrushin, R. (1959). General formulation of Shannon's main theorem in information theory. *Transacrions of the American Mathematical Society*, 33(1):323–438.

Dornhege, G., Blankertz, B., Krauledat, M., Losch, F., Curio, G., and Muller, K.-R. (2006). Optimizing spatiotemporal filters for improving BCI. In *Advances in Neural Information Processing Systems (NIPS 05)*, volume 18, pages 315–322. MIT Press.

Eckmann, J., Kamphorst, S., and Ruelle, D. (1987). Recurrence plots of dynamical systems. *Europhysics Letter*, 4(9):973–977.

Ehlers, C., Havstad, J., Prichard, D., and Theiler, J. (1998). Low doses of ethanol reduce evidence for nonlinear structure in brain activity. *The Journal of Neuroscience*, 18(18):7474–7486.

Ellis, D. and Bilmes, J. (2000). Using mutual information to design feature combinations. In *Proceedings of the International Conference on Spoken Language Processing, 1620 October*, Beijing.

Faes, L., Porta, A., and Nollo, G. (2008). Mutual nonlinear prediction as a tool to evaluate coupling strength and directionality in bivariate time series: Comparison among different strategies based on *k* nearest neighbors. *Physical Review E*, 78(2):026201.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27:861–874.

Feder, J. (1988). *Fractals*. Plenum Press, New York.

Federico, P., Abbott, D., Briellmann, R., Harvey, A., and Jackson, G. (2005). Functional MRI of the pre-ictal state. *Brain*, 128:1811–1817.

Feldmann, U. and Bhattacharya, J. (2004). Predictability improvement as an asymmetrical measure of interdependence in bivariate time series. *International Journal of Bifurcation and Chaos*, 14(2):505–514.

Fisher, K. (1935). Statistical tests. *Nature*, 136(3438):474.

Fisher, R. (1973). *Statistical Methods and Scientific Inference (3rd Edn.)*. Macmillan, New York.

Fraser, A. and Swinney, H. (1986). Independent coordinates for strange attractors from mutual information. *Physical Review A*, 33:1134–1140.

Freedman, D. and Diaconis, P. (1981). On the histogram as a density estimator: L2 theory. *Zeitschrift fur Wahrscheinlichkeitstheorie und verwandt Gebiete*, 57:453–476.

Gersch, B. and Goddard, G. (1970). Epileptic focus localisation: Spectral analysis method. *Science*, 169:701–702.

Govindan, R., Wilson, J., Eswaran, H., Lowery, C., and Prebl, H. (2007). Revisiting sample entropy analysis. *Physica A*, 376:158–164.

Granger, C. and Lin, J. (1994). Using the mutual information coefficiente to identify lags in nonlinear models. *Journal of Time Series Analysis*, 14(4):371–384.

Granger, C., Maasoumi, E., and Racine, J. (2004). A dependence metric for possibly nonlinear processes. *Journal of Time Series Analysis*, 25:649–669.

Granger, J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Acta Physica Polonica B*, 37:424–438.

Grassberger, P. (1988). Finite sample corrections to entropy and dimension estimates. *Physics Letter A*, 128(6–7):369–373.

Grassberger, P. and Procaccia, I. (1983). Measuring the strangeness of strange attractors. *Physica D*, 9(1–2):189–208.

Grassberger, P., Schreiber, T., and Schaffrath, C. (1991). Nonlinear time sequence analysis. *International Journal Bifurcation and Chaos*, 1:512–547.

Gray, R. (1990). *Entropy and Information Theory*. Springer Verlag, New York.

Green, D. M. and Swets, J. A. (1966). *Signal Detection Theory and Psychophysics*. John Wiley and Sons, New York.

Gudmundsson, S., Runarsson, T., Sigurdsson, S., Eiriksdottir, G., and Johnsen, K. (2007). Reliability of quantitative EEG features. *Clinical Neurophysiology*, 118:2162–2171.

Hamilton, J. (1964). *Time Series Analysis*. Princeton University Press, NJ.

Hanus, P., Goebel, B., Dingel, J., Weindl, J., Zech, J., Dawy, Z., Hagenauer, J., and Mueller, J. (2007). Information and communication theory in molecular biology. *Electrical Engineering*, 90(2):161–173.

Harrold, T., Sharma, A., and Sheather, S. (2001). Selection of a kernel bandwidth for measuring dependence in hydrologic time series using the mutual information criterion. *Stochastic Environmental Research and Risk Assesment*, 15(4):310–324.

Henon, M. (1976). A two dimensional mapping with a strange attractor. *Communications in Mathematical Physics*, 50(1):69–77.

Hilbert, D. (1953). *Grundzuge einer allgemeinen Theorie der linearen Integralgleichungen*. Chelsea Pub. Co.

Hinich, M. and Patterson, D. (1995). Detecting epochs of transient dependence in white noise.

Hirsch, E., Andermann, F., Chauvel, P., Engel, J., Lopes da Silva, F., and Luders, H. (2006). Generalized seizures: from clinical phenomenology to underlying systems and networks.

Hively, L. and Protopopescu, V. (2003). Channel-consistent forewarning of epileptic events from scalp EEG. In *IEEE Transactions on Biomedical Enginnering*, volume 50, pages 584–593.

Hively, L., Protopopescu, V., and Gailey, P. (2000). Timely detection of dynamical change in scalp EEG signals. *Chaos*, 10(4):864–875.

Hlavackova-Schindler, K., Palus, M., Vejmelka, M., and Bhattacharya, J. (2007). Causality detection based on information-theoretic approaches in time series analysis. *Physics Reports*, 441(1):1–46.

Hope, A. (1968). A simplified Monte Carlo test procedure. *Journal of the Royal Statistical Society B*, 30:582–598.

Hoph, E. (1937). *Ergodentheorie*. Springer-Verlag, Berlin.

Hutter, M. and Zaffalon, M. (2005). Distribution of mutual information from complete and incomplete data. *Computational Statistics and Data Analysis*, 48(3):633–657.

Iasemidis, L., Pardalos, P., Sackellares, J., and Shiau, D. (2001). Quadratic binary programming and dynamical system approach to determine the predictability of epileptic seizures. *Journal of Combinatorial Optimization*, 5:9–26.

Iasemidis, L., Sackellares, J., Zaveri, H., and Williams, W. (1990). Phase space topography of the electrocorticogram and the Lyapunov exponent in partial seizures. *Clinical Neurophysiology*, 116:187–201.

Iasemidis, L., Shiau, D., Pardalos, P., Chaovalitwongse, W., Narayanan, K., Prasad, A., Tsakalis, K., Carney, P., and Sackellares, J. (2005). Long-term prospective

on-line real-time seizure prediction. *Computational Statistics and Data Analysis*, 116(3):532–544.

Jasper, H. (1958). The ten-twenty electrode system of the international federation. *Electroencephalography and Clinical Neurophysiology*, 10:371–375.

Jevrejeva, S., Moore, J., and Grinsted, A. (2003). Influence of the arctic oscillation and el nino-southern oscillation (ENSO) on ice conditions in the baltic sea: The wavelet approach. *Journal of Geophysical Research*, 108(D21):4677.

Jones, M., Marron, J., and Sheather, S. (1996). A brief survey of bandwidth selection for density estimation. *Journal of American Statistical Association*, 91(433):401–407.

K., F. (1972). *Introduction to Statistical Pattern Recognition*. Academic Press, New York.

Kalitzin, S., Parra, J., Velis, D., and Lopes da Silva, F. (2002). Enhancement of phase clustering in the EEG/MEG gamma frequency band anticipates transitions to paroxysmal epileptiform activity in epileptic patients with known visual sensitivity. In *IEEE Transaction on Biomedical Engineering*, volume 49, pages 1279–1286.

Kaminski, M. (2005). Determination of transmission patterns in multichannel data. *Philosophical Transaction of the Royal Society B, Biological Sciences*, 360(1457):947–952.

Kaminski, M. and Blinowska, K. (1991). A new method of the description of the information flow. *Biological Cybernetics*, 65:203–210.

Kantz, H. and Schreiber, T. (1997). *Nonlinear Time Series Analysis*. Cambridge University Press, Reading, Massachusetts.

Kantz, H. and Shürmann, T. (1996). Enlarged scaling ranges in entropy and dimension estimates. *Chaos*, 6(2):167–171.

Kaplan, J. and Yorke, J. (1979). In *Functional Differential Equations and Approximations of Fixed Points: Proceedings, Bonn, July 1978*, page 204, Berlin. Springer-Verlag.

Kendall, M. (1938). A new measure of rank correlation. *Biometrika*, 30(1–2):81–89.

Kennel, M., Brown, R., and Abarbanel, H. (1992). Determining embedding dimension for phase-space reconstruction using a geometrical construction. *Physical Review A*, 45:3403.

Khan, S., Bandyooadhyay, S., Ganguly, A., Saigal, S., Erickson, D., Protopopescu, V., and Ostrouchov, G. (2007). Relative performance of mutual information estimation methods for quantifying the dependence among short and noisy data. *Physical Review E*, 76(2):026209.

Knuth, K. (2006). Optimal data-based binning for histograms.

Kraskov, A., Stögbauer, H., Andrzejak, R. G., and Grassberger, P. (2005). Hierarchical clustering using mutual information. *Europhysics Letters*, 70(2):278–284.

Kraskov, A., Stögbauer, H., and Grassberger, P. (2004). Estimating mutual information. *Physical Review E*, 69(6):066138.

Kreuz, T., Mormann, F., Andrzejak, R., Kraskov, A., Lehnertz, K., and Grassberger, P. (2007). Measuring synchronization in coupled model systems: A comparison of different approaches. *Physica D*, 225(1):29–42.

Krug, D., Osterhage, H., Elger, C., and Lehnertz, K. (2007). Estimating nonlinear interdependences in dynamical systems using cellular nonlinear networks. *Physical Review E*, 76:041916.

Kugiumtzis, D. (1996). State space reconstruction parameters in the analysis of chaotic time series - the role of the time window length. *Physica D*, 95:13–28.

Kugiumtzis, D. (2002). Statistically transformed autoregressive process and surrogate data test for nonlinearity. *Physical Review E*, 66:025201.

Kugiumtzis, D. and Larsson, P. (2000). In *Proceedings of the 1999 Workshop, 'Chaos in Brain?'*, Singapore.

Kullback, S. (1959). John Wiley and Sons, NY.

Kullback, S. and Leibler, R. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86.

Kus, R., Kaminski, M., and Blinowska, K. (2004). Determination of EEG activity propagation: Pair-wise versus multichannel estimate. *IEEE Transactions on Biomedical Engineering*, 51(9):1501.

Le Van Quyen, M., Soss, J., Navarro, V., Robertson, R., Chavez, M., Baulac, M., and Martinerie, J. (2005). Preictal state identification by synchronization changes in long-term intracranial EEG recordings. *Journal of Clinical Neurophysiology*, 116:559–568.

Lehnertz, K., Andrzejak, R., Arnhold, J., Kreuz, T., Mormann, F., Rieke, C., Widman, G., and Elger, C. (2001). Nonlinear EEG analysis in epilepsy: Its possible use for interictal focus localization, seizure anticipation, and prevention. *Journal of Clinical Neurophysiology*, 18:209–222.

Lehnertz, K., Arnhold, J., Grassberger, P., and Elger, C. (2000). World Scientific, Singapore.

Lerner, D. (1996). Monitoring changing dynamics with correlation integrals: Case study of an epileptic seizure. *Physica D*, 97(4):563–576.

Li, X. (2006). Wavelet spectral entropy for indication of epileptic seizure in extracranial EEG. In *International Conference on Neural Information Processing (ICONIP)*, volume 3, pages 66–73.

Lorenz, E. (1963). Deterministic nonperiodic flow. *Journal of Atmospheric Science*, 20:130–141.

Lungarella, M., Ishiguro, K., Kuniyoshi, Y., and OTSU, N. (2007). Methods for quantifying the causal structure of bivariate time series. *Journal of Bifurcation and Chaos*, 17(3):903–921.

Mackey, M. and Glass, L. (1977). Oscillation and chaos in physiological control systems. *Science*, 197(4300):287–289.

Mandelbrot, B. (1974). Intermittent turbulence in self-similar cascades: Divergence of high moments and dimension of the carrier. *Science*, 197(4300):287–289.

Manzan, S. and Diks, C. (2002). Tests for serial independence and linearity based on correlation integrals. *Studies in Nonlinear Dynamics and Econometrics*, 6(2):1005.

Marschinski, R. and Kantz, H. (2002). Analysing the information flow between financial time series. *European Physical Journal B*, 30:275281.

McSharry, P., Smith, L., and Tarassenko, L. (2003). Comparison of predictability of epileptic seizures by a linear and a nonlinear method. *IEEE Transactions on Biomedical Engineering*, 50(5):628–633.

Moddemeijer, R. (1988). The variance of the mutual information estimator. Technical Report AM.080.88.13, Enschede (NL).

Moon, Y., Rajagopalan, B., and Lall, U. (1995). Estimation of mutual information using kernel density estimators. *Physical Review E*, 52(3):2318–2321.

Mormann, F., Andrzejak, R., Elger, C., and Lehnertz, K. (2007). Seizure prediction: the long and winding road. *Brain*, 130:314–333.

Mormann, F., Elger, C., and Lehnertz, K. (2006). Seizure anticipation: from algorithms to clinical practice. *Current Opinion in Neurology*, 19:187–193.

Mormann, F., Kreuz, T., Rieke, C., Andrzejak, R., Kraskov, A., David, P., Elger, C., and Lehnertz, K. (2005). On the predictability of epileptic seizures. *Clinical Neurophysiology*, 116(3):569–587.

Mormann, F., Lehnertz, K., David, P., and Elger, C. a. (2000). Mean phase coherence as a measure for phase synchronization and its application to the EEG of epilepsy patients. *Physica D*, 144:358–369.

Muller-Gerking, J., Pfurtscheller, G., and Flyvbjerg, H. (1999). Designing optimal spatial filters for single-trial EEG classification in a movement task. *Clinical Neurophysiology*, 110:787–798.

Naa, S., Jina, S.-H., Kima, S., and Hamb, B.-J. (2002). EEG in schizophrenic patients: Mutual information analysis. *Clinical Neurophysiology*, 113(12):1954–1960.

Neyman, J. and Pearson, E. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society A*, 231:289–337.

Nicolaou, N. and Nasuto, S. (2005). Mutual information for EEG analysis. In *4th IEEE Postgraduate Conference on Biomedical Engineering and Medical Physics*, pages 23–24.

Nunez, P. (1995). Oxford University Press, Oxford.

Olbrich, E. and Kantz, H. (1997). Inferring chaotic dynamics from time-series: On which length scale determinism becomes visible. *Physics Letters A*, 232(1–2):63–69.

Osowski, S., Swiderski, B., Cichocki, A., and Rysz, A. (2007). Epileptic seizure characterization by Lyapunov exponent of EEG signal. *International Journal for Computation and Mathematics in Electrical and Electronic Engineering*, 26(5):1276–1287.

Osterhage, H., Mormann, F., Wagner, T., and Lehnertz, K. (2008). Detecting directional coupling in the human epileptic brain: Limitations and potential pitfalls. *Physical Review E*, 77(1):011914.

Palus, M. (1993). Identifying and quantifying chaos by using information-theoretic functionals. In Weigend, A. and Gershenfeld, N., editors, *Time Series Prediction: Forecasting the Future and Understanding the Past*, volume XV of *Santa Fe Institute Studies in the Sciences of Complexity*, pages 387–413. Addison-Wesley, Reading.

Palus, M. (1995). Testing for nonlinearity using redundancies: Quantitative and qualitative aspects. *Physica D*, 80(1):186–205.

Palus, M. (1996). Nonlinearity in normal human EEG: Cycles, temporal asymmetry, nonstationarity and randomness, not chaos. *Biological Cybernetics*, 75:389–396.

Palus, M., Komarek, V., Prochazka, T., Hrncir, Z., and Sterbova, K. (2001a). Synchronization and information flow in EEGs of epileptic patients. 20(5):65–71.

Palus, M., Korarek, V., Hrncir, Z., and Sterbova, K. (2001b). Synchronization as adjustment of information rates: Detection from bivariate time series. *Physical Review E*, 63(4):046211.

Palus, M. and Stefanovska, A. (2003). Direction of coupling from phases of interacting oscillators: An information-theoretic approach. *Physical Review E*, 67:055201.

Palus, M. and Vejmelka, M. (2007). Directionality of coupling from bivariate time series: How to avoid false causalities and missed connections. *Physical Review E*, 75:056211.

Paninski, L. (2003). Estimation of entropy and mutual information. *Neural Computation*, 15(6):1191–1253.

Pardalos, P., Yatsenko, V., Sackellares, J., Shiau, D.-S., Chaovalitwongse, W., and Iasemidis, L. (2003). Analysis of EEG data using optimization, statistics, and dynamical system techniques. *Computational Statistics and Data Analysis*, 44:391–408.

Pardo, J. (1995). Some applications of the useful mutual information. *Applied Mathematics and Computation*, 72(1):33–50.

Pawelzik, K. and Schuster, H. (1987). Generalized dimensions and entropies from a measured time series. *Physical Review A*, 35(1):481484.

Pereda, E., Quiroga, R., and Bhattacharya, J. (2005). Nonlinear multivariate analysis of neurophysiological signals. *Progress in Neurophysiology E*, 77(1–2):1–37.

Perrin, F., Pernier, J., Bertrand, O., and Echallier, J. (1989). Spherical splines for scalp potential and current density mapping. *EEG and Clininal Neurophysiology*, 72:184–187.

Pijn, J. and Lopes Da Silva, F. (1993). *Propagation of Electrical Activity: Nonlinear Associations and Time Delays between EEG Signals*. Birkauser, Boston.

Pikovsky, A., Rosenblum, M., and Kurths, J.

Pincus, C. (1991). Approximate entropy as a measure of system complexity. *Proceding National Academy of Sciences, USA*, 88:2297–2301.

Pompe, B. (1993). Measuring statistical dependencies in a time series. *Journal of Statistical Physics*, 73:587–610.

Porter, R. (1993). *Classification of Epileptic Seizures and Epileptic Syndromes*. Churchill Livingstone, London. A Textbook of Epilepsy.

Priness, I., Maimon, O., and Ben-Gal, I. (2007). Evaluation of gene-expression clustering via mutual information distance measure. *BMC Bioinformatics*, 8(6):111.

Quiroga, Q. R., Arnhold, J., Lehnertz, K., and Grassberger, P. (2000a). Kulback-leibler and renormalized entropies: Applications to electroencephalograms of epilepsy patients. *Physical Review E*, 62:8380–8386.

Quiroga, R., Arnhold, J., and Grassberger, P. (2000b). Learning driver-response relationships from synchronization patterns. *Physical Review E*, 61(5):5142–5148.

Quiroga, R., Kreuz, T., and Grassberger, P. (2002). Event synchronization: A simple and fast method to measure synchronicity and time delay patterns. *Physical Review E*, 66(4):041904.

Renyi, A. (1959). On measures of dependence. *Acta Mathematica Academiae Scientiarum Hungaricae*, 10:441–451.

Renyi, A. (1961). On measures of entropy and information. In of California Press, U., editor, *4th Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 547–561.

Richman, J. and Moorman, J. (2000). Physiological time-series analysis using approximate entropy and sample entropy. *American Journal of Physiology. Heart and Circulatory Physiology*, 278(6):H2039–2049.

Rissanen, J. (1992). *Stochastic Complexity in Statistical Inquiry*. World Scientific, Singapore.

Roberts, S. and Everson, R. (2001). *Independent component analysis: Priciples and Practice*. Cambridge University Press, Cambridge.

Rogowski, Z., Gath, I., and Bental, E. (1981). Brain electrical activity analysis using wavelet-based informational tools on the prediction of epileptic seizures. *Biological Cybernetics*, 42:9–15.

Romano, M., Thiel, M., Kurths, J., and Grebogi, C. (2007). Estimation of the direction of the coupling by conditional probabilities of recurrence. *Physical Review E*, 76:036211.

Rosenblum, M. and Pikovsky, A. (2001). Detecting direction of coupling in interacting oscillators. *Physical Review E*, 64(4):045202.

Rosso, O., Martin, M., and Plastino, A. (2002). Brain electrical activity analysis using wavelet-based informational tools. *Physica A*, 313:587.

Roulston, M. (1997). Significance testing of information theoretic functionals. *Physica D*, 110(1–2):62–66.

Schelter, B., Winterhalder, M., Eichler, M., Peifer, M., Hellwig, B., Guschlbauer, B., Lucking, C., Dahlhaus, R., and Timmer, J. (2006). Testing for directed influences in neuroscience using partial directed coherence. *Journal of Neuroscience Methods*, 152:210–219.

Schiff, S., Aldroubi, A., Unser, M., and Sato, S. (1994). Fast wavelet transformation of EEG. *Electroencephalography and Clinical Neurophysiology*, 91:442.

Schiff, S., So, P., and Chang, T. (2000). Detecting dynamical interdependence and generalized synchrony through mutual prediction in a neural ensemble. *Physical Review E*, 54(6):6708–6724.

Schmid, M., Conforto, S., Bibbo, D., and D'Alessio, T. (2004). Respiration and postural sway: Detection of phase synchronizations and interactions. *Human Movement Science*, 23(2):105–119.

Schonwald, S., Gerhardt, G., Santa-Helena, E., and Chaves, M. (2003). Characteristics of human EEG sleep spindles assessed by gabor transform. *Physica A: Statistical Mechanics and its Applications*, 327(1–2):180–184.

Schreiber, T. (1998). Constrained randomization of time series data. *Physica Review Letter*, 80.

Schreiber, T. (2000). Measuring information transfer. *Physical Review Letters*, 85(2):461–464.

Schreiber, T. and Schmitz, A. (1996). Improved surrogate data for nonlinearity tests. *Physica Review Letter*, 77:635–638.

Schuster, H. (1988). *Deterministic Chaos (2nd Edition ed.)*. Physik Verlag, Weinheim.

Schweizer, B. and Wolff, E. (1981). On nonparametric measures of dependence for random variables. *Annals of Statistics*, 9(4):879–885.

Scott, D. (1979). On optimal and data-based histograms. *Biometrika*, 66(3):605–610.

Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423 and 623–656.

Sheather, S. and Jones, M. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society B*, 53(3):683–690.

Shiau, D., Iasemidis, L., Suharitdamrong, W., Dance, L., Chaovalitwongse, W., Pardalos, P., Carney, P., and Sackellares, J. (2003). Detection of the preictal period by dynamical analysis of scalp EEG. *Epilepsia*, 44(S9):233–234.

Silverman, B. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.

Smirnov, D. and Andrzejak, R. (2005). Detection of weak directional coupling: Phase-dynamics approach versus state-space approach. *Physical Review E*, 71(3):036207.

Smirnov, D. and Bezruchko, B. (2003). Estimation of interaction strength and direction from short and noisy time series. *Physical Review E*, 68(4):046209.

Song, L. and Epps, J. (1998). Improving separability of EEG signals during motor imagery with an efficient circular laplacian. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '06)*, volume 2, pages 1048–1051.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology*, 15:72–101.

Srinivasan, R., Nunez, P., and Silberstein, R. (1998). Spatial filtering and neocortical dynamics: Estimates of EEG coherence. In *IEEE Transactions on Biomedical Engineering*, volume 45, pages 814–826.

Staniek, M. and Lehnertz, K. (2008). Symbolic transfer entropy. *Physical Review Letters*, 100(15):158101.

Steuer, R., Kurths, J., Daub, C., Weise, J., and Selbig, J. (2002). The mutual information: Detecting and evaluating dependences between variables. *Bioinformatics*, 18(2):S231–S240.

Sturge, H. (1926). The choice of a class interval. *Journal of the American Statistical Association*, 21(1):65–66.

Takens, F. (1980). Deteting strange attractors in turbulence. *Dynamical Systems and Turbulence, Lecture Notes in Mathematics*, 898:366–381.

Terrell, G. and Scott, D. (1985). Oversmooth nonparametric density estimates. *Journal of the American Statistical Association*, 80:209–214.

Theiler, J. (1986). Spurious dimension from correlation algorithms applied to limited time-series data. *Physical Review A*, 34:2427–2432.

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

Thenevaz, P. and Unser, M. (2000). Optimization of mutual information for multiresolution image registration. *IEEE Transactions on Image Processing*, 9:2083–2099.

Tourassi, G., Frederick, E., Markey, M., and Floyd, C. (2001). Application of the mutual information criterion for feature selection in computer-aided diagnosis. *Medical Physics*, 28(12):2394–2402.

Trappenberg, T., Ouyang, J., and Back, A. (2006). Input variable selection: Mutual information and linear mixing measures. In *IEEE Transactions on Knowledge and Data Engineering*, volume 18, pages 37–46.

Trulla, L., Giuliani, A., Zbilut, J., and Webber, C. (1996). Recurrence quantification analysis of the logistic equation with transients. *Physics Letters A*, 223:255–260.

Tsallis, C. (1988). Possible generalization of boltzmann-gibbs statistics. *Journal of Statistical Physics*, 52:479–487.

Tucker, D. (1993). Spatial sampling of head electrical fields: The geodesic sensor net. *Electroencephalography and Clinical Neurophysiology*, 87:154–163.

Tukey, J. and Mosteller, F. (1977). *Data Analysis and Regression*. Addison-Wesley, Reading, MA.

Tykierko, M. (2008). Using invariants to change detection in dynamical system with chaos. *Physica D*, 237:6–13.

Vejmelka, M. and Palus, M. (2008). Inferring the directionality of coupling with conditional mutual information. *Physical Review E*, 77(2):026214.

Viglione, S. and Walsh, G. (1975). Proceedings: Epileptic seizure prediction. *Electroencephalography and Clinical Neurophysiology*, 39:435–6.

Walter, P. (1975). *Ergodic Theory - Introductory Lectures Notes*. Springer, Berlin.

Wand, M. and Jones, M. (1993). Comparison of smoothing parameterizations in bivariate kernel density estimation. *Journal of the American Statistical Association*, 88(422):520–528.

Wand, M. and Jones, M. (1995). *Kernel Smoothing*. Chapman and Hall, London.

Weinand, M., Carter, L., El-Saadany, W., Sioutos, P., Labiner, D., and Oommen, K. (1997). Cerebral blood flow and temporal lobe epileptogenicity. *Journal of Neurosurgery*, 86(2):226–232.

Wicks, R., Chapman, S., and Dendy, R. (2007). Mutual information as a tool for identifying phase transitions in dynamical complex systems with limited data. *Physical Review E*, 75(5):051125.

Winterhalder, M., Schelter, B., Hesse, W., Schwab, K., Leistritz, L., Klan, D., Bauer, R., Timmer, J., and Witte, H. (2005). Comparison of linear signal processing techniques to infer directed interactions in multivariate neural systems. *Signal Processing*, 85(11):2137–2160.

Winterhalder, M., Schelter, B., Hesse, W., Schwab, K., Leistritz, L., Timmer, J., and Witte, H. (2006). Detection of directed information flow in biosignals. *Biomedizinische Technik*, 51:281–287.

Yao, J. (1993). Gabor transform: Theory and computations. In *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, volume 2034, pages 137–148.

Zografos, K. (1993). Asymptotic properties of phi-divergence statistic and its applications in contingency tables. *International Journal of Mathematical and Statistical Science*, 2(1):5–22.