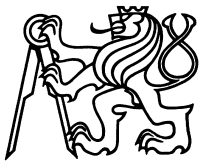
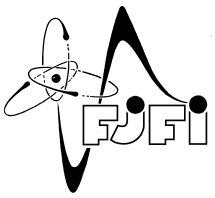


CZECH TECHNICAL UNIVERSITY IN PRAGUE
Faculty of Nuclear Sciences and Physical Engineering



Biomarker Analysis of Psychiatric Patients using EEG Signal Analysis and Machine Learning

Analýza biomarkerů psychiatrických pacientů pomocí analýzy EEG signálu a strojového učení

Diploma thesis

Author: **Bc. Miroslav Kovář**

Supervisor: **M.Sc. M.A. Sebastián Basterrech, Ph.D.**

Academic year: 2018/2019

- Zadání práce -

- Zadání práce (zadní strana) -

Acknowledgment:
Some acknoledgment here.

Author's declaration:
I declare that this research project is entirely my own work and I have listed all the used sources in the bibliography.

Prague, April 7, 2019 Bc. Miroslav Kovář

Název práce:

Analýza biomarkerů psychiatrických pacientů pomocí analýzy EEG signálu a strojového učení

Autor: Bc. Miroslav Kovář

Obor: Aplikace přírodních věd

Zaměření: Matematická informatika

Druh práce: Diplomová práce

Vedoucí práce: M.Sc. M.A. Sebastián Basterrech, Ph.D., Artificial Intelligence Center, FEE, CTU Prague

Abstrakt:

Klíčová slova:

Title:

Biomarker Analysis of Psychiatric Patients using EEG Signal Analysis and Machine Learning

Author: Bc. Miroslav Kovář

Abstract: Major depressive disorder has high population prevalence and significant impact on quality of life, and in its severe form may result in suicide. Hence, substantial amount of suffering may be alleviated by facilitating early diagnosis and accurate prediction of treatment response. In this thesis, we apply techniques of nonlinear dynamical analysis and iterative optimization of convolutional neural network models to the problem of prediction of depression severity and treatment response on an original dataset of EEG recordings. Using nonlinear measures, we obtain classification accuracy of approximately 75% on both tasks, and provide evidence that the Largest Lyapunov exponents may be relevant for treatment response prediction. Moreover, we find that larger, more diverse datasets may be needed to develop more complex machine learning models for this problem. In spite of the difficulties of the studied problems, we provide new insights on mental disorder diagnostics and show the potential of nonlinear measures for analysing EEG signals of depresive patients.

Key words: machine learning, nonlinear dynamical analysis, convolutional neural networks, electroencephalography, major depressive disorder

Contents

Acronyms

AAFT Amplitude Adjusted Fourier Transform.

ADFD Average Displacement from Diagonal.

AFN Average False Neighbors.

BCI Brain Computer Interface.

CD Correlation Dimension.

CNN Convolutional Neural Network.

CS Cosine Similarity.

CSP Common Spatial Patterns.

DFA Detrended Fluctuation Analysis.

DMI Delayed Mutual Information.

ECT Electroconvulsive Therapy.

EEG Electroencephalography, Electroencephalogram.

FBCSP Filter Bank Common Spatial Patterns.

fBm fractional Brownian motion.

fGn fractional Gaussian noise.

FNN False Nearest Neighbors.

FTPR Fourier Transform Phase Randomization.

GAF Gramian Angular Field.

GPU Graphical Processing Unit.

HD Higuchi fractal Dimension.

HE Hurst Exponent.

iAAFT improved Amplitude Adjusted Fourier Transform.

ILD Integral Local Deformation.

LLE Largest Lyapunov Exponent.

LR Logistic Regression.

PCA Principal Components Analysis.

RP Recurrence Plot.

SE Sample Entropy.

SVD Singular Value Decomposition.

SVM Support Vector Machine.

List of Figures

List of Tables

Introduction

Depression and its Diagnosis

Depression is one of the most common brain disorders - it affects 121-300 million people worldwide, and this number is expected to increase in the future [1, 2]. It is estimated that over 20 million people in the United States alone have this mood disorder, but only 50% have been diagnosed [3]. Although effective treatments are known, World Health Organization (WHO) estimates that fewer than half of affected individuals receive those treatments. According to WHO, major barriers include insufficient resources, lack of properly trained practitioners, inaccurate assessment and misdiagnosis [2]. Moreover, the diagnosis is further complicated by the fact that depressive symptoms often mimic other disorders, and coexisting conditions may confound diagnosis [3]. Indeed, self-assessed questionnaires are often inaccurate, and structured or semi-structured interviews (SDIs) require time and expertise of trained professionals.

For these reasons, it is important that affordable, noninvasive, portable and easy to use tools to aid its diagnosis are developed. Electroencephalography (EEG)[1], a method of recording spatiotemporal evolution of electrical activity in the cortical regions near the scalp, may be one such tool thanks to its comparatively low cost and easy recording process. Unlike SDIs and laboratory methods such as dexamethasone suppression test, it assesses ongoing activity in the responsible organ itself. Moreover, unlike measurements of glucose utilization or blood oxygenation in the brain, it captures the electrical activity directly [4]. These properties make EEG technology potentially useful as a practical tool to aid depression treatment, which may, for example, improve diagnosis accuracy, predict treatment response, or track development of depression severity over time. Fully realizing this potential would likely require more detailed understanding of brain dynamics [5]. At present, however, EEG signal analysis still remains relatively unpopular method of depression diagnosis in contrast with SDIs [3]. This may be due to insufficient standardization of research, overly positive interpretation of findings, and relatively small volume of limited datasets, together impeding the possibility of meaningful objective conclusions or meta-studies [4].

Given this situation, the following areas were previously identified as central to improvement of the state of the art of EEG analysis [5, 4]:

1. development of more effective EEG signal analysis techniques,

2. identification of reliable EEG-based biomarkers[2] of depression,

3. understanding of the brain dynamics with focus on differences between healthy and diseased states,

[1] In this work, we will use the same abreviation for electroencephalography (recording method) and electroencephalogrpham (the recorded data) where the distinction is apparent from the context.

[2] Biomarker is defined as objectively measurable indicator of the biological state of the organism [4].

4. standardization of recording conditions, preprocessing, result analysis and other research procedures,

5. enrollment of larger, hypothesis-driven cohort studies, and

6. successful clinical applications to estabilish trust amongst mental health practitioners.

Research Objectives

In this work, our aim is to contribute positively to this state of the art of this important area by analyzing the limited, but relatively large dataset of depressed patients using two approaches previously advocated as effective EEG analysis techniques [6, 7, 8, 9, 10, 11, 12, 13, 5]: nonlinear dynamical analysis and deep learning. In the process, our objective is to address the areas mentioned above by:

1. providing evidence to or against the effectiveness of these techniques,

2. providing evidence to or against the status of nonlinear dynamical measures as reliable biomarkers of depression,

3. discovering possible relationships of depression severity or future responsiveness to treatment to distributions of these nonlinear measures across brain regions,

4. finding an effective procedure for computing the nonlinear measures and evaluate these procedures,

5. performing this analysis on relatively large dataset, and, eventually,

6. implementing a diagnosis-aid software tool which will use the techniques to help asses depression severity and predict treatment responsiveness.

Problem Statement

The dataset was provided by the Czech National Institute of Mental Health[3]. It comprises of 266 multivariate EEG signal recordings of various durations and sampling rates obtained from 133 depressed patients, each recorded on two ocassions: at the beginning of the treatment and, 4 weeks later, after drug adminstration. On both of those ocassions, patient's depression score (a number quantifying patient's depression severity) was measured by a trained professional using a standardized questionnaire. For more details about the dataset, see Section 2.1. Using these recordings and depression scores, we aim to solve the following problems:

diagnosis: predict the patients depression score range measured *at the time of the recording* based on an EEG signal sample obtained *at the time of the recording*, and

prognosis: predict the patients depression score range measured *during the first visit* based on an EEG signal sample obtained *during the first visit*.

For the purposes of abstraction and tractability, we define these problems as classification tasks. In order to analyze the EEG signals and extract relevant features, we separately use and evaluate two approaches previously evidenced in the literature to be effective at distinguishing healthy and diseased patients with depression and other psychiatric disorders:

[3] http://www.nudz.cz/en/

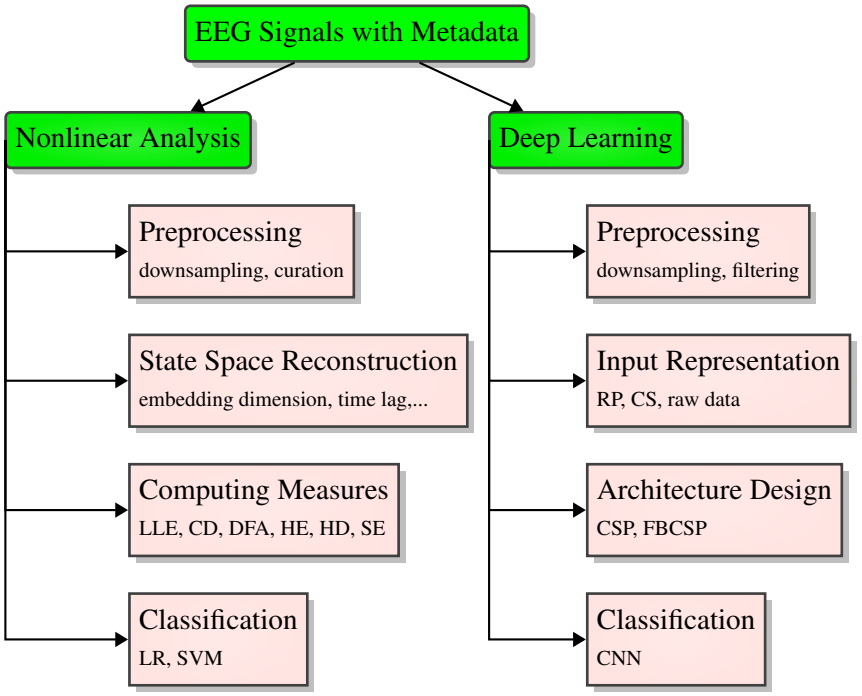


Figure 1: Diagram of the thesis.

nonlinear analysis approach: compute selected nonlinear measures using nonlinear dynamical system analysis of the EEG signals, and use machine learning techniques to perform classification, and

deep learning approach: use convolutional neural networks to both extract relevant features from an input representation of the EEG signals, and perform the classification task.

I think I am just duplicating myself multiple times here, but supervisor thinks something "more friendly" description is missing.

The overall structure of the approaches is depicted in Figure 1. In each approach, we first downsample the signals to common sampling frequency. In the **nonlinear analysis approach**, we also shorten the recordings to common recording length. Then, we proceed with computing selected nonlinear measures. Since this involves (partial) reconstruction of the state space of the dynamical system in which to embed an attractor, we evaluate, analyze and compare the results of multiple embedding parameter estimation algorithms. After selecting the input parameters and computing the nonlinear measures, the differences in values of those measures across brain regions between various subgroups of the dataset are inspected, as well as the correlations with depression score and treatment response. Finally, the classification task is performed and its results evaluated.

In the **deep learning approach**, each recording is optionally filtered, and divided into multiple windows to increase the number of samples. Because neural networks are capable of processing various modalities, an appropriate input data representation is selected. Then, we design and evaluate a number of neural network architectures, and finally perform the classification task and evaluate the results.

Content Organization

Chapter 1: Nonlinear Time Series Analysis

The topic of nonlinear dynamical analysis in relation EEG analysis is presented. The concepts of

a dynamical system and its state spaces are introduced, along with methods of describing them used further. State space reconstruction using the method of time delays is discussed in detail, followed by description of algorithms for time delay and embedding dimension selection. Nonlinear dynamical measures and algorithms for computing them are described further, along with algorithms to used to validate the results. This chapter finishes with review of the current state of the literature on the topic of depression diagnosis using both linear techniques and nonlinear dynamical analysis, followed by limitations of this approach to EEG signal analysis.

Chapter 2: Nonlinear Analysis Approach

Aspects of the dataset and its metadata are described in detail, followed by discussion of the preprocessing step. The process of state space reconstruction is carried out using the algorithms described in the previous chapter, and the results are analyzed and discussed in detail. The procedure of computing each nonlinear measure is described and compared with existing literature. For each nonlinear measure, the obtained results based on input parameters are evaluated, and reasoning behind selection of input parameters used to obtain the final results is provided. Differences between the obtained results in various subgroups of the dataset are analyzed. Finally, the experimental settings for the classification tasks are described, and the obtained results are presented.

Chapter 3: Deep Learning Approach

Convolutional neural networks are put into historical context and their relevant features are described. A signal processing technique which inspired our final selection of neural network architecture is presented. Possible approaches to input representation are discussed. The most successful architectures are introduced, and the chapter concludes with description of the obtained classification resuts.

Conclusion

In the final chapter, the contributions to scientific knowledge made in this thesis are discussed, and recommendations for future work are made.

Appendix

The appendix includes figures and tables which were left out of the main text for clarity.

Chapter 1

Nonlinear Time Series Analysis

The nature is constantly undergoing change. Around us, we can observe many processes evolving in time. Some of the aspects of these processes, we can measure, and attempt to discover apparent patterns in those measurements. The simplest of those patterns are periodicities, probably best exemplified, and first noticed by humans, are the motions of the sun and the moon. Weather, on the other hand, is an example of processes seemingly defying any simple description. Those examples represent two classes of processes existent before the rise of nonlinear dynamics:

Deterministic processes are processes for which there is a unique consequent to every state [14]. Before advent of nonlinear dynamics, these were thought necessarily periodic (or quasi-periodic), fully describable by its Fourier spectrum [15].

Stochastic process are processes for which there is a probability distribution of possible to each state consequents [14], i.e. they contain pure randomness as driving forces [15].

Nonlinear dynamical analysis studies a third class of processes, which are irregular, nonperiodic, yet still deterministic. Every nonperiodic, deterministic process is nonlinear (but not necessarily the other way around). Existence of these processes was known already in mid-19th century to J. C. Maxwell, but the field began fully developing only with the rise of feasibility of complex enough numerical simulations, peaking in 1980s. Majority of new insights was generated precisely using numerical simulations [15].

1.1 Connection to Electroencephalography

Electroencephalography (EEG) is a noninvasive method of measuring fluctuations of electric potentials near the skull caused by synchronized firing of neurons in the upper cortical layers. Electroencephalogram is a record of these fluctuations measured over a period of time measured by 8-256 electrodes placed on patient's head [16].

Although EEG has significantly lower spatial resolution in comparison with other functional neuroimaging techniques such as functional magnetic resonance sampling (fMRI) and magnetoencephalography (MEG) [17], and although it enables measuring only neural activity near the cortical surface, as a depression diagnostic tool, it has numerous benefits. Importantly, its significantly lower costs [18, 19], high portability, and ease of operation imply increased availability to the patients [20]. Moreover, it is perfectly noninvasive, which means less complications such as claustrophopia or anxiety [21].

Due to the phenomenon of neural oscillations, patterns may appear in multiple frequency bands, from slow cortical potentials of δ-waves at 0.5-4 Hz, to high γ frequency band at 70-150 Hz. Patterns of oscillatory activity in various frequency bands have been linked to various mental states [22, 23]

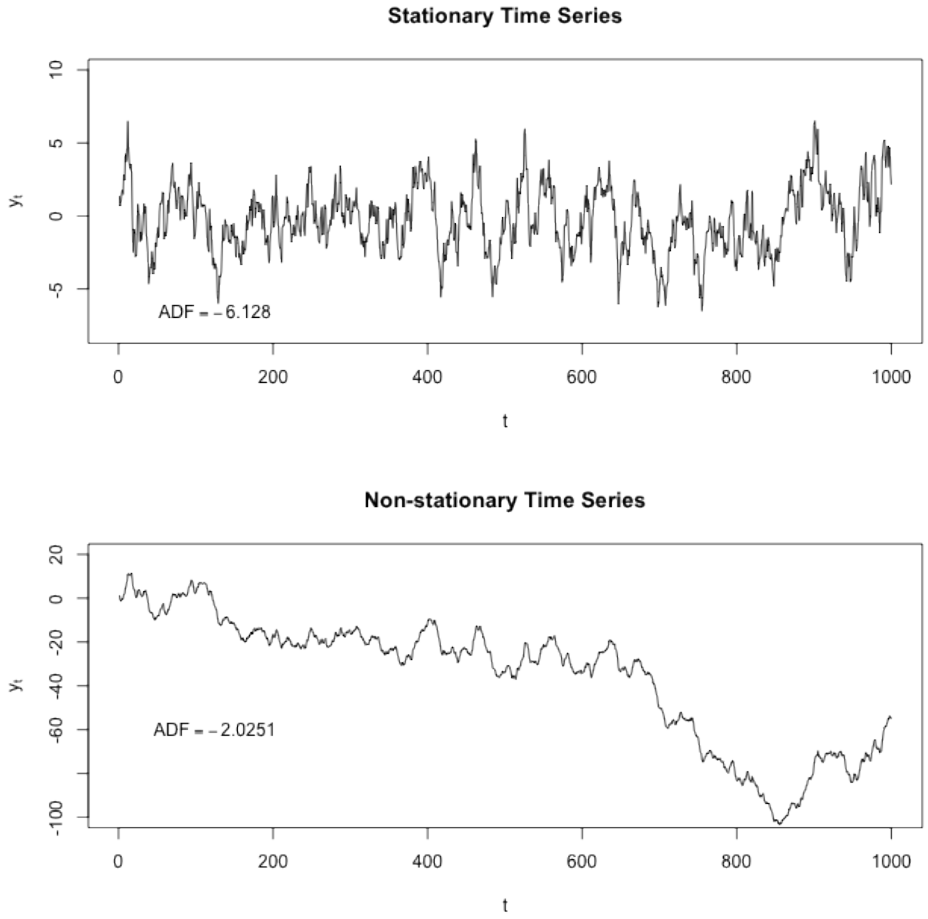


Figure 1.1: A comparison of stationary and nonstationary time series. ([28])

and diseases such as epilepsy [24], tremor [25], Parkinson's disease and depression [26]. Many of the diseases, including depression, share common oscillatory patterns known as thalamocortical dysrythmia, characterized by descrease in normal resting-state α (8-12 Hz) activity slowing down to θ (4-8 Hz) frequencies, accompanied by increase in β and γ (25-50 Hz) activity [27].

The science of EEG signal analysis as a diagnostic tool brings compelling clinical promise as a result of the aforementioned benefits. However, it also presents multiple technical and conceptual challenges. In the following, we will define two of properties of EEG important in terms of its analysis - nonstationarity and nonlinearity.

Definition 1 ([28]). *A series $\{x_t\}_{t\in\mathbb{Z}}$ is called **stationary**, if $\{x_t\}_{t\in\mathbb{Z}}$ for any set of times $t_1, t_2, \ldots, t_n$ and any $k \in \mathbb{N}$, $P[x_{t_1}, x_{t_2}, \ldots, x_{t_n}] = P[x_{t_1+k}, x_{t_2+k}, \ldots, x_{t_n+k}]$, i.e. the joint probability distribution of $\{x_t\}_{t\in\mathbb{Z}}$ is not a function of time. It is called **nonstationary**, if it is not stationary.*

The fundamental objective of time series analysis is to unveil the probability law which underlies the observed time series. A popular approach for doing this is to constrain the law to a class of the models, and then to find the most plausible model within this class. Two large distinct classes are linear and nonlinear models, and there are many different subclassifications of both. Historically, nonlinear models started to fully develop after it became apparent that some time series posses "nonstandard" (today called

nonlinear) features, such as nonlinear relationship between expectations of temporally delayed variables, variation of predictability over state space and sensitivity to initial conditions (chaoticity). These features are beyond the scope of standard linear models, such autoregressive (AR), moving average (MA) models and their derivations [29], and time possessing them are called **nonlinear time series**.

In a nonlinear system therefore, not only randomness, but also unmeasurable perturbations to the system can lead to apparent irregularity.

EEG signals are known to be *nonstationary and nonlinear* [30, 5]. Moreover, they are prone to be infected with *noise* due to imperfect isolation from the surrounding environment and patients involuntary movements, such as blinking or heartbeat. Since some patterns do not activate relative to a stimulus, a successful classifier must be able to detect a pattern regardless of its starting time, or find one. This is further complicated by the fact that EEG records are relatively high dimensional - typical headsets containing 8-256 electrodes and sampling at 256 Hz result in 2048-65536 data points per second. In spite of this, EEG has low spatial resolution compared to the complexity of the brain. In other words, the electrodes are recording the cumulative activity of large number of neurons. Furthermore, the transmission through the scalp may blur the signals [15].

1.2 Dynamical Systems

1.2.1 Definitions

Definition 2 ([15])**.** *Assume that state of a system can be fully described by a finite set of d variabes, such that each state corresponds to a point $\xi \in M$, where M is a d-dimensional differentiable manifold. Then we will call M a (true)* ***state space*** *or, equivalently, a (true)* ***phase space****, and d its (true)* ***dimension****.*

Although in this study, we will only consider Euclidean state space M, the true state space is needs not necessarily be Euclidean. For example, if some of the state variables are angles, the state space exhibits toroidal topology. However, any topological manifold is locally Euclidean [31] and, since, in EEG signal analysis both M and d are unknown, we have no other alternative than to work in Euclidean state space M.

Definition 3 ([15])**.** *Let $\xi : \mathbb{R} \to \mathbb{R}^d$ be an $d \in \mathbb{N}$ dimensional state (phase) space vector dependent on time, and $\mathbf{F}$ a smooth vector field in $\mathbb{R}^d$. A* ***deterministic dynamical system***[1] *is described by a set of d first-order differential equations*

$$\frac{d}{dt}\xi(t) = \mathbf{F}(\xi(t)), \qquad t \in \mathbb{R}_0^+,$$

such that there exists a unique[2] *diffeomorphic*[3] *function $f^t : M \to M$ satisfying*

$$\xi(t) = f^t(\xi(0))$$

for any initial condition $\xi(0)$. We will call this mapping ***state evolution function****, and vector field $\mathbf{F}$* ***dynamics of the system****. We call the system linear if $\mathbf{F}$ is a linear vector field.*

[1] In this work, we are going to assume that the brain is a deterministic dynamical system, and that any stochastic component is small and does not change nonlinear properties of the system. Thus, by the term dynamical system, we will always mean a deterministic dynamical system. This assumption is necessary for nonlinear dynamical analysis. On the other hand, nonlinear dynamical analysis also provides techniques (see Section 1.5) which can partially address (yet not fully answer) justification for this assumption. The question whether the brain is truly deterministic is open [5, 32], but it is often thought to be probable that the brain is a nonlinear, deterministic, dissipative (i.e. exchanging energy with its environment) system [5].

[2] In other words, we assume that $\mathbf{F}$ satisfied the conditions of the uniqueness theorem of differential equations.

[3] This means that f^t is smooth, and has smooth inverse.

In late 1800s, H. Poincaré developed a geometric approach to analyzing the stability (asymptotic evolution) of these systems via examination of the solution $(\xi_1(t), \xi_2(t), \ldots, \xi_d(t))$ as a *trajectory* in the phase space M (assuming the solution is known, e.g. measured). These ideas were later extended into deeper understanding of chaos in dynamical systems [33].

In general, any system with temporally changing state is dynamic. A *deterministic* dynamical system is describable by a model giving precise transition of a system from one state to another in time. This means that total description of system's evolution in its phase space (its *trajectory*) is given by the initial state and a set of equations $\mathbf{F}$ (if $\mathbf{F}$ satisfies certain reasonable properties given by the uniqueness theorem). With *stochastic* dynamical systems, such mapping is not possible, since these transitions are not given precisely.

A nonlinear dynamical system is a system where the differential equations describing its dynamics are nonlinear. Unlike in a linear system, changes in the initial state of a non-dynamical system are allowed to have a nonlinear relationship to the state space trajectory of the system [30].

It is important to note the obvious fact that in the case of EEG signal analysis, it is not possible to measure the true state of the system $\xi(t)$. In fact, the observed variables are only a function of the true state of the system. To capture this, we will define a measurement function $s : \mathbb{R}^d \subset M \to \mathbb{R}^{d'}$, $d' \ll d$, as

$$s(\xi(t)) = \mathbf{x}(t) + \eta(t),$$

where $\eta(t)$ is measurement noise, which encompasses the measurement error and the noise coming from the measurement conditions. In the following text, we will usually disregard the measurement function and assume we have direct access to $\xi(t)$ and thus assuming the observed state equals the true state $\mathbf{x}(t) \equiv \xi(t)$, since is usually neglected in the explained theory. Nevertheless, it is important to remember that the noise term have affect the results we obtain using the theory explained theory.

1.2.2 Attractor

Depending on the properties of $\mathbf{F}$, there are several possibilities of how the system might evolve when as $t \to \infty$. In the following, we will focus on so called dissipative dynamical systems (of which brain is considered a member [5]).

Definition 4 ([34])**.** *A dynamical system is called dissipative, when it is the case that*

$$E[|\det \mathbf{J}_\mathbf{F}|] < 1, \tag{1.1}$$

where $\mathbf{J}_\mathbf{F}$ is the Jacobian of vector field $\mathbf{F}$ and the expectation is taken over the state space M. In other words, average state space volume of a set of initial conditions of non-zero measure is contracted as the system evolves.

For these systems, after sufficient passage of time, all future states will continue evolving on a bounded, time-invariant subset of M. This subset is a geometrical object called an **attractor**. Example of four basic attractors, point, limit, torus and chaotic attractors, can be seen in Figure 1.2. Examples of several chaotic attractors are shown in Figure 1.4 [33, 35, 36].

Statistical methods can be used to analyze observations of a complex system. Another branch of mathematics providing us with powerful tools to study systems with apparently complex behavior is chaos theory, which studies so called chaotic systems. These systems exhibit dynamics which extend volumes of clusters of initially nearby states in some directions. Although there is mixed evidence on low dimensional chaos in the brain and in the biological systems in general, its techniques have found many successful applications in their analysis [37].

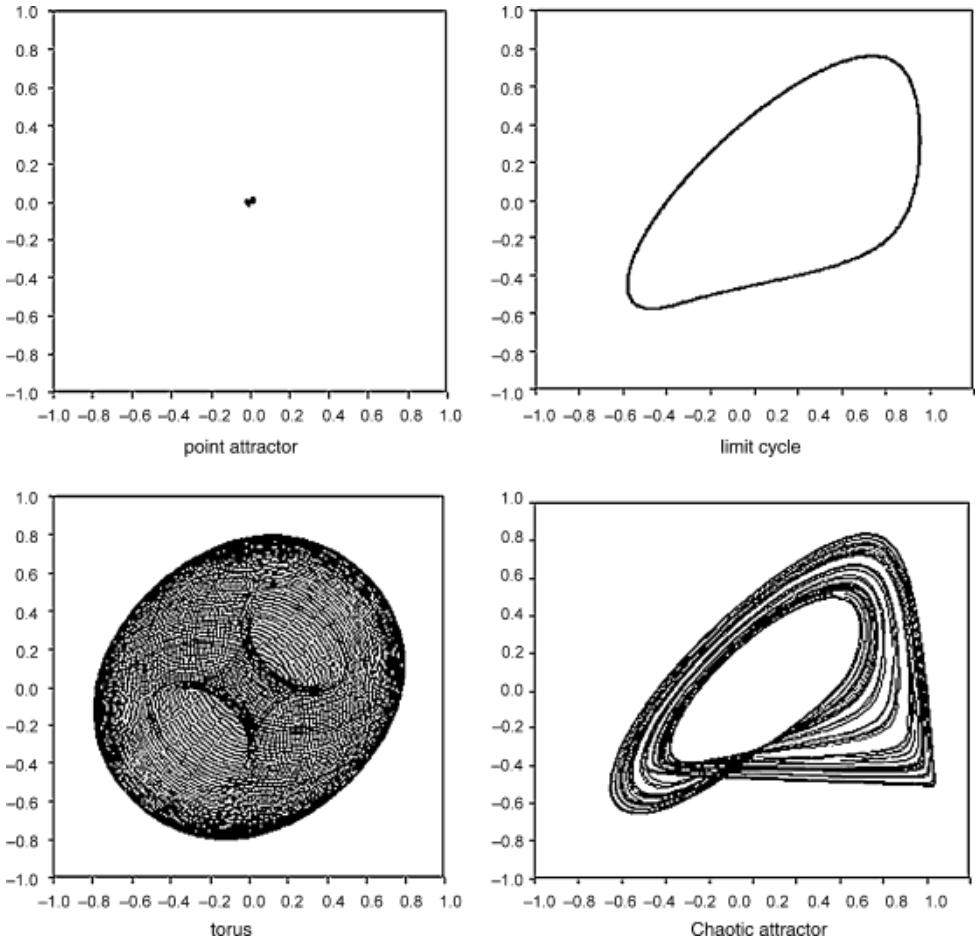


Figure 1.2: Visualization of four common attractor types (units are arbitrary). Left to right, top to bottom: **Point attractor** is the only type of attractor of linear deterministic dissipative systems. It consist of a single final state to which all points from the corresponding region of attraction evolve to. **Limit cycle** corresponds to a periodic dynamical system. It is formed by set of states visited periodically, consituting a trajectory through the state space. **Torus attractor** corresponds to a quasi-periodic dynamical system, resulting (in this example) from a superposition of two periodic oscillations. **Chaotic** (strange) **attractor**, characteristic of dynamical systems with extending (instead of shrinking) volumes in *some* directions. Corresponding dynamical system may appear stochastic, yet still be completely deterministic [15]. ([5])

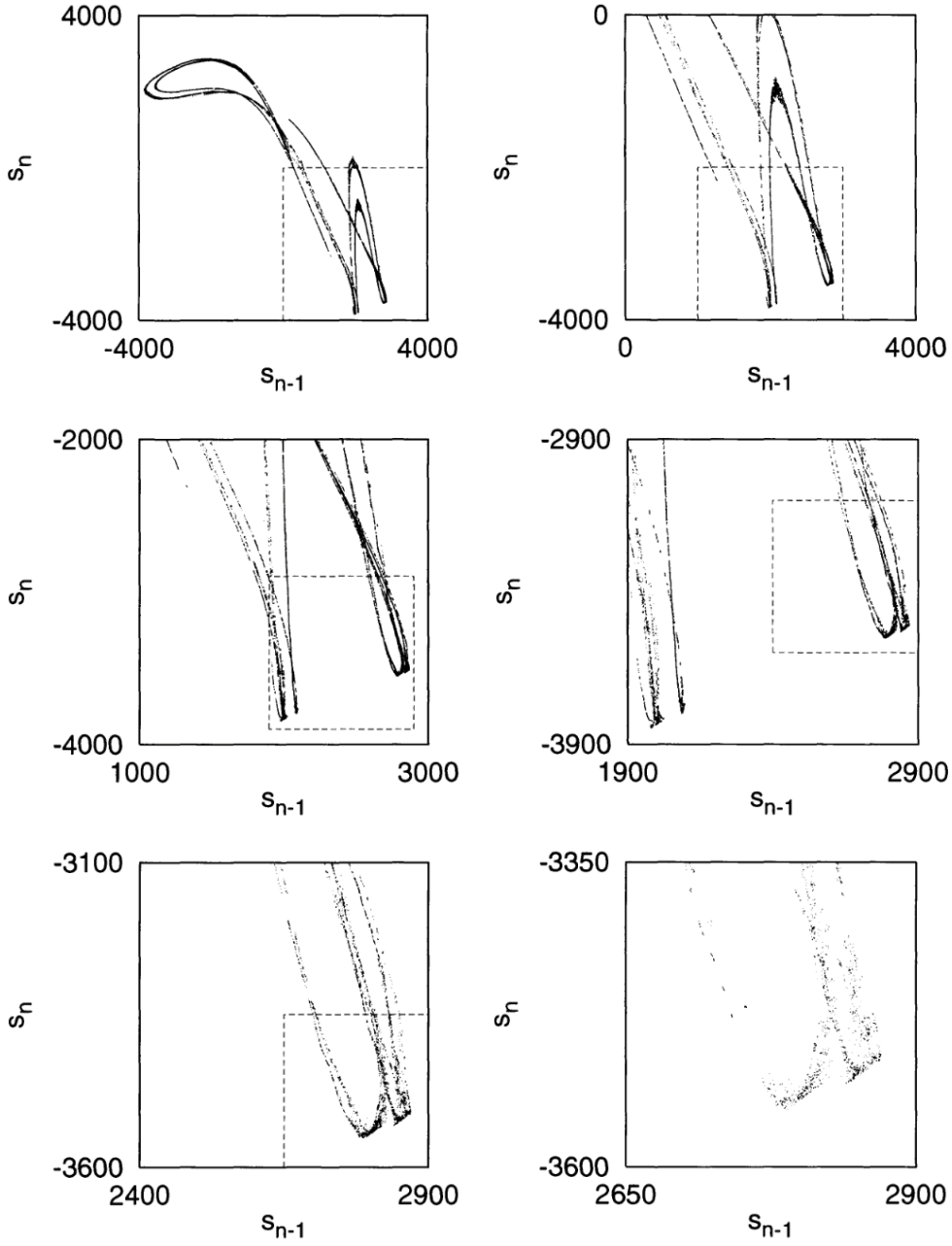


Figure 1.3: Noise-reduced visualization of successive enlargements of highly self-similar attractor [34].

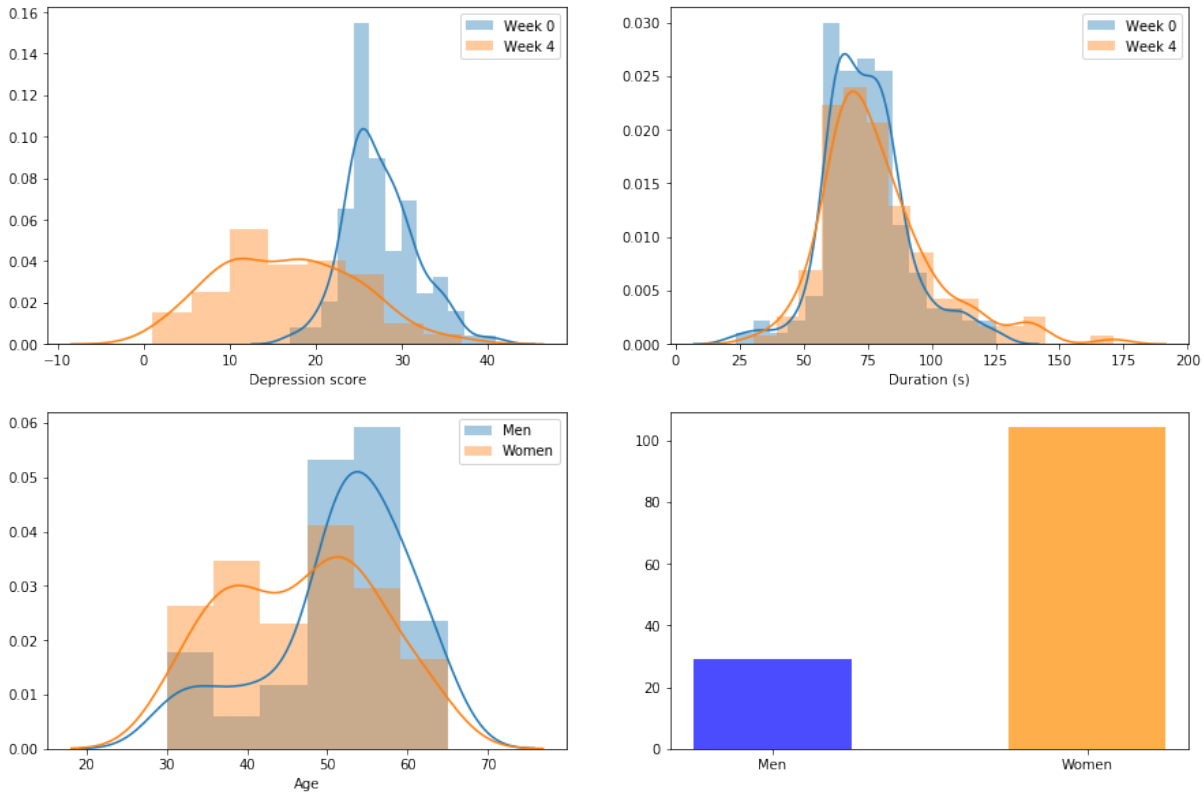


Figure 1.4: Lorenz, Roessler and Mackey-Glass attractors generated from corresponding systems of differential equations [33]. These systems are known to exhibit chaotic dynamics for certain parameter values [35].

Attractors of chaotic systems, coined by Ruelle and Takens in 1971 *chaotic* (strange) *attractors* [38], possess interesting properties. Since, as mentioned, attractors are bounded, the divergence of nearby states due to chaos eventually stops and the two trajectories fold together [34]. This continuous expansion and folding creates a "self-similar", *fractal* object. An example of a strange attractor can be seen in Figure 1.3.

This self-similarity can be quantified by a class of scalar measures called *fractal dimensions*. Indeed, we will use one of the members of this class - correlation dimension - in our experiments, and will treat it in detail. In addition, let us give another example of a fractal dimension, called box-counting dimension, be useful for understanding the implications of Taken's embedding theorem (1) in Section 1.3.1:

Definition 5 ([39])**.** *Let F be any non-empty bounded subset of $\mathbb{R}^n$, and let $N_\epsilon(F)$ be the smallest number of sets of diameter at most ϵ ("mesh cubes") which can cover F. Then, the* ***box-counting dimension*** *(also known as Minkowski–Bouligand dimension) is defined as*

$$d_0(F) = \lim_{\epsilon \to 0} -\frac{\log N_\epsilon(F)}{\log \epsilon}, \tag{1.2}$$

if it exists.

Intuitively, the number of mesh cubes of side ϵ intersecting F gives an indication about how irregular the set is when inspected at scale ϵ, and the box-counting dimension reflects "how rapidly" the irregularities develop as $\epsilon \to 0$ [39].

1.2.3 Stationarity

Nonstationarity is a phenomenon which considerably complicates practical analysis of dynamical systems. All the techniques presented in this text assume stationary process, since this assumption is a prerequisite to deterministic chaos [40]. We will call system **nonstationary** if the dynamics of the system are influenced by causes lying outside of them (and **stationary** if the opposity is true). In ergodic theory (study of the invariant measures of dynamical systems), the concept of stationarity is defined more rigorously. However, these definitions are not suited for numerical applications [15]. Nevertheless, a relevant subset of nonstationary systems can be defined more explicitly:

Definition 6 ([15])**.** *A dynamical system is called* ***nonautonomous*** *if its dynamics* $\mathbf{F}$ *are explicitly dependent on time:*

$$\frac{d}{dt}\xi(t) = \mathbf{F}(\xi(t), t), \qquad t \in \mathbb{R}_0^+.$$

No reliable tests for nonstationarity in this strong sense exist. There is another common definition of a stationary process (sometimes referred to as weak stationarity). A process is called **weakly stationary**, if all statistical second-order quantities (like mean, variance, and power spectrum) are independent of the absolute time, and at most function of relative times [40].

This weaker definition employs only linear quantities, and is therefore not strictly suitable for nonlinear time series analysis. On the other hand, statistical tests of this property exist. In our study, we use the following test for weak stationarity discussed by H. Isliker and J. Kurths in [40].

This technique attempts to approximate a projection of so called *physical invariant measure* ρ defined in [41] as the time average of Dirac δ-distributions along a trajectory:

$$\rho(\xi) \coloneqq \lim_{T \to \infty} \frac{1}{T} \int_0^T \delta_{\xi(t)} \, \mathrm{d}t \,.$$

Roughly speaking, this measure quantifies "how often" are different subsets of the state space visited over infinite time. In other words, it gives a probability that a randomly chosen point on a trajectory will happen to belong to a given subset "after enough time passed". It is a statistical description of a system in the state space which contains information about all statistical moments [40], which should be indepent of the trajectory length for a stationary process.

This measure is related to computation of correlation dimension. Mention it in corresponding section.

Let x_1 represent the time series of the measured quantity, and N be the length of the time series. The algorithm then computes the projection $\rho(x_1)$ as follows. The range of the time series is divided into K "equiprobable" intervals $[x_1^{(k)}, x_1^{(k+1)}]$, $k = 1, 2, \ldots, K$, such that the interval boundaries are K-quantiles of the distribution of the values of the time series (i.e. application of the quantile function of the distribution to the values $1/K, 2/K, \ldots, (K-1)/K$), and the number of values falling into each of those intervals is counted:

$$\begin{aligned} n_k &:= \#\{x_1^{(k)} \leq x_1 \leq x_1^{(k+1)}\} \\ &\approx \sum_{x_1} \int_{x_1^{(k)}}^{x_1^{(k+1)}} \delta(x - x_1)\,\mathrm{d}x \\ &= \sum_{x_1} \chi_{[x_1^{(k)}, x_1^{(k+1)}]}(x_1), \end{aligned}$$

where $\chi_{[a,b]}$ is the characteristic function of the set $[a, b]$. The density over the entire series is then approximated by a histogram with K bins as

$$p_k^{\mathrm{all}} = \frac{n_k^{\mathrm{all}}}{\sum_k n_k^{\mathrm{all}}}.$$

If the system is stationary, then the probability distribution for the first half of the time series should be the same as for the entire time series. Hence, this distribution (with the same intervals) is computed for the first half of the time series (n_k^{half}). Then, the two probability distributions are compared using the χ^2 statistical test with the null hypothesis of stationarity. The corresponding Pearson's cumulative test statistic then is

$$\chi^2 := \sum_k \frac{(n_k^{\mathrm{half}} - Z p_k^{\mathrm{all}})^2}{Z p_k^{\mathrm{all}}},$$

where $Z = \lceil N/2 \rceil = \sum_k n_k^{\mathrm{half}}$. Under the null hypthesis, this quantity is expected to have χ^2 distribution [40], and thus the Pearson's test can be used to potentially reject the hypothesis of stationarity of the observed time series.

1.2.4 Recurrence Plot

When presented with the task of finding regularities in measurements obtained from nonlinear dynamical systems, one possible approach is analysing at least approximate repetitions of simple patterns, which can be further used for reconstruction of more complicated rules. Recurrence plot, proposed by Eckmann in [42], is a method of visualizing obtained state-space trajectory segments in relation to each other in order to achieve this goal. Furthermore, it can be used to test necessary conditions for validity of dynamical parameters derivable from a nonlinear time series such as the correlation dimension, entropies and Lyapunov exponents [43]. The property which makes them especially interesting is that the information contained in recurrence plots is not easily obtainable by other known methods [42].

Definition 7 ([42]). *Let N be the length of given time series, $\mathbf{x}_i$ for $i \in \{1, 2, \ldots, N\}$ be a i-th delay vector of any integer embedding dimension, $\|\cdot\|$ a norm, $\Theta(\cdot)$ a Heaviside step function, and $\epsilon \in \mathbb{R}_0^+$ a tolerance parameter. Then,* ***recurrence plot*** *is the matrix*

$$M_{ij} = \Theta(\epsilon - \|\boldsymbol{x}_i - \boldsymbol{x}_j\|)\,. \tag{1.3}$$

In other words, M_{ij} is a symmetric[4] binary $N \times N$ matrix, where $M_{ij} = 1$ when i-th and j-th points of the reconstructed trajectory enter each other's ϵ neighborhood. Since those points are, in fact, times, recurrence plots are a way of visualizing subtle time correlation information.

In [42], J. Eckmann et al. analyzed patterns typically observed in recurrence plots and distinguished between large-scale *typology* and small-scale *texture*. Moreover, they were able identify multiple different patterns easily distinguishible by the human eye typical of dynamical systems with distinct properties. This work was furhter extended in [44].

The essential drawback of recurrence plot is their size - it is quadratic in the length of the time series. A simple way of reducing its dimension is to partition the time series into disjoined segments, and let M_{ij} represent the distance between those two segments. This technique, introduced in [43], is known as **meta-recurrence plot**. The measure for "dynamical closeness" between two segments is based on the correlation integral (1.11) we will introduce in Section 1.4.2.

A more objective approach to analyzing recurrence plots is an ensemble of techniques group under the term Recurrence Quantification Analysis (RQA). Using these techniques, a number of scalar measures can be used to quantify properties of recurrence plots, such as determinism, periodicity, chaos and stationarity [45]. An important ingredient for computation of these measures is the distribution of lengths of diagonal lines in the plot. It can be shown that this distribution is directly related to correlation dimension, which we will cover in Section 1.4.2 [44].

1.3 State Space Reconstruction

In the previous section, we have introduced a concept of state space of a dynamical system. In the case of EEG analysis, however, our observations do not directly form a state space object, but a set of time series', one for each electrode. Moreover, it is necessary to deal with the fact that our data, however rich, rarely represent complete information about the studied system. In the case of EEG signals, the complete state of the system at any moment is determined by many variables, and the sensors are only able to collect traces of their cumulative effects (and noise). So we are confronted with a problem: how to convert this data into state space trajectories? This procedure is called *state space reconstruction*.

However, it is not necessary to reconstruct the entire original state state space of the system. In general, we are only interested in the subspace where the attractor of the system evolves. Thus, one possible approach to nonlinear time series analysis consists of the following steps [5]:

Saying dynamics is not true. We are not reconstructing the vector field **F**.

1. reconstruction of the attractor of given system from recorded data,

2. characterization of the reconstructed attractor,

3. checking validity of the results with surrogate data testing.

[4] Although this is true for our definition, it may not be true for an alternative definition using a more general topology instead of a norm. For example, each point may have been assigned its own ϵ-neighborhood.

1.3.1 Embedding

To the goal of state space reconstruction, let $\mathbf{x}_n$ be the reconstructed vector we are trying to find, and let us have a time series x of scalar measurements $x_1, x_2, \ldots, x_N$ of a quantity dependent on the current state of the system. As mentioned in Section 1.2.1, we will discard the noise term η_n in (1.1) and assume that the observed value corresponds to the true value.[5] Furthermore, let us consider a function $\Phi : \mathbb{R}^d \subset M \to \mathbb{R}^m$, such that $\mathbf{x}_n = \Phi(\xi(n\Delta t))$. Such function is called an **embedding**, and $m \in \mathbb{N}$ is called the embedding dimension. What properties does Φ have to satisfy so that it provides useful information about the true state space trajectories?

Firstly, note that we assume the studied dynamical system to be deterministic. If our reconstructed embedded space is to represent the true state space, evolution of any state on every trajectory we observe in the embedded space should depend only on its current state. Therefore, we may reasonably require Φ to be one-to-one, i.e. contain no intersections [47]. This will be relevant in Sections 1.3.4 and 1.3.4.5.

Secondly, since many of the attractor properties we care about (such as correlation dimensions, Lyapunov exponents, etc.) are only invariant under smooth non-singular transformations [47], in order to preserve these properties in the embedded space, we require Φ to preserve the differential structure of the state space M (which corresponds to the tangent map gradΦ, which is a $m \times d$ matrix constant for every ξ) also being a one-to-one mapping [15]. The proof of Taken's Theorem 1 mentioned later also requires this property.

These two properties together are equivalent of Φ being a diffeomorphism, and form necessary and sufficient conditions for Φ being and embedding [15] between the state space M and the embedding space $\Phi(M)$. The dynamical system $\mathbf{F}$ on M then induces unique dynamical sytem on $\Phi(M)$ [15].

As we have stated in Section 1.2, our observations are formed by application of noninvertible measurement function $s : \mathbb{R}^d \to \mathbb{R}^{d'}$, $d' \ll d$, to the true states of the system. Aside from being a projection, s may be also be a distortion. With those properties of s, it might seem impossible to reconstruct the true state space trajectory and this indeed may be the case in some situations. On the other hand, there are quantities invariant under distortion which may be preserved [15]. Moreover, if our goal was to study only the attractor properties, perfect reconstruction may not even be desirable in the case that the attractor dimension is smaller than the dimension of the original space [34].

1.3.2 Method of Time Delays

There are two approaches to the problem of state space reconstruction for EEG time series data:

Time delay embedding state space is reconstructed separately for each time series.

Spatial embedding each electrode represents a dimension of the embedding space, and each vector in the reconstruction contains amplitudes measured at a particular time.

It has been demonstated, however, that spatial embedding, when applied to EEG, does not reliably reconstruct the complexity of state space dynamics. Instead, in this case, is rather a measure of cross-correlation between individual channels [48]. It seems to remain relatively obscure approach to embedding, being used in a minority of groups.[6] On the other hand, time delay embedding is widely used, and, as we will show in this section, has a long history and relatively strong theoretical basis.

It had been already known since 1936, that every n-dimensional differentiable manifold can be embedded in $\mathbb{R}^{2n+1}$, and that the set of such embeddings is open and dense in the space of generic smooth

[5] For theoretical implications of the noise term one may consult [46].

[6] Which, however, strongly advocate for its use [49, 50, 51].

maps, which is known as Whitney's theorem [52].[7] In other words, $2n+1$ independent measurements of a n-dimensional system can be uniquely mapped to a $2n+1$ dimensional space, hence each such $2n+1$ dimensional vector identifies identifies state of the system perfecly, thus reconstructing the true state space.

Time delay embedding is a technique of state space reconstruction, which achieves the same goal, but with a single measured quantity. It was first introduced into the field of nonlinear dynamical system analysis by N. H. Packard in 1980 (although it was already being used in different fields in 1950s [15]). Studying the Rossler system, Packard noticed that by sampling a single coordinate, he was able to obtain a faithful phase-space representation of the original system by simply using a value of a coordinate with its values at two previous times [53]. In other he demonstrated numerically that past and future measurements of one variable contain information about the unobserved variables and can be used to define the present state.

In particular, for each time t, we define an embedding window τ_w, and use measurements obtained at times t' for $t \leq t' \leq t + \tau_w$. To this goal, we use m measurements, τ elements apart. Here, τ is called *lag* or *time delay*, and is measured in number of samples.[8] Using the notation in the Section 1.3.1, the time delay reconstruction is then formed by the following vectors

$$\mathbf{x}_n = (x_n, x_{n+\tau}, x_{n+2\tau}, \ldots, x_{n+(m-2)\tau}, x_{n+(m-1)\tau}), \tag{1.4}$$

for $n > (m-1)\tau = \tau_w$ [34].

A year after Packard's discovery, in [54], F. Takens has proved theoretically that the attractor reconstructed using this method may have the same dynamical properties (entropy, dimension, Lyapunov spectrum) as attractor of the original system under some conditions. Takens delay embedding theorem is an important result of nonlinear time series analysis and can be stated as follows:

Theorem 1 ([54])**.** *Let M be a compact[9] smooth manifold specifying the state space of a deterministic dynamical system of dimension $d \in \mathbb{N}$, $s : M \to \mathbb{R}^n, s \in C^2$ a smooth measurement function, $f^t : M \to M, f \in C^2$ a set smooth diffeomorphic state evolution functions for $t \in \mathbb{R}$. Then the set of maps $\phi_{(s,f^t)} : M \to \mathbb{R}^{2d+1}$, defined by*

$$\phi_{(s,f^t)}(x) = (s(\xi), s(f^{-\tau}(\xi)), \ldots, s(f^{-2d\tau}(\xi))), \tag{1.5}$$

for which Φ is an embedding is an open and dense set in the space of maps satisfying the assumptions above.

This idea has a simile in the existence theorems in the theory of differential equations, which say that a unique solution exists for each $x(t), \dot{x}(t), \ddot{x}(t), \ldots$. For example, in many body dynamics under Newtonian gravitation, knowledge of a body's position and momentum is sufficient to uniquely determine its future dynamics [55].

Taken's theorem, although of theoretical importance, is not necessarily useful in practice, since even dense sets can have measure zero. Moreover, it is restricted to smooth manifolds. An add came ten years later, when T. Sauer both generalized Takens' result as follows (in a simplified form):

Theorem 2 (Sauer, [56])**.** *Let A be a compact fractal with box-counting dimension d_A (see Definition 5), and let A be a subset of a m-dimensional manifold. Then a member of the set*

$$\{\Phi : A \to \mathbb{R}^m | \Phi \in C^1, m > 2d_A\} \text{ is an embedding with probability } 1.$$

[7] The second part of the theorem is a consequence of the fact that two hyperplanes with dimensions d_1 and d_2 in m-dimensional space are likely to intersect if $d_1 + d_2 \geq m$.

[8] Some authors use the time units $\tau\Delta t$, where $\Delta t = t_s = 1/f_s$ is the sampling period.

[9] This theorem can be proved for M non-compact provided less restrictions are imposed on s.

Theorem 1 and Theorem 2 together ensure that when m is chosen such that $m > d_A$ (which may be a considerable reduction in dimension compared to $m \geq 2d+1$), then Φ a true embedding of the underlying attractor for almost any τ (note only sufficiency of the result - $\mathbf{x}_n$ may be an embedding even for smaller m).

A fascinating consequence of Theorem 2 when applied to a sequence of measurements recorded from a physical system is that a successfully reconstructed attractor does not describe the time series, but the system itself. In the words of Theiler: "If one believes that the brain (say) is a deterministic system, then it may be possible to study the brain by looking at the electrical output of a single neuron. This example is an ambitious one, but the point is that the delay-time embedding makes it possible for one to analyze the self-organizing behavior of a complex dynamical system without knowing the full state at any given time" [57].

1.3.3 The Effects of Noise

Although these theoretical results are important to know about, they all make practically unrealistic assumptions, such as infinite amount of data and infinite measurement precision, and absence of noise. Moreover, practical applications present further challanges, such as presence of noise. Several factors complicate successful reconstruction from real-world, experimental data [58]:

Observational noise. Given a reconstructed vector $\mathbf{x} \in \mathbb{R}^m$, there is a (approximatly Gaussian shaped in natural scenarios) distribution $p(\mathbf{x})$ in the reconstruction space due to the noise term in equation (1.1) [15].

Dynamic noise (nonstationarity). External influences perturb the system, which consequently appears nondeterministic.

Estimation error. Estimation of the dynamics of the system is performed using only limited amount of data.

Quantization error. The measured analogue quantity is converted and stored as a number with only finite number of bits.

Moreover, different reconstructions can amplify the already present noise to varying degree. In [58], Casdagli et al. provide a quantitative way of analyzing this amplification, and, by extension, of insight into selection of embedding parameters so that the noise amplification is minimized.

1.3.4 Time Delay Selection

Note that the results of theorems in Section 1.3.2 do not depend on the value of the delay τ.[10]. Embeddings with the same value of the embedding dimension m, but different values of τ are theoretically equivalent. In practice, however, some theoretically sound time delay reconstructions may fail to be embeddings. Although some researchers propose that the only important parameter is the length of the embedding window $\tau_w = \tau(m-1)$ [46], as we will see, the choice of time delay has effects independent of the choice of embedding dimension, and vice versa. Some of the reasons a reconstruction may fail to be an embedding are as follows:

1. The embedding may fail to be a one-to-one map due to finite precision, or presence of noise in the data [15].

[10] This is because of the fact that the measurements are infinitely precise [58].

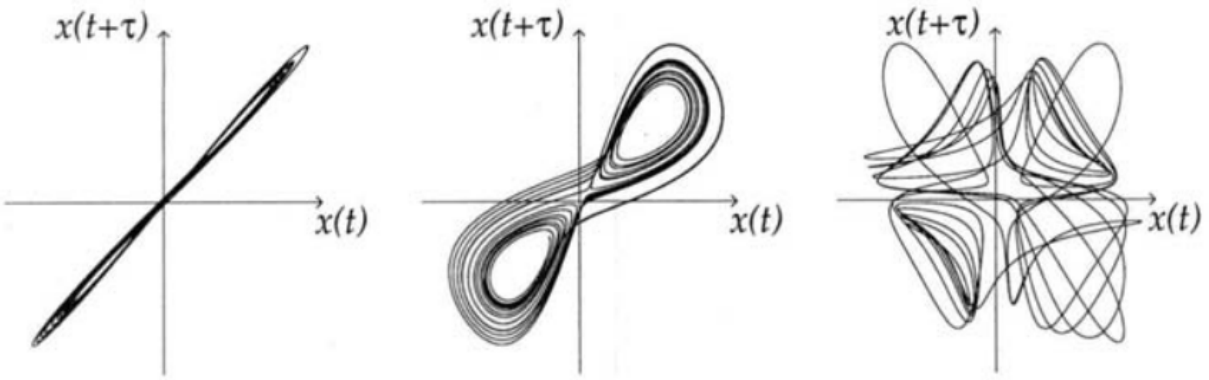


Figure 1.5: Time delay reconstructions of the Lorenz attractor for different values of τ. Figure on the left hand side shows choice of small τ and represents the case of redundance - the states concentrate along the main diagonal. Figure in the middle shows a successful reconstruction (although not an embedding, for which $m \geq 3$ is required). Figure on the right hand side shows a choice of large τ and represents the case of irrelevance - the reconstruction lacks apparent structure [15].

2. For some highly chaotic systems with large Lyapunov exponents (see Section 1.4.1) and large dimension, projection to a low dimensional time series causes amplifies the effects noise. As a result, this imposes limits on short term predictablity and state space reconstruction may become impossible. Such systems should be treated as operationally stochastic [58].

3. It was shown that increasing τ leads to rise in entropy [59].

4. Deterministic behavior can be observed only when τ_w is smaller than the time scale of the foldings naturally produced as result of time embedding [58].

5. If the values of τ are *too small* in comparison to the typical time scales of the series (measured e.g. by mean period), then the successive elements of reconstructed state space vectors become almost equal. This effect is often called *redundance*. Since $x_t \approx x_{t+\tau}$, the reconstructed attractor will concentrate along the main diagonal (see Figure 1.5, left hand side). Moreover, in this case, the effect of noise is amplified [58].

6. If the values of τ are *too large*, the successive elements in the reconstructed vector are almost independent. This effect, called *irrelevance* or *overfolding* is even magnified if the underlying attractor is chaotic, since deterministic correlations between states are lost even at very small time scales, i.e. even measurements performed at time t and $t + \tau$ for very small τ may be already unrelated. The reconstructed attractor will form a seemingly random clound in $\mathbb{R}^m$ - thus the reconstructed attractor may appear complex, even if the true attractor is simple (see Figure 1.5, right hand side).

In summary, picking the proper value of τ is a balancing act between redundance and irrelevance. It is important to minimize excessive foldings, and extreme closeness between adjacent points on the trajectory (ideally, the distances between points is same in the reconstructed as in the true space).

1.3.4.1 Autocorrelation

From the above, we understand that each of a successful reconstruction should provide valuable information about the state of the system. This may mean "reasonably" low statistical correlation between

values of coordinates of the reconstructed vectors $\mathbf{x}_n$. Thus, a natural method of estimating the optimal time delay is studying the *autocorrelation function* A, and picking the first τ where $A(\tau)$ decays below a threshold value - commonly used are $A(0)/e$ [5], $1-A(0)/e$ [34], or even the first local minimum [60, 61] or the first 0 crossing [34].

Definition 8 ([34]). *Let x be a scalar time series of measurements $x_1, x_2, \cdots, x_N$.* ***Autocorrelation*** *$A : \mathbb{N} \to \mathbb{R}$ for time delay τ is given by*

$$A(\tau) = \frac{1}{\mathrm{var}(N-\tau)} \sum_{i=1}^{N-\tau} (x_i - \overline{x})(x_{i+\tau} + \overline{x})$$

where $\overline{x}$ is the mean of the time series, and var *is its variance.*

Note that var normalizes autocorrelation function to $A(0) = 1$. White noise produced $A(\tau) = 0$ for all $\tau \neq 0$. Filtered noise and chaotic time series can be expected to have $A(\tau)$ slowly decaying to zero with increasing time delay [15].

Computing the autocorrelation function is not only useful for examining the stationarity of the time series, but it also gives a geometrical insight into the shape of the attractor: if we approximate the cloud of reconstructed vectors $\mathbf{x}_n \in \mathbb{R}^m$ by an ellipsoid, lengths of its semi-axis are given by the square root of the eigenvalues of its auto-covariance matrix. In two dimensions, zero of the covariance matrix corresponds to those eigenvalues being equal, i.e. x_t and $x_{t+\tau}$ being completely uncorrelated [34]. An obvious objection is that correlation between x_t and $x_{t+\tau}$ says nothing about correlation between x_t and $x_{t+2\tau}$, etc. Indeed, this method computes correlations only between two successive coordinates, but since it neglects possible correlations between other pairs of coordinates, it is mostly useful for low dimensional systems [62].

Autocorrelation also provides a lower bound for τ in the following sense. If the data is noisy, vectors formed by time delay embedding procedure are practically meaningless, if the variation of the signal in the time covered in the time window $\tau_w = (m-1)\tau$ is less than the variation of noise. This means that τ should be selected such that $A(\tau) > A(0) - \mathrm{var}_{\mathrm{noise}}/\mathrm{var}_{\mathrm{signal}}$ [34]. Of course, in practice, it may be difficult to estimate variance of noise in the data.

1.3.4.2 Delayed Mutual Information

Another commonly used method is to use the first minimum of the *time delayed mutual information* [63].

Definition 9 ([34]). *Let probability density of the values of a time series be split into ϵ-wide histogram bins. Let p_i be the probability that a signal assumes value in i-th bin of the histogram, and let $p_{ij}(\tau)$ be the the probability that x_t is in a bin i and $x_{t+\tau}$ is in a bin j.* ***Delayed mutual information*** *$\mathcal{I}_\epsilon$ for time delay τ is defined as*

$$\mathcal{I}_\epsilon(\tau) = \sum_{i,j} p_{ij}(\tau) \ln p_{ij}(\tau) - 2 \sum_i p_i \ln p_i.$$

In other words, time delayed mutual information is the average mutual information between measurements obtained by the original time series and its τ-shifted (time delayed) counterpart. The optimal τ is usually selected as $\arg\min_\tau \mathcal{I}_\epsilon(\tau)$.

Although this approach yields coordinates independent in a more general sense than simple linear independence provided by the autocorrelation function, the same criticism applies: minimum dependence between x_t and $x_{t-\tau}$ says nothing about dependencies between other coordinates. Again, using

this method is justifiable only for two-dimensional reconstructions. However, delayed mutual information has been generalized for multiple dimensions by its proponent A. M. Fraser using multidimensional distributions into a concept he called *redundancy*, which basically measures the degree to which the reconstructed vectors accumulate around the bisectrix of the embedding space [47]. Another criticism of delayed mutual information that some systems exhibit slowly decaying mutual information which has no minima [64].

1.3.4.3 Average Displacement from Diagonal

Average displacement from diagonal is a simple technique which simply measures the average distance of the embedding vectors from their original location:

$$\mathrm{ADFD}(m,\tau) = \frac{1}{N_{(m,\tau)}} \sum_{i=1}^{N_{(m,\tau)}} \|\mathbf{x}_i^{(m,\tau)} - \mathbf{x}_i^{(m,0)}\|,$$

where $\mathbf{x}_i^{(m,\tau)}$ is the i-th vector of time delay embedding with embedding dimension m and time delay τ.

Rosenstein et al. presented multiple methods for quantifying expansion from the diagonal (the identity line of the attractor), and found ADFD to be the most computationally efficient, robust to noise, and accurate [65]. They also experimentelly identified optimal τ as the one for which the slope of ADFD drops below 40% of its initial value.

1.3.4.4 Singular Values Analysis

All the approaches described so far address the issue of redundance by attempting to make the coordinates less correlated or expanding the reconstruction from the identity line. The issue of irrelevance however, needs to be addressed as well. In fact, based mostly on empirical, rather than theoretical grounds, most time delay estimation techniques optimize for the following criteria [46]:

1. The reconstructed attractor must be expanded from the diagonal.

2. The components of the reconstructed vector $\mathbf{x}_n$ must be uncorrelated.

Those criteria are similar, and bias towards larger estimates of τ. This leads many authors to suggest more advanced techniques, such as generalized delayed mutual information mentioned above, or some of those introduced in the following text.

Principal component analysis, in particular, can be used to measure the volume occupied by the reconstructed attractor. Both overfolded and redundant attractors may be marked by low volume [15].

Given a fixed embedding dimension m, the corresponding m singular values as a function of τ contain information about the degree of extension of the embedded vectors in the m directions in the reconstructed space. Rapid increase followed by rapid decrease of some singular values accompanied by the opposite behavior of others indicate a collapse of the attractor. Also, high number of large singular values is an indicator of volume of the reconstructed attractor.

Is this so?

If we assume, without loss of generality, that the time series is standardized and denote

$$\mathbb{X} := \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \cdots \\ \mathbf{x}_{N_{(m,\tau)}}^T \end{pmatrix},$$

then

$$(\mathbb{C})_{ij} \coloneqq (\mathbb{X}^T\mathbb{X})_{ij} = A\left((i-j)\tau\right).$$

This matrix is symmetric and thus diagonalizable, and also at least non-negative definite. Its eigenvalues are called the singular values, and correspond the the magnitude of variance of projections of the embedded vectors into individual directions of the principal components.

If the time delay is too small, then all the elements of matrix $\mathbb{C}$ will have similar value $(\mathbb{C})_{ij} \approx A(0)$, and thus there will be one dominant singular value, while others will remain close to zero. This singular value then corresponds to the main diagonal of the attractor.

If the time delay is too large, then the diagonal elements will approach average of the squared time series $(\mathbb{C})_{ii} \approx \langle x^2 \rangle$, while the remaining elements will converge to zero due to decay of the autocorrelation function, $\mathbb{C} \approx c\mathbb{I}$ for some constant c. This corresponds to the reconstruction forming a featureless noise [15].

One drawback of this method is that it requires evaluation either by a human or a more complicated algorithm. Moreover, it was suggested that although this method is effective noise reduction technique, its effectiveness at delay estimation is less clear - the number of large singular values is sensitive to noise [66].

1.3.4.5 Integral Local Deformation

In Section 1.2.1, we require our dynamical systems to be deterministic, which means that no trajectories in the state space should intersect themselves. Moreover, in real physical systems, it may be reasonable to assume that it is highly unlikely to find closeby trajectories of opposite or orthogonal directions. This property is maintained by a successful embedding, and (if the assumption holds) can occur only in an improper reconstruction.

T. Buzug and G. Pfister presented a quantitative measure of these close trajectory intersections by comparing the the evolutions of reference trajectories with centroids of points on the neighboring trajectories [67]. For the optimal embedding, divergence between these trajectories should be minimized.

First, multiple random reference points are chosen. Let $\mathbf{x}_i(0)$ be such a reference point at time 0. Then, either a fixed number of nearest neighbors or all neighbors within a given radius and their centroid $\mathbf{x}_i^{\mathrm{COM}}(0)$ (where COM stands for "center of mass") are found. Then, the absolute growth of the distance between the centroid of those originally neighboring points and the reference point after qt_{ev} time steps is found as:

$$\Delta(q, m, \tau) = \|\mathbf{x}^{\mathrm{COM}}(qt_{ev}) - \mathbf{x}_i(qt_{ev})\| - \|\mathbf{x}_i^{\mathrm{COM}}(0) - \mathbf{x}_i(0)\|.$$

The values $\Delta(q, m, \tau)$ are discretely integrated from $q = 1$ to $q = q_{max}$:

$$\mathcal{D}(m, \tau, i) = \frac{t_{ev}}{2} \sum_{q=1}^{q_{max}} \left(\Delta(q-1, m, \tau) - \Delta(q, m, \tau)\right).$$

This expression, called **integral local deformation**, is then averaged over N_{ref} reference points and normalized:

$$\mathrm{ILD}(m, \tau) = \langle \mathcal{D}(m, \tau, i) \rangle_i = \frac{t_{ev} \sum_{i=1}^{N_{ref}} \sum_{q=1}^{q_{max}} \left(\Delta(q-1, m, \tau) - \Delta(q, m, \tau)\right)}{2N_{ref}\Delta t \left(\max_{i\in 1,2,\ldots,N} x_i - \min_{i\in 1,2,\ldots,N} x_i\right)}$$

Finally, we obtained a measure of non-homogeneity of the average flow in the neighborhood of the points in the reconstructed embedding space as a function $\mathrm{ILD}_m(\tau)$ of the time delay τ and parameterized by m. According to our assumption about the reasonable property of physical dynamical systems, the optimal τ for each m is the minimum $\arg\min_\tau \mathrm{ILD}_m(\tau)$.

The ILD algorithm provides the detailed information about the flow of the reconstruction, and is arguably the most powerful out of the algorithms we described, since it is the only one which measures the *dynamical* properties of the reconstruction, not only topological ones [15]. Moreover, since we may expect that for a sufficiently high m, the $\mathrm{ILD}_m(\tau)$ curves will converge [67], this technique allows for simultaneous estimation of both the embedding dimension m and the time delay τ. However, one considerable drawback is much larger computational cost, since for each m and τ, closest neighbors from the entire reconstruction have to be determined for each point.

1.3.5 Embedding Dimension Selection

1.3.5.1 False Nearest Neighbors

Since the dynamics $\mathbf{F}$ are assumed to be a *smooth* vector field and the attractor A is a *compact* set in the phase space, its members acquire near neighbors, which should be subject to similar evolution. Therefore, these neighbors should remain close to each other after a short interval of time (even though chaos may introduce exponential divergence between them). This is a useful fact, which can be used, for example, to predict future evolution of a trajectory, or a computation of Lyapunov exponents. The **false nearest neighbors** algorithm uses them for estimation of embedding dimension [68].

The main idea is to use the transition from dimension m to dimension $m+1$ in the embedding procedure to differentiate between "true" and "false" neighbors. If the embedding dimension m is too small, some members of A that are close to each other may not be neighbors in the true state space, simply because the true state space is projected down to a smaller space. These members are *false neighbors*, all other neighbors are *true*. When the attractor is fully unfolded into large enough dimension and is properly embedded, all neighbors are true.

Add the figure from Kennel.

Let $n(j, r, m, m)$ denote the index of of r-th nearest neighbor of the m dimensional embedding vector $\mathbf{x}_j$. Then, let $R_m(j, r, m)$ denote the Euclidean distance between $\mathbf{x}_j$ and its neighbor:

$$R_m(j,r,m) = \|\mathbf{x}_j - \mathbf{x}_{n(j,r,m)}\|_2 = \sqrt{\sum_{k=0}^{m-1} [x_{j+k\tau} - x_{n(j,r,m)+k\tau}]^2}$$

Then, any near neigbor for which the distance increase after transition from dimension m to dimension $m+1$ is large in comparison to the initial distance is marked as false:

$$\left[\frac{R_m^2(j,r,m) - R_{m+1}^2(j,r,m)}{R_m^2(j,r,m)}\right]^{1/2} = \frac{x_{j+k\tau} - x_{n(j,r,m)+k\tau}}{R_m(j,r,m)} > R, \tag{1.6}$$

where $R \in \mathbb{R}$ is some threashold. The m for which the relative proportion of false neighbors to all neighbors reaches zero is the embedding dimension suggested by this criterion.

This criterion, by itself, is not sufficient for determining proper embedding dimension. When applied to limited amount of white noise data, it erroneously suggested embedding the noise into a low dimensional attractor. This happens because even though a state may be a nearest neigbor, it is not necessarily temporally close, and thus the assumptions above do not hold. The experiments performed by Kennel et al. show for such states it is usually $R_m(j, r, m) \approx R_A$, where R_A is radius of the attractor. Furthermore, for increasing amount of data, the embedding dimension suggested by this criterion also increased - behavior not observed for relatively small dimensional attractors [68].

Therefore, Kennel et al. propose another criterion in addition to the one above. Since false neighbors which are near, but temporally distant, are usually stretched to the extremeties of the attractor with transition from m to $m + 1$, they suggest marking all near neighbors satisfying

$$\frac{R_{m+1}(j, r, m)}{R_A} > A \tag{1.7}$$

as false, where R_A may be computed as, for example

$$R_A = \frac{1}{N} \sum_{j=1}^{N} \left[x_j - \overline{x} \right]^2 .$$

Although this technique is commonly used, it is not without its drawbacks. An obvious point is that altough it is true that distance between neighbors in unfolded attractor should not grow with increase in dimension, the inverse is not necessarily true, i.e. stable distance between near neighbors with increase in dimension does not guarantee that these neighbors are true.

The authors suggest some values of the tolerance parameters they found useful in their experiments, but, in general, the results of this technique may depend on the choice of R and A. Their selection is subjective and somewhat arbitrary. The best course of action is to evaluate the technique for multiple values of R and A and select those with the most "reasonable" results.

In practice, it has been found that the results of this method are sensitive not only to the tolerance parameters R and A, but also to the lag as well [46].

Also, this method tends to underestimate m for very small τ. Small τ forces the attractor to lie near the diagonal in $\mathbb{R}^m$ and further increasing m imposes very little effect on the geometry of the attractor. In effect, most points will appear as true neighbors leading to a wrong conclusion [46].

Lastly, in presence of measurement noise, the proportion of false neighbors may increase after transition to a higher dimension, since even identical vectors will diverge [34].

1.3.5.2 Average False Neighbors

This technique by Cao [69] addresses one of the drawbacks of false nearest neighbors mentioned in the previous section - the variance of results based on subjective choice of embedding parameters. It does so by defining two parameter free functions dependent only on the embedding parameters.

The first function measures the variation of average ratio of distance of two neighbors in one dimension to the distance of the same neighbors in a higher dimension. More precisely, let

$$E(m) = \frac{1}{N_{(m,\tau)}} \sum_{i=1}^{N_{(m,\tau)}} \frac{\|\mathbf{x}_i^{(m+1)} - \mathbf{x}_{n(i,1,m)}^{(m+1)}\|_\infty}{\|\mathbf{x}_i^{(m)} - \mathbf{x}_{n(i,1,m)}^{(m)}\|_\infty},$$

where $n(i, 1, m)$ denotes the nearest neighbor of vector $\mathbf{x}_i$ in dimension m, and $\|\cdot\|_\infty$ denotes the Chebyshev norm[11]. Then, the first statistic is defined as

$$E_1(m) = \frac{E(m + 1)}{E(m)}.$$

In principle, $E_1(m)$ saturates and stops increasing after some threshold m for systems with finite embedding dimension.

[11] This norm is suggested by the author, but another norm can be used.

For systems with infinite embedding dimensions it may be difficult in practice to resolve whether E_1 indeed stopped increasing or is still slowly increasing. Alternatively, it may still saturate because of limited amount of data. For this reason, Cao introduces another statistic, whose purpose is to distinguish stochastic from deterministic sources of data.

Let

$$E^*(m) = \frac{1}{N - m\tau} \sum_{i=1}^{N-m\tau} |x_{i+m\tau} - x_{n(i,1,m)+m\tau}|.$$

Then, similarly to above, the second statistic is defined as

$$E_2(m) = \frac{E^*(m+1)}{E^*(m)}.$$

Since, for random time series, the future values are independent of the present ones, the ratio $E_2(m)$ is expected to be close to 1 for all m.

1.4 Nonlinear Measures

In this section, we will study quantities invariant under embedding. These can be further use to characterize the dynamics of deterministic dynamical systems.

1.4.1 Lyapunov Exponents

The characteristic property of chaotic systems is their sensitivity to initial conditions - similar causes need not have similar effects. Consequently, even small uncertainty in the current state of the system (due to, at best, with limited storage space) results in virtual impossibility of predicting future state of the sytem more than a short amount of time into the future, since uncertainty in the initial state is expanded at exponential rate with passage of time by the chaotic dynamics for the predicted future states.

Lyapunov exponents can be used to quantify this sensitivity [34]. Consider a small sphere of initial conditions $B_r(\mathbf{x})$ for a state $\mathbf{x}$ in the phase space, r infinitesimal, and $\mathbf{x}_n \in B_r(\mathbf{x})$. To study the evolution of states in this ball, we can use a linear approximation of $\mathbf{F}$. Let us assume that $\mathbf{x}_{n+1} = \mathbf{F}(\mathbf{x}_n)$. Then for infinitesimal divergences $\delta\mathbf{x}_n$, $\delta\mathbf{x}_{n+1}$, we have

$$\delta\mathbf{x}_{n+1} = T^{(n)}\delta\mathbf{x}_n,$$

for a tangent map $T^{(n)}$ defined as

$$(T^{(n)})_{ik} = \frac{\partial F_i(\mathbf{x}_n)}{\partial x_{n+k}}.$$

Product of these tangent maps for subsequent states along a trajectory can be written as a product of two rotations and a diagonal matrix [70]:

$$\prod_{n=1}^{N} T^{(n)} = R_d T_{diag} R_b,$$

where $(T_{diag})_{ij} = (T)_{ij}$ for $i = j$ and $(T_{diag})_{ij} = 0$ for $i \neq j$.

Then, the Lyapunov exponents can be defined as [70]

$$\lambda_i = \lim_{n\to\infty} \frac{1}{N} \log(T_{diag})_{ii}.$$

In other words, as the system evolves, $B_r(\mathbf{x})$ expands (or contracts) exponentially in m directions defining semiaxes of a sphere, where length of each semiaxis corresponds to the rate of expansion (or contraction) in the corresponding direction. The average lengths of these semiaxis for $\mathbf{x}$ over the entire state space are exactly Lyapunov exponents. Hence, m dimensional system has exactly m Lyapunov exponents, collectively called its *Lyapunov spectrum*.

Computation of the Lyapunov spectrum for analyticial given $\mathbf{F}$ is straightforward using the definition above. But for dynamics given implicitly in a time series is difficult (although some algorithms, e.g. the one introduced by Eckmann in 1986 [71]). It is commonly agreed that estimating Lyapunov exponents is even more difficult than esimating correlation dimension [15], although they have been successfully employed in EEG analysis [72, 12, 10, 5]. It has been claimed by P. Grassberger et al. that any application of these measures to physical systems should be interpreted with caution, mainly because all physical measurements are corrupted by noise, and reliable separation of signal is not always possible [70]. They suggest that when emplying these techniques, the goal should not be to estabilish to strongest form of determinism, but to use them to ask whether determinism can be ruled out at all.

Since the direction of the largest Lyapunov exponent dominates growth, we can say that the average rate of separation between two points in the phase space with similar initial conditions can be characterized by the largest Lyapunov exponent. As a consequence, it is unnecessary to compute the entire Lyapunov spectrum - which would require identifying appropriate Lyapunov directions - if our goal is to find a global property of the system characterizing the degree of average instability and unpredictability. It is sufficient to measure the average rate of separation [73].

Hence, let us define $\|\mathbf{s}_{n_1} - \mathbf{s}_{n_2}\| = d(0) \ll 1$ as an initial distance between two nearby points in the state space, and $d(i) = \|\mathbf{s}_{n_1+i} - \mathbf{s}_{n_2+i}\|$. Then, the largest Lyapunov exponent λ_1 can be approximated as

$$d(i) = d(0)e^{\lambda_1(i\Delta t)}, \quad d(i) \ll 1, \quad i \to \infty, \quad d(0) \to 0, \tag{1.8}$$

where Δt is sampling time of the time series.

The Lyapunov exponents carry the units of an inverse time - $1/\lambda_1$ gives a typical time scale for the divergence or convergence of nearby trajectories [34]. Equivalently, $1/\lambda_1$ is (on average) an upper bound on predictability in the system [15]. Also equivalently, they also can be seen as quantification of the degree of chaos in the system; a sigle positive exponents is a sufficient indication of presence of chaos [73].

Say what different values of λ_1 say about the system.

1.4.1.1 Rosenstein's algorithm

In the following, we will describe *Rosenstein's algorithm* for computation of the largest Lyapunov exponent [73]. This algorithm was found to be relatively robust to noise, values of the embedding parameters and limited amount of data.

First, state space is reconstructed using time delay embedding (see Section 1.3.1). The suggested method of time delay selection is the autocorrelation method (see Section 1.3.4.1).

For given embedding dimension m and each point on the trajectory $\mathbf{x}_j$, the algorithm locates the nearest neighbor $\mathbf{x}_{n(j,m)}$, such that their distance in the embedded space is minimized:

$$d_j(0) = \|\mathbf{x}_j - \mathbf{x}_{n(j,m)}\|.$$

As an approximation, we want to assume $\mathbf{x}_j$ and $\mathbf{x}_{n(j,m)}$ to be nearby initial conditions, but at the same time, we know they lie on the same trajectory. Hence, we may impose a condition on their minimal temporal separation, called a *Theiler window*. In the original paper [73], Rosenstein suggests

$$\frac{1}{4}\text{ time series length} > |j - n(j,m)| > \text{mean period of the time series.}$$

Then, assuming the j-th pair of nearest neighbors diverge exponentially at a rate given by the largest Lyapunov exponent, we have

$$d_j(i) \approx d_j(0) e^{\lambda_1 (i\Delta t)}.$$

By taking logarithm of both sides, we obtain

$$\ln d_j(i) \approx \ln d_j(0) + \lambda_1 (i\Delta t).$$

This represents a set of lines, one for each point on the reconstructed trajectory, each with a slope roughly proportional to λ_1. So, the algorithm approximates the largest Lyapunov exponent by least squares fit to the average line

$$d(i) = \frac{1}{\Delta t} \langle \ln d_j(i) \rangle_{j=1,2,\ldots,N_{(m,\tau)}},$$

usually evaluated for values $i \in \langle 0, t_e \rangle$, where t_e is called the evolution time.

Note that the user may decide to set $\Delta t = 1$ and work with units of time series indeces instead of seconds. It is well known that the results of the largest Lyapunov exponent may vary drastically based on input parameters [74]. Moreover, we can even rescale or shift the data, since Lyapunov exponents are invariant under any smooth invertible map.

There are many other algorithms to compute the larest Lyapunov exponents, such as Kantz's algorithm [75], Eckmann's algorithm [71], Wolf's algorithm [76] (relatively unstable, it is impossible to distinguish exponential divergence [34]), and Sato's algorithm (produces spurius results in certain cases [73]). The main competetive advantage of Rosenstein's algorithm is its easy implementation, low computational cost, and robustness to noise (due to averaging in the last step) and applicability to small datasets [73].

As we have mentioned already, the projection involved in the measurement may make distances shrink apparently for short times, although they grow in the true state space [34]. Moreover, in the true state space distances do not grow everywhere on the attractor with the same rate, and locally they may even shrink. LLE is average of those local divergence rates. Influence of noise can be minimised by using an appropriate averaging statistics.

1.4.1.2 Dataset Size Requirements

The minimum dataset requirements was estimated by Eckmann and Ruelle in [77] by imposing requirements on the distances and number of neighbors for each point. If $\Gamma(r) \gg 1$ is the average number of neighbors withing radius r, we may approximate it as

$$\Gamma(r) \approx \text{const.} \times r^m,$$

and we also know that $\Gamma(d) \approx N$, where d is the diameter of the attractor. Therefore, we obtain

$$\Gamma(r) \approx N \left(\frac{r}{d}\right)^m \gg 1 \implies N > \left(\frac{d}{r}\right)^m.$$

For example, if we require the ratio of the average distance to the nearest neighbor to the extent of the attractor to be $r/d \leq 0.1$, we have $N > 10^m$ as the minimum time series length requirement.

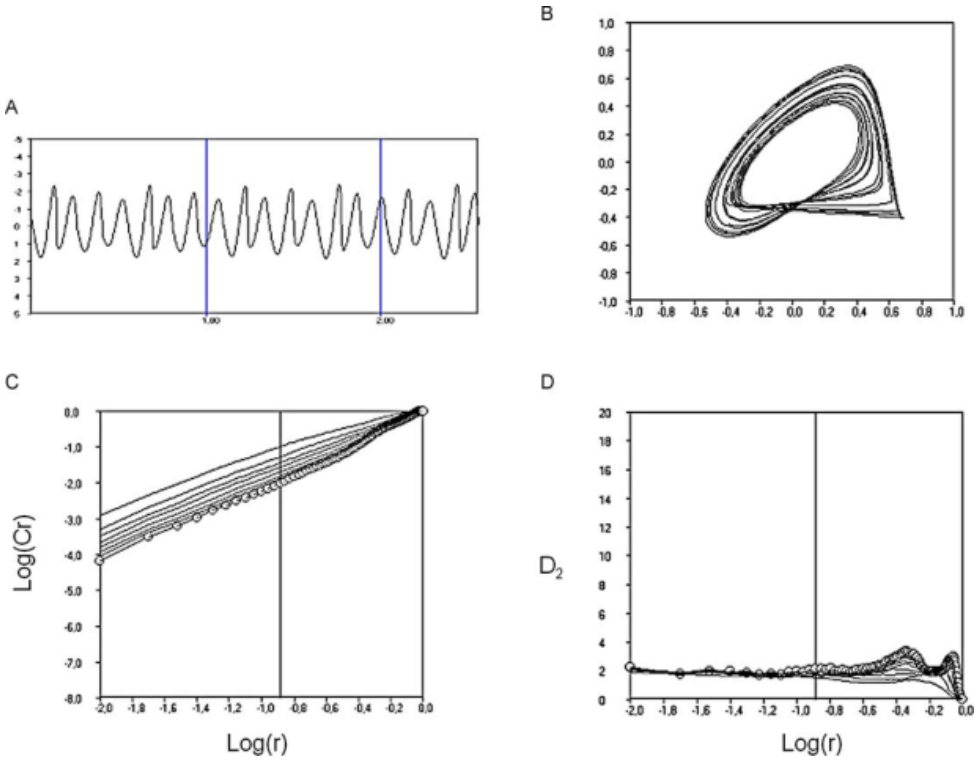


Figure 1.6: Example computation of the correlation dimension [5]. The axes are dimensionless. In the clockwise direction starting from the upper left hand side, the figures show the original time series, the reconstructed attractor, logarithmic plot of the correlation integrals $C(r)$ for different values of the embedding dimension m (starting with $m = 2$ in the uppermost line, and increasing by one with each line below), and their derivatives, corresponding to the correlation dimension D_2. In the derivatives plot, the vertical line signifies the cutoff of $\log r$ after which the values become imprecise due to numerical instability. We can see that the derivatives converge to approximately 2 with decreasing radius r.

1.4.2 Correlation Dimension

The world of mathematics offers numerous definitions of dimension (box-counting dimension (1.2), Hausdorff dimension, information dimension, etc.) and similar quantities, but many of them can be regarded as variations of the following, simple and intuitive analogy: [57]

$$\text{bulk} \approx \text{size}^{\text{dimension}} \implies \text{dimension} = \lim_{\text{size}\to 0} \frac{\log \text{bulk}}{\log \text{size}}. \tag{1.9}$$

In other words, dimension can be loosely defined as scaling of "bulk" (corresponds to mathematical concept of measure) as a function of its linear "size". Of course, dimensions of different definitions may not be equal to each other, but for our purposes, we are interested in the most computationally accessible.

Unlike Lyapunov exponents, which measure dynamical properties of the system, (correlation) dimension is a purely geometrical property of the attractor, independent of the ordering of the reconstructed vectors.

Is this true?

In this thesis, we are interested in dimension estimation for the following reasons:

1. Even a system with high number of degrees of freedom, such as a brain, may actually evolve in a much lower-dimensional subspace. The number of active degrees of freedom may provide a measure of complexity of the observed system. This information is available in the attractor of the system and it can be shown that this property is preserved by state space reconstruction [15].

2. It can help distinguish stochastic and deterministic processes, since stochastic processes, after sufficient passage of time, use all available state space dimensions.

Of course, although these expectations can be justified theoretically, the numerical reality may be different.

Most definitions of dimension are based on first covering the studied object in the state space with the smallest possible balls (using a given metric). Correlation dimension is a special case of generalized box-counting dimension (which is a generalization of box-counting dimension already introduced in Definition 5), defined as

$$D_\kappa(A) = \lim_{r\to 0} \frac{1}{\kappa} \frac{\log \int_M (\mu(B_r(\mathbf{x})))^\kappa \, \mathrm{d}\mu(\mathbf{x})}{\log r},$$

where the integration is over the whole state space M and μ is measure concentrated on A. If we define μ as

$$\mu(\mathbf{x}) \coloneqq \int_M \Phi(r - \|\mathbf{x} - \mathbf{y}\|) \, \mathrm{d}\mu(\mathbf{y}) \tag{1.10}$$

The we can write the, "bulk" of A, so called generalized correlation integral as

$$C(\kappa, r) = \left(\int_M (\mu(B_r(\mathbf{x})))^\kappa\right)^{\frac{1}{\kappa}} = \left[\int_M \left(\int_M \Phi(r - \|\mathbf{x} - \mathbf{y}\|) \, \mathrm{d}\mu(\mathbf{y})\right)^\kappa \mathrm{d}\mu(\mathbf{x})\right]^{\frac{1}{\kappa}}$$

It can indeed be shown that $C(\kappa, r) \propto r_\kappa^d$.

In the continuous case, correlation dimension than takes to form

$$D_2(A) = \lim_{r\to 0} \frac{\log C(r, 2)}{\log r}.$$

As explained in Section 1.2.4, correlation dimension is closely related to the distribution of lengths of diagonal lines on recurrence plots. Intuitively, we can see that both methods are measuring temporal correlations in the original time series.

1.4.2.1 Grassberger-Procaccia Algorithm

There are essentially three ways of computing correlation dimension: box-counting algorithms, pairwise distance algorithms, and nearest neighbors algorithms. Grassberger-Procaccia algorithm, which we use to compute correlation dimension, is a variant of a pairwise distance algorithm.

This class of algorithms, used in discrete cases with limited amount of data, estimates the measure of a box centered on point $\mathbf{x}_i$ in the reconstructed space as

$$\mu_i = \frac{1}{N_{(m,\tau)}}$$

and zero everywhere else.

Thus, in the discrete case, the correlation sum $C(r)$ can be computed as

$$C(r) \coloneqq C(r,2) = \frac{2}{N_{(m,\tau)}(N_{(m,\tau)}-1)} \sum_{i<j} \Phi(r - \|\mathbf{x}_i - \mathbf{x}_j\|). \tag{1.11}$$

which corresponds to the fraction of pairs of points in the phase space whose distance is smaller than r. Under certain reasonable conditions, correlation sum is an unbiased estimator of the correlation integral [78]. In our application, we also use a lower bound on the distance of pairs of points, which is called a Theiler window, noted w_t.

Typical behavior of the correlation sum is shown in Figure 1.7. We can see that the curves are forced to meet at the same point for all m - for high enough r, all points are counted and $C(r) = 1$ (or $C(r) = \binom{N_{(m,\tau)}}{2}$ not normalized). As the lines shift to the right with increasing m and stay parallel in the proper scaling region, the slope near that point necessarily increases with m. For high enough m, the scaling region disappears. Moreover, the values of $C(r)$ are inaccurate for small r due to noise and for small $C(r)$ due to statistical fluctuations (corresponding to horizontal lines). Thus, there is only a limited interval of r and limited set of embedding dimensions m for which an accurate estimation of D_2 can be made [15].

In our experiments, we used *local slopes approach* to estimating the correlation dimension, which is based on the idea of assigning a dimension estimate to each value of r by defining

$$D_2(r) = \frac{\partial \log C(r)}{\partial \log r}. \tag{1.12}$$

In our implementation, we perform a least squares fit of values $(\log r, \log C(r))$ for a window of 6 neigboring points for each sampled r. Expected behavior of the resulting function in a favorable case can be seen in Figure 1.8.

1.4.2.2 Dataset Size Requirements

There are multiple estimations of the minimum dataset size. Most of them are based on an attempt to avoid so called *edge effect*. It can be shown that the correlation dimension for a hypercube in m-dimensions of unit edge length the local correlation dimension is

$$D_2^{(m)}(r) = m - \frac{mr}{2-r} \approx m(1 - \frac{r}{2}).$$

For large enough r, $D_2^{(m)}(r)$ converges to zero. This result, which can be generlized to any finite object, is a consequence of the discontinuity of the measure (1.10) at the boundaries of the hypercube. Theiler,

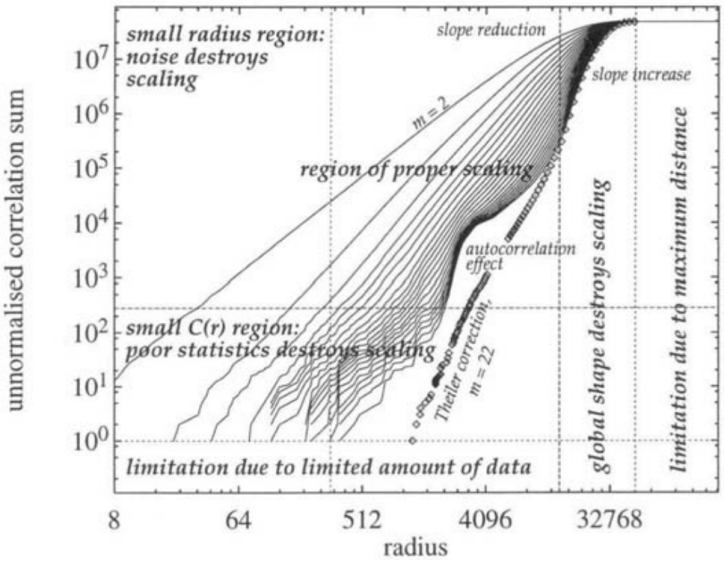


Figure 1.7: Plot of typical behavior of the non-normalized corretion sum $C(r)$ with regions relevant to D_2 estimation (both axes are logarithmically scaled) [15]. It is important to observe that either too low or too high values of r lead to poor estimation of the derivative: the former caused by statistical fluctuations, and the latter by the fact that the maximum pairwise distance is bounded. Using too high embedding dimension may also lead to poor estimations due to autocorrelation effects (an umbrella term for effects due to sampling and nonstationarity). To compare with our results, see Figure 2.14.

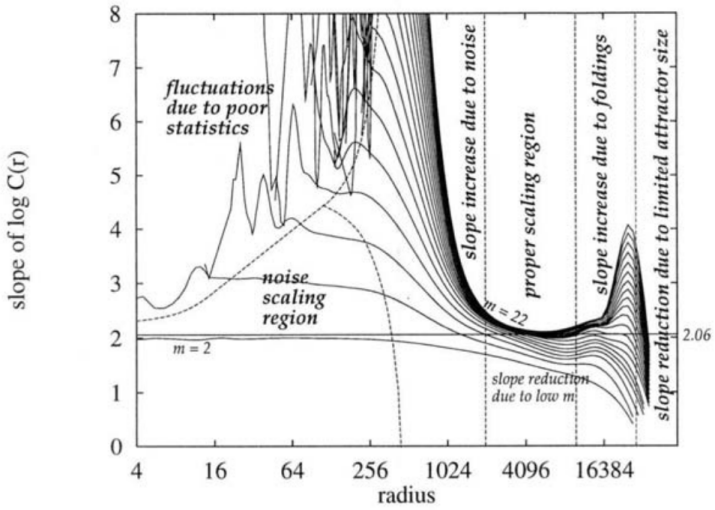


Figure 1.8: Plot of a typical local correlation dimension estimates for embedding dimensions from $m = 2$ to $m = 22$ (from bottom to top) in favorable case [15]. Note mainly the scaling region where the slope estimates converge to the value of 2. To compare with our results, see Figure 2.15. The results has been generated using time series with 1600 datapoints.

assuming evalution of the local correlation dimension for radius where each point has on average one neighbor (such that $C(r) = 1/N_{(m,\tau)}$), derived an estimate for the minimum data set size as

$$N_{(m,\tau)} = \frac{1}{(4\rho)^m},$$

where ρ is the maximum error. This implies an exponential increase of minimum required dataset size with embedded dimension. For example, $N_{(m,\tau)} = 5^m$ for $\rho = 5\%$ [15].

1.4.3 Detrended Fluctuation Analysis

Physiological time series, such as EEG, may exhibit so called statistical self-affine properties. Self-affinity is a special case of self-similarity, which occurs when one or more small parts of a fractal object is exactly or approximately similar to itself. When self-similarity is expressed in terms of statistical properties (e.g. mean value and variance of a part of time series are scaled version of its overall mean and variance), then the object is called statistically self-similar.[12] Self-similarity, in turn, differs from self-affinity in that self-affine objects witness similarity anisotropically, i.e. after applying an anisotropic affine transformation.[13] [79] Stated more formally:

Definition 10 ([80])**.** *A time series X given by $x_1, x_2, \ldots, x_n$ is said to be **statistically self-affine** if*

$$\mathrm{std}(X, Lt) \approx L^H \mathrm{std}(X, t),$$

where $\mathrm{std}(X, k)$ *is the standard deviation of the process X calculated over windows of length k, H is the Hurst parameter, and L a window length factor.*

The Hurst parameter, which behaves similarly to the Hurst exponent (see Section 1.4.4), ranges between 0 and 1. Higher values of H describe smoother signals, with high values followed by low, whereas low values of H indicate radical oscillations between high and low values. Note that since a stationary process has constant variance across time scales, Definition 10 applies only to nonstationary processes. However, even stationary processes may exhibit scale-free behavior. These are modelled as so called fractional Gaussian noise (fGn), whereas nonstationary processes are modelled as fractional Brownian motion (fBm). These processes are related by the fact that increments in fBm can be modelled as a fGn process with same H. This relationship allows us to generalize Definition 10 for nonstationary processes [79].

DFA is a method of estimating H without making prior assumptions about stationarity of the process by exploiting the relationship between fGn and fBm procesess. First, a so called signal profile, i.e. integral of the de-meaned signal is computed as

$$y_k = \sum_{i=1}^{k} (x_i - \langle x \rangle).$$

The resulting time series y is then divided into segments of varying length n (each value of n representing a time scale). A local linear least-squares fit is applied to each of these segments. Let us designate the

[12] An example of statistically self-similar object are are naval coastlines.

[13] The measured property of a self-similar process (e.g. the size of a flower on Romanesco cauliflower) do not follow normal distributions, but power law distributions. Hence, mode and mean of provide a poor representation of this representation. These processes do not have a scale at which to measure these statistics to characterize them, and are therefore called scale-free.

resulting piecewise linear fit as $y_n(k)$. The integrated time series is then detrended by subtracting the local linear fit. The root mean square error is then given by

$$F(n) \coloneqq \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - y_n(i))^2}. \tag{1.13}$$

Finally, if a log-log plot $F(n)$ as a function n shows a linear scaling region (i.e. the original time series exhibits self-similar, scale-free properties described above), the slope of this line approximates H and represents the result of DFA analysis [13].

Maybe add some more intuitive explanation.

The importance of DFA for EEG analysis comes from the observation that it can reveal so called long-range temporal correlations (LRTC) in neuronal activity, and so even for nonstationary time series [13]. This is especially important because EEG signals are nonstationary in some cases, and, as we will see in Section 2.3.4.1, some other frequently used measures, such as LLE and CD, theoretically depend on stationarity property.

Long-range temporal correlations, or long-range dependence, is a phenomenon which occurs when the average rate of decay of statistical dependence between increasingly (temporally) distant points in the time series is slower than exponential. Large-scale patterns in EEG activity may be characteristic of baseline processing during eyes-closed wakeful condition in healthy human brain. These parameters computed from the theta amplitudes were shown to be negatively correlated with (Hamilton) depression score, thus suggesting that depressed patients display abnormally small autocorrelations on large scale [81]. In our study, we observed similar results, see Section 2.4.2.

1.4.4 Hurst Exponent

As mentioned in the previous section, similarly to the Hurst parameter in DFA, Hurst exponent is a measure of presence of long-range temporal dependencies in the time series. It was developed from Edwin H. Hurst observation when researching the optimal (or minimal required) storage sizing of river dams. Supposing there is a constant reservoir outflow equal to the mean annual water discharge, required storage size corresponds to the range (i.e. the difference between the maximum and the minimum value) of a cumulative sums of deviations from the mean annual discharge. We shall call this value, as a function of the number of years, $R(n)$ [82]. After manually analyzing about a hundred records of natural phenomena, Hurst was able to demonstrate this value, on average and after normalizing by the standard deviation of the original time series, follows the following trend:

$$R(n)/\mathrm{std}(n) \propto (n/2)^K. \tag{1.14}$$

In this equation, $R(n)/\mathrm{std}(n)$ is called the rescaled range, and K is called the Hurst exponent [83]. Obviously, it is always the case that $0 \leq K \leq 1$.

The algorithm we used for computing estimation of the Hurst exponent is as follows [84]. Let us have time series x, with values $x_1, x_2, \ldots, x_N$. The time series is split into d non-overlapping subseries $x^{(m)}$, $m \in 1, 2, \ldots, d$ of fixed length n. Sample mean $\langle x^{(m)} \rangle$ and standard deviation $\mathrm{std}(m)$ is computed for each. Each subseries $x^{(m)}$ is then normalized, and cumulative time series is computed as

$$z_k^{(m)} = \sum_{i=1}^{k}(x_i^{(m)} - \langle x^{(m)} \rangle) \qquad \text{for } k = 1, \ldots, n.$$

Range $R(m)$ is then computed for each subseries as

$$R(m) = \max_{i=1,\dots,n} z_i^{(m)} - \min_{i=1,\dots,n} z_i^{(m)}.$$

Then, the mean value of the rescaled range

$$(R/S)_n = \frac{1}{d}\sum_{m=1}^{d} R(m)/S(m) \tag{1.15}$$

is computed. This procedure is performed for chosen values of n. As observed by Hurst [82], and later proven by Mandelbrot [85], the rescaled range asymptotically follows the relation

$$(R/S)_n \propto cn^K.$$

The Hurst exponent K then can be obtained using linear regression as slope of the line

$$\log(R/S)_n = \log c + K \log n. \tag{1.16}$$

Interestingly, if the measured quantites resulted from mutually completely independent events (i.e. white noise, with its corresponding cumulative sum, random walk), the relationship in equation (1.14) is replaced with

$$R(n)/\mathrm{std}(n) \propto 1.25\sqrt{N},$$

as can be easily verified by flipping a set of coins.[14] [83] This allows us to recognize stochastic processes with mutually uncorrelated values. The value of Hurst exponent for white noise is $K = 1/2$, and many natural processes, such as rainfalls, river water level heights, temperatures and pressures, annual growth of tree rings, and even financial markets have $K > 1/2$, suggesting long-term temporal correlations in the processes. Values of $0 \leq K < 1/2$, on the other hand, suggest long-time negative correlations, i.e. high values being often followed by low values in the future [82].

1.4.5 Higuchi's fractal Dimension

In this section, it will be beneficial to change our usual notation for the purpose of readability. Let us have time series $x(1), x(2), \dots, x(N)$. We select a $k \in \mathbb{N}$ and construct k new time series, denoted x_k^m for $m = 1, 2, \dots, k$, as

$$x(m), x(m+k), x(m+2k), \dots, x(m + \lfloor \frac{N-m}{k} \rfloor \cdot k), \qquad m = 1, 2, \dots, k, \qquad m, k \in \mathbb{N},$$

where m is represents the initial time, and k the interval time. In this way, we sample k subseries where the delay between successive points, or size in equation (1.9), is precisely k for each. Then, we define normalized average length L_m^k of each x_k^m, or its bulk in equation (1.9), as follows

$$L_m^k = \left(\sum_{i=1}^{\lfloor \frac{N-m}{k} \rfloor} |x(m+ik) - x(m+(i-1)k)| \right) \frac{N-1}{\lfloor \frac{N-m}{k} \rfloor \cdot k},$$

where $\frac{N-1}{\lfloor \frac{N-m}{k} \rfloor \cdot k}$ is a normalization factor. Length of the original curve as a function of k is then defined as the average $L(k) \coloneqq \langle L_m^k \rangle_m$ over k values of L_m^k. If $L(k) \propto k^{-D_H}$ for some value of D_H, then the curve

[14] Hurst himself actually made experiments, tossing 10 coins 1000 times. It took him almost 6 hours [86].

can be considered fractal with fractal dimension D_H, which can be estimated by least-squares fitting the logarithm of the length $\log L(k)$ as a function of $\log k$ [87].

In summary, Higuchi's fractal dimension can be defined, in analogy with correlation dimension and equation (1.12), as

$$D_H = -\frac{\partial \log L(k)}{\partial \log k}.$$

In other words, one may see Higuchi's fractal dimension as measure of irregularity, which in turn is measured as logarithmic rate of average variation of successive points [8]. Comparing with equation (1.9), can see that bulk is corresponds to the mean curve length over multiple subsampled time series $L(k)$ as a function of their size, which is the delay length k.

In comparison with correlation dimension, one of the benefits of Higuchi's fractal dimension is its relatively fast computation.

1.4.6 Sample Entropy

Understood in the context of dynamical systems, entropy is the rate at which a given system produces information. It is equal to the sum of all positive Lyapunov exponents of the system's attractor, and positive entropy indicates presence of chaotic dynamics [15]. Computing entropy from a physiological time series directly using the information-theoretical definition, is, however, problematic. The time series produced during measurements on biological systems are often short and noisy. Moreover, in EEG analysis, the impact of noise is especially severe. To combat this issue, many methods of computing entropy for such time series has been devised. Sample entropy represents an improvement [88] on other entropy measure popular in clinical settings, called approximate entropy, which has been successfully applied on EEG to classify diseases such as schizophrenia, epilepsy, and addiction [89].

For a given embedding dimension m and tolerance parameter r, sample entropy can be defined as the negative natural logarithm of the conditional probability that two subseqences of the time series of length m which are similar, i.e. their distance is less than r (excluding self-matches[15]), will remain similar after including the next point, i.e. when their respective lengths are increased to $m+1$. Thus, sample entropy is a measure of predictability, lower value of sample entropy indicates more self-similarity, or, in a certain sense, less complexity. More concrete definition may be stated as follows:

Definition 11 ([88])**.** *Let us have a time series $x_1, x_2, \ldots, x_N$, and let $\mathbf{x}_i^{(m)} = (x_i, x_{i+1}, \ldots, x_{i+m-1})$ be the i-th vector in the embedding of dimension m, and $\|\cdot\|_\infty$ be the Chebyshev metric[16]. Then,* ***sample entropy*** *is defined as*

$$\mathrm{SE}(m, r) = -\ln \frac{A^m(r)}{B^m(r)},$$

where

$A^m(r)$ *is the number of vector pairs in the embedding satisfying $\|\mathbf{x}_i^{(m+1)} - \mathbf{x}_j^{(m+1)}\|_\infty < r$, $i \neq j$, and*

$B^m(r)$ *is the number of vector pairs in the embedding satisfying $\|\mathbf{x}_i^{(m)} - \mathbf{x}_j^{(m)}\|_\infty < r$, $i \neq j$.*

We call the value r tolerance*.*

[15] This is one of the differences of sample entropy from approximate entropy.

[16] Any metric can be used, but Chebyshev metric is recommended by the original authors [88].

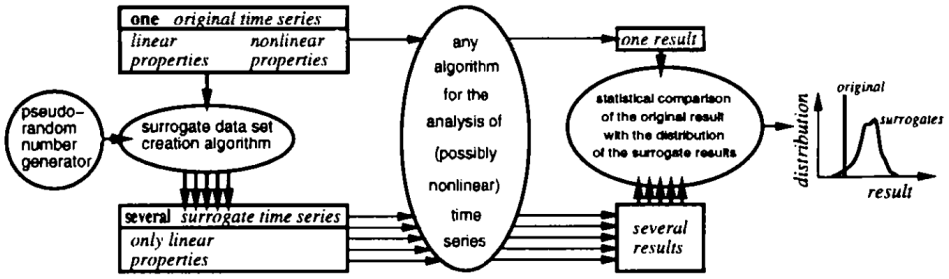


Figure 1.9: A schematic depiction of surrogate data testing process for the null hypothesis of a linear process [15].

Obviously, it is always $A^m(r) \leq B^m(r)$, hence sample entropy is always non-negative. If $A^m(r) = B^m(r) = 0$, no regularity has been detected, and $B^m(r) \neq 0$ with $A^m(r) = 0$ corresponds to (the above mentioned) conditional probability of 0, and infinite value of sample entropy. The value of tolerance r recommended by the authors is $0.2 * \text{std}(x)$, where $\text{std}(x)$ is the standard deviation of the original time series [88].

The algorithm for computing sample entropy first constructs the embedding vectors for embedding dimensions m and $m + 1$. For each embedding dimension, it counts the number of pairs of vectors $\mathbf{x}_i, \mathbf{x}_j$ where $i \neq j$, for which $\|\mathbf{x}_i - \mathbf{x}_j\|_\infty < r$. The negative natural logarithm of the ratio of number of those pairs for dimension $m + 1$ and the number of pairs for dimension m is the result.

Sample entropy has been successfully employed for diagnosing depression [7, 90]. It was shown to be significantly different between middle aged and elderly women during sleep [91].

Extend this section a little bit.

1.5 Surrogate Data Testing

It has been shown that, for example, fitered noise can mimic low-dimensional chaotic attractors when examined by Grassberger-Procaccia algorithm described above In the following, we will describe a method for answering this question.

To this end, we construct a Monte Carlo hypothesis test of nonlinearity. We choose a null hypothesis of a model for the process creating obtained data which denies the property we assume to measure. For each time series, we create so called *surrogate data* which deliberately capture only properties consistent with chosen null hypothesis, and compute the estimates using the same method as for the original data. If the result for the original time series is significantly different from the surrogate estimates, we reject the the null hypothesis. In the opposite case, we fail reject the null hypothesis. A schematic depiction of the process can be seen in Figure 1.9.

We may also want to know whether the process is deterministic or not. There are tests for that, but we are not using them in this thesis.

Here, we use two sided test, and measure of significance is defined as

$$\mathcal{S} \equiv \frac{|Q_{\text{orig}} - \mu_{\text{surr}}|}{\text{std}_{\text{surr}}}, \tag{1.17}$$

where Q_{orig} is the statistic computed for the original time series, and μ_{surr}, std_{surr} are the mean and standard deviation of the statistic computed for the surrogate time series [92]. If we assume that distribution of the generated is Gaussian, than $\mathcal{S} \geq 2$ is required for 95 % significance level. However, validity of this assumption is not always guaranteed. For non-Gaussian distributions, we may require larger $\mathcal{S}$, or, alternatively, use a rank based test, as follows [34].

Using rank-based test, we want to test if Q_{orig} is smaller or larger than the expected value of estimates produced by the null hypothesis model. If we generate n_s surrogate estimates, then, we have n_s estimates

following the null hypothesis, each having a probability $2/n_s$ of being the smallest or largest. A false rejection will happen if Q_{orig} happens to also follow the null hypothesis and is either the smallest or the largest, which happens with probablity $1 - \alpha := 2/(n_s + 1)$, where α is the confidence level. Hence, for confidence level $\alpha = 95\%$, the number of surrogates should be $n_s = 38$ [15]. Note that, since many of the algorithms used for estimating nonlinear measures are relatively computationally expensive, surrogate analysis with high confidence levels, especially on large datasets of multichannel signals such as EEG, is even more so.

1.5.0.1 Generating Improved Amplitude Adjusted Surrogates

For our purposes, since we assume that the data are produced by a nonlinear process, a reasonable null hypothesis may be that the data are produced by a Gaussian linear stochastic process AR(p)

$$x_{t+1} = \mu + \sum_{j=0}^{p-1} a_j x_{t-j} + \sigma e_t, \tag{1.18}$$

with unknown parameters $a_j, e_t, \mu, \sigma \in \mathbb{R}$ [92].

If the computed nonlinear statistic depends on the free parameters in AR(p) (1.18) (which is not true, e.g. for D_2), then one may try to estimate these parameters from the original time series. Alternatively (and this is the approach we use in our analysis), one may exploit the fact that AR(p) can be also perfectly described by its power spectrum [92].[17] Hence, to obtain a surrogate, one may simply perform a Fourier transform of the original time series, randomize phases, and apply inverse Fourier transform. This way, the amplitudes (composing the power spectrum) are preserved. This procedure has been named *Fourier transform phase randomization* (FTPR).

e.g. or i.e. here?

However, there is a drawback of FTPR. It has been shown that if the amplitudes of AR(p) are not Gaussian (as in (1.18)), e.g. nonlinear,then the surrogates created using this method show nonlinear behavior [34]. Rarely do the amplitudes of an experimental process follow a Gaussian distribution. Hence, we change our model to correspond a nonlinear, time independent filter applied to the output of AR(p). Surrogate creation algorithm for this model was described by Theiler in [92]: rescale the values of the original time series so that they are Gaussian, apply FTPR described above, rescale the values back to follow the same distribution of the original time series. This surrogate creation method is called *amplitude-adjusted Fourier transform* (AAFT), and has been successfully applied to EEG signal [93].

Even this method is not without its drawbacks: due to the final reordering, the original power spectrum is slightly distorted in the surrogate. In [93], it was proposed how to mitigate this effect. The amplitudes of Fourier transform of AAFT surrogates are replaced by the amplitudes of the original time series. The power spectrum is now correct, but the distribution is wrong. So, the original time series is reordered to according to ranks of values in this surrogate. This results in precisely the desired distribution of values, but again, slightly deviant power spectrum. These steps are then iterated and, experimentally, they results seem to converge. Hence, the final procedure, called *improved (iterated) amplitude-adjusted Fourier transform* (iAAFT) can be summarized as follows: [15]

Maybe talk about the problems, e.g. endpoint mismatch? We will need to refer to them later.

1. Compute and store the moduli of the original time series.

2. Create an AAFT surrogate as follows:

[17] This is due to Wiener-Khinchin theorem, which states, roughly, that spectral decomposition of autocorrelation of a stationary process is the power spectrum of the process.

2.1. Create a set of random numbers with Gaussian distribution.

2.2. Rank order the original time series (or the one obtained in step 5.), and reorder the random numbers created in the previous step such that they achieve the same ordering as the original time series.

2.3. Randomize the phases Fourier transform of the time series obtained in previous step and apply inverse Fourier transform.

2.4. Find the rank ordering of the time series obtained in the previous step, and reorder the original time series so that it assumes the same rank ordering.

3. Replace the moduli of these surrogates by those of the original time series and apply inverse Fourier transform.

4. Find the rank ordering of the time series obtained in the previous step, and reorder the original time series so that is assumes the same rank ordering.

5. Apply step 2. to time series obtained in the previous step, or stop if stopping criterion is reached.

1.6 Practial Applications

1.6.1 Applications in Depression Diagnosis

1.6.1.1 Nonlinear Measures

Although nonlinear dynamical analysis of EEG signal has been successfully applied to many psychological and psychiatric conditions, such as insomnia, schizophrenia, epilepsy, dementia, Alzheimer's disease, the number of studies applying methods of nonlinear time series analysis for clinical depression diagnosis is still relatively limited [5, 1]. Summary of studies reviewed in this section along with the obtained results can be seen in Table 1.1. Hereby we present the studies in chronological order.

It has been found that the EEG dynamics of depressed patients exhibit more predictability, and therefore less complexity, than those of non-depressed ones, with this indicator receding after treatment [11, 94]. Instead of measuring compexity using correlation dimension, the authors applied method of nonlinear forecasting of dynamical systems based on the local approximation of neighborhood evolution [11].

Another study analyzed sleep EEGs of depressed and control subjects, and found significantly decreased values of Lyapunov exponents in a sleep stage IV in depressed patients relative to control [12].

In 2007, Lee et al. found significant differences between the values of DFA in 11 depressed patients matched with the same number of healthy controls. Moreover, they found signiciant correlation between the values of DFA and depression scores in almost all recorded channels [13].

In 2012, Ahmadlou et al. decomposed 5 EEG channels recorded from frontal lobes of healthy and depressed patients using wavelet filter banks, measured their complexity using Higuchi's fractal dimension, subsequently used ANOVA to discover the most meaningful differences between the groups, and trained a probabilistic neural network classifier, achieving 91.3% classification accuracy on limited amount of data. This research suggested potential of frontal lobe signal assymetry as a measure for depression [8].

A year later, Hosseinifard et al. extracted Correlation Dimension (CD), the Largest Lyapunov Exponent (LLE) and Higuchi's fractal Dimension (HD) from 4 EEG channels of 90 patients split evenly between depressed and non-depressed subjects, achieving 90% accuracy using a logistic regression classifier and 3-fold leave-one-out cross validation [10].

In the same year, Bachmann et al. compared linear measure called Spectral Assymetry index (SASI) and HD for depression diagnosis on 34 subjects split evenly between depressed and control group. SASI achieved true detection rate in 88% in depressives and 82% in the controls, while HD provided true detection rate of 94% in the depressives and 76% in the controls [9], thus both giving comparable results.

In 2014, sample entropy, approximate entropy, Renyi entropy and bispectral phase entropy were used by Faust et al. in [6] to achieve overall 99.5% accuracy using 10-fold cross validation on sample of 30 healthy and 30 depressed patients. This study reports the highest accuracy out of the reviewed studies.

A year later, Acharyia et al. invented a new biomarker based on nonlinear measures called Depression Diagnosis Index (DDI) and also achieved considerable accuracy, but on a small sample of 15 depressed patients matched with healthy controls [7].

In comparison with depression diagnosis, less research has been dedicated to the problem of treatment outcome prediction using nonlinear measures. The only study we are aware of is [95], where the authors computed LLE using false nearest neighbors and another nonlinear measure called Lempel-Ziv Complexity (LZC) in order to investigate depressed patients' response to treatment with repeated Transcranial Magnetic Stimulation (rTMS). They observed a significant decrease in LZC in nonresponding patients and significant increase in responding patients and healthy controls.

1.6.1.2 Other Techniques

Most of the results of applications of quantitative methods to EEG analysis of depressed patients were obtained using linear methods, such as extracting band power features to detect assymetries between hemispheres of the brain [1]. For example, in [96], the researchers were able to distinguish between 23 healthy and 70 depressed patients using only linear measures[18]. Statistically significant differences between depressed patients and healthy controls using anterior α assymetry were found in [97], followed by [98] with more linear assymetry features and more substantial sample size and similar results. Moreover, statistically significant differences in relative frontal power were found in [99] between depressive patients responding and nonresponding to treatment. Puthankattil et al. were able to obtain high 98.1% accuracy using relative wavelet energy and neural networks on a sample of 30 healthy and 30 depressed patients. Each patient was recorded multiple times to increase the size of the training and test sample, which may have introduced correlations between the training and test sets. Nonetheless, these studies, supported by neuroscientific theory, sparked interest in linear assymetry measures in depression diagnosis.

However, many studies cast doubts over validity of these measures, such as [100, 101, 102] finally followed by [103] in 2013 with more substantial sample size, all either finding no statistically significant differences between depressed patients and healthy controls, or no statistically significant correlations of assymetry measures with depression scores. Moreover, when compared directly to nonlinear measures, linear techniques seem to perform at most comparably to nonlinear ones, or worse. As already mentioned, in [9], nonlinear measure HD was comparably discriminative to linear measure of spectral assymetry. In [10], Hosseinifard et al. obtained lower classification accuracy using linear band power features in comparison to CD, LLE, and HD.

In attempt to discriminate between MDD patients responding and nonresponding to rTMS treatment, in [104], the researchers identified increased frontal θ power and lower peak α frequency as potential biomarkers for patient's failure to respond to the treatment.

Considerably more research has been devoted to the problem of treatment outcome prediction using linear measures in comparison with nonlinear measures. However, most of the studies include only small

[18] Note however the class imbalance in the study.

number of samples and tend to limit their focus on a specific biomarkers [4]. For a comprehensive review, see [4].

1.6.1.3 Potentially Related Diseases

Sleep disorder diagnosis may also relevant to this work for the very close connection of depression with disturbed sleep and insomnia [105]. The first study emplying techniques of nonlinear analysis on human EEG was published in 1985 and dealt with sleep recordings [106]. This early success sparked intensive research focus on applying nonlinear analysis to sleep data, thus generating relatively large amount of results.

Many studies focused on extracting Lyapunov exponents of EEGs measured during various sleep stages. The general pattern that emerged was that deep sleep stages exhibit lower complexity evidenced by lower fractal dimensionality and lower values of the largest Lyapunov exponent [5].

Recurrence plots, and RQA in particular, have been demonstrated to be effective at decoding neuroscientific physiological time series. For example, they have been suggested as a method of lowering signal-to-noise ratio in analysis of event related potentials in response to a surprising stimulus, where repeated exposure would influence the outcome (and thus classical averaging methods are not viable) [44]. Moreover, they have been successfully employed in detecting epileptic seizures using intracranial recordings [107]. Simple K-nearest neighbors classifiers achieved surprisingly high accuracies at emotion recognition tasks [108], and convolutional neural networks used recurrence plots for activity recognition [109]. Most importantly for our study, recurrence plots of signals in the left hemisphere were observed to qualitatively differ between healthy baseline and depressed patients. The authors suggested that this area is worth further exploration [110].

Measure	Analysis method	Accuracy or obtained results	Test sample		# channels	Rec. length	Reference	Note
			Healthy	Depressed				
SE and other entropies	WPD+PNN	0.99	30	30	9	5 min	[6]	Evaluated using 10f-CV.
DDI	SVM	0.98	15	15	-	8 s	[7]	-
HD	EPNN	0.91	3	3	19	1 min	[8]	-
HD	LDA	0.85	17	17	18	2 s	[9]	Evaluated on training set.
LLE, CD, DFA, HD	LR	0.9	45	45	19	5 min	[10]	-
NL forecasting	Measuring predictability	Increased predictability in MDD patients	8	16	12	40 s	[11]	-
LLE	Statistical analysis	Found s.s. diff.	13	15	4	2.44 min	[12]	-
DFA	Statistical analysis	Found s.s. diff.	11	11	5	5 min	[13]	-
RWE	NN	0.98	40	40	9	5 min	[111]	-
Power, assymetry, coherence measures	Statistical analysis	0.91	23	70	21	20 min	[96]	-
Anterior α-assymetry	Statistical analysis	Found s.s. differences	15	22	19	2 s	[97]	-
Power and assymetry measures	Statistical analysis	Found s.s. differences	86	53	8	-	[98]	-
SASI	Statistical analysis	No s.s. diff. found	18	18	9	2.5 s	[100]	-
FAA	Statistical analysis	No s.s. corr. found	-	30	25	1 min	[101]	FAA is stable in depressed patients, but uncorrelated with depression level.
STARC	Statistical analysis	No s.s. diff. found	20	22	19	1 s	[102]	Focus is on classification of male and female patients.
FAA	Statistical analysis	No s.s. corr. found	-	79	32	2 s	[103]	Measured correlations with depression scores.

Table 1.1: Overview of reviewed studies analyzing EEG in relation to depression, ordered by obtained classification accuracy (if applicable), otherwise chronological order was used. The first part corresponds to studies using nonlinear dynamical measures.
Abbreviations: DDI - Depression Diagnosis Index, (EP)NN - (Enhanced Probabilistic) Neural Network, FAA - Frontal Alpha Assymetry, RWE - Relative Wavelet Energy, SASI - Spectral Assymetry Index, STARC - Spatiotemporal Analysis of Relative Convergence, WPD - Wavelet Power Decomposition, kf-CV - k-fold Cross Validation.

1.6.2 Applications in Depression Prognosis

In terms of EEG-based treatment outcome prediction, substantial effort has been devoted to examination of frontal θ frequencies, especially relative to α frequencies. This rationale is supported by several neuroscientific theories [112]. For example, in [99], the authors developed Antidepressant Treatment Index (ATR), a nonlinear combination of α and θ powers to predict treatment outcome with 63% accuracy. The same research group had been examining relative θ powers for multiple years prior to this study, with the best results achieving 71% accuracy on a dataset of 52 MDD patients [112].

In [113], was used frequency principal component analysis derived from current source density waveforms to extract EEG α aplitude, achieving also 63% accuracy. However, the highest accuracy of all reviewed studies was achieved using more advanced techniques in [114], where the authors extracted log Power Spectral Densities (PSD) levels at all frequencies of interest for each electrode, spectral coherence, mutual information, log ratios of left and right hemisphere powers and log ratios of anterior and posterior powers, and then Fisher discriminant ratio for feature selection and mixture of factors machine learning technique based on maximum likelihood for classification, finally achieving 88% accuracy. Nevertheless, this result was obtained on a relatively small dataset of 22 patients.

Measure	**Analysis method**	**Accuracy or obtained results**	**Test sample**		**# channels**	**Rec. length**	**Reference**	**Treatment**
			Responding	**Nonresponding**				
Power, coherence, mutual information measures	MFA	0.88	7	15	16	60 s	[114]	SSRI
Frontal θ power	ATR	0.63	45	37	4	2 s	[99]	SSRI
Average α amplitude	Measure above median of controls	0.63	28	13	67	2 min	[113]	SSRI

Table 1.2: Overview of reviewed studies analyzing EEG in relation to response, ordered by obtained classification accuracy (if applicable), otherwise chronological order was used.
Abbreviations: ATR - Antidepressant Treatment Index, MFA - Mixture of Factors

1.6.3 Limitations

Some authors suggests that the since most plausible research target for explaining the brain dynamics are the assemblies of coupled and synchronously active neurons, and since majority of those assemblies are describable by nonlinear differential equations, principles derived from nonlinear dynamics are applicable to characterization of these neuronal systems [30]. The approach of estimating a finite embedding dimension (see Section 1.3), however, has been doubted by some of the most prominent figures in the field of nonlinear dynamical analysis, such as the originators of Grassberger-Procaccia algorithm (see Section 1.4.2) [115, 116], followed by many others later [37, 117, 15]. With the exception in the case of epileptic seizures [5], there is very little evidence for the seemingly improbable hypothesis that such complex system with many extrinsic influences and interactions, such as the brain, would exhibit a level of comlexity comparable to e.g. a Lorenz system. The observed estimates of low dimension may due to artifacts or limited data size [115, 116].

On the other hand, as we will see in Section 1.6.1, the techniqes derived from these theories still provide some useful information and are successfully applied in many practical situations. Therefore, it seems to be the case that indeed, brain dynamics are much more complex than we are forced to assume based on the theory, but nonlinear dynamical analysis still manages to capture some of its important aspects. Some researchers argued that these observations ask for reinterpratation of those results [60], whereas others call for development of new measures which reflect the ongoing progression in understanding of brain dynamics [5].

Chapter 2

Nonlinear Analysis Approach

2.1 Dataset

The dataset was provided by the Czech National Institute of Mental Health, and the recording was performed by the institution's local specialists. It comprises of 133 subjects, 104 women and 29 men, ranging in age from 30 to 65 (47.7 ± 9.58).[1] Handedness was not recorded. Montgomery-Asberg Depession Rating Scale (MADRS) [118] questionnaire assessed by a trained psychologist was used to measure depression severity. This psychometric measurement results in a depression score ranging from 0 (normal) to 40 (severe depression), usually with the following cutoff points [119]:

0 - 6 : symptom absent,

7 - 19 : mild depression,

20 - 34 : moderate depression,

34 - 40 : severe depression.

The experiment lasted 4 weeks. At the beginning of week 1, each subject's depression score was measured, their EEG signal was recorded, and, based on the measurement and patient's history, prescription of up to 4 treatments (drugs or rTMS) was made. After 4 weeks, depression score was remeasured and EEG signal recorded again.

During the EEG recording, 19 electrodes were placed on the scalp in accordance with the Internation 10-20 system (FP1, FP2, F3, F4, C3, C4, P3, P4, O1, O2, F7, F8, T3, T4, T5, T6, Fz, Cz, Pz), see Figure 2.1 for reference. EEG signals of 99 subjects were recorded at sampling frequency f_s of 250 Hz, while 1000 Hz was used for the remaining 34 patients. The patients were not told to close their eyes for the duration of the recording, resulting in unwanted artifacts in the signal. Some of the artifacts were removed manually by the researchers by omitting those parts from the recording, and concatenating the remaining parts. Durations of the resulting measurements range from 23.5 s to 170 s (75.6 ± 20 s) for f_s = 250 Hz, and from 48.8 s to 140.4 s (79.5 ± 18.4 s) for f_s = 1000 Hz. The distributions of depression scores, EEG recording durations, patients' ages and sexes are visualized in Figure 2.2. A typical recoding can be seen in Figure 2.3.

We should recognize limitations of this dataset:

- That the patients were not randomly selected - all the patients entered the study because they were experiencing problems negatively impacting their lives. Thus, as a study of depression biomarkers,

[1] We use the notation mean ± std.

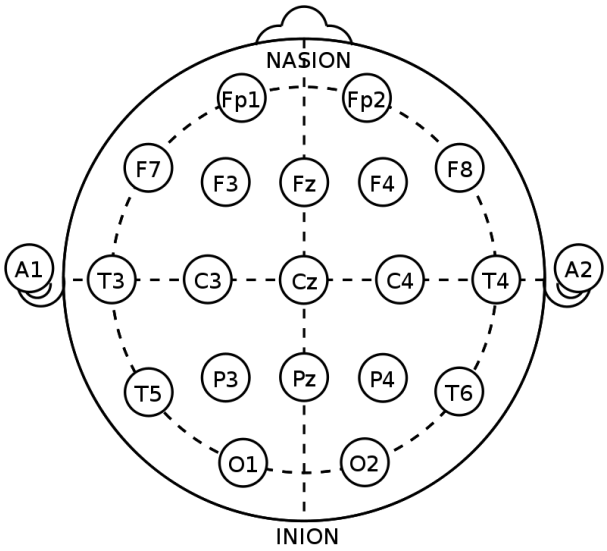


Figure 2.1: The International 10-20 system for placement of EEG electrodes used in our dataset. ([120])

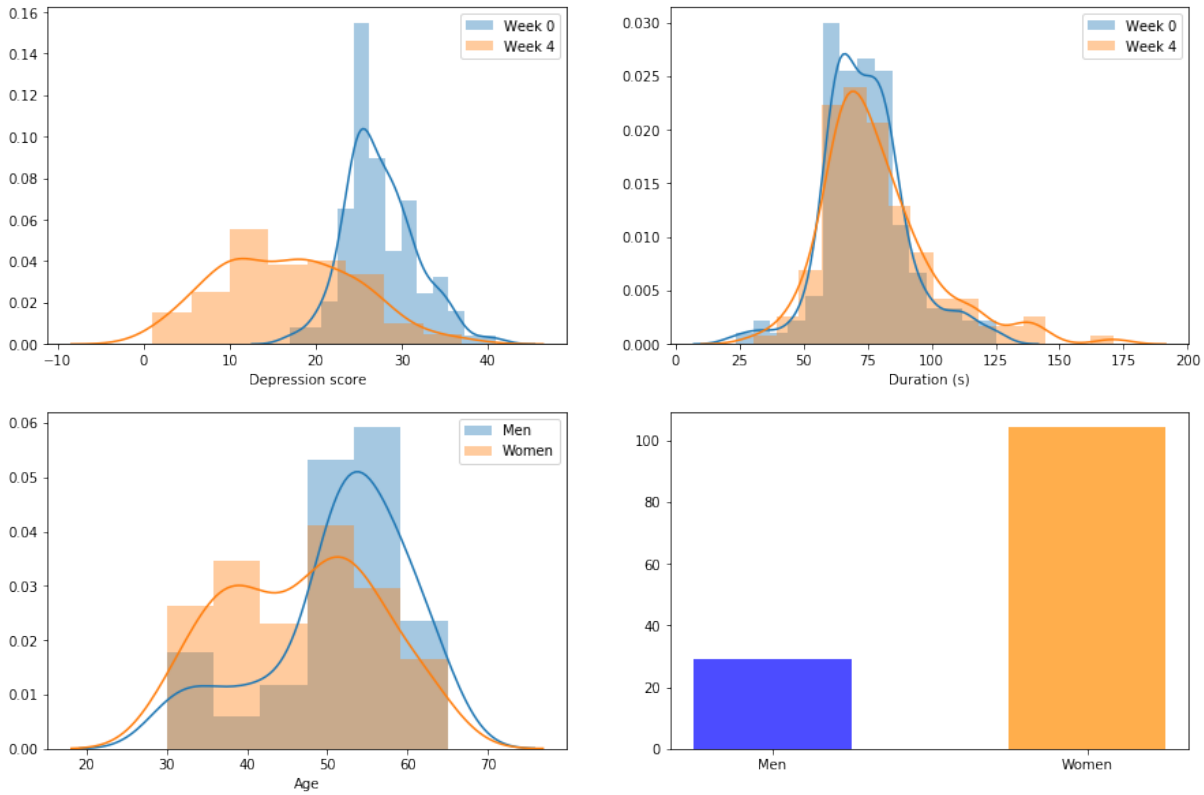


Figure 2.2: Visualization of the main dataset features. Starting in the upper left hand side and continuing in the clockwise direction, the first figure shows the distributions of depression scores measured on the first and second patient visit respectively. The second figure shows the distributions EEG recording durations on the first and second week respectively. The third figure shows distributions of ages for male and female participants, and the last figure visualizes the number male and female participants.

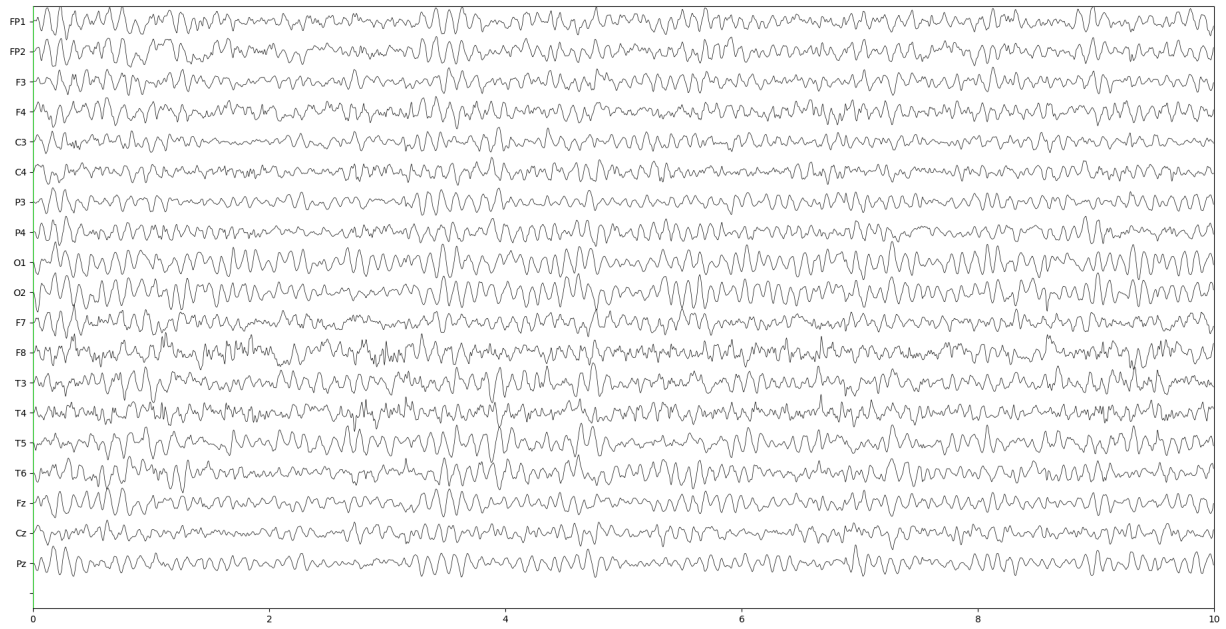


Figure 2.3: An example recording for patient 1, first session. Horizontal line shows seconds, vertical line shows voltage scaled for purpose of visualization.

the experiment lacks truly symptom absent group. However, the patients did differ significantly in severity of the disease.

- For a study of brain regions associated with depression, this study lacks data on patients' handedness, which may be relevant for distribution of activity in the hemispheres.
- For a study of response of patients to treatments, this dataset lacks a control group given no treatment (or placebo).
- For a study of treatment effects, patients were assigned different cominations of drugs, making an attempt of finding the singular cause of any observed effects impossible.
- Durations of the recordings do not allow us to put sufficiently low bound on the maximum error we can theoretically achieve for computation of the lagest Lyapunov exponent and correlation dimension (see Sections 1.4.1.2 and 1.4.2.2).

2.2 Preprocessing

Recordings of $f_s = 1000$ Hz were downsampled (decimated) by factor 4 to 250 Hz using the Fourier method (also known as trigonometric interpolation), i.e. by performing discrete Fourier transform on the original series, dividing it into $2 * 1000/250 = 8$ intervals, removing all but the first and the last intervals (thus removing the highest positive and negative frequencies, corresponding to low-pass filtering), and performing inverse discrete Fourier transform. This procedure assumes that the signal is periodic, and may have some influence on the obtained results. However, it was observed that this effect is almost negligible, even for considerably higher decimation factors [121].

In further analysis, unless otherwise specified, recordings were shortened to fixed length of 60 s (15 000 timesteps). This measure was taken in order to achieve the the following characteristics on the data:

1. minimize the stationarity effect (see Section 1.2.3), and at the same time

2. include as many timesteps in the recordings (i.e. as high time series length as possible) as possible, while

3. including as many recordings as possible.

This resulted in exclusion of 26 recordings from the total of 266. We have to recognize that, as we saw in Sections 1.4.1.2 and 1.4.2.2, this number of timesteps limits us to relatively small embedding dimensions to achieve theoretically low bounds on the maximum error for computation of multiple nonlinear measures.

In [10], band-pass filtering was used to remove frequencies which are physiologically impossible to produce by neural oscillations (e.g. high-pass filtering with 0.5 Hz threshold or lowpass filtering with 70 Hz threshold). The activity in the delta band (1-4 Hz) may exhibit less differences between healthy and depressed subjects [96]. Furthermore, it may lead to reduction of the noise present due to effects of blinking and cardiac artifacts [122]. It has been suggested to notch filter at power line frequencies (40 Hz or 50 Hz) [123]. However, some authors suggest that linear filtering may adversely affect the results of nonlinear analysis [15]. Others, on the other hand, observed that simple linear filtering does not influence the reconstruction of embedding space considerably [124]. If quality of the data is sufficient, then filtering is not necessary [125].

To determine the effects of filtering on our datasets, we filtered the data to the 0.5 Hz - 70 Hz range using Butterworth filter of order 3, and then we performed the preliminary analysis described in Section 2.4. However, we did not find any significant improvement, and therefore we decided to avoid the impact of filtering on the final results.

2.3 Estimation of Nonlinear Measures

2.3.1 Our Procedure

It is well known that the results of algorithms for estimation of nonlinear measures presented in Section 1.4 depend not only on the amount of noise in the data, the preprocessing and the choice of the algorithm, but also, to a considerable degree, on the embedding parameters and other input parameters [74, 126]. Therefore, selection of the algorithm and its parameters is of substantial importance.

Our procedure for their selection was as follows. For each nonlinear measure, we created a list of feasible parameters which we subsequentially evaluated in a preliminary analysis. Because most of the algorithms are relatively computationally expensive, length of the list should be limited. Then, we proceeded to label the data for each item on the list and evaluated it. In Section 2.5 we present the results for the most discriminative parameters. For creating the list, we used the following methods:

Literature review: Review the literature on use of the nonlinear measure in question on EEG data, and search what parameters were used.

(Algorithm-assisted) manual estimation: Estimate the optimal embedding parameters by analyzing the algorithm outputs for multiple samples. Use algorithms for estimating the embedding parameters and analyze their results.

Automatic pipeline: Create automatic pipeline, which will select the optimal parameters for each sample. This step was performed only for correlation dimension and the largest Lyapunov exponent.

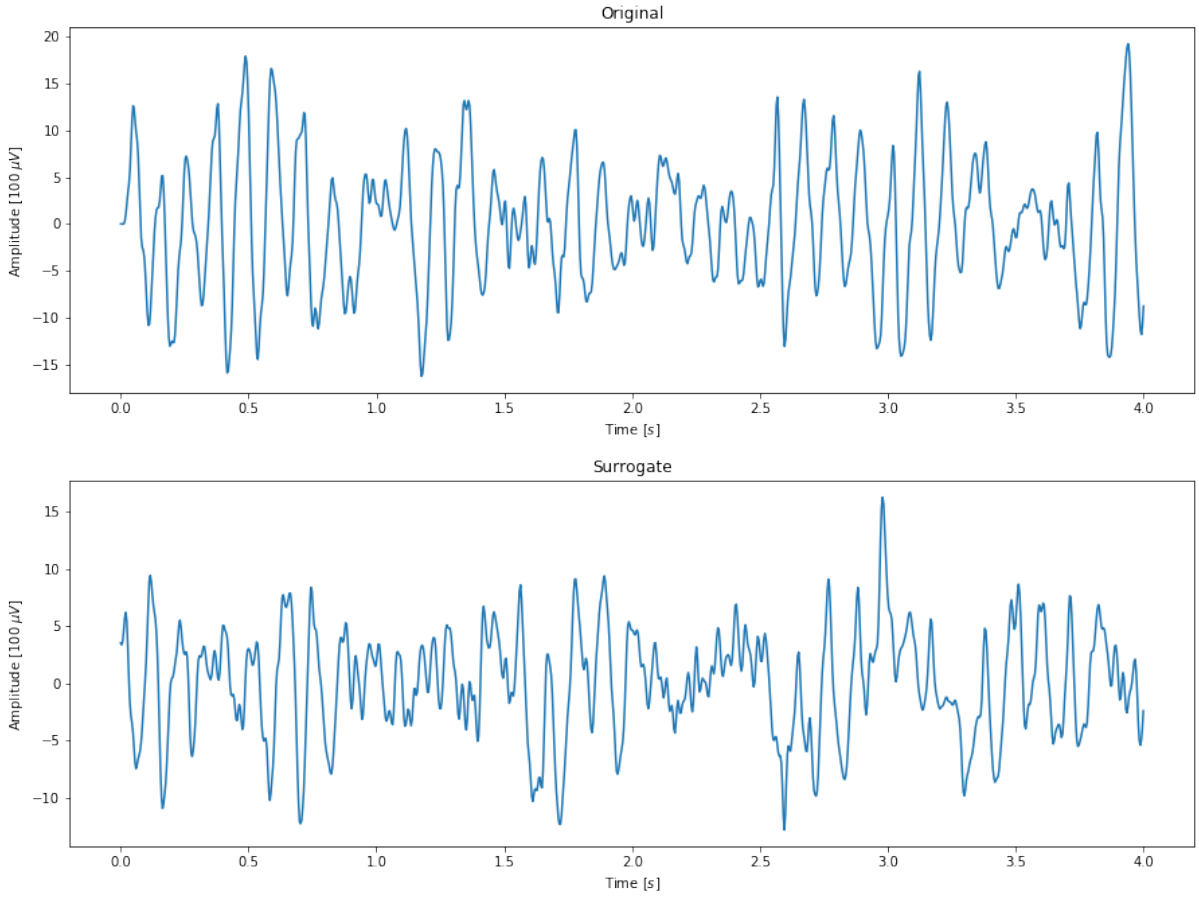


Figure 2.4: A comparison of the first 4s of a time series and its AAFT surrogate.

2.3.2 State Space Reconstruction

2.3.2.1 Time Delay

In this section, we present an analysis of the estimation of time delay using several techniques. For illustrative purposes, we explain how these techniques can be used in analysis of, and present the results of these techniques for, the signal obtained from the FP1 electrode of recording of patient 75, second visit. As described in Section 2.1, the time series, shown in Figure 2.4, was clipped to 60 s (15 000 data points). The selected techniques of time delay analysis were:

1. Reconstruction plots
2. Autocorrelation $A(\tau)$ (see Section 1.3.4.1)
3. Delayed mutual information $\mathcal{I}(\tau)$ (see Section 1.3.4.2)
4. Average displacement from diagonal (ADFD) (see Section 1.3.4.3)
5. PCA reconstructions comparison (see Section 1.3.4.4)
6. Integral local deformation (ILD) (see Section 1.3.4.5)

Figure 2.5 shows reconstructed trajectories for the first 4 s (1000 data points) of the recording, for varying time delay τ. As expected, the reconstructed attractors for small delays cluster along the main diagonal, expand, and then become increasingly chaotic with larger τ. However, it is impossible to judge objectively on the degree of folding in the attractor from these plots (even for shorter time series), which highlights the importance of qualitative measures for EEG signals.

Typical plots of autocorrelation and delayed mutual information can be seen in Figure 2.6. First local minima of DMI and first τ for which $A(\tau) \leq 1/e$, respectively $A(\tau) \leq 1 - 1/e$ are marked by yellow dots. For this channel, these are $\tau_{\mathrm{DMI}} = 10$ and $\tau_A = 4$, respectively $\tau_A = 6$. It is immediately obvious that estimates of these techniques differ considerably. However, the variance of the estimated parameters is small both across channels and across patients for both techniques. To illustrate, we computed the estimates all channels of this recording, and their distribution for both DMI and autocorrelation can be seen in Figure 2.7. For the selected patient, autocorrelation shows less variance and lower suggested time delays. This behavior was observed across patients.

Figure 2.8 shows singular values of the PCA reconstruction as functions of time delay τ. The two largest singular values corresponding to the main axes of the attractor clearly stand out. Besides the dominant collapse at $\tau = 14$, upon closer observation, one may notice multiple smaller collapses, such as one at $\tau = 7$. We can see how the attractor expands in the third and fourth dimension for $\tau = 3$ (similar behavior is also visible in Figure 2.5), which may suggest $\tau_{\mathrm{SVD}} \in \{3, 4\}$ as optimal. However, one may also choose $\tau_{\mathrm{SVD}} = 6$ as optimal, since all the attractor seems mostly unfolded in all the available directions. This highlights how subjective is evaluation of results of this technique. Thus, for automatic evaluation, it is preferrable to use other method.

The results obtained by ADFD for embedding dimensions 5, 10 and 15 can be seen in Figure 2.9; the green dashed lines represent derivatives of the respective curves, and the points mark the minimum value of τ for which derivative ADFD drops below 40% of its initial value, as discussed in Section 1.3.4.3. The average displacement tends to increase with m, and saturates for relavely small values of τ - thus, the estimated time delays are (consistently) lower than those obtained by most other techniques. Moreover, ADFD requires prior selection of m, while the algorithms for selection of m we use (FFN and AFN), require estimation of τ, making this technique largely impractical.

	Estimated optimal time delay
Reconstruction plot	-
Autocorrelation	6, 4
Delayed mutual information	7
SVD analysis	6
Average displacement	2, 3
Integral local deformation	4

Table 2.1: Estimated optimal time delay values of the individual examined techniques for patient 75, second session. The results vary widely based on used technique, but have relatively low variance across patients.

The most powerful algorithm for embedding parameters we used is ILD. We implemented the algorithm based on Buzug's original description [67]. Figure 2.10 shows multiple curves ILD(m, τ) as functions of time delay τ for fixed values of the embedding dimension m. There is a clear minimum at $\tau_{\text{ILD}} = 4$ for all the curves, and the curves become increasingly similar. Interestingly, the convergence is slower near the minimum. Various classes of behavior were observed across channels and patients; however, since this is highly computationally expensive algorithm, it is impractical to analyze them on datasets of the size of the one used in this study.

As explained in Section 1.3.4, these techniques should be used only as inspection tools, not as reliable guides for selection of τ. The ultimate goal of the reconstruction is to obtain as accurate values of the nonlinear parameters as possible, and thus selection of the optimal embedding parameters may differ for each of them. Thus, for example, in order to select the proper embedding parameters for computation of the largest Lyapunov exponent, we inspected the scaling regions for multiple values of m, τ, Theiler window and other parameters, and picked those with the longest scaling regions (since the length of the scaling regions is proportional to the certainty of the estimate [68]).

Table 2.1 shows an overview of estimated values of τ. Autocorrelation, DMI, and singular values analysis report lower values than ADFD and ILD. However, Rosenstein notes that the best estimates of largest Lyapunov exponents were obtain for the autocorrelation threshold of $1 - 1/e$. For this threshold, the autocorrelation suggests $\tau_A = \tau_{ILD} = 4$ as optimal (and the distributions shift accordingly), thus in agreement with ILD.

For comparison of obtained In the Section 2.3.3, we will show the effects of increasing τ on the average divergence.

2.3.2.2 Embedding Dimension

For estimating the embedding dimension, we used combination of *False Nearest Neighbors* (FNN) and Average False Neighbors (AFN) algorithms, described in Sections 1.3.5.1 and 1.3.5.2 respectively. The convergence of ILD curves and saturation of correlation dimension also provides insight into optimal choice of embedding dimension, but, as mentioned, is impractical due to high computational cost. As explained in Section 1.4.2, correlation dimension is expected to saturate for high enough choices of embedding dimension. However, we found that instead of saturating, it tended to decrease after reaching a maximum as a function of m, see Figure 2.17. This may be because the attractor is not represented adequately in high embedding dimensions with limited amount of data. Moreover, the computational costs of this method are also considerable.

As expected, the percentage of reported false neighbors depends strongly on the selected values of R and A from equations (1.6) and (1.7). This is illustrated in Figure 2.11, showing the percentage of false

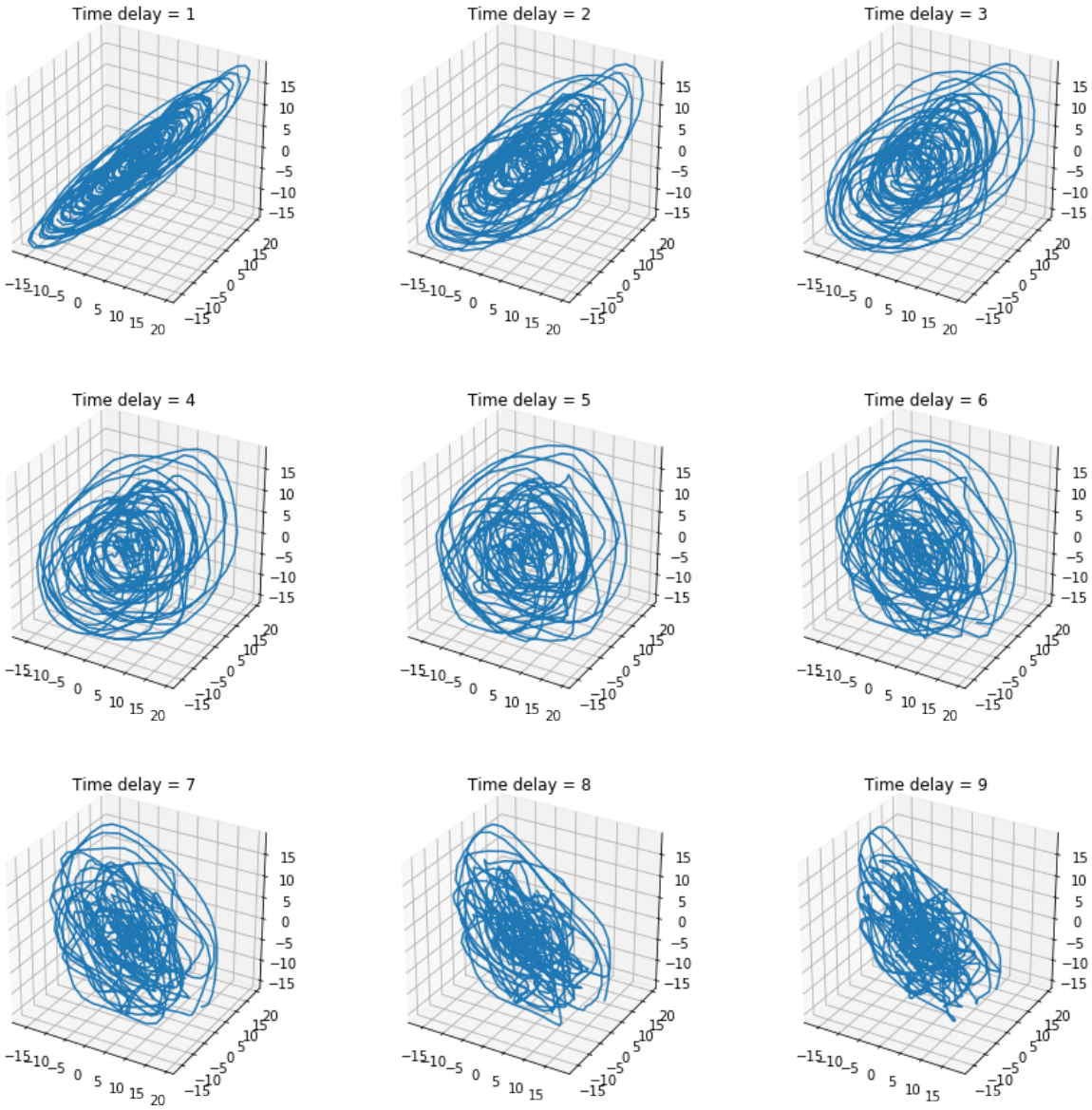


Figure 2.5: State space reconstruction for embedding dimension $m = 3$ for various values of time delay τ. Only first 40 seconds of the recording used for purpose of visualization. The axes represent the state vector coordinates. One may observe the increasing complexity and slowly progressing expansion of the attractor from a line to a random cloud of points.

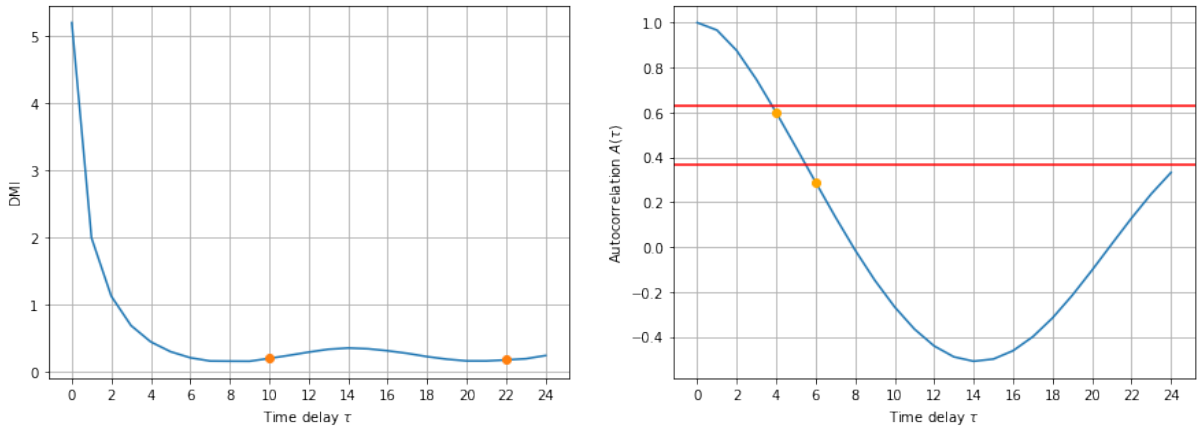


Figure 2.6: Delayed mutual information (DMI) and autocorrelation as functions of τ. The red line shows threshold values $1 - 1/e$ and $1/e$ respectively. The plots of surrogate data are equivalent. For this time series, DMI estimates optimal $\tau = 10$, and autocorrelation $\tau = 4$ or $\tau = 6$. We can see that the estimates vary significantly across those techniques, but judging by Figure 2.5, the autocorrelation estimates seem more reasonable.

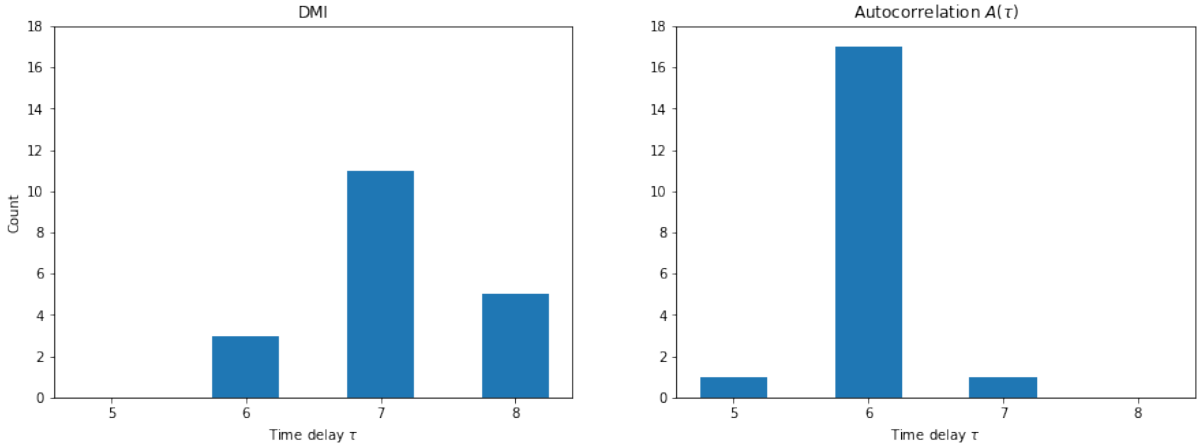


Figure 2.7: Distributions of time delays across channels computed using delayed mutual information and autocorrelation for threshold $1/e$.

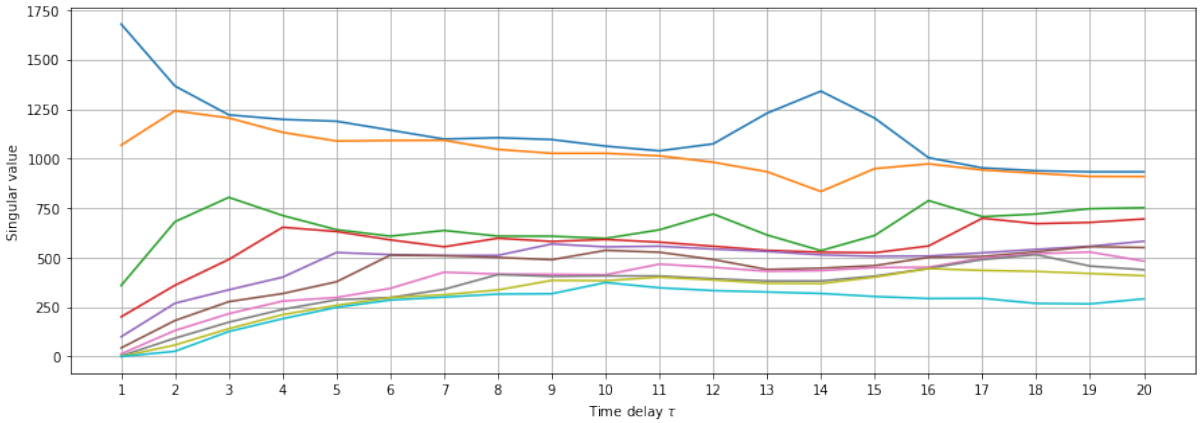


Figure 2.8: Plot of singular values as functions of τ for $m = 10$. The two largest singular values corresponding to the main diagonals of the attractor are clearly visible. The singular values are approximately for $\tau = 12$ before a collapse in the reconstruction immediately for the following values of τ.

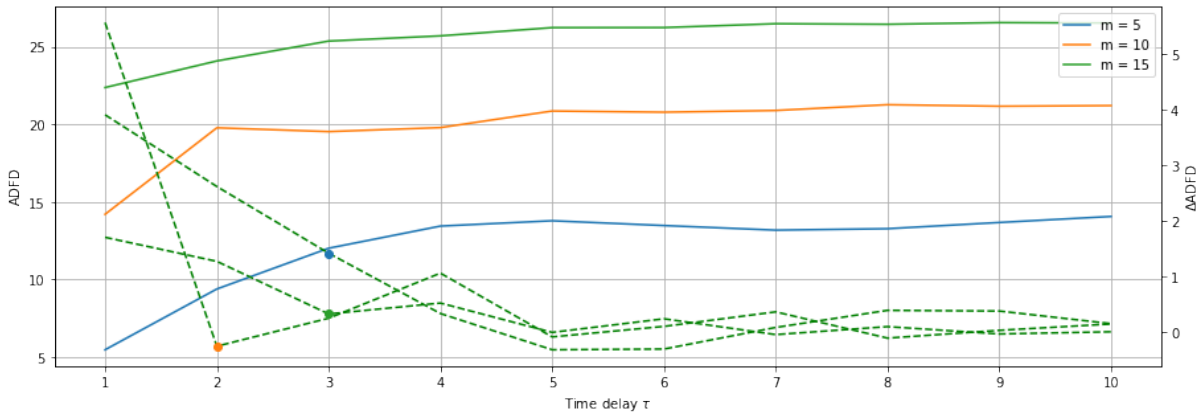


Figure 2.9: Plot of average displacement from diagonal for embedding dimensions 5, 10, and 15. The dashed green lines represent derivatives or respectives curves, the dots mark the minimal values of τ for which the derivative of ADFD(τ) reaches 40% of its initial value.

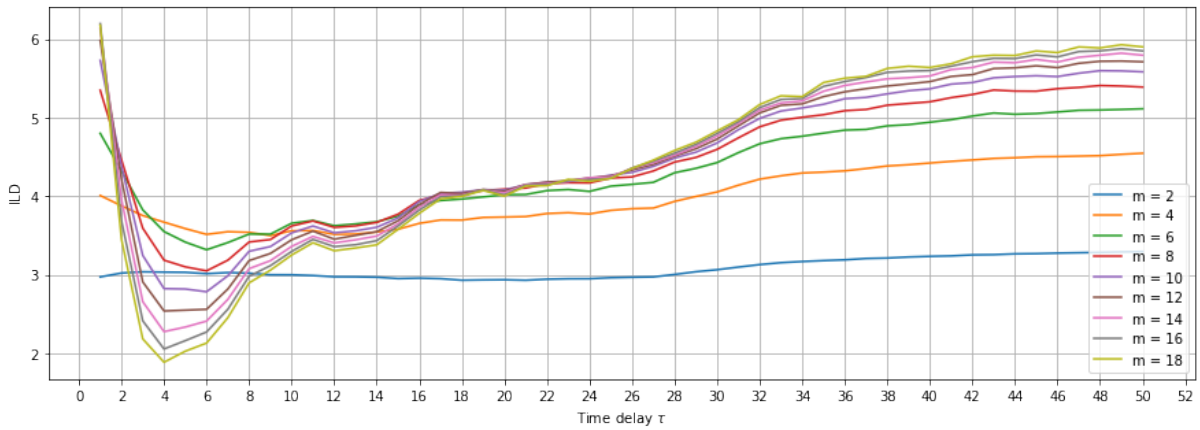


Figure 2.10: Plot of integral local deformation for varying values of the embedding dimension m. The individual curves converge with clear minimum at $\tau = 4$. The parameters used for this computation are $q_{\max} = 10$, $t_e = 3$, $N_{\text{ref}} = N_v$, $k = 20$ and $w_t = 10$ (see Section 1.3.4.5).

neighbors reported by the respective criteria for varying values of A and R, and for several values of time delay τ. The percentages reported by the criterion I are almost independent of τ, whereas increasing τ tends to increase the percentage reported by criterion II. For high enough τ, criterion II will report all neighbors as false.

The apparent independence of the results of the criterion I on τ indicates that, regardless of τ, the same percentage of near neighbors changes their distance proportionally with increase in m. As explained in Section 1.3.5.1, this behavior that can be expected of randomly generated uniformly distributed sequence of numbers. Indeed, behavior of the criterion II is consistent with this hypothesis - it eventually increases to 100% for all values of A, essentially indicating infinite dimension. By selecting proper parameters and using both criteria cojointly, however, FNN can still be used to obtain reasonable results, consistent with estimates obtained by ILD, AFN, and the literature. We will use this fact in our procedure of automatic selection of embedding parameters.

Actually explain it there - nearest ≠ close, etc. . . , [68]

The E_1 statistic of AFN usually stops increasing for approximately the same value as reported by criterion I of FNN for $R = 2.5$, see Figure 2.12. The E_2 statistic, tends to oscillate in small neighborhood of value 1, which is an indication of nondeterminism [69].

Report average m computed by ANN and FNN, $R = 2.5$, $A = 2.0$, $\Delta E_1 \leq 0.005$ for this patient using a histogram.

2.3.3 Largest Lyapunov Exponents

2.3.3.1 Manual Analysis

For all computations of the largest Lyapunov exponent, we used the Rosenstein's algorithm [73] described in Section 1.4.1.1, with Theiler window w_t length of 50 (200 ms). We found that the results were similar for values w_t of 10, 50, 100 and 1000. Similar to Section 2.3.2.1, here we present an illustrative analysis of the results for the FP1 electrode of the same arbitrarily selected patient.

Figure 2.13 shows divergence plots for different values of the embedding dimension m and time delay τ. Longer scaling regions correspond to higher certainty of the estimate. The short scaling regions and high slopes for small embedding dimension appear because, when the attractor is not unfolded, near neighbors are not actually close in the phase space and thus their trajectories diverge quickly. With increasing embedding dimension the scaling region clearly lengthens, but the slope also slowly approaches zero, and scaling region gradually disappears. This is because the average divergence cannot exceed the

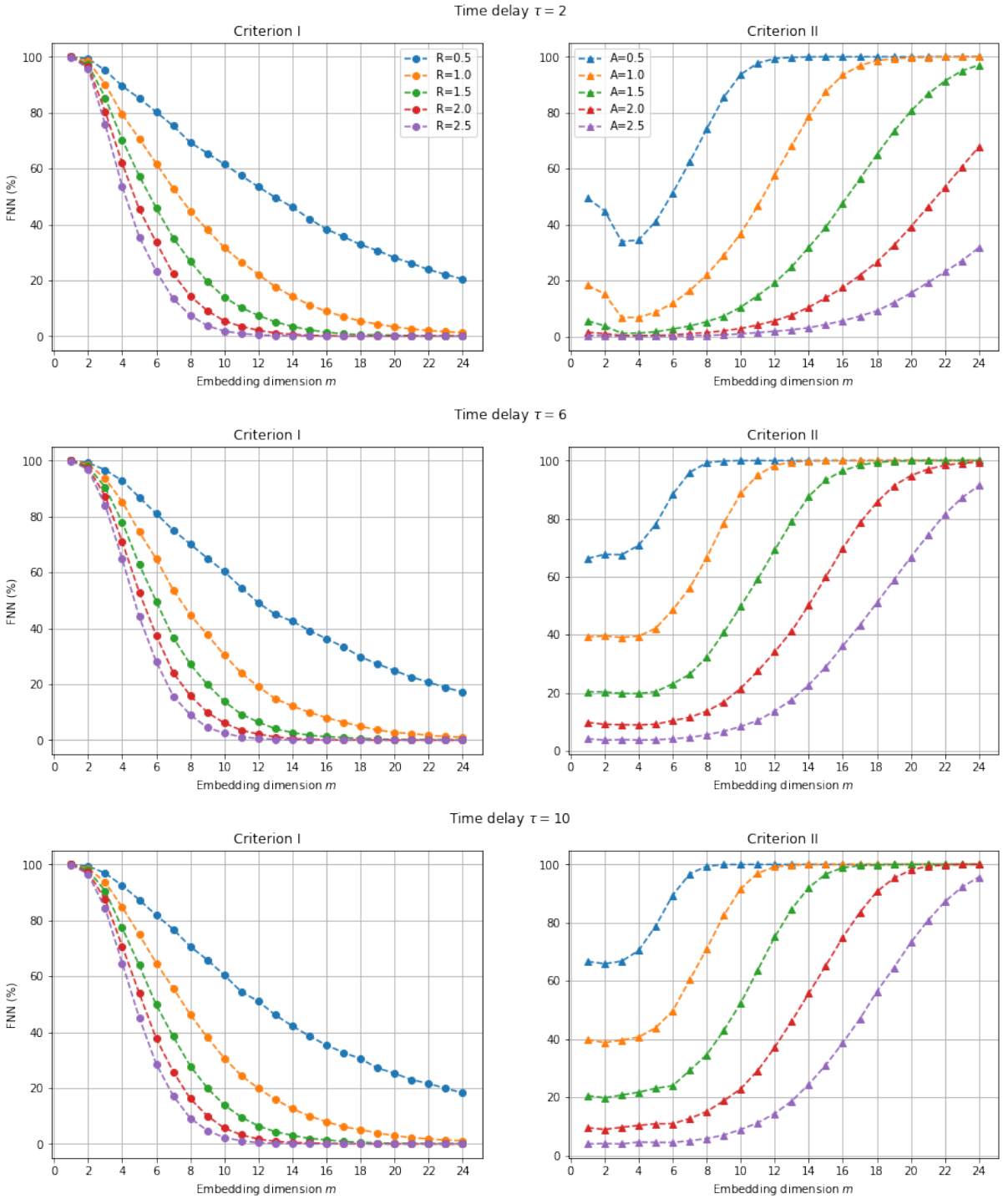


Figure 2.11: The effect of values of the tolerance parameters on the percentage of false neighbors reported by I. criterion (1.6) and II. criterion (1.7), Theiler window $w_t = 50$.

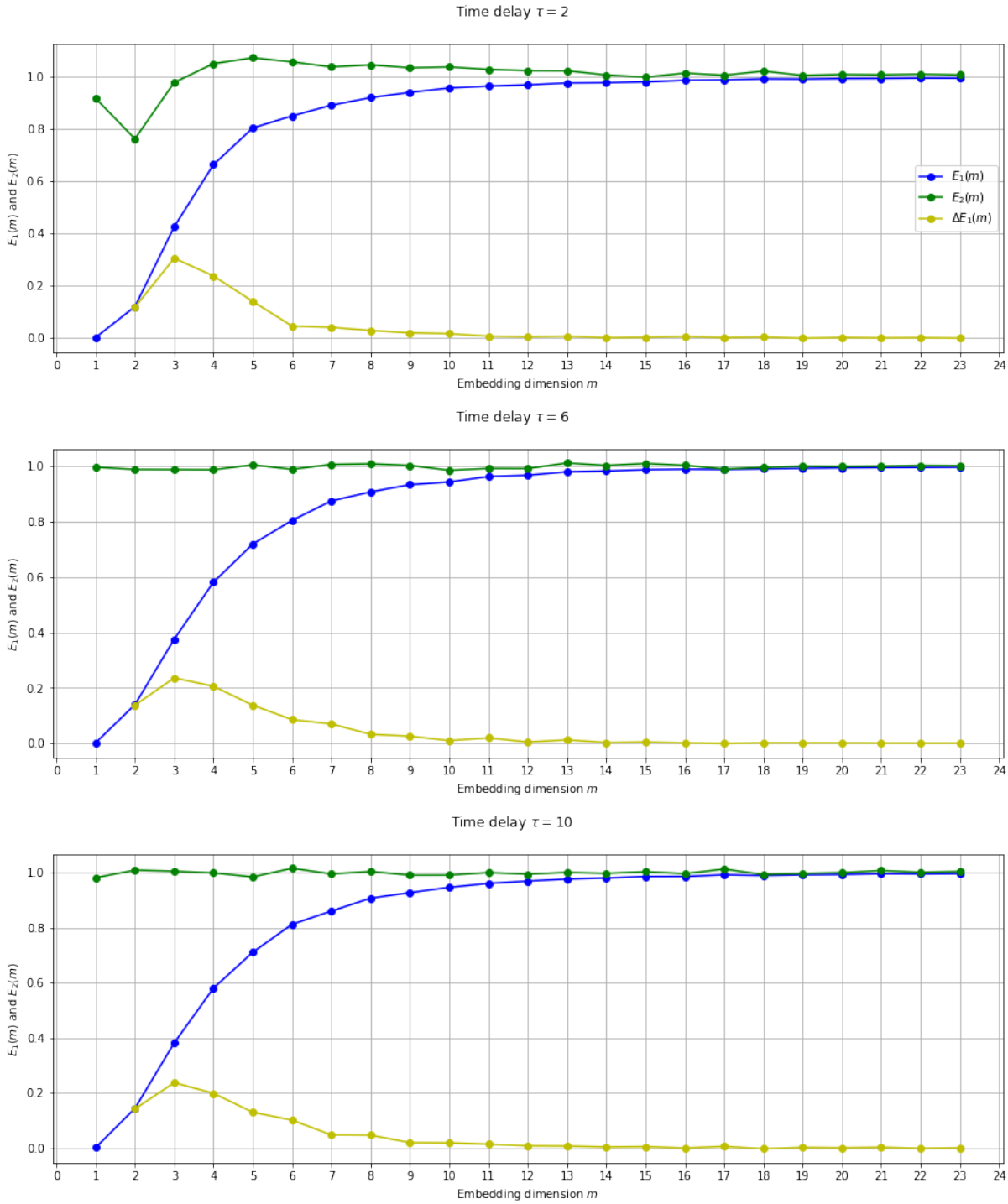


Figure 2.12: The results of AFN for varying values of time delay τ, Theiler window $w_t = 50$.

diameter of the attractor, which is finite, since the attractor is bounded in the phase space. Therefore, selecting proper embedding dimension based on divergence plots involves compromising between those two effects. Moreover, notice that the length of the scaling region is approximately $m\tau$.

How to explain this?

With increasing time delay τ, we observe gradually damped oscillation-like behavior with period τ and amplitudes also increasing with τ. Average divergence we computed using Kantz' algorithm also exhibits this behavior. Oscillation-like behavior was observed for white noise data in [73], and for periodic data with period equal to the dominant period of the system in [34]. One possible explanation is as follows [74]. Let $x_1, x_2, \ldots, x_N$ represent sampled time series, and $y_i \in \mathbb{R}^m$ an embedded point in the reconstructed orbit. Then

$$\begin{aligned} y_i &= \begin{pmatrix} x_i & x_{i+\tau} & \ldots & x_{i+(m-2)\tau} & x_{i+(m-1)\tau} \end{pmatrix} \\ y_{i+\tau} &= \begin{pmatrix} x_{i+\tau} & x_{i+2\tau} & \ldots & x_{i+(m-1)\tau} & \mu_1 \end{pmatrix} \\ &\ldots \\ y_{i+(m-1)\tau} &= \begin{pmatrix} x_{i+(m-1)\tau} & \mu_1 & \ldots & \mu_{m-2} & \mu_{m-1} \end{pmatrix}. \end{aligned}$$

This means that for a given y_i, possible values of $y_{i+\tau}$ are restricted to a line parallel to the direction of the m-th basis vector. Analogously, $y_{i+2\tau}, y_{i+3\tau}, \ldots, y_{i+(m-1)\tau}$ are restricted to $2, 3, \ldots, m-1$ dimensional hyperplanes in the m dimensional embedding space. As explained in Section 1.4.1.1, Rosenstein's algorithm finds pairs of vectors y_i and $y_{n(i,m)}$ with certain properties, and computes the evolution of their distances over time. This means that the possible values of $y_{i+\tau}$ and $y_{n(i,m)+\tau}$ are restricted to lie on two hyperplanes parallel to the m-th basis vector. Therefore, if $d_i(0) = \|y_i - y_{n(i,m)}\|$ is their initial distance, then the maximum possible distance after evolution by τ timesteps is $\|y_{i+\tau} - y_{k+\tau}\| = \sqrt{(d_i(0))^2 + A_m^2}$, where A_m is the maximum amplitude $\max_{i \in N(m,\tau)} |x_i|$. However, generally, the maximum distance is $A_m \sqrt{m}$. Thus, we may expect the average distances fluctuate with period τ.

There are several methods for mitigating this effect, but it cannot be evaded completely, since it is a join property of the time delay embedding and the data [74]. One may choose smaller τ, choose the evolution time $t_e < \tau$ or $t_e \geq \tau m$. Alternatively, one may choose different time delays τ_i for individual vector coordinates. The benefit of chossing smaller time delay or bounding the evolution time is that these choices still enable using hardware acceleration provided by vectorized operations in the implementation of the algorithm. Lastly, some algorithms, such as a modification of Wolf's algorithm [12], attempt to minimize the effect implicitly.

Can this occur due to measurement projection? Also, even if the largest Lyapunov exponent is positive, in dissipative systems (i.e. those possessing an attractor, see Section 1.2.2) the sum of all Lyapunov exponents is negative, and thus, even on average, states will diverge in some directions. These effects can be compensated for by using proper averaging statistics [34].

2.3.3.2 Automatic Selection Procedure

To estimate the LLE values using automatic selection of proper embedding parameters, we proceeded as follows. Selection of time delay was done using autocorrelation function with threshold $1 - 1/e$. Results ranged from 2 to 5, depending on the channel. The selected τ was used to compute the embedding dimension with smallest FNN percentage from embedding dimensions in range from 1 to 20, i.e. $m_1 = \arg\min_{m' \in \{1,\ldots,20\}} \mathrm{FNN}(m')$. The tolerance parameters were $R = 2.5$, $A = 2.0$. Moreover, we found the first embedding dimension m_2 for which $E_1(m_2) - E_1(m_2 - 1) < 0.008$. The estimates m_1 and m_2 computed in this manner were usually similar, and ranging from 8 to 11, depending on the channel. The final embedding dimension m was selected as their average $m = \lceil m_1 + m_2)/2 \rceil$. The length of the scaling region $t_e = m\tau$ and the Theiler window, as mentioned, $w_t = 50$.

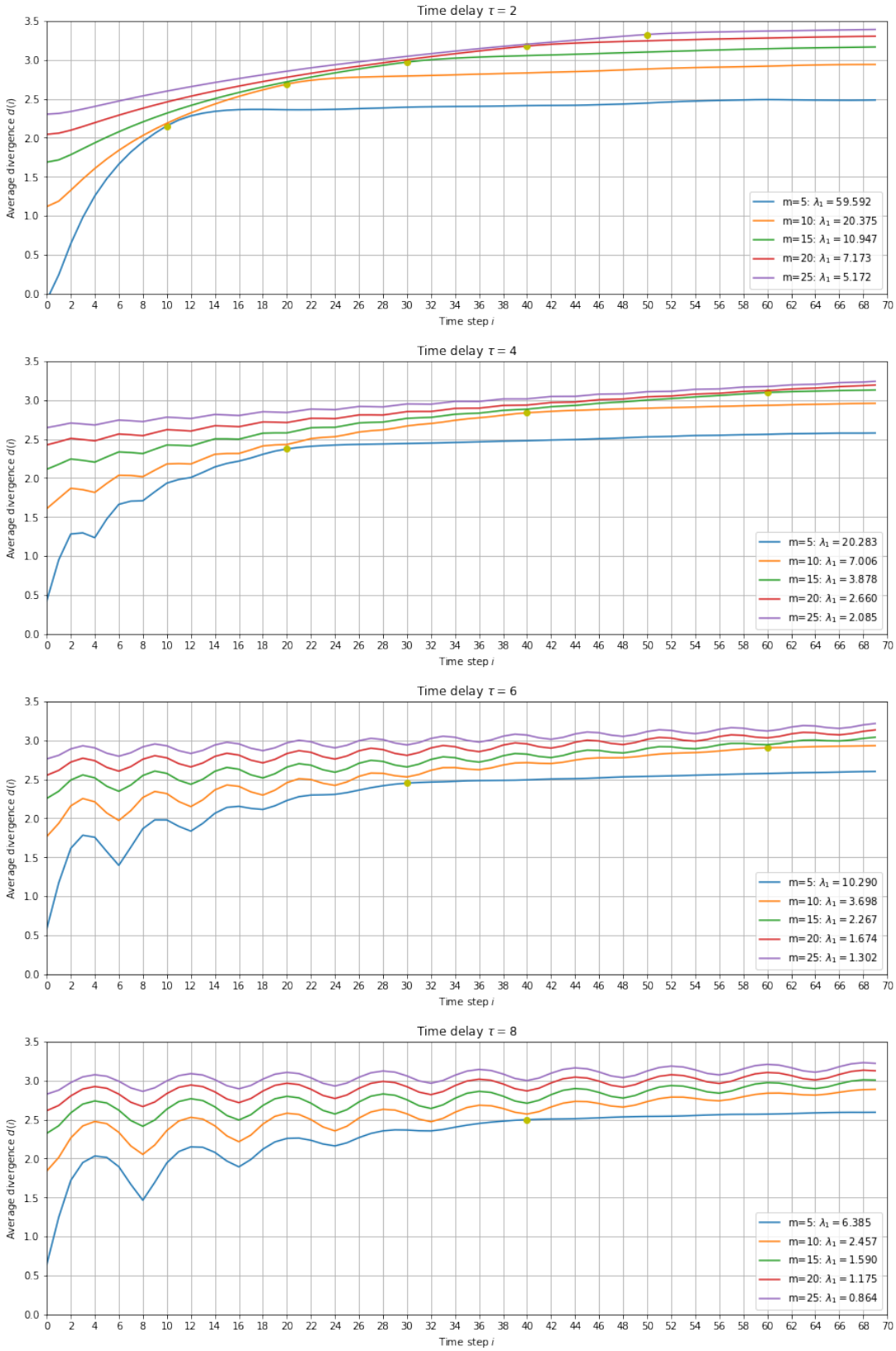


Figure 2.13: Average divergence plots for varying values of m and τ.

m	Method of selection of m	τ	Application	Reference
5	D_2 saturation	-	epileptic seizures	[127]
7	D_2 saturation	5	stationarity estimation	[128]
6-8	D_2 saturation	DMI	measuring variance in D_2 during sleep	[129]
7	-	10	measuring regularity during ECT seizures	[130]
10	-	3	emotion recognition	[133]
10	-	zero-crossing of $A(\tau)$	sleep in schizophrenia	[72]
10	-	zero-crossing of $A(\tau)$	sleep	[131]
10	-	random	depression	[12]
10	-	3	depression	-
11-15	FNN, AFN average	$1 - 1/e$ crossing of $A(\tau)$	depression	-

Table 2.2: Embedding dimension m and time delay τ choices found across available literature, along with methods of their selection and particular use case. All the studies used sampling frequency in range $250 - 256$ Hz. The range results we obtained using various algorithms and observed in literature highlights the fact that the embedding parameters selection algorithms, as well as visual inspection of the divergence plots largely are unreliable, and that the optimal values depend on particular dataset and problem.

2.3.3.3 Literature Review

Using the correlation dimension saturation, the following estimates have been obtained: 5 [127], 7 [128], 6-8 (depending on sleep stage) [129]. We found similar estimates using correlation dimension maximum (as mentioned in Section 2.3.2.2, we observed no saturation) with time delay $\tau = 1$.

By analyzing records ECT seizures, manually separating them into more and less regular, and inspecting multiple values of LLE as a measure of separation between these classes, 7 was selected as optimal embedding dimension in [130].

Remaining studies we evaluated, including some analyzing depression, used embedding dimension 10 [12, 131, 132]. Especially relevant is [12], where depression data were analyzed, including the dependence of LLE on the embedding dimension. The authors decided to use various values of time delay depending on the embedding dimension corrdinate (for rationale, see Section 2.3.3.1). Emotion recognition using LLE from EEG signals was performed in [133] with embedding dimension $m = 10$ and time delay $\tau = 3$. No reasoning behind this choice was provided. Summary of the reviewed studies can be seen in Table 2.2.

In summary, we obtained a wide range of results using traditional parameter selection techniques, with the most powerful algorithm, ILD, indicating sligthly lower values of time delay. Most of the embedding dimension estimation algorithms, including ILD, agree with the literature that the optimal embedding embedding dimension should be set around 10. We proceeded with computing multiple sets of LLE labels for the dataset, each with a different member of the following set of input parameters (m, τ): $(7, 3)$, $(7, 6)$, $(10, 3)$, $(10, 6)$, $(15, 4)$, automatic (as described in Section 2.3.3.2). Then, we analyzed the label distributions between studied groups, as presented in Section 2.4, where we present results only for the most discriminative pairs of parameters, which was $m = 10$, $\tau = 3$.

2.3.4 Correlation Dimension

2.3.4.1 Manual Analysis

For computation of correlation dimension, we used Grassberger-Procaccia algorithm described in Section 1.4.2, using Chebyshev metric (as suggested in [15]), Theiler window $w_t = 50$, for values of r

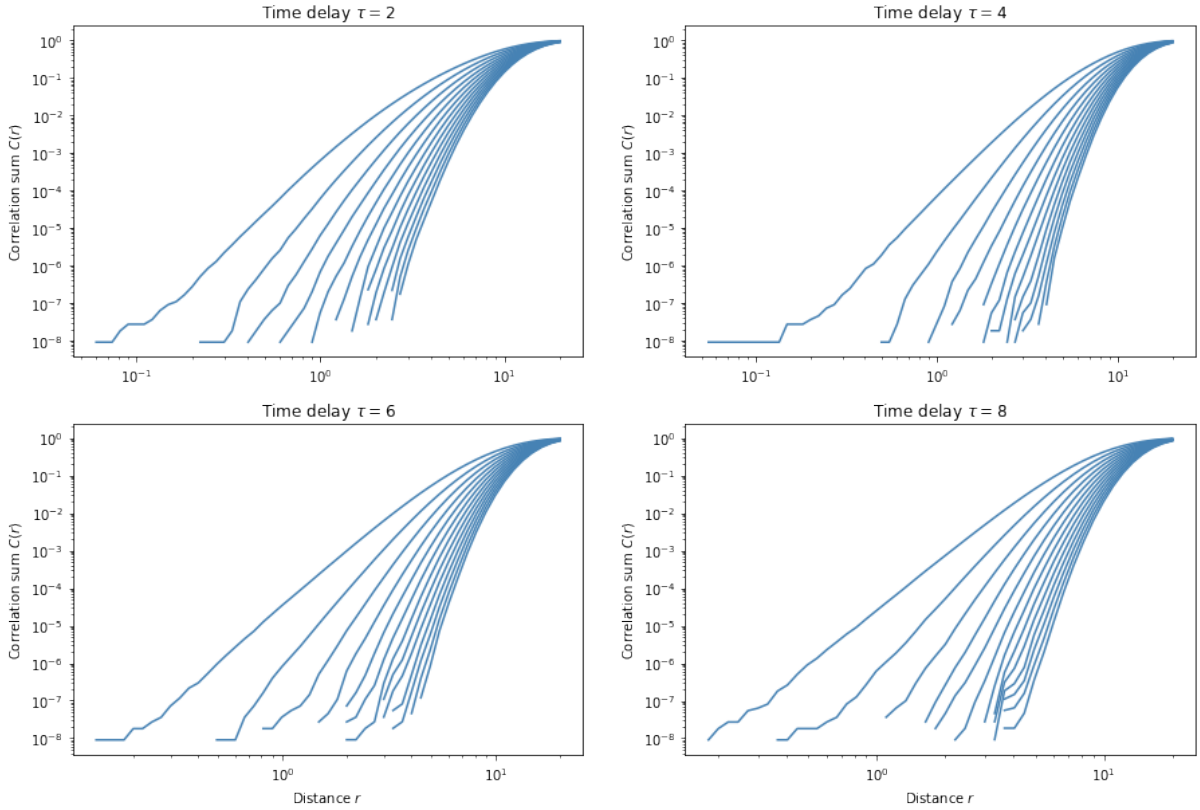


Figure 2.14: Normalized correlation sum $C(r)$ as a function of radius r for embedding dimensions $m = 5, 7, 9, \ldots, 29$ (lowest m at the bottom, highest at the top) and time delays $\tau = 2, 4, 6, 8$.

either in geometrical progression of 100 values from 0.05 to 10. Hereby we analyze the results for patient number 75, second session, FP1 electrode.

Plots of normalized correlation sums $C(r)$ as functions of radius r (both axis are logarithmically scaled) for embedding dimensions $m = 5, 7, 11, \ldots, 29$ can be seen in Figure 2.14. Different time delay τ has been used for each plot. The reader may want to compare these results with Figure 1.7. There are clear straight lines indicating expected relationship $C(r) \propto r^{D_2}$. Also, as expected, we can see that the lines shift to the right, increasing their slopes with m. The effect of time delay is noticeably weaker than in LLE estimation.

Figure 2.15 shows the local slope of $\log C(r)$ as a function of r (axis logarithmically scaled). The local slope has been approximated for each value of r_i by considering its 6 neighbors $\log r_{i-3}$, $\log r_{i+3}$, and fitting a line through the seven points $(\log r_{i-3}, \log C(r_{i-3})), \ldots, (\log r_{i+3}, \log C(r_{i+3})$ and minimizing least squares error. In contrast with the ideal case presented in Figure 1.8, there are no apparent scaling regions at all, which means we cannot provide theoretically meaningful finite estimate of D_2 using this method. Morever, by comparing with the same plot for iAAFT surrogate of the same time series (see Figure 2.16), we cannot even reject the hypothesis of a linear stochastic process. These results are not unique for this sample - we obtained similar results for all other examined samples.

On the other hand, as explained in Section 1.5, even rejecting the null hypothesis is not a sufficient proof of nonlinearity. In addition, this effect is known to happen due to noise, and many studies have failed to significantly distinguish EEG data from surrogates [15].

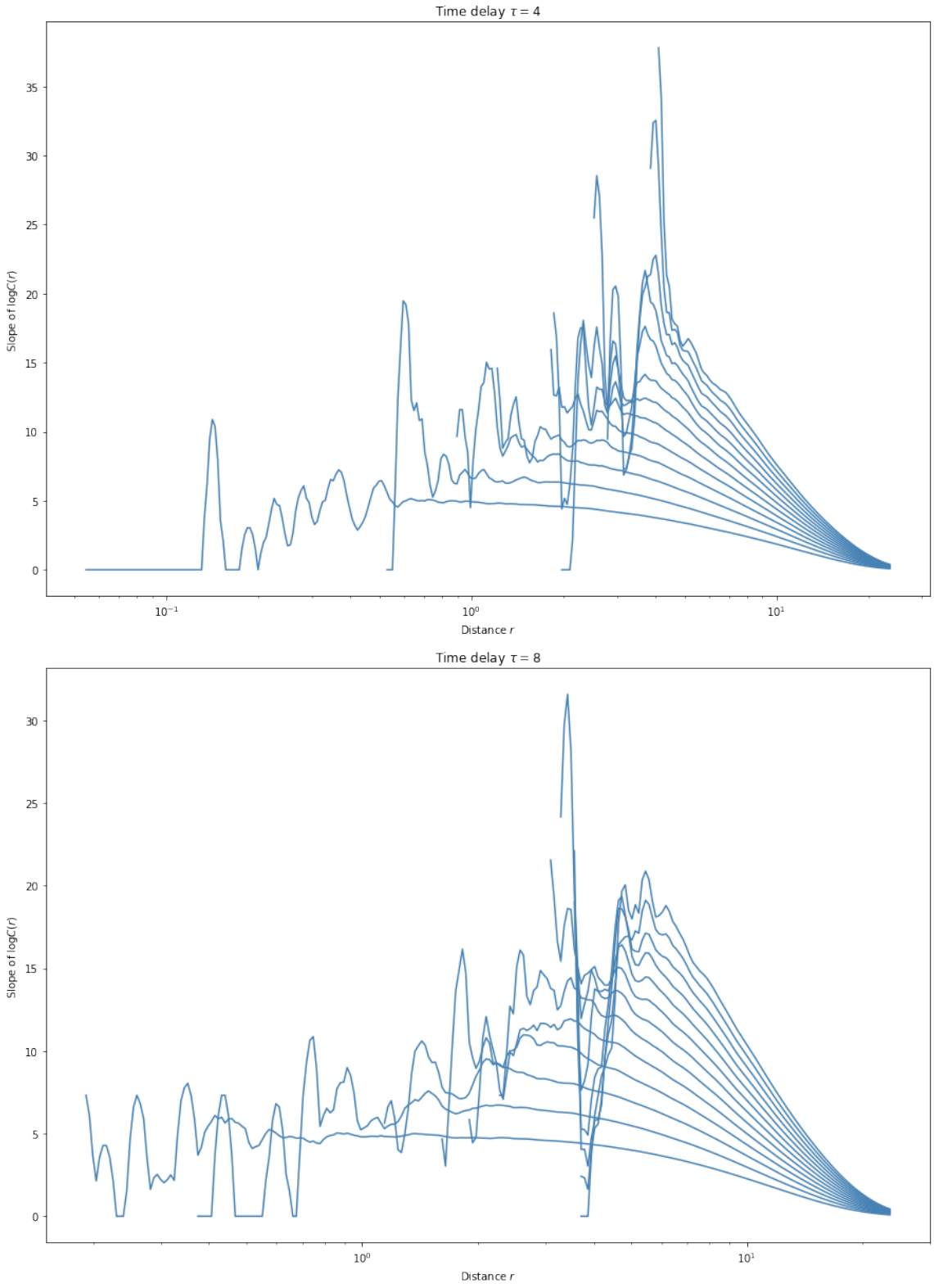


Figure 2.15: Local correlation dimension D_2 as a function of radius r for embedding dimensions $m = 5, 7, 9, \ldots, 29$ (lowest m at the bottom, highest at the top) and time delays $\tau = 4$ and $\tau = 8$.

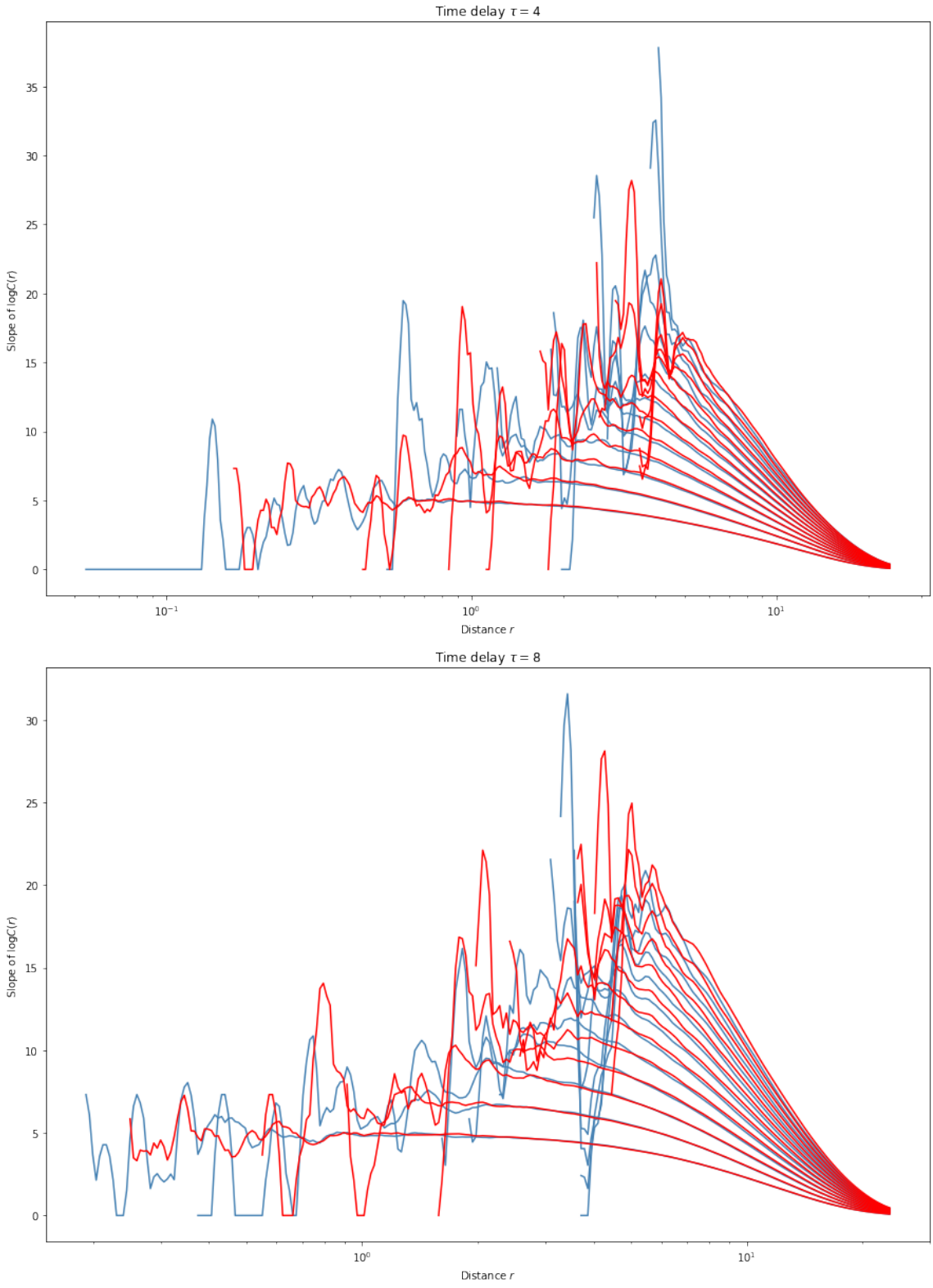


Figure 2.16: Local correlation dimension D_2 as a function of radius r for embedding dimensions $m = 5, 7, 9, \ldots, 29$ (lowest m at the bottom, highest at the top) and time delays $\tau = 4$ and $\tau = 8$ for the original series (blue) and its surrogate series computed using iAAFT (red).

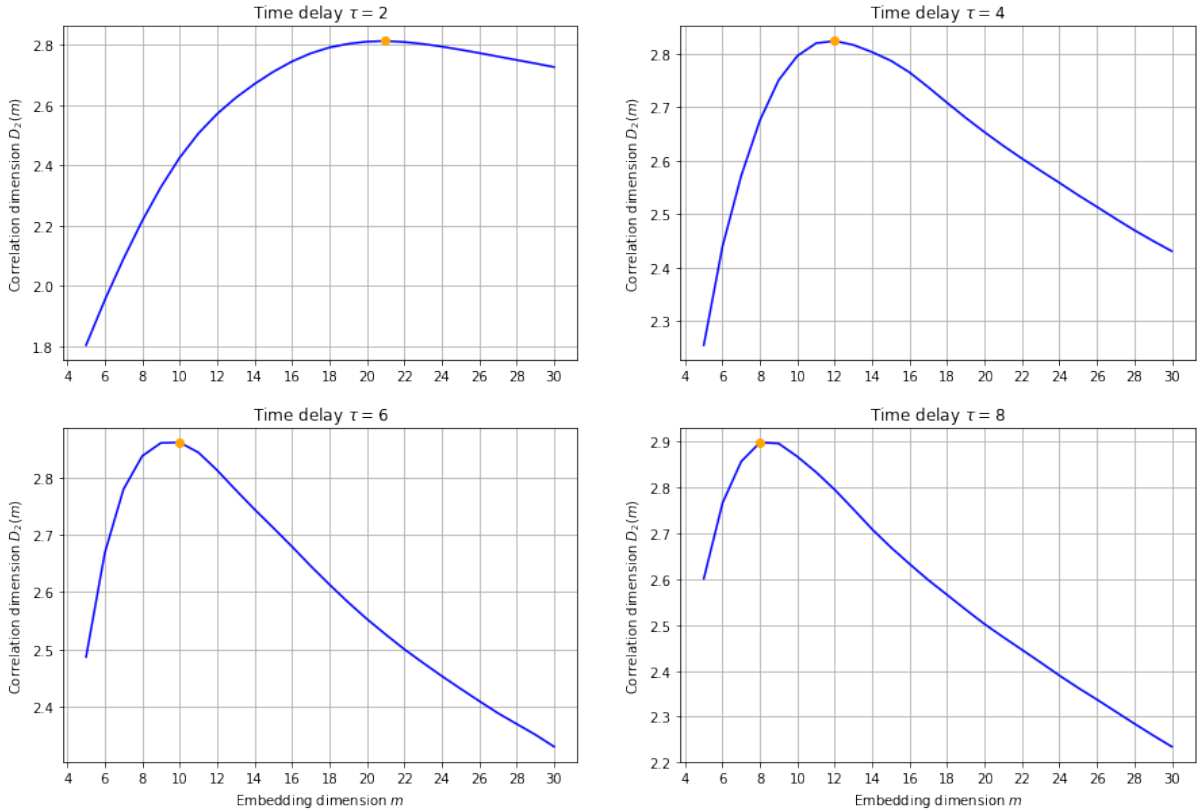


Figure 2.17: Correlation dimension as function of the embedding dimension m.

2.3.4.2 Automatic Selection Procedure

To compute correlation dimension automatically, we proceeded as follows. We create multiple embeddings with incresing embedding dimensions in range from 2 to 30 with the optimal time lag selected according to the autocorrelation function with threshold $1 - 1/e$. For each embedding, we evaluate the slope of $\log C(\log r)$ on the interval $[r_{\text{lower}}, r_{\text{upper}}]$, where r_{lower} corresponds to the average nearest neighbor distance on the reconstructed attractor, r_{upper} is given by

$$\log r_{\text{upper}} = \log r_{\text{lower}} + \frac{1}{10}\left(\log r_{\max} - \log r_{\text{lower}}\right),$$

where $r_{\max}$ denotes the largest ocurring pairwise distance on the attractor. This approach of automatic selection radius bounds for evaluation of D_2 is borrowed from [15]. Then, we computed the slope of the correlation integral $C(r)$ over this scaling region for each of those embeddings using the method explained in previous section, and select the global maximum as the final value.

Figure 2.17 shows D_2 computed this way as a function of the embedding dimension m for varying values of the embedding dimension τ. There are no signs of saturation, correlation dimension reaches a global maximum and then starts to decrease. This may be because of insufficient data. The reached maximum value of D_2, however, is reasonably consistent with results obtained by various methods in the literature [134, 135, 136].

m	Method of selection of m	τ	Application	Reference
3-30	saturation	DMI	depression	[10]
-	"unfolding dimension" (see text)	objective function minimization	intraindividual classification	[137]
2-20 (scaling region 16-20)	saturation	$A(\tau)$, threshold $1/e$	awake / sleep classification	[134]
12	-	8	mental state classification	[135]
2-12	saturation	5	schizophrenia	[136]
15	-	4	depression	-
2-30	saturation	$1 - 1/e$ crossing of $A(\tau)$	depression	-

Table 2.3: Embedding dimension m and time delay τ choices found across available literature, along with methods of their selection and particular use case. All the studies used sampling frequency in range 200-256 Hz.

2.3.4.3 Literature Review

In [137], the authors layed out fully automatic, albeit complicated, algorithm for selecting the embedding parameters. They suggest modelling correlation dimension biparametrically as a function of embedding dimension m as $D_2(m) = b_0(1 - e^{-b_1 m})$, where b_0, b_1 are parameters. They call $m^* := 1/b_1$ the unfolding dimension, since "it represents the embedding dimension at which the attractor has unfolded up to $1/e$ of its full extent, namely, the asymptotic D_2 value: b_0" [137]. The optimal embedding dimension m is selected as the next integer greater than the embedding dimension at which the exponential fit has approached 95% of its expontial value b_0. The remaining studies used either threshold crossing of the autocorrelation function $A(\tau)$ or fixed value for selection of time delay τ. Although [134] used $1/e$ as the threshold for the autocorrelation function, we found that the threshold $1 - 1/e$ results in better discrimination between studied groups using correlation dimension.

In summary, our experiments show lack of scaling regions in local D_2, indicating that no theoretically meaningful interpretation of correlation dimension is possible. Similar behavior is observed in surrogate data, hence even the null hypothesis of a linear stochatic process cannot be rejected. Moreover, our automatic procedure fails to saturate - however, this may be due to limited amount of data. Many studies reviewed in [15] failed to reveal finite embedding dimension on EEG data using the local slopes approach. The studies we examined all omitted the step of searching for optimal scaling region, and still succeeded in using D_2 for classification, simply by fitting the $C(r)$ curve. Table 2.3 shows a summary of of the studies we examined.

For these reasons, we decided to use the same approach (in addition to our automatic approach). We selected fixed pairs out of all combinations of the following values of $m = 10, 15, 20, 25$ and $\tau = 4, 8$ and computed correlation dimension for each sample with each pairs of embedding parameters by fitting the $\log C(r)$ against $\log r$ curve with a line using least squares fo values of r geometrically progressing from 0.05 to 10. The most discriminative pairs of parameters were $m = 15$, $\tau = 4$, in accordance with our ILD algorithm. Moreover, we also computed correlation dimension for each sample using our automatic approach described above. For both, we used Theiler window $w_t = 50$.

Another approach we evaluated, following the suggestion in [15], is to cut each sample into a number of small (overlapping) "moving windows", and perform the steps of estimating embedding parameters and computing correlation dimension for each of those windows. As we noted in 1.4.2.2, increasing the time series length is theoretically assumed to improve the estimates. On the other hand, shortening the time series may ameliorate the issue of apparent non-stationarity of EEG signal we observed in Section 2.3.4.1, since for a short time intervals, the signal may be assumed approximately nonstationary [15].

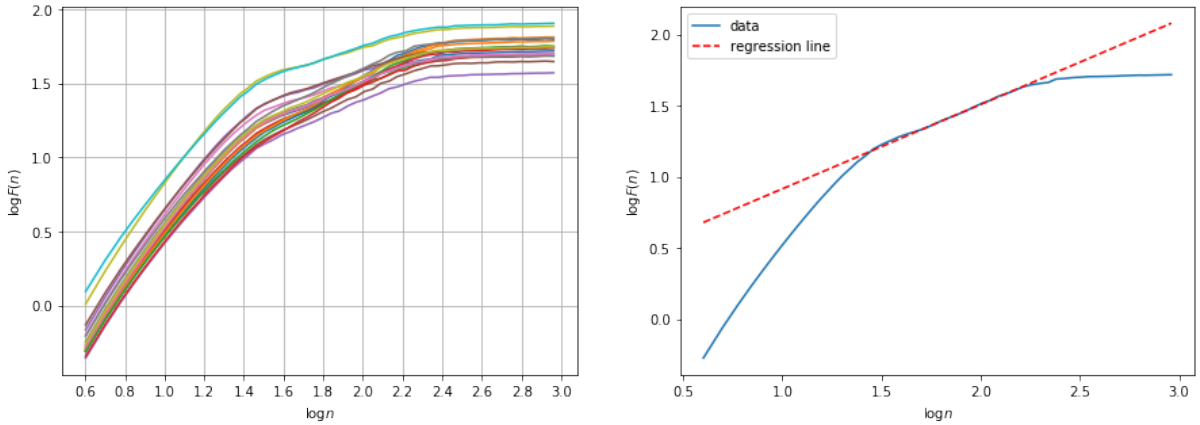


Figure 2.18: These figures show logarithm of root mean squared errors $\log F(n)$ (see equation (1.13)) plotted against logarithm of the segment length n. On the left hand side, each curve corresponds to one electrode for patient 75, second session. We can observe clear common scaling region for values of $n \in (50, 320)$. On the right hand side, an example regression line fitted to the scaling region of one curve is shown.

2.3.5 Detrended Fluctuation Analysis

The concept of Detrended Fluctuation Analysis (DFA) was described in Section 1.4.3. To compute it, we used the method of fitting a scaling region on curve of log-fluctuations $\log F(n)$ against logorarithm of segment lengths $\log n$ (see (1.13)), as explained in Section 1.4.3. Following the suggestions in [79], we defined the set of segment lengths n to be spread equidistantly on a logarithmic scale (by multiplying by factor 1.1), with the lower bound of 4 (fitting a line through less points may be prone to error). The upper bound was gradually increased, until a scaling region appeared on the curve of $\log F(n)$ against $\log n$ (see Figure 2.18). This curve was then plotted across multiple channels and patients, to determine the interval where the scaling region usually appeared, and if an automatic procedure for finding a scaling region is necessary. However, the interval was approximately same for all samples, and appear for values of $n \in (50, 320)$.

Moreover, following the suggestions in [79], we evaluated DFA computed from the envelope of the signal band-pass filtered to each one of α (8-13 Hz), β (16-37 Hz), and θ (4-7 Hz) frequencies, which were all found to be associated with depression [13, 9, 81]. The bounds of the scaling region were modified using the method above to $r \in (4, 100)$. However, for DFA estimates obtained using this method, we obtained no significant differences between the studied groups using the tests performed in Section 2.4.

2.3.6 Hurst Exponent

Hurst Exponent (HE) was described in Section 1.4.3. As far as we know, the literature regarding HE estimation omits description of the chosen values n for which to calculate the scaled range $(R/S)_n$. However, we may expect that higher values of subinterval lengths n will result in smaller number of subsequences d, and thus less precise estimate of the mean $(R/S)_n$ in (1.15). This effect will also be increasingly pronounced with decreasing original time series length N. Indeed, in Figure 2.19, which shows the mean rescaled ranges $(R/S)_n$ as a function of the subsequence length n (both axes are logarithmically scaled) for all electrodes, we may observe larger fluctuations for larger values of n, likely because of increasing uncertainty due to this effect. It has been observed that small values of n also lead to large deviations from the linear slope [84], because the relationship (1.14) is asymptotic, and thus valid

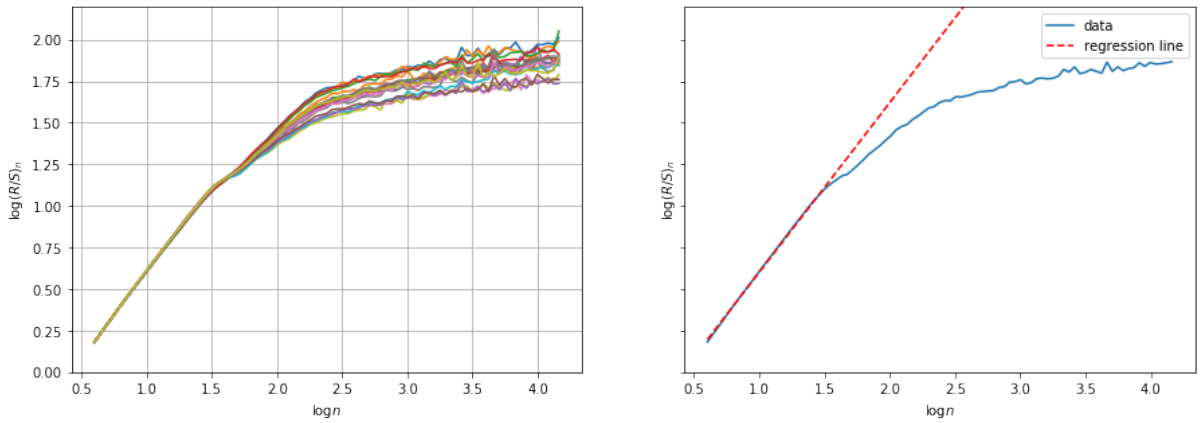


Figure 2.19: These figures show logarithm of the scaled range (see equation (1.16)) plotted against logarithm of the segment length n. On the left hand side, each curve corresponds to one electrode for patient 75, second session. We can observe clear common scaling region for values of $n \in (0, 20)$. On the right hand side, an example regression line fitted to the scaling region of one curve is shown.

$\mathbf{k}_{\max}$	**Sampling Frequency** (Hz)	**Recording Time**	**Application**	**Reference**
30	256	5 min	depression	[10]
50	256	5 s	depression	[9]
48	169.55	5 min	depression	[138]

Table 2.4: Summary of choices for the $k_{\max}$ parameter for Higuchi's fractal dimension applied to EEG we found across literature.

only for large n. Nonetheless, in our case, we have not encountered this phenomenon. Therefore, we have decided to compute the Hurst exponent as the slope of the log R/S curve for the values if $n \in (0, 20)$ spaced equidistantly on the logarithmic scale.

2.3.7 Higuchi's Fractal Dimension

The algorithm for estimating Higuchi's fractal Dimension (HD), as described Section 1.4.5, requires selecting the values k for which $L(k)$ is to be computed. Since this can be done relatively fast, we do this by selecting values equidistant on logarithmic scale $k_1, k_2, \ldots, k_{\max}$, where $k_{\max}$ is the single input parameter. Indeed, by using this approach, we follow most of the studies we evaluated.

It has been suggested that the parameter $k_{\max}$ can be selected by plotting values of HD as a function of $k_{\max}$, and selecting the value of $k_{\max}$ where the values of HD plateau. Our results can be seen in Figure 2.20.

In [9], the authors computed HD for each patient as follows. They used 5 second sliding windows with 0.5 second shift and computed local HD for each of those windows using $k_{\max} = 50$, thus obtaining 591 values per recording. The final HD for each electrode was obtained by averaging all local HDs. We tried using the same approach. However, the labels obtained in this way were less discriminative in our study. Finally, we computed labels from the entire recording for values $k_{\max} = 7, 30, 50, 100, 200$, where $k_{\max} = 50$ was shown to be the most discriminative.

Maybe few more studies.

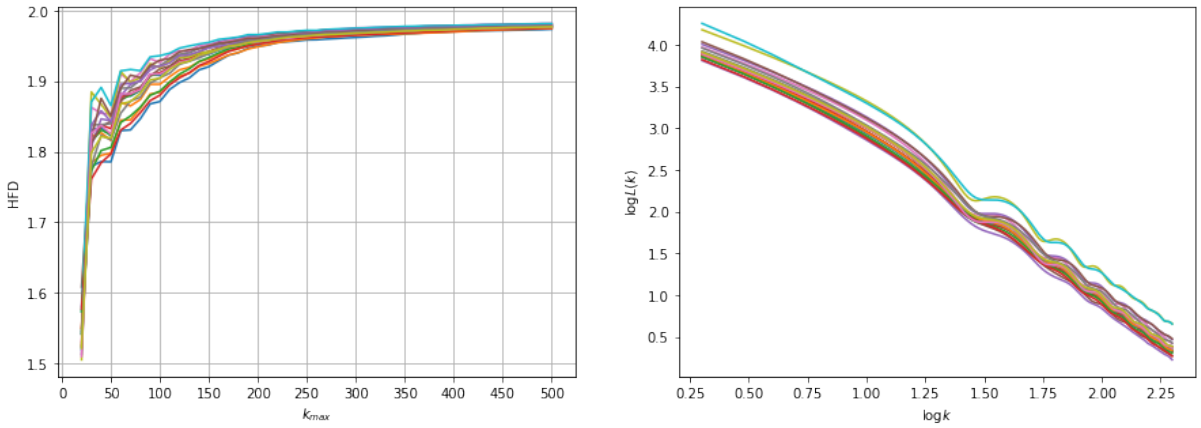


Figure 2.20: Figure on the left hand side shows values of Higuchi's fractal dimension as a function of $k_{\max}$ parameter for all channels of patient number 75, second session. We can see that HD values plateau for $k_{\max} > 200$ for all channels. On the other hand, variance among channels disappears. Figure on the right hand side shows logarithm of mean curve length $L(k)$ as a function of $\log k$. There is a clear linear trend with slight fluctuations for higher values of k.

2.3.8 Sample Entropy

The algorithm for computing sample entropy, described in Section 1.4.6, requires selection of the embedding dimension m and the tolerance parameter r. Embedding dimension m is usually set to 1 or 2 and tolerance parameter r is usually set at $0.1 * \text{std}(x)$ to $0.25 * \text{std}(x)$ [139]. As can be seen in Table 2.5, in all of the studies we evaluated, in spite of various applications and time series lengths, the same values were selected for m and τ, in particular, $m = 2$ and $r = 0.2 * \text{std}(x)$, where $\text{std}(x)$ denotes standard deviation of the original time series. In Figure 2.21, we can see why this is a reasonable choice. It shows histograms of pairwise distances in the embedding space for two successive embedding dimensions. With increasing embedding dimension, the histograms gradually overlap more, and for very high embedding dimensions, sample entropy becomes small for any r. Moreover, the histograms shift to the right, which requires higher value of r, but r is a measure of the degree to which vectors are judged "similar", and thus should be kept low - for $r \geq \max_{i \in \{1,2,\ldots,N\}} x_i$, all vectors will be counted in both embedding spaces and sample entropy is (approximately) zero.[2] For these reasons, we decided to follow the general concensus, and selected $m = 2$ and $r = 0.2 * \text{std}(x)$ [90, 7, 139, 140, 141, 142].

2.3.9 Frequency Band Amplitudes

Although our focus in this study is to examine the relationship between depression scores, treatment response and nonlinear measures, we also computed the mean frequency amplitudes in α, β, γ, δ and θ frequency bands using discrete fast Fourier transform by averaging the amplitudes corresponding to the frequencies in the respective bands. Figure 2.22 shows an example result for FP1 amplitude of a randomly selected patient. Then, we performed the analysis of differences between various subgroups described in Section 2.4. No significant differences between any two of the studied subgroups were found, and the mean amplitude values exhibited large variance between individual patients. Therefore, we decided to not pursue this approach of frequency band analysis further in this study.

[2]The lowest value sample entropy can reach is $-\ln(2/((N-m-1)(N-m)))$ [88].

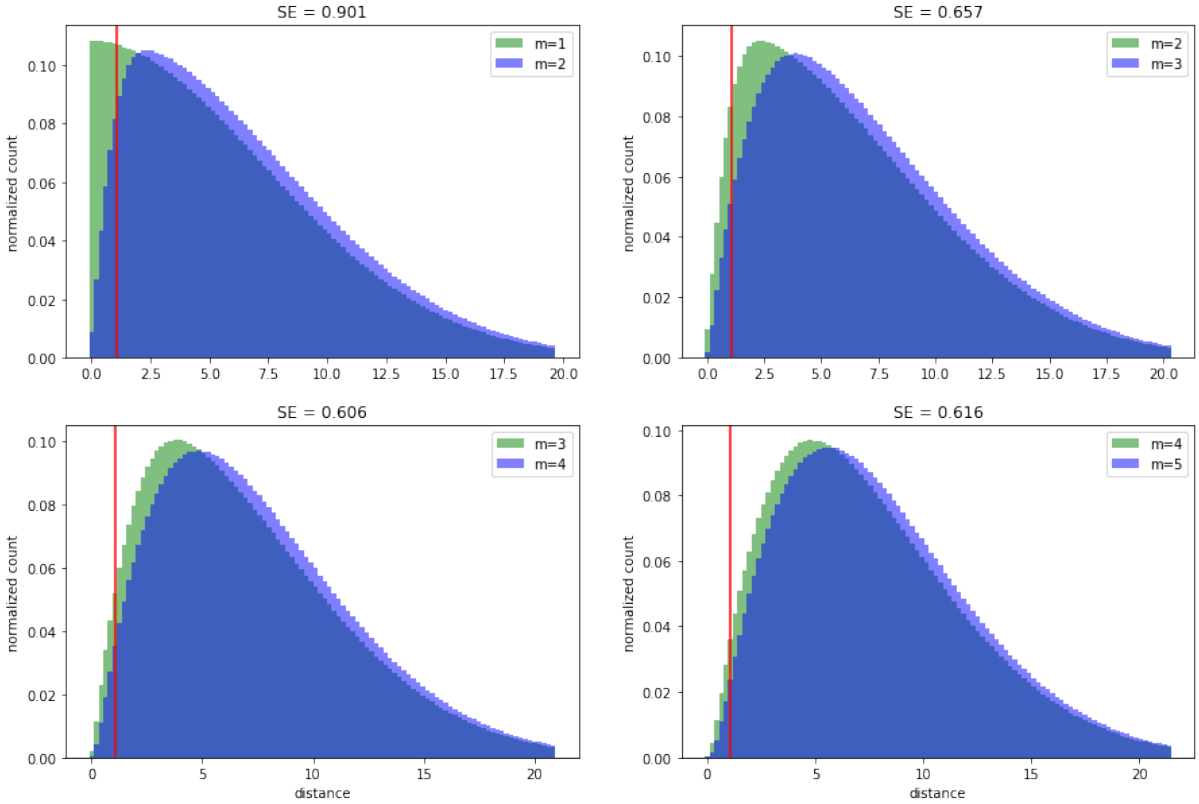


Figure 2.21: Histograms of pairwise distances between vectors in embedding space for gradually increasing embedding dimension m. The vertical red line indicates the value of the tolerance parameter $r = 0.2 * \text{std}(x)$. We can see histograms shifting to the right with increasing embedding dimension and progressively overlapping more. This means that for high embedding dimensions, sample entropy will be small for any value of tolerance parameter r.

r	m	Sampling Frequency	Recording Time	Application	Reference
$0.2 * \text{std}(x)$	2	-	256 min	depression	[90]
$0.2 * \text{std}(x)$	2	256 Hz	5 min	depression	[7]
$0.2 * \text{std}(x)$	2	512 Hz	63 s	emotion recognition	[139]
$0.2 * \text{std}(x)$	2	173.6 Hz	23.6 s	epilepsy	[140]
$0.2 * \text{std}(x)$	2	173.6 Hz	23.6 s	epilepsy	[141]
$0.2 * \text{std}(x)$	2	128 Hz	30 s	denoising	[142]
$0.2 * \text{std}(x)$	2	250 Hz	60 s	depression	-

Table 2.5: Summary of choices for the embedding dimension m and tolerance r parameters for sample entropy applied to EEG signals we found across literature. Note that all choices for m and r are the same.

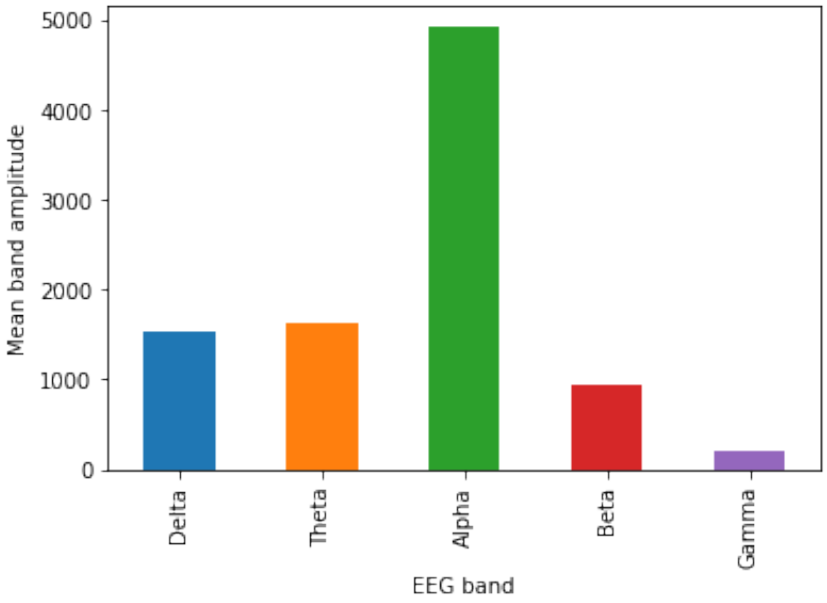


Figure 2.22: An example distribution of mean band amplitudes in channel FP1 of a randomly selected recording.

2.3.10 Summary of Parameters

In this section, we described our process for selecting the parameters used for computation of the nonlinear measures. In the rest of this chapter, all presented results were obtained using the values of nonlinear measures computed in manner described above. For the purpose of clarity, in Table 2.6, we include the summary of the selected parameters.

2.3.11 Surrogate Analysis

As mentioned in Section 1.5, surrogate analysis for high condifence levels requires generating surrogate dataset tens of times larger than the original dataset, and computing corresponding nonlinear measures for these surrogate signals. For instance, to achieve confidence level $\alpha = 95\%$ for our dataset,

Measure	Parameter	Value
LLE	m	10
	τ	3
	w_t	50
	t_e	30
CD	m	15
	τ	4
	w_t	50
	metric	Chebyshev
	r	GP on $(0.05, 10)$

Measure	Parameter	Value
DFA	n	GP on $(4, 320)$
	offset	50
	segment overlap	50%
HE	n	GP on $(0, 20)$
HD	$k_{\max}$	50
SE	m	2
	r	$0.2 * \text{std}(x)$

Table 2.6: Summary of parameters used for computation of the nonlinear measures. The abbrevation GP refers to geometric progression.

this translates to generating $266 * 19 * 38 = 192052$ surrogate samples and computing on them each nonlinear measure considered in our study. Since nonlinear algorithms considered (described in the preceding subsections) are relatively computationally expensive, performing surrogate analysis for all measures and all patients is computationally infeasible. Thus, we analyzed each algorithm (with varying parameters) on a single recording (patient number 75, second session), using 19 surrogate samples. For generating the surrogate data, we used the iAAFT algorithm described in Section 1.5.0.1. Note that to perform this as a test of nonlinearity, we have to assume that the choice of embedding parameters for the algorithms is correct.

An example of a result of such analysis for the largest Lyapunov exponent (embedding dimension 10, time delay 3) can be seen in Figure 2.23; the results for other measures were similar. First, we can observe that the distribution of the values computed for the surrogate data does not seem normal for all channels. As mentioned in Section 1.5, this increases the required value sigma to achieve the same confidence, or requires performing a rank based test. It can be easily observed, then, that based on the rank based test, the hypothesis of a linear stochastic process cannot be rejected on (admittedly relatively low) confidence level $\alpha = 1 - 2/(19 + 1) = 90\%$ for all channels except FP2, C4, T4, Pz. Obviously, does not necessarily imply that the process underlying corresponding time series is stochastic, because, for example, there still may be other nonlinear measures (or different choice of embedding parameters) which can discriminate between the original time series and the surrogate data. Neither does it suggest that the choice of embedding parameters is incorrect, because the process underlying corresponding time series may be stochastic. It is simply a failed attempt at disproving the null hypothesis of a stochastic linear process. Moreover, all our analyses concerned only a single patient.

2.3.12 Nonstationarity

To reduce the effects of possible nonstationarity, we attempted to find the most stationary window of the desired length using the stationarity test described in Section 1.2.3. However, we found that selecting the least stationary window using this test did not improve the results as measured by the surrogate data analysis in Section 2.3.11. This may be beacause the optimal effectiveness of AAFT and iAAFT the first and last point of the window should have the same value [15]. Moreover, selecting a different time window for each channel may result in inaccurate representation of the mental state by the vector of measures computed across channels. Therefore, it may be beneficial to use the same time window for all channels. If the results are dependent on the time during recording, then it would be advisable to select a fixed time window for all samples. Thus, we decided to skip this window selection step and pick a fixed time window for all channels and all recordings.

2.4 Analysis of Measure Distributions between Groups

2.4.1 Before and After Treatment Groups

As the first step of our analysis, we conducted an investigation of the differences in the nonlinear measures computed from the signals obtained before and after treatment. The purpose of this inquiry is to determine brain regions and measures affected by treatment. This is warranted by the fact that the patterns in EEG signals tend to be relatively stable over time. On the other hand, we realize the limitations of this attempt in the case of this study, since each patient recieved personalized method of treatment, and the methods may have differing impact.

This should be cited!

For each group, we performed two-sided Kruskal-Wallis test for the null hypothesis that the distributions of values computed for measurements before and after treatment are the same. No significant

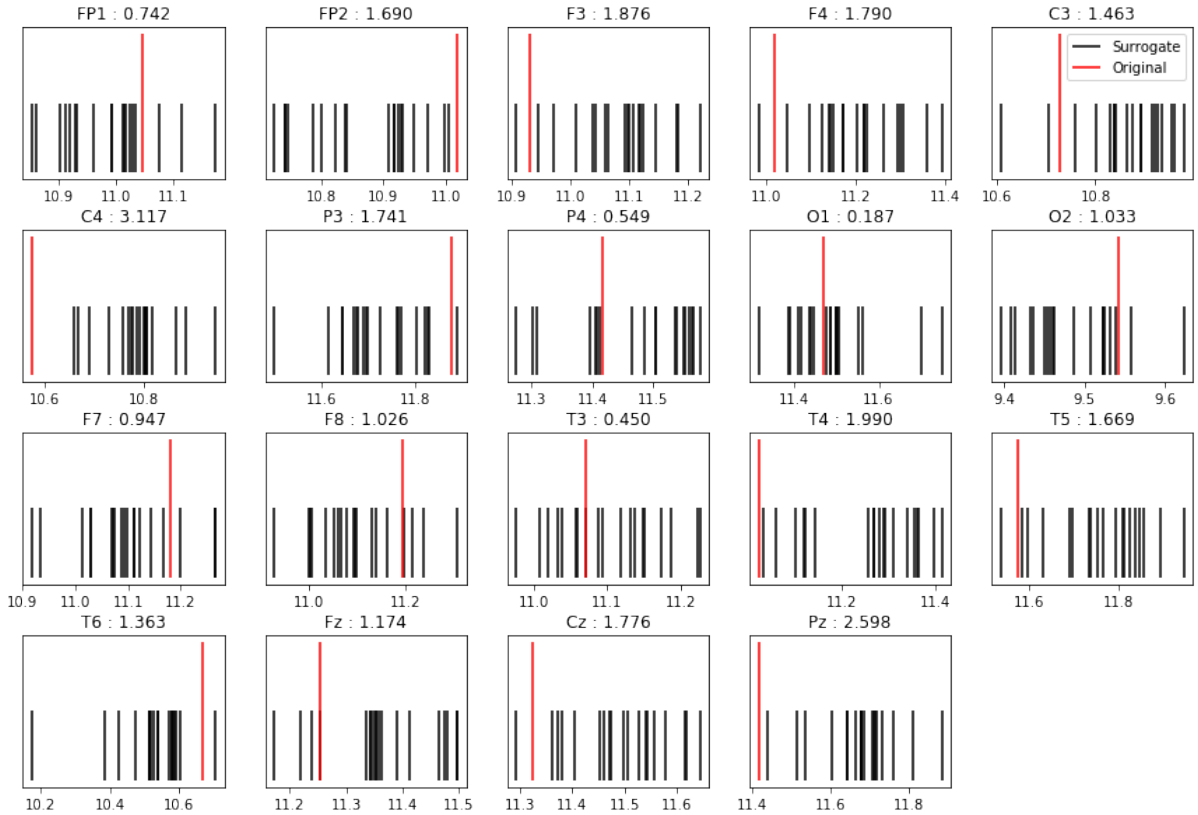


Figure 2.23: Example distribution of the largest Lyapunov exponent (embedding dimension 10, time delay 3) for 19 surrogate samples and the original for all channels. The number next to each channel name represents the confidence in sigma, computed as in equation 1.17.

differences in distributions were found for DFA, HE, and HD, D_2 and λ_1 computed using automatic selection of embedding parameters. Therefore, we include only results for the remaining measures. The results can be seen in Table 4 in the Appendix.[3] No significant differences between the groups are observed in DFA and HE values, and HD and CD show differences only in left temporal areas (T3, T4). On the other hand, LLE shows signficant differences in frontal (F3, F4), "central" (C3, C4) and left temporal (T3) areas, with similar, yet less signficant differences in those areas are observed for SE values.

Differences in the mean values of each measure in each electrode before and after treatment are visualized in Figure 12. Length of the error bars corresponds to two standard deviations. One can notice, for example, that values of LLE tend to be higher before treatment. As we will comment more in Section 2.4.3, this effect seems especially pronounced in the group responding positively to the treatment.

Finally, we analyzed the differences in the before and after treatment recordings in the groups specifically responding and nonresponding to the treatment in Section 2.4.2. Moreover, we performed unsupervised analysis of before and after groups using PCA in 2,3, and 4 dimensions, and compared centroids and mean distances between before and after treatment recording for each group. However, the resulting plots and heatmaps are featureless and thus we will leave them out. The mean distances are also uninformative.

One notable outcome of these analyses is how relatively stable nonlinear measures are across time for individual patients. In Figure 16, which shows relative differences between before and after treatment recordings for each patient and mean of each measure across channels, one can see that the relative changes for each measure remain comparatively small. The smallest relative change is observed for LLE and CD, whereas the largest are observed in DFA and SE. The intraindividual specificity of nonlinear measures, especially correlation dimension, has been noted in the literature, and has been suggested as a viable method of personal identification [137, 143, 144].

2.4.2 Low and High Depression Score Groups

As mentioned in Section 2.1, studied dataset lacks symptom absent group, which makes the task of creating a generally applicable classifier for depression diagnosis more difficult. The patients, however, still vary in severity of their symptoms, which allows us to study correlation between symptom severity (which may, in turn, inform the task of finding a classifier). To this goal, we explored the differences in distributions of computed nonlinear between groups of the "healthiest" and most depressed patients visually and using statistical tests, and in this section, we present some of the results.

With the goal of analyzing the differences between the lightest and most severe symptoms, we selected two classes of recordings for analysis in this section as follows. The first class, called *healthy*, of 50 recordings with reported depression score ≤ 16, and the second class, called *depressed*, of 50 recordings with depression score ≥ 28. We should recognize that including after treatment recordings does not control for possible effects of treatment not reflecting in the depression scores but reflecting in the signal, or the inverse. Indeed, all the healthy recordings were made after treatment, and most of the depressed recordings were made before treatment.

First, we inspected histograms of computed measures between the two groups. There were striking trends in the means of the two distributions in almost all channels for all measures except correlation dimension. Means of depressed recordings are typically shifted to the left of the mean of healthy recordings for all measures except for largest Lyapunov exponent, for which the means are shifted to the right. For correlation dimension, the distributions in both groups are similar. Figure 17 in the Appendix, which shows the distributions of the largest Lyapunov exponents for both groups, exemplifies the differences.

[3] In all tables, the p-value cutoffs for significance ratings are 0.05, 0.01, 0.005.

Next, we investigated at the correlations between the individual measures and correlation scores. Figure 2.24 shows visually clear negative correlation for DFA, and Figure 2.25 shows positive correlation for the largest Lyapunov exponent. Correlation dimension shows no significant correlations between depression scores on any channel, whereas the remaining measures are significantly correlated in almost all channels, and all those correlations are negative. Trends similar to the one observed for DFA were observed for Hurst exponent and Higuchi's fractal dimension, where, however, DFA shows slightly more significant correlations. Sample entropy values were significantly negatively correlated only on F4 and T6 channels. We also visualized correlations of individual measures to the the depression score and treatment response label on topographic scalp in Figure 2.26, and the corresponding p-values in Figure 2.27. Interestingly, the patterns in p-values for depression score correlations seem almost centrally symmetrical for DFA, HE, and CD, and centrally symmetrical for LLE and SE.

Finally, we compared the distributions using Kruskal-Wallis test. Table 5 in the Appendix show the results. The mean values across all channels are significantly different for DFA, LLE, SE, and HD. The differences in DFA seem to be the most significant over parietal and occipital regions, and right parietal region (P4) in particular. Although LLE also exhibits differences in those areas, in addition, it shows slight differences in frontal (F4) area, and it is the right temporal region (T6) which shows the most pronounced differences. Similar behavior of significant differences across frontal, occipital, parietal, and right temporal region may be observed in SE and HD.

In contrast to previously observed lower EEG complexity (higher predictability) in depression [11], we find find that depressed patients exhibit slightly higher higher LLE and CD, indicating higher chaoticity and complexity, which implies lower predictibility. However, we used different method from the mentioned study to reach these conclusions.

2.4.3 Low and High Response Groups

Neurocorrelates of remission, or, in other words, positive response to a treatment, are interesting apart from the neurocorrelates of depression itself. Instead of indicating whether a treatment should be prescribed in the first place, the effects of various drugs on the brain may help in designing more individualized treatments, or in developments of new drugs, even for other conditions. However, as noted in Section 2.1, in our dataset, different kinds of treatments (including rTMS) are mixed for most patients, thus make the task of distinguishing the singular causes of any observed changes challenging. Nevertheless, we may still attempt to find discrepancies between the responding and stagnant patients. If we assume that any prescribed treatment was beneficial, we may be able identify traits of patients who are difficult to treat. Indeed, medical literature recognizes entire categories of such patients [145].

For each patient, we computed a measure of change in depression scores (and, by extension, of effectivity of the treatment) we call *response* (RES). If DS_0 denotes the depression score measured on the first visit and DS_4 denotes the depression score measured on the second visit after 4 weeks, then response is computed as their ratio $RES = DS_4/DS_0$. Mean response is 2.47, mode 1.66, standard deviation 3.14. Most values range from 1 to 5, with a few outliers improving their symptoms 14 and 16-fold respectively. Only 9 patients stagnated exactly or slightly worsened their symptoms (response ≤ 1). Subsequently, we sampled 50 patients with the lowest value of response we call *nonresponding* (nonresponders) and 50 patients with the highest value of response we call *responding* (responders).

We performed Kruskal-Wallis test to see the differences in the computed measures between the two groups in individual channels, considering only the before treatment recordings. The most significant differences found were in Lyapunov exponent, especially in frontal, parietal, and right temporal areas. Aside from the largest Lyapunov exponent, Higuchi's fractal dimension then also showed significant differences in frontal areas.. The results are shown in Table 6 in the Appendix. In comparison with

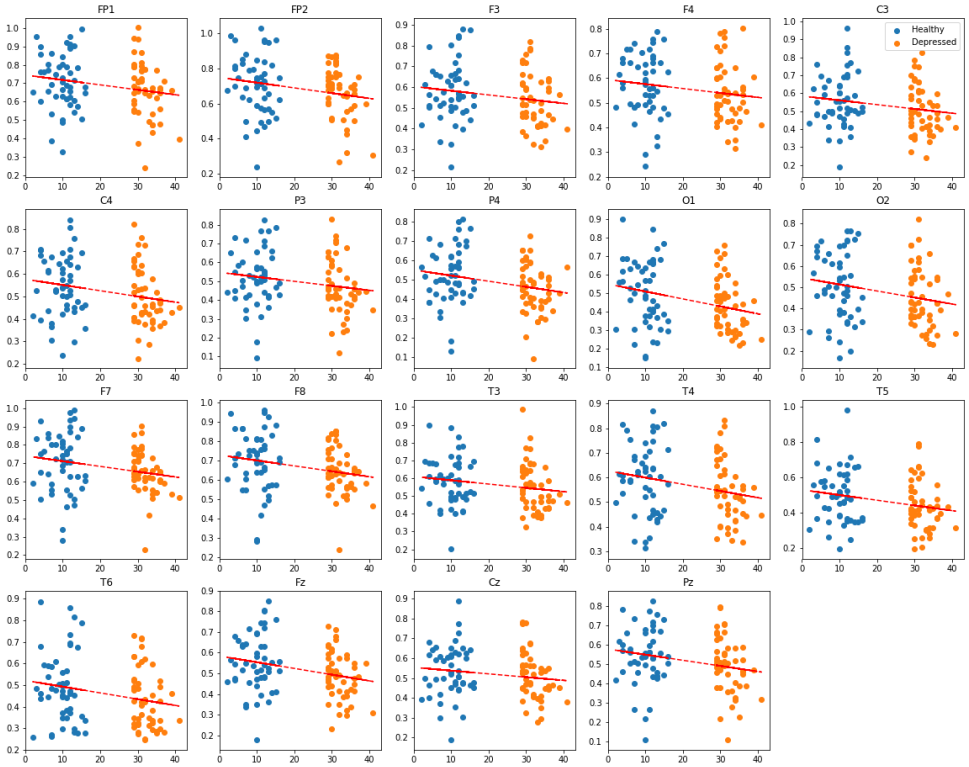


Figure 2.24: Trend of values of DFA as a function of depression score. The correlation is significantly ($p < 0.05$) negative for all channels with exception of F3, F4, C3, P3, T3, Cz. Similar trend is present in Hurst exponent, Higuchi's fractal dimension and sample entropy. No correlation of correlation dimension with depression score is significant in any channel.

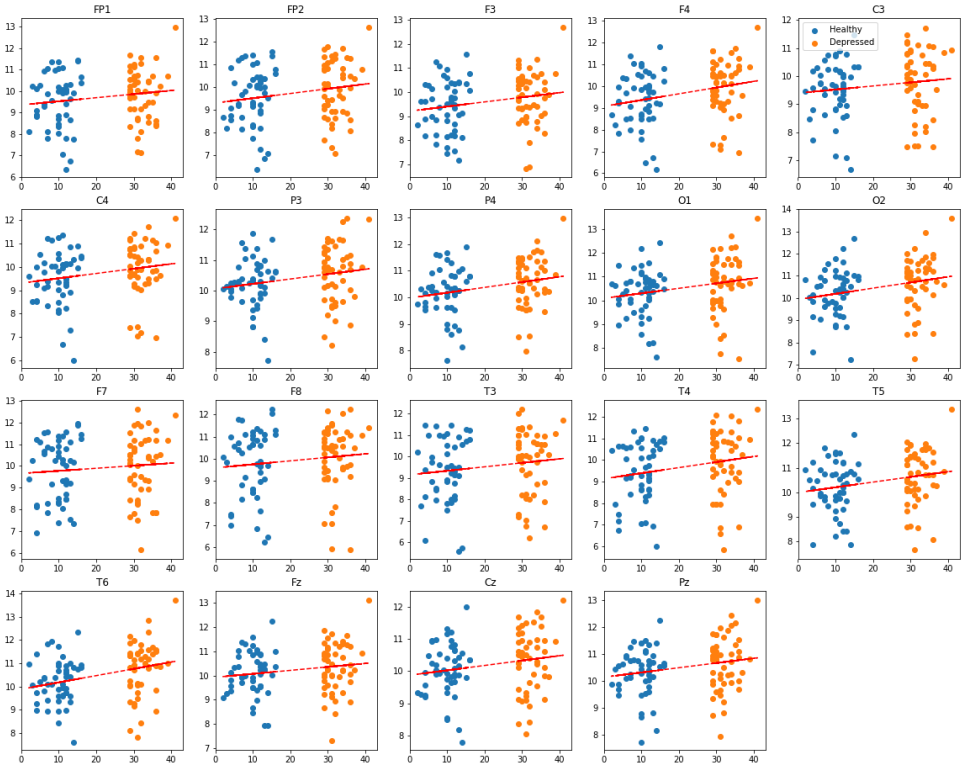


Figure 2.25: Trends of values of largest Lyapunov exponent as a function of depression score. The correlation is significantly ($p < 0.05$) positive for all channels with exception of FP1, FP2, C3, F7, F8, T3.

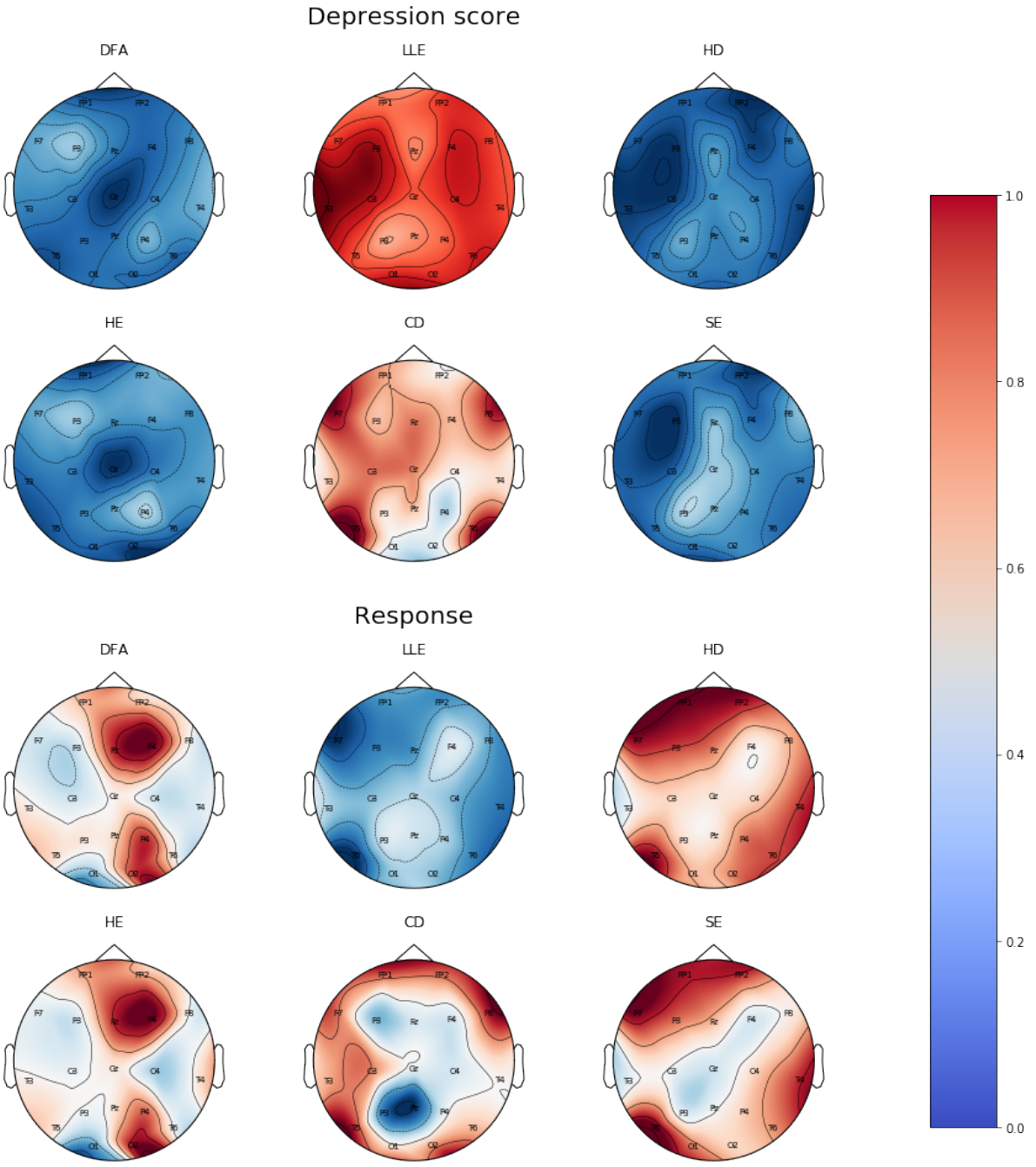


Figure 2.26: Topographic map of correlation coefficients on a scalp for depression diagnosis and response prognosis respectively. Color at each channel marks the value of Pearson's correlation coefficient between the value of corresponding measure and either depression score or treatment response label. The colors are then interpolated for smooth transitions. For depression score, all patients were included, whereas for response, only before treatment recordings were included.

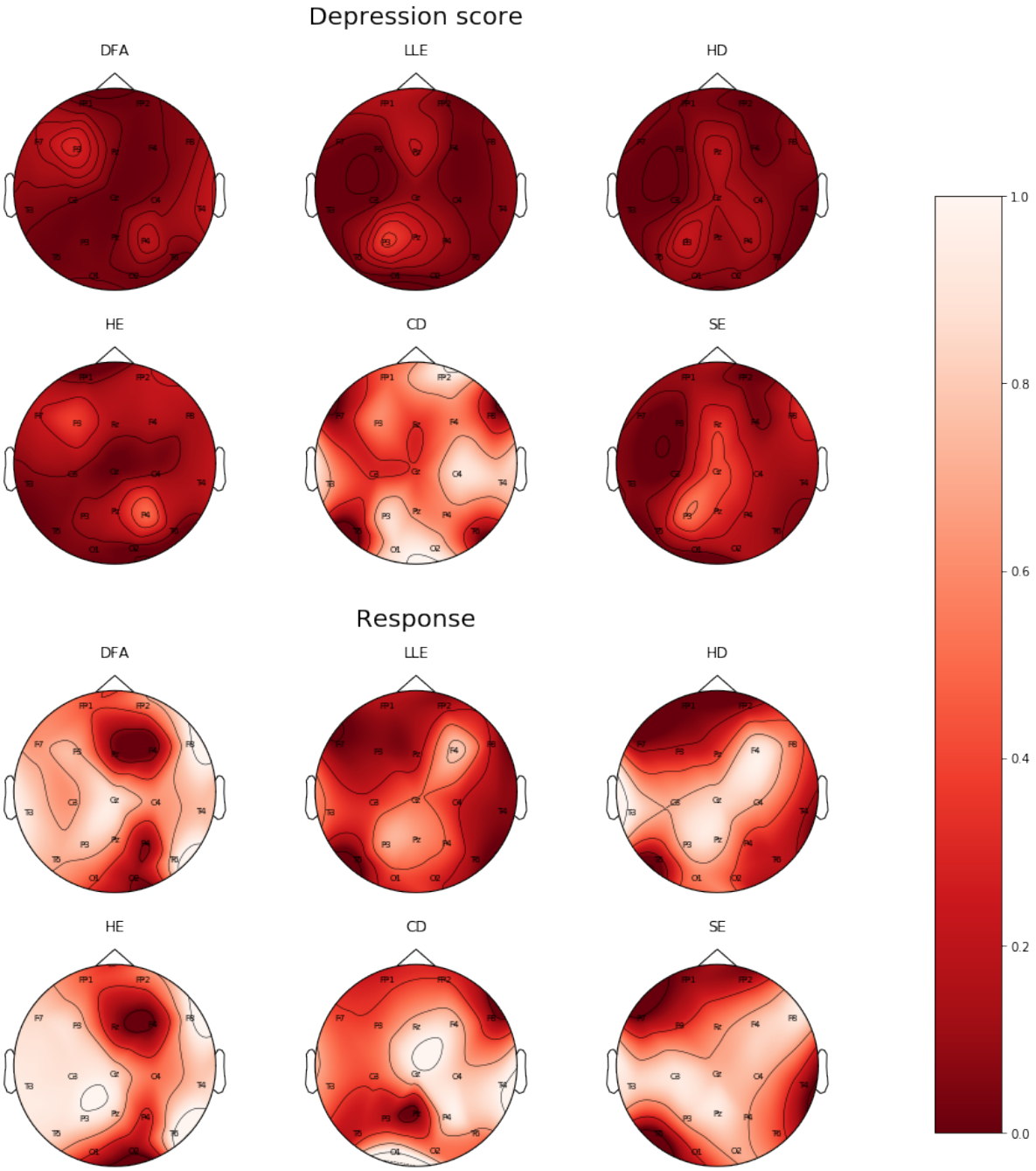


Figure 2.27: Topographic map of correlation coefficients' p-value on a scalp for depression diagnosis and response prognosis respectively. Color at each channel marks the value of Pearson's correlation coefficient's p-value between the value of corresponding measure and either depression score or treatment response label across all patients. The colors are then interpolated for smooth transitions. For depression score, all patients were included, whereas for response, only before treatment recordings were included.

Section 2.4.2, these results suggest that, with exception of LLE, differences in nonlinear measures between responders and nonresponders in the recordings obtained before treatment are considerably less significant than differences in depression score. On the other hand, it is possible that the differences are less significant due to smaller dataset size, since the test is performed on smaller number of samples. For completeness, we also decided to include comparison which includes both before and after treatment recordings, which can be seen in Table 7.

Moreover, we also compared before and after treatment recordings of responding, and before and after treatment recordings of nonresponding patients. Significant differences were found only in LLE, CD and SE, for which the results can be seen in the Appendix in Tables 8, 9, and 10 respectively. These differences are also visualized in Figures 13, 14, and 15 in the Appendix. We can see, for example, that the mean values of LLE across channels are higher before treatment for the responding group, but this effect is less pronounced for the nonresponding group, where depression score also slightly descreased. Since this group also received treatment, the observed higher depression score accompanied by LLE increase is not due to the effects of treatment. Together with Table 6, this result further supports the hypothesis that it is the decrease in depression score which results in higher values of LLE.

Of course, analyzing effectivity of treatment is difficult problem, and we realize the many limitations of this analysis. For example, many variables, including age, sex, starting depression score, behavioral changes occuring in the interim period and the kind of treatment, were not accounted for. Moreover, many of the patients selected as nonresponding actually improved their symptoms slightly. In fact, symptoms of only 9 patients of the whole dataset worsened or stayed stagnant. No significant differences between before and after treatment recordings of these "strictly" nonresponding patients, were found. Thus, one possibility of strengthening the differences between the responding and nonresponding groups may be to include the after treatment recordings for the "strictly" nonresponding group to the nonresponding group. However, we decided against doing this, because some differences in the before and after treatment groups for the "strictly" nonresponding patients may have remained insignificant due to small size of this group.

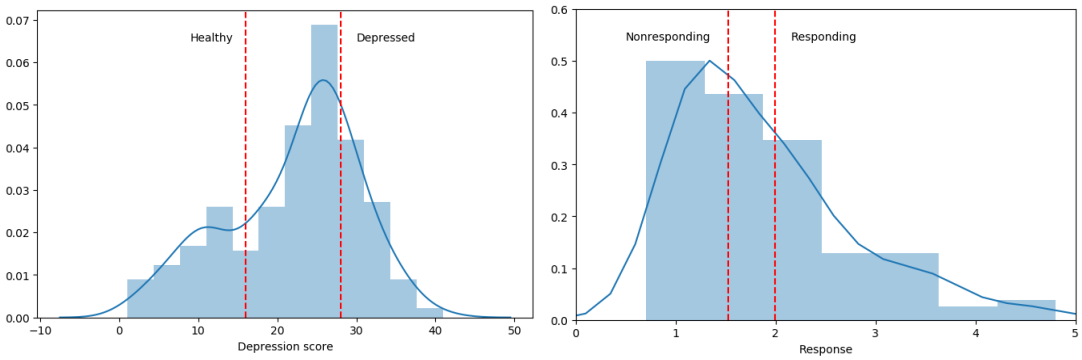
2.5 Results

2.5.1 Methodology

We used two classifiers: Logistic Regression (LR) and Support Vector Machine (SVM). One third of randomly selected samples was held out as a test set, the rest was used for training and cross validation. The following feature selection algorithms were evaluated with LR (regularization strength 1) and SVM (regularization strength 1 and linear kernel, i.e. $k(\boldsymbol{x_1}, \boldsymbol{x_2}) = \boldsymbol{x_1} \cdot \boldsymbol{x_2}$):

- recursive feature elimination with 3-fold cross validation based on coefficients of the linear model,
- elimination of features with below-mean coefficients of the linear model,
- selection of 5 features with the highest χ^2 statistics between values of the feature and corresponding class,
- genetic algorithm with 5-fold cross validation (scoring models based on ROC AUC, population size 80, 80 generations, crossover probability 0.8, mutation probability 0.2, and tournament size 5),
- manual selection of channels with significantly different means of corresponding features between the two considered classes, as reported by the Kruskal-Wallis test.

Am I justified in using χ^2?



(a) Label thresholds for depression diagnosis.

(b) Label thresholds for response prognosis.

Figure 2.28: Label selection for diagnosis and prognosis tasks. The thresholds were selected such that enough samples was present in each class, and such that the classes remained balanced.

Out of those feature selection techniques, genetic algorithm was by far the most effective. However, most of the best performing and thus reported classifiers were found by combination of the last two techniques, i.e. by applying genetic selection algorithm to the subset of channels marked as having differing means between the two groups.

Evaluation was performed using 5-fold cross-validation. The best performing classifiers (based on accuracy, precision, recall, f-score, and number of features) were selected for each measure and for all measures by varying the maximum number of features considered by the genetic algorithm from 3 to the 1/10 of the corresponding training set size. Then, a brute force grid search with 5-fold cross validation was performed on each classifier to select

- the optimal regularization stength, and norm for LR, and
- the optimal regularization strength and kernel type (linear, polynomial, or radial basis function with coefficients $\gamma = 1/n_f$, where n_f is the number of selected features) for SVM.

Am I justified in changing the kernel type?

This resulted in slight improvement in accuracy, and correspondingly slight bias of the reported classifiers.

2.5.2 Depression Diagnosis

The recordings were separated into two classes as follows:

Healthy : 50 recordings with associated depression score at most 16.

Depressed : 50 recordings with associated depression score at least 28.

These thresholds are visualized on the distribution of depression scores in the dataset in Figure 2.28a. The results are shown in Table 2.7. The best performing classifiers in this section were SVMs. The largest Lyapunov exponent was the most predictive out of all considered nonlinear measures, both achieving the highest accuracy out of the single-measure classifiers (0.72 ± 0.04), and being one of the measures in majority of the best performing combined-measure classifiers. It was followed, perhaps surprisingly considering the results obtained in Section 2.4.2, by correlation dimension (0.71 ± 0.05). Although the accuracy of the remaining classifiers, whose features were obtained using the Kruskal-Wallis test from Section 2.4.2, was slightly lower (with higher variance), they are also simpler in terms of the number of selected channels.

All the channels in the combined-measure classifiers were found using the genetic algorithm, as described in the opening to this section. The best overall accuracy was achieved by combination of the largest Lyapunov exponent and sample entropy (0.75 ± 0.10). However, second to it was a combination of the largest Lyapunov exponent and correlation dimension, which has lower variance (0.74 ± 0.04). Other measures performing well together with the largest Lyapunov exponent are the Hurst exponent, and sample entropy together with DFA. The best combination not inluding the largest Lyapunov exponent is correlation dimension and Higuchi's fractal dimension.

In Section 2.4.2, we have seen that correlations of LLE with depression score are the most significant of all measures, which may explain why in this section, we found LLE to be the most discriminative measure. Moreover, we have seen that the differences between depressed and healthy patients are the most significant in the temporal areas for all measures, which may explain the relatively high rate of selection of the T3 and T6 channels.

Consistency in the selected channels for a fixed measure might indicate a relationship between activity captured by the nonlinear measure in the brain region local to the channel and corresponding label. However, the selected features seem inconsistent across classifiers using mixed measures or a single measure included in the mix. One possible explanation is that different measures complement themselves in such a way that different channels are relevant when classification is performed based on single measure as opposed to a combination of measures.

Although the obtained accuracies are notably lower than those obtained in studies we reviewed in Section 1.6.1, these results are not directly mutually comparable. Most importantly, all the reviewed studies employed a control group of symptomless, healthy patients, whereas our dataset composed entirely of patients suffering from depression with various degree of symptom severity. Nevertheless, the obtained accuracies are substantial, suggesting possible nonsaturating relationship between depression severity and nonlinear measures recorded in EEG signals obtained from various cortical regions.

2.5.3 Response Prognosis

As we mentioned in Section 2.4.3, response is defined as the ratio of the depression score reported on the second session (after administration of drugs) to the depression score reported on the first session (before administration of drugs). For the task of predicting response to treatment, recordings were separated into two classes as follows:

Nonresponding : 50 recordings made before administration of drugs with the lowest response.

Responding : 50 recordings made before administration of drugs with the highest response.

The results can be seen in Table 2.8. Again, the largest Lyapunov exponent was the most predictive nonlinear measure. Unlike for the depression diagnosis, sample entropy and correlation dimension were the only other nonlinear measures which were able to achieve accuracy above 70% both in combination with other measures and as stand-alone features. Interestingly, F3 channel seems to be considerably more relevant for response prognosis than for depression diagnosis.

In Section 2.4.3, we have observed the most significant differences between responders and nonresponders in the recordings obtained before treatment for LLE (especially F4 and T6 channels), which may explain that LLE is the most discriminative nonlinear measure for treatment response prediction. Although the differences for other measures remained relatively insignificant, CD, HD, and SE achieve accuracy of approximately 67%. This may be explained by the fact that in Section 2.4.3, we tested only for the differences in distributions in values recorded on individual channels, whereas in this section, we perform classification based on combination of multiple channels.

Measure	Classifier	Accuracy		Precision		Recall		F-score		Channels
		Mean	**Std**	**Mean**	**Std**	**Mean**	**Std**	**Mean**	**Std**	
LLE, SE	SVM (lin.)	**0.75**	0.10	0.77	0.09	0.75	0.10	0.75	0.10	*LLE*: C4, T3, T6, Pz *SE*: C3, P4
LLE, CD	SVM (lin.)	0.74	0.04	0.76	0.04	0.74	0.04	0.74	0.05	*LLE*: F3, F7, T6 *CD*: O1, O2, T5
LLE, HE	SVM (lin.)	0.73	0.06	0.74	0.06	0.73	0.06	0.73	0.06	*LLE*: P3, T3, T6, Pz *HE*: C3, T3
LLE, SE, DFA	SVM (lin.)	0.73	0.09	0.74	0.10	0.73	0.09	0.73	0.09	*LLE*: T6, Fz *SE*: T6 *DFA*: P4
CD, HD	LR	0.73	0.10	0.74	0.11	0.73	0.10	0.73	0.10	*CD*: F3, Fz *HD*: P3, Cz
LLE	SVM (lin.)	0.72	0.04	0.73	0.04	0.72	0.04	0.72	0.04	T3, T5, T6, Pz
CD	SVM (lin.)	0.71	0.05	0.72	0.05	0.71	0.05	0.71	0.05	F3, C4, P3, F8, T5, T6, Fz, Cz
SE	LR	0.68	0.12	0.69	0.12	0.68	0.12	0.68	0.12	C4, O2, T6
HD	SVM (rbf)	0.67	0.11	0.67	0.12	0.67	0.11	0.67	0.11	C3, C4, P4, T6, Cz
DFA	LR	0.67	0.16	0.68	0.17	0.67	0.16	0.67	0.16	F8, O2
HE	LR	0.67	0.17	0.68	0.18	0.67	0.17	0.67	0.17	O2, T4

Table 2.7: Evaluation of depression classification. The two classes consist of 50 / 50 recordings with the smallest / highest associated depression score out of recordings performed both before and after administration of drugs.

Measure	Classifier	Accuracy		Precision		Recall		F-score		Channels
		Mean	**Std**	**Mean**	**Std**	**Mean**	**Std**	**Mean**	**Std**	
LLE, SE	SVM (lin.)	**0.75**	0.10	0.77	0.09	0.75	0.10	0.75	0.10	*LLE*: FP2, F3, O1, T4, T6 *SE*: F3, C3, T6
LLE, CD	SVM (lin.)	**0.75**	0.11	0.76	0.11	0.75	0.11	0.75	0.11	*LLE*: F3, O2, T5, T6 *CD*: FP2, F4, O2
LLE	LR	0.71	0.08	0.73	0.08	0.71	0.08	0.70	0.09	F3, F4, T5, T6
CD	LR	0.67	0.09	0.70	0.11	0.67	0.09	0.65	0.10	F3, F4, O2, Pz
HD	LR	0.66	0.05	0.72	0.08	0.66	0.05	0.64	0.05	F3, F8
SE	LR	0.66	0.09	0.66	0.09	0.66	0.09	0.65	0.10	FP1, F3, P3, Cz
DFA	SVM (lin.)	0.64	0.15	0.65	0.15	0.64	0.15	0.63	0.15	T3, T4, Cz
HE	SVM (rbf)	0.63	0.09	0.64	0.10	0.63	0.09	0.62	0.09	C3, T6

Table 2.8: Evaluation of response classification. Only recordings obtained before drug administration were considered. The two classes consist of the 50 patients with the highest and least improvement in depression score after the drug administration (as measured by ratio of the two depression scores).

By comparing our results to results of studies reviewed in Section 1.6.2, we can see that our results, obtained using nonlinear dynamical analysis and LLE in particular, is relatively competetive with some results in published literature focusing on relative frontal θ powers. Our results may therefore encourage effort in examination of nonlinear measures and LLE in relation to treatment response.

2.6 Implementation

All the code used for this section is publicly available online, as well as on the CD accompanying this text. Used programming language was Python. The plots were produced using Matplotlib library. Other used libraries include NumPy for vectorized computations, SciPy for signal processing algorithms, Pandas for data storage and manipulation, and Jupyter for development.

For computation of nonlinear measures, we used slightly modified versions of nolds library by C. Schölzel and published under MIT Licence, and nolitsa library by M. Mannatil and published under 3-clause BSD licence.

Hereby we provide a summary of used software:

Public availability: online, accompanying CD

Main programming language: Python

Plotting library: Matplotlib

Vectorized computations: NumPy

Signal processing algorithms: SciPy

Data storage and manipulation: Pandas

Execution environment: Jupyter

Nonlinear Dynamical Analysis: modifications of nolds and nolitsa

Chapter 3

Deep Learning Approach

3.1 Convolutional Neural Networks

In this section, we introduce the Convolutional Neural Networks (CNNs), a class of neural network architectures, which we use in this chapter to perform the depression diagnosis and prognosis classification tasks.

3.1.1 Mathematical Background

In order to make some of the essential terms used in this chapter more concrete, we will provide their definitions in the following text. The first term requiring a definition is the term convolution, which gives the name to convolutional neural networks.

Definition 12. *Let I be an image function, K a kernel. A (discrete)* ***convolution*** *of I and K is a linear operation defined as*

$$(I * K)(i, j) = \sum_m \sum_n I(m, n)K(i - m, j - n)\,. \tag{3.1}$$

Note that some machine learning libraries (such as TensorFlow) implement **cross-correlation** instead of convolution, but preserve the term convolution for the operation [146]. Cross-correlation corresponds to convolution with kernel rotated by 180 degrees:

$$(I * K)(i, j) = \sum_m \sum_n I(m, n)K(i + m, j + n)\,. \tag{3.2}$$

Unlike convolution, cross-correlation is not commutative, but this property is not necessary for neural network applications [147].

When discussing the distinguishing properties of CNNs, we will use the the terms invariance and equivariance.

Definition 13 ([148])**.** *Let $f : X \to Y$ be arbitrary function, G a group of transformations on X. We say f is an* ***invariant*** *function under group of transformations G if*

$$f(T(x)) \equiv f(x), \qquad \forall T \in G,\ \forall x \in X.$$

Moreover, if G is a group and X, Y its G-sets, then $f : X \to Y$ is an ***equivariant*** *function under G if*

$$f(T(x)) \equiv T(f(x)), \qquad \forall T \in G,\ \forall x \in X.$$

Invariance to a transformation means that if the transformation is applied to the input of the function, the output remains the same. Equivariance, on the other hand, means that if the transformation is applied to the input, the output is transformed in the same way. For example, as we will see in Section 3.1.3.3, CNNs are invariant to local translation, but are not invariant to rotation, for example. Furthermore, convolution is an operation equivariant to translation [147].

A neural network can be viewed as a function approximation model based on an iterative optimization algorithm. There is a number of such algorithms in use and many of them are variations of **gradient descent** algorithm, introduced by Werbos in 1974 [149]. Gradient descent is a first order iterative method of finding an extremum of a differentiable function $f : \mathbb{R}^m \to \mathbb{R}^n$, $f \in C^1$, based on continually moving a point in its domain space in the opposite direction of its gradient at that point, until the absolute value of the gradient is below a certain threshold.

Algorithm 1 A basic version of gradient descent algorithm. The parameter α is called the learning rate.

Initialize random $x_0 \in \mathbb{R}^n$,
$n \leftarrow 0$
step_size $\leftarrow 1$
while step_size < threshold and $n <$ iters_limit **do**
    $x_{n+1} = x_n - \alpha * \text{grad} f(x_n)$
    step_size $\leftarrow |x_{n+1} - x_n|$
    $n \leftarrow n + 1$
    Optionally update α
end while

The family of first order optimization methods is vast, including techniques such as stochastic variations of gradient descent [150], gradient descent with momentum [151], RMSProp (Root Mean Squared Propagation) [152], and Adam (Adaptive Moment estimation) [153]. In Algorithm 1, we can see that gradient descent consists of two main steps: computation of the gradient $\text{grad} f(x_n)$, and the parameter update. Most of the variations modify one at least one those steps, i.e. the step direction and its magnitude.

Another family of iterative optimization algorithms, second order methods, is based on second derivatives. In many cases the second derivative is computed using quasi-Newton techniques for approximating the Hessian matrix. In the case of CNNs, the most common optimization methods used are based only on the first derivative due to high computational costs of computing both gradient and Hessian approximations.

3.1.2 History

The classical approach to image pattern recognition consists, in general, of the following stages:

Preprocessing: Main focus is on supressing undesirable distortions and noise. In some cases, enhancements beneficial for further processing may be applied, such as modification of contrast.

Segmentation: Partitioning the image into disjoint sets of pixels sharing a property, such as color or texture. For example, distinguishing disparate objects from the background.

Feature extraction: Gathering relevant information about the properties of the objects, removing irrelevant variations.

Classification: Categorizing segmented objects based features obtained in the previous step into classes.

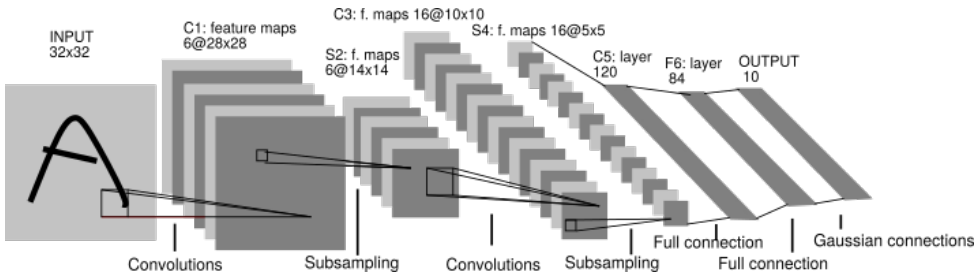


Figure 3.1: Architecture diagram of LeNet-5, one of the first convolutional neural networks successfully applied to image classification [157].

The preprocessing step may require additional assumptions about the data or further processing, which are potentially too restrictive or too broad. Getting around this limitation requires dealing with complications such as high dimensionality of the input (number of pixels) and desirability of invariance to a number of allowable distortions and geometrical transformations.

Artificial neural networks in combination with gradient-based learning are one possible solution to the problem. By gradually optimizing a set of weights based on a training data set using a differentiable error function, they provide a framework for learning a suitable set of assumptions automatically from the data.

One of the oldest neural network architectures, fully connected multi-layer perceptron (FC-MLP), can be theoretically used for image pattern recognition. However, it has at least the following drawbacks [154]:

parameter explosion: the number of parameters of such network is exponential in the number of layers, increasing the capacity of the network and therefore need for more data,

no invariance: no invariance even with respect to common geometrical transformation such as translation, rotation and scaling,

ignoring input topology: natural images exhibit strong local structure and high correlation between intensities of neighboring pixels, but FC-MLPs are unstructured - inputs can be presented in any order.

Although the main idea of CNNs dates back to 1980, when K. Fukushima introduced neocognitron [155], the back-propagation algorithm was not widely known at the time. The first convolutional architecture successfully applied on an image pattern recognition problem by attempting to solve the aforementioned problems, dubbed LeNet-5 (see Figure 3.1), was proposed in 1998 by Y. LeCun, L. Bottou, Y. Bengio and P. Haffner [156].

3.1.3 Properties

Being inspired by visual processing systems in biological organisms[1], LeNet-5 proposed the following design principles to enforce *shift, scale and distortion invariance* [157]:

local receptive fields: each neuron in a layer receives input from a small neighborhood in the previous layer,

[1] As early as in 1968, D. H. Hubel and T.N. Wiesel discovered that some cells (called simple cells) in cat's primary visual cortex (V1) with small receptive fields (shared by neighboring neurons) are sensitive to straight lines and edges of light of particular orientation, and other cells (called complex cells) with larger receptive fields further in the visual cortex also respond to straight lines and edges, but with invariance to translation [158].

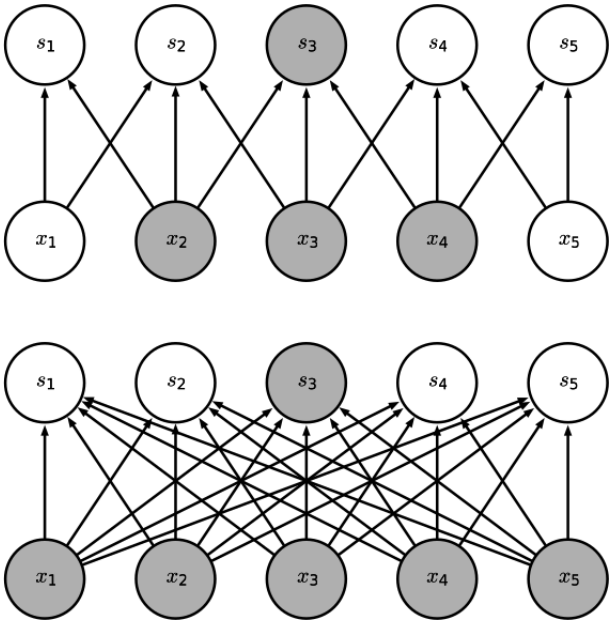


Figure 3.2: Visual depiction of the concept of a receptive field [147]. The upper diagram shows the receptive field of s_3 unit in case of kernel width of 3, whereas the bottom diagram shows the same receptive field in case of fully connected layers. We can see that the second case will require substantially more parameters stored in memory.

shared weights: each layer is composed of neurons organized in 2D planes (matrices) stacked into a 3D volume (tensor), where neurons within each plane share the same weight vector (feature map),

spatial subsampling: adding subsampling layers, which reduce the resolution of the previous layer by averaging or by taking the maximum value of neighboring pixels in the previous layer.

In the following, we will describe each of these properties and their implications in more detail.

3.1.3.1 Local Receptive Fields

Local receptive fields enable the network to synthesize filters that produce strong response to elementary salient features in the early layers (such as lines, edges and corners in a visual input, and their equivalents in other modalities), and then learn to combine them in the subsequent layers to produce higher-order feature detectors. This property enables the models to induce the previously hardcoded filters from data.

For a visual explanation of the concept of receptive field, see Figure 3.2. In the layers closer to the input, locality of receptive fields implies locality of "influence" of each input unit on the output. On the other hand, units in the deeper layers can be indirectly connected to some or all units of the input, thus enabling them to achieve the aforementioned effect of combining more complex features from simpler ones. This means that increasing number of layers enables CNNs to form more complex spatial representations efficiently. Recently, it has been proven that the number of parameters needed for approximating a function in the case of shallower networks (less layers) is much larger in comparison to the number of neurons in a deeper architectures (more layers) [159].

Furthermore, locality of receptive fields implies sparser connectivity, and hence more efficient computations in comparison with fully connected neural networks. A fully connected neural network with

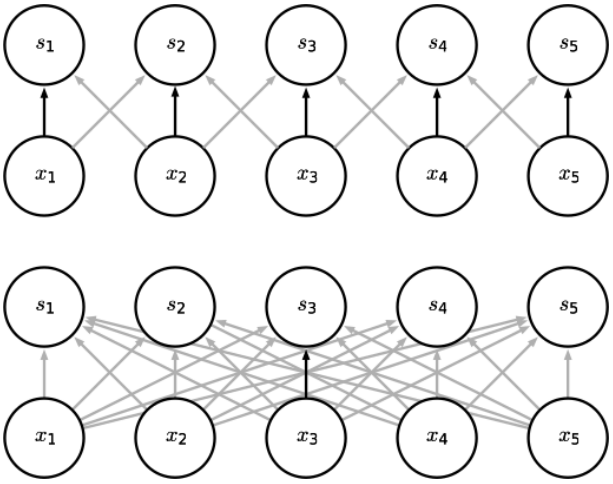


Figure 3.3: Visual depiction of the concept of shared weights [147]. Black arrows in both the upper and lower models correspond to connections which use a single shared weight parameter. In the upper model, where the black arrow marks the central unit of a kernel of size 3, a single parameter is reused at all output locations. In the lower, fully connected model without shared weights, each parameter is used only at a single location.

no hidden layers with m inputs and n outputs has $m \times n$ weight parameters, and the correspoding feed forward pass (matrix multiplication) is of $O(m \times n)$ time complexity per input. If the number of connections per output unit is limited to $k < m$, the achieved runtime is $O(k \times n)$, where k is usally in practice several orders of magnitude smaller than m [147].

3.1.3.2 Shared Weights

With *shared weights*, neural units in a layer with differing receptive fields have the same feature map and the same feature detecting operation (convolution with feature map kernel followed by additive bias and a application of a nonlinear function) is performed on different parts of the image (see Figure 3.3). A single convolutional layer is composed of multiple feature detecting planes stacked onto each other to form a volume.

Shared weights principle exploits the fact that in natural images, a function of small number of neighboring pixels can be useful in multiple parts of the image. For example, an edge detector can be used accross the entire image to detect edges in the first layer, an object detector can then be used to detect presence of edges in particular arrangements in the next layer, etc.

Although it does not reduce the time complexity of the feedforward pass, it does reduce the memory requirements. If k is the kernel size, m the number of inputs, n the number of outputs, the number of parameters per layer is k instead of $m \times n$ (per feature detecting plane) in a fully connected case. Since k is usually in practice several orders of magnitude smaller than m, and usually m and n are comparable in size, the memory savings are substantial [147].

One of the drawbacks of classical CNNs is that although convolution in combination with weight sharing causes layer output to be equivariant to translation of the input, this is not the case for scaling and rotation [147]. Moreover, equivariance to input may not be always desirable. Consider a case of face detection, where all training and test images are centered. Then, the relative positions of individual features are important, and it may be favorable to fix feature detectors (and thus weights) to certain locations in the image [147].

3.1.3.3 Pooling Operation

The final output activations of a convolutional layer followed by a pooling layer are computed in subsequent stages:

1. linear unit activations are computed via the convolution operation,
2. a nonlinear activation function is applied to the activations,
3. a spatial subsampling (pooling) operation is applied to the activations.

The rationale behind applying a nonlinearity is that it makes the network capable of modelling nonlinear functions.[2] Common activation functions include rectified linear $\max(0, x)$, exponential linear units [161], sigmoid $\frac{1}{1+\exp(-x)}$, hyperbolic tangent tanh, and many others. They have varying properties making them useful in different situations.

Pooling operation splits the neural units into sets of multiple adjacent activations and computes a summary statistic, such as the maximum element (max pooling) or the average (average pooling), per such set and outputs the result. If the stride between the sets is greater than one, the spatial dimension of output is decreased relative to input (subsampling). The purpose of spatial subsampling is to ensure scale and distortion invariance by reducing the precision at which a feature is encoded in a feature map by reducing its resolution - when scale and distortion invariance is assumed, the exact location of a feature becomes less important and is allowed to exhibit slight positional variance - roughly speaking, an "approximate" translation invariance [147].

Although the combination of convolution and pooling performs well in many practical situations, it has multiple drawbacks. For example, the learned representations are not rotation invariant and thus, to mitigate this, the capacity of the network has to be increased and the training dataset must be enhanced to contain examples of rotated features, often extending the amount of data necessary and training time. A number of alternative approaches were suggested in the litarature.[3] For another example of a limitation, see Figure 3.4.

3.2 Common Spatial Patterns

The method of Common Spatial Patterns (CSP) was originally proposed for people with impeded motor control (e.g. disabled people) in context of brain-computer interfaces, and thus most studies focus on its use in classification of motion performed or visualized by the subject. In our study, we will apply convolutional neural network architectures inspired by Filter Bank Common Spatial Patterns (FBCSP, see Section 3.2.2) for depression diagnosis and prediction of future remission of the disease.

As mentioned in Section 1.1, the task of finding patterns in EEG signal associated with particular mind state or motor action presents us with numerous challenges. The CSP algorithm, and its extension FBCSP in particular (see Section 3.2.2), are methods devised in attempt to overcome mainly two of them. Firstly, information about different temporally overlapping brain activities is conveyed in parallel in multiple frequency bands. For example, resting wakeful state comprises distinct idle rhythms over different cortical areas (such as α-rhythm characteristic of idling visual cortex in the occipital area), which are overlapping with μ-rhythms produced in sensorimotor areas both during imagined and performed

[2] In fact, a feedforward neural network with finite number of units and "proper" activation function can approximate any Borel measurable function on a finite dimensional compact subset of $\mathbb{R}^n$ [160].

[3] For instance, Hinton's *CapsNet*, described in [162], is an attempt to transform the manifold of images of similar shape (which is highly nonlinear in the space of pixel intensities) to a space where it is globally linear by the way of using so called capsules instead of traditional convolutional layers.

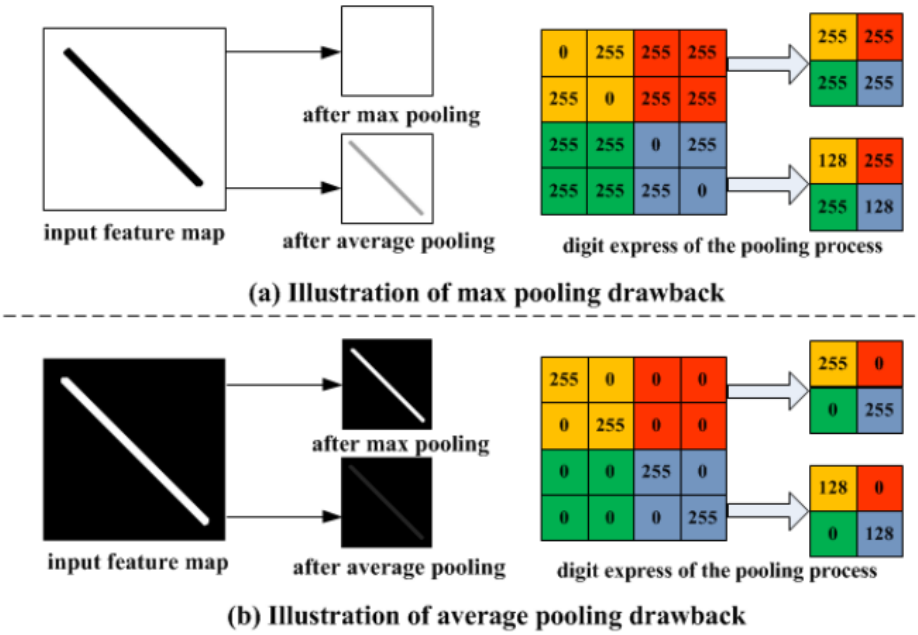


Figure 3.4: Examples of drawbacks of the pooling operation [163]. Max pooling discards all except the maximum element, and valuable information may thus be lost. Average pooling considers all the values, and the information about their contrast is reduced. Moreover, extreme values may have undesired effects on the result.

movement [164]. Secondly, the spatial origin of those signals is important for associating them with said mind states or motor actions. For example, different parts of the sensorimotor cortex over the central sulcus map directly to movements of distinct bodyparts. This is further complicated by the fact that EEG apparatus has, as we discussed in Section 1.1, inherently low spatial resolution due to small number of electrodes and poor volume conduction.

Spatial filtering, then, is process of addressing this second challenge by accentuating signals from some areas, while attenuating others. And CSP analysis is a data-driven approach of achieving this by mutually maximizing the variance of spatially filtered signal associated with one activity, while minimizing the variance of filtered signal associated with other activity, thus making the signals independent (as Gaussian random processes) [164]. In the following section, we will explain the process in detail.

3.2.1 Algorithm

In the description of the CSP algorithm, we follow the notatition from [164]. Let C be the number of channels, and $\mathbf{x}(t) \in \mathbb{R}^C$ be a band-passed, de-meaned and scaled multichannel EEG recording. CSP analysis yields a projection of $\mathbf{x}(t)$ of the signal from the original signal space to $\mathbf{x}_{\text{CSP}}(t) \in \mathbb{R}^C$ by finding a matrix $W \in \mathbb{R}^{C \times C}$, where

$$\mathbf{x}_{\text{CSP}}(t) = W^T \mathbf{x}(t).$$

Each column vector of W is referred to as spatial filter. Thus, CSP decomposes the original signals into additive subcomponents, column vectors of $A := (W^{-1})^T$, referred to as spatial patterns, giving name to the technique.

The matrix W is found under optimization criteria, which we will describe in the following text. Firstly, let $\Sigma^{(+)} \in \mathbb{R}^C$ and $\Sigma^{(-)} \in \mathbb{R}^C$ be estimates of the inter-channel covariance matrices, corresponding

to signals recorded in the two conditions c we aim to distinguish, + and −:

$$\Sigma^{(c)} = \frac{1}{|I_c|} \sum_{i \in I_c} X_i X_i^T, \qquad c \in \{+, -\},$$

where I_c is the set of time indeces matching the two conditions.[4] Since variance of band-pass filtered is the power present in the frequency band, the diagonal elements of $\Sigma^{(c)}$ represent the fraction of the total band power in each channel, and the off-diagonal elements represent the fractional covariance [165]. CSP then performs simultaneous decomposition

$$W^T \Sigma^{(+)} W = \Lambda^{(+)},$$
$$W^T \Sigma^{(-)} W = \Lambda^{(-)},$$

under the conditions that both $\Lambda^{(+)}$ and $\Lambda^{(-)}$ are diagonal, and $\Lambda^{(+)} + \Lambda^{(+)} = I$. This is equivalent to solving the generalized eigenvalue problem

$$\Sigma^{(+)} \mathbf{w} = \lambda \Sigma^{(-)} \mathbf{w}$$

for generalized eigenvectors $\mathbf{w}$ and their eigenvalues λ. The resulting eigenvectors $\mathbf{w}_j$, $j \in \{1, \ldots, C\}$ then are the column vectors of W, and corresponding eigenvalues $\lambda_j^{(c)} = \mathbf{w}_j^T \Sigma^{(c)} \mathbf{w}_j$ are the diagonal elements of $\Lambda^{(c)}$. Then, $\lambda_j = \lambda_j^{(+)} / \lambda_j^{(-)}$, and $\lambda_j^{(+)} + \lambda_j^{(-)} = 1$. This means that high variance in the direction of $\mathbf{w}_j$ of signal in class + results in small variance in signal in class −, and vice versa (see Figure 3.5) [164].

This method, although loosely based on PCA, is better suited for supervised classification, since, unlike PCA, it is guaranteed to find components which are responsible for the maximum differences in variance between the two classes. These eigenvectors are an orthonormal set which spans $\mathbb{R}^C$, and are optimal for the amount of variance they account for in the least squares sense [165].

3.2.2 Filter Bank Common Spatial Patterns

Although CSP usually yields good performance when the signals have been filtered in frequency range carefully tuned for the particular subject and classification problem at hand, its performance rapidly decreases when measurements are either unfiltered or filtered in inappropriate frequency range [166]. Thus, an improvement has been developed under the name of Filter Bank Spatial Patterns (FBCSP). It comprises of four stages: frequency filtering, spatial filtering, feature selection and classification (see Figure 3.6). In stage 1, multiple band-pass filters are applied to split the signal into distanct filter banks. Then, in stage 2, spatial filters are found for each of the respective filter banks using CSP analysis, as described in Section 3.2.1. These filters characterize features present in the signal specific to the corresponding frequency band. In stage 3, a feature selection algorithm is employed to select the most discriminative features of all the filters found in previous step. Finally, in stage 4, a classification algorithm uses the selected features for classiying the input signal into a class.

Using the fact that magnitudes of the CSP eigenvalues are proportional to the amount of variance explained by corresponding signal in the direction corresponding to the eigenvector, the CSP algorithm in stage 2 is slightly modified to order the eigenvectors according to the magnitude of their eigenvalues, and only the first and last m filtered signals are selected for further classification. This means selecting only the first and last m rows (channels) from the matrix $X_{\text{CSP}} = W^T X$, yielding a matrix $Z \in \mathbb{R}^{(2m) \times T}$ with row vectors of signals Z_j, $j = 1, \ldots, 2m$ with the maximum and minimum variances in the projected

[4] Here we suppose that two separate events happened during a single recording to simplify notation.

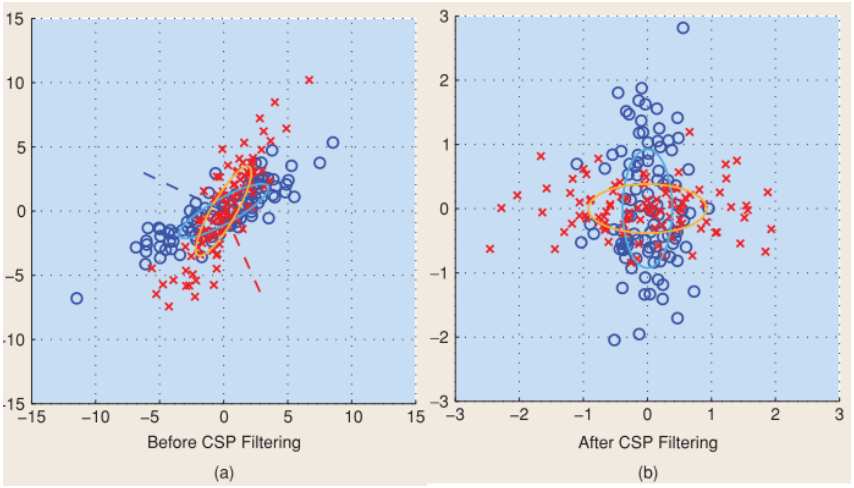


Figure 3.5: A demonstration of the CSP algorithm on a two classes of 2-dimensional samples generated randomly using different distributions [164]. The dashed lines on the right hand side correspond to the CSP projection vectors w_1, w_2. Note that although there is a strong correlation for both classes in the original space, correlation for one class is maximized and minimized for the other class in the projected space.

space. The final feature vector $\mathbf{f}$ is composed as logarithm of the contribution of variance of each row vector to the total variance as follows [166]:

$$f_j = \log\left(\frac{\text{var}(Z_j)}{\sum_{i=1}^{2m}\text{var}(Z_i)}\right). \tag{3.3}$$

3.3 Dataset

For experiments in this chapter, we used the same dataset as in Chapter 2, described in Section 2.1. However, to increase the number of samples, we decided to use the entire recordings, in contrast to our approach in Chapter 2, where we used only the beginning in each recording. This is mainly because the classification algorithms we used in this chapter have larger variance, and thus are easily overfit on small datasets. Thus, each of the recordings, after downsampling to 250 Hz (see Section 2.1) was segmented into disjoint subrecordings of length 256, each subrecording forming a data sample. The subrecording length was selected as a tradeoff between

- the number of obtained samples,
- the amount of information contained withing each sample, and
- the size of data batch transferred to the GPU as input to the CNN.

As we will see in Section 3.4, we decided to represent the input to the CNNs either as image-encoded signal, where the resulting size of the input is $O(N^2)$ for subrecording length N, or using raw data, where the input size is $O(C \times N)$ for constant number of selected channels C. Thus, the raw data input representation allows for longer subrecording length for the same data batch size compared to the image-encoded signal representation. However, we decided to use the same subrecording length for both input

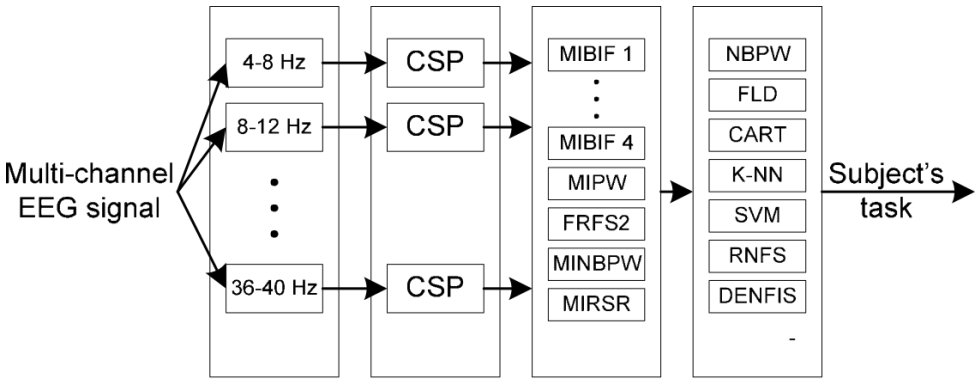


Figure 3.6: A schematic overview of the FBCSP [166]. First, the signal is filtered into various bands. Then, CSP algorithm is applied to each band, and thus can maximize variance for class-specific features present in each band. After selecting features from subset of the signals, a classifier uses the features from all bands to perform the classification task.

representations in order to make the results comparable. For the raw data, we did separate experiments increasing the length to 512 or 1024 using overlapping windows to keep the resulting number of samples constant, and obtained similar performance.

For the selected sampling frequency, this subrecording length corresponds to approximately a second of recording, which was shown to contain enough information to identify individuals with satisfactory accuracy using CNNs [167]. Moreover, memory operations on data batches power of two bytes in size are easier to coalesce among multiple threads, thus improving runtime [147]. Finally, the reason for keeping the subrecordings disjoint was to minimize the amount of correlation between the samples.

After splitting the recordings, we assigned positive, neutral or negative label to each subrecording in order to split the dataset into three groups based on depression score of the subject at the time of the recording for depression classification, or based on the subject's before to after treatment depression score for response classification (in which case only before treatment recordings were further used). The neutral class was then removed and not further considered.

The threshold values separating these classes were selected such that the classes remained relatively balanced and that enough samples were present in each class to train and evaluate a model of moderate capacity. In the case of depression classification, the amount of inter-class variance is inherently limited by nature of the provided data - patients were not randomly sampled, but visited the institution to seek professional help. In attempt to partially remedy this issue, the depression score threshold was set such that 71 patients remained in each depressed and healthy classes, leaving 124 neutral subjects. This corresponds to depression score ranges $\langle 0, 17\rangle$ for healthy, $(17, 27)$ for neutral, and $\langle 27, 34\rangle$ for depressed. In the case of response classification, our ability to potentially increase inter-class variance in this way is more limited due the amount of available data, since only before treatment recordings are used. Thus, we removed only 14 neutral subjects, leaving 59 nonresponding and 60 responding subjects.

3.4 Input Representation

Before applying any machine learning technique to the classification problem at hand, the question of optimal input representation needs to be answered. To this end, multiple factors need to be considered.

- What is the dimensionality of the input relative to the resulting number of samples, and does it allow construction of sufficiently complex architectures compared to complexity of the classification problem?
- Does each input sample contain enough information to perform successful classification?
- Is the input representation appropriate for the kind of data, i.e. does it help or hinder successful classification?

In our case, for all the methods considered, answer to the first question is a function of recording slice length used to generate the input. Answering the remaining questions, however, is difficult without prior experiments on similar datasets. For this reason, research on applying known techniques to new problems is useful.

Since well designed neural networks are able to learn feature maps given enough data, one obvious possibility is to use raw data. This approach has multiple benefits. First one, as we will see, is relatively low dimensionality. Global, and local features, no prior bias. Our first choice, then, was then to segment each recording into subrecordings of fixed length $l = 256$, and, for each subrecording, order all its $C = 19$ channels into rows of a $C \times l$ matrix. These matrices were then used as input samples.

To incorporate the results from the previous chapter, we also evaluated the models using only subsets of all the available electrodes. Those subsets were selected from the electrodes in which the distributions of nonlinear measures in Section 2.4 were shown to be significantly different between studied groups, or which were selected during the feature selection step in 2.5. However, we found either no notable improvements, or slight deterioration in performance.

In Chapter 1, we assumed that the brain can be modelled as a nonlinear deterministic dynamical system. Another possibility, therefore, is to represent the input data as recurrence plots, which have considerable potential for representing properties of dynamical systems, as mentioned in Section 1.2.4. The obvious drawbacks of recurrence plots are redundancy due to symmetry and high dimensionality, which is quadratic in the trajectory length. On the other hand, recurrence plots are known to capture properties of the system which are difficult to obtain using other methods in some cases [42]. Morever, they have already been applied with success to classification of physical activites using convolutional neural networks [109], and even some qualitative differences in recurrence plots have been observed between depressed and healthy patients [110]. We have discussed more applications in 1.6.1.

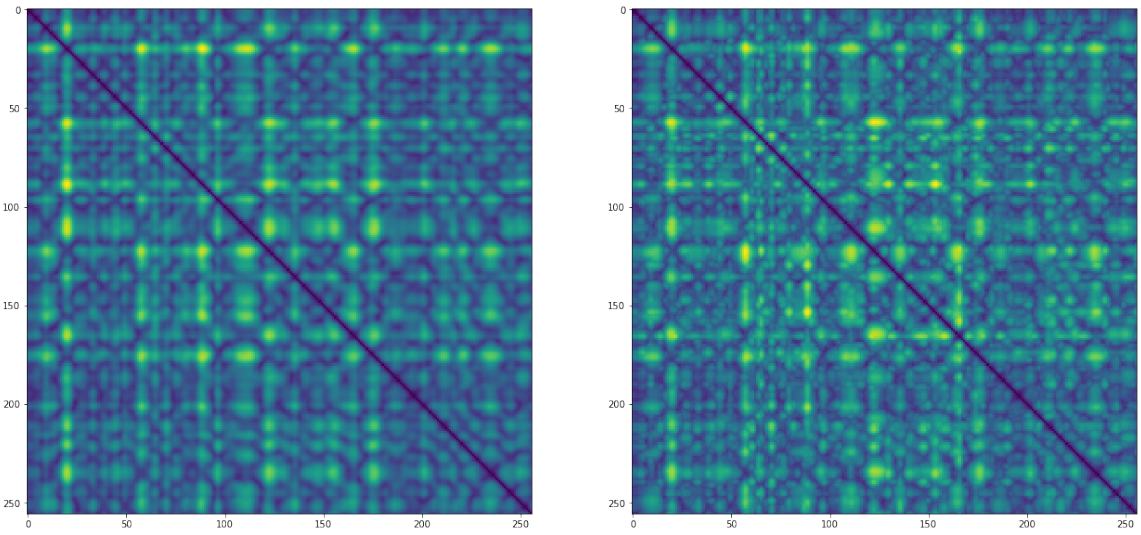
For our computation of recurrence plot matrices, we used the Chebyshev norm, which has multiple benefits. The most relevant ones are relatively low computational cost and distances independent of embedding dimension. Moreover, we observed subtler patterns on matrices computed using Chebyshev norm as opposed to those computed using Euclidean L_2 norm. For comparison, see Figure 3.7.

Our last method of input representation is inspired by success of Gramian Angular Fields (GAFs) for sequence classification [168] using convolutional neural networks. To obtain GAF matrix from a scalar time series $x_1, x_2, \ldots, x_N$, one first scales the time series into interval $(-1, 1)$, and then each value x_i of the time series is converted into complex number with mode and radius given as

$$\begin{aligned} \phi_i &= \arccos(x_i), \\ r_i &= i/N. \end{aligned}$$

This way, temporal dependencies are conserved through the radius. Then, instead of scalar product, an operation $\oplus$ is defined as

$$x_i \oplus x_j = \cos(\phi_i + \phi_j)$$



(a) Euclidean norm.

(b) Chebyshev norm.

Figure 3.7: Recurrence plots computed using different norms. We can see that the figure on the right hand side has sligthly crisper patterns.

and a quasi-Gram $N \times N$ matrix G is computed as

$$G = \begin{pmatrix} \cos(\phi_1 + \phi_1) & \cos(\phi_1 + \phi_2) & \dots & \cos(\phi_1 + \phi_N) \\ \cos(\phi_2 + \phi_1) & \cos(\phi_2 + \phi_2) & \dots & \cos(\phi_2 + \phi_N) \\ \vdots & \vdots & \dots & \vdots \\ \cos(\phi_N + \phi_1) & \cos(\phi_N + \phi_2) & \dots & \cos(\phi_N + \phi_N) \end{pmatrix}.$$

Since GAFs are defined only for single channel time series, we modify this approach and use spatial embedding (see Section 1.3.2), thus obtaining a multi-channel time series $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N$. Then, we compute cosine similarities between each pair of those vectors as

$$G = \begin{pmatrix} \frac{\mathbf{x}_1 \cdot \mathbf{x}_1}{\|\mathbf{x}_1\| \|\mathbf{x}_1\|} & \frac{\mathbf{x}_1 \cdot \mathbf{x}_2}{\|\mathbf{x}_1\| \|\mathbf{x}_2\|} & \dots & \frac{\mathbf{x}_1 \cdot \mathbf{x}_N}{\|\mathbf{x}_1\| \|\mathbf{x}_N\|} \\ \frac{\mathbf{x}_2 \cdot \mathbf{x}_1}{\|\mathbf{x}_2\| \|\mathbf{x}_1\|} & \frac{\mathbf{x}_2 \cdot \mathbf{x}_2}{\|\mathbf{x}_2\| \|\mathbf{x}_2\|} & \dots & \frac{\mathbf{x}_2 \cdot \mathbf{x}_N}{\|\mathbf{x}_2\| \|\mathbf{x}_N\|} \\ \vdots & \vdots & \dots & \vdots \\ \frac{\mathbf{x}_N \cdot \mathbf{x}_1}{\|\mathbf{x}_N\| \|\mathbf{x}_1\|} & \frac{\mathbf{x}_N \cdot \mathbf{x}_2}{\|\mathbf{x}_N\| \|\mathbf{x}_2\|} & \dots & \frac{\mathbf{x}_N \cdot \mathbf{x}_N}{\|\mathbf{x}_N\| \|\mathbf{x}_N\|} \end{pmatrix}.$$

Since both recurrence plot and cosine similarity matrix are symmetric, we applied the following procedure for computing them. For a given subseries length l_s, we computed recurrence plot of subseries $2l_s$, and considered only lower left quadrant. This way, the inherent redundancy was completely removed, while preserving some of the information - the lower left quadrant contains relationships (i.e. distances or similarities) between time states occuring in the previous subseries of length l_s.

Another possibility is to learn on flattened scalp images of topographical distributions of different band powers. However, as explained in [169], this presents two main challanges. As we have verified in the previous chapter (see Section 2.3.9), the relevant variance is probably spatially global in nature, and not hierarchically compositional to make use of CNN. On the other hand, the temporal patterns are more likely to be hierarchically compositional.

3.5 Preprocessing

Signals were preprocessed before either image-encoding or direct classification. First, the electrode voltages were converted to mV to improve numerical stability. Then, optionally, a high-pass Butterworth

filter of order 3 with 4 Hz cutoff frequency was applied. As we have discussed in Section 2.2, although there may be negative effects in case of nonlinear dynamical analysis, it has been suggested that activity in the delta band (1-4 Hz) exhibits less differences between healthy and depressed subjects [96], and may possibly reduce the effects of blinking and cardiac artifacts [122]. In image-encoded case, the signals were encoded after the filtering step. Finally, Welford's algorithm for running mean and standard deviation was used to compute the mean and variance over the whole dataset, and the dataset was then centered and normalized according to the found values.

3.6 Architecture

The most successful CNN architectures we evaluated were presented in [169]. These architectures, and in particular the second one (in order of description below), were designed by the authors to be analogous to the FBCSP pipeline described in detail in Section 3.2.1.

The first architecture, called *deep* (see Figure 3.8), is more generic of the two architectures used, bearing resemblence to the architectures which proved successful in traditional computer vision tasks. It consists of four convolutional blocks with batch normalization ($\epsilon = 10^{-5}$, momentum = 0.1) and ELU nonlinearity [161], followed by max pooling and dropout ($p = 0.5$). The batch normalization was applied before the activation function. The convolution was performed only along the temporal dimension, with kernel size $(1, 3)$, stride 1. The pooling operation was also performed only along the temporal dimension, with kernel size $(1, 3)$, stride 3. For image input, we used traditional 2D convolution with kernel size $(3, 3)$. The first convolutional layer is an exception - to explicitly seperate the linear transformation into combination of temporal and spatial convolution, this layer is split into two layers with no activation function in between. First, a temporal convolution with kernel size $(1, 10)$ is performed, followed by a spatial convolution across all channels with kernel size $(19, 1)$. Note that the first operation can be seen as analogue of band-pass filtering, and the second as spatial filtering, as performed by CSP algorithm, with the difference that the filters are "constructed" by gradient descent. Batch normalization and pooling operations are also performed as described above.

The second architecture, called *shallow* (see Figure 3.9), is more specialized, tailored to mimic the transformations performed by the FBCSP pipeline. The first and only convolutional layer is split in the same way as in the deep architecture described above, and batch normalization is also applied before the activation. However, squaring nonlinearity was used as activation function for the layer instead, followed by average pooling. This can be seen as approximation of computing mean power. Moreover, following the recommendations mentioned in [169], larger kernel size $(1, 25)$ is used for the temporal filtering in this network. Then, logarithm nonlinearity is applied, analogous to the mean log-variance computation in FBCSP, see (3.3). One of the advantages of this architecture over FBCSP that it can learn the structure of temporal changes in the representation of "band-powers".

In both deep and shallow architectures, the classification is performed by classification layer with 2D convolution of kernel size of the last layer, 2 filters, and logistic activation to produce probability estimate for each class. For optimization, we used stochastic gradient descent with Nesterov momentum 0.99, decay 10^{-5}, learning rate 0.01 for batch size 128 (which we used for raw data), and learning rate 0.001 for batch size 64 (which we used for image data due to hardware limitations). This last change was made because lower batch size leads to more updates per epoch.

We also evaluated different configurations: increasing or decreasing the number of layers in the deep network, increasing or decreasing the kernel sizes, ReLU activation functions, and Adam or RMSProp optimizers. However, we found any of these changes leading to decrease in accuracy.

For image-encoded data, i.e. recurrence plots and cosine similarities, we also implemented and evaluated the architecture (along with the same hyperparameters) used in [109], which resulted in overall

accuracy of 0.942 and 0.804 recall on classification task of 6 activies using recurrence plots and CNNs (as mentioned in Section 1.6.1), evaluated using 10-fold cross validation on over 10 000 samples. However, this architectured was not more effective than the deep and shallow architectures on our dataset. This may be because of the difference in input image sizes, number of used input channels - the authors had only 4 electrodes available, and used all of them as input channels, whereas we used spatial embedding (see Section 1.3.2).

Moreover, we also evaluated multiple simpler architectures. The highest accuracy was achieved (both on image-encoded and raw data) using a VGG-like model with 3 convolution-pooling modules (convolution kernels $(3, 3)$, pooling kernels $(2, 2)$, ReLU activation functions) with 8, 16 and 16 filters respectively [170]. These were followed by dropout ($p = 0.5$), and fully connected sigmoid classification layer. This model, optimized by the RMSProp algorithm, achieved 73% accuracy on stand-out test set on raw data, and below 60% on the image-encoded data. All attempts of modifying capacity and regularization, i.e. adding batch normalization, adding or removing layers, increasing or decreasing the number of filters, as well as adding weight normalization or changing the optimizer, lead to deterioration of accuracy.

3.7 Results

3.7.1 Methodology

In this section, we describe the results obtained for the deep and shallow architectures (see Section 3.6). In order to train and evaluate the models, we used 5-fold cross validation performed as follows. Before training and evaluation the models, all the samples were shuffled and split into two parts. The first part, containing 80% of the samples, was used for training (80%) and validation (20%). The second part, containing the remaining 20% of the samples, was used as test set to evaluate the trained model. This procedure was again performed 5 times, using different initial splits into training with validation and test set. The results in Table 3.2 report the mean and standard deviation of accuracies obtained on the test sets. Figure 3.10 shows evolution of the accuracy and loss of the shallow model on the depression classification task during one cross validation cycle.

The models were trained for 200 epochs. In each epoch, the current iteration of the model was evaluated on the validation set. The model which achieved the highest accuracy on the validation set over all iterations was then selected for evaluation on the test set. This procedure, performed on the dataset described in Section 3.3, results approximately in the number of samples for each of training, validation and test sets shown in Table 3.1.

In Table 3.2, where we use SHAL to denote the shallow architecture and DEEP to denote the deep architecture (see Section 3.6). The unfiltered input is denoted $0 - f_{\text{fin}}$, and $4 - f_{\text{fin}}$ signifies that high pass filter with cutoff frequency of 4 Hz was applied to the input (see Section 3.5). The results present the following characteristics:

Architecture design relevance: The shallow architecture performs noticeably better on the prognosis task than the deep architecture.

Filtering impact: Filtering exhibits only minor effect. It seems to improve the results slightly for the prognosis task, but not on the diagnosis task.

Input pattern representation effect: Although it was suggested in [110] that features of recurrence plots may reflect depression and GAFs were shown to improve time series classification [168], both recurrence plots and cosine similarities do not seem to be effective encoding techniques for convolutional neural networks we evaluated on this dataset.

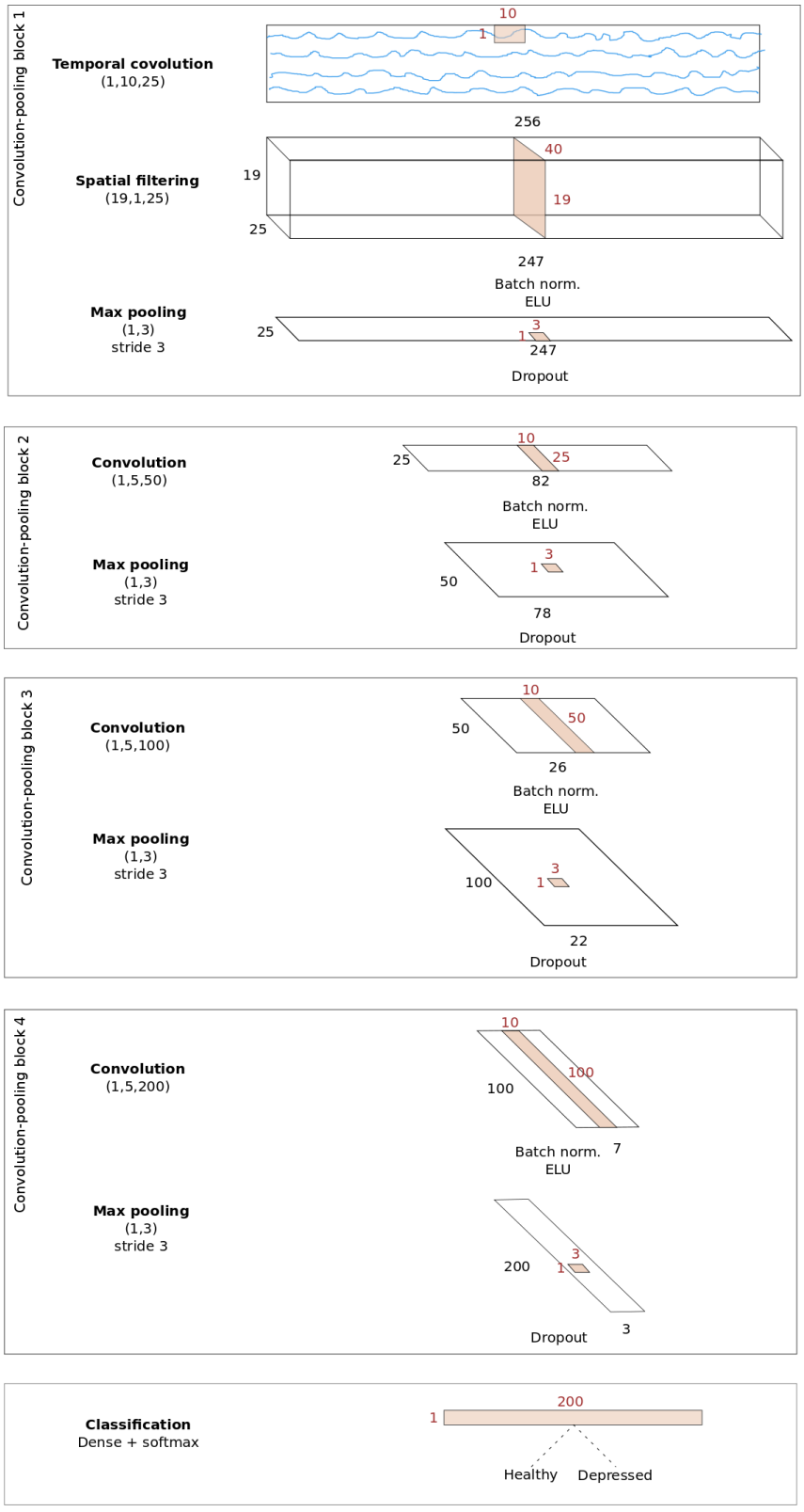


Figure 3.8: Deep architecture for evaluation on the raw data. For evaluation on image-encoded data, the kernel sizes were changed. For details, see Section 3.6.

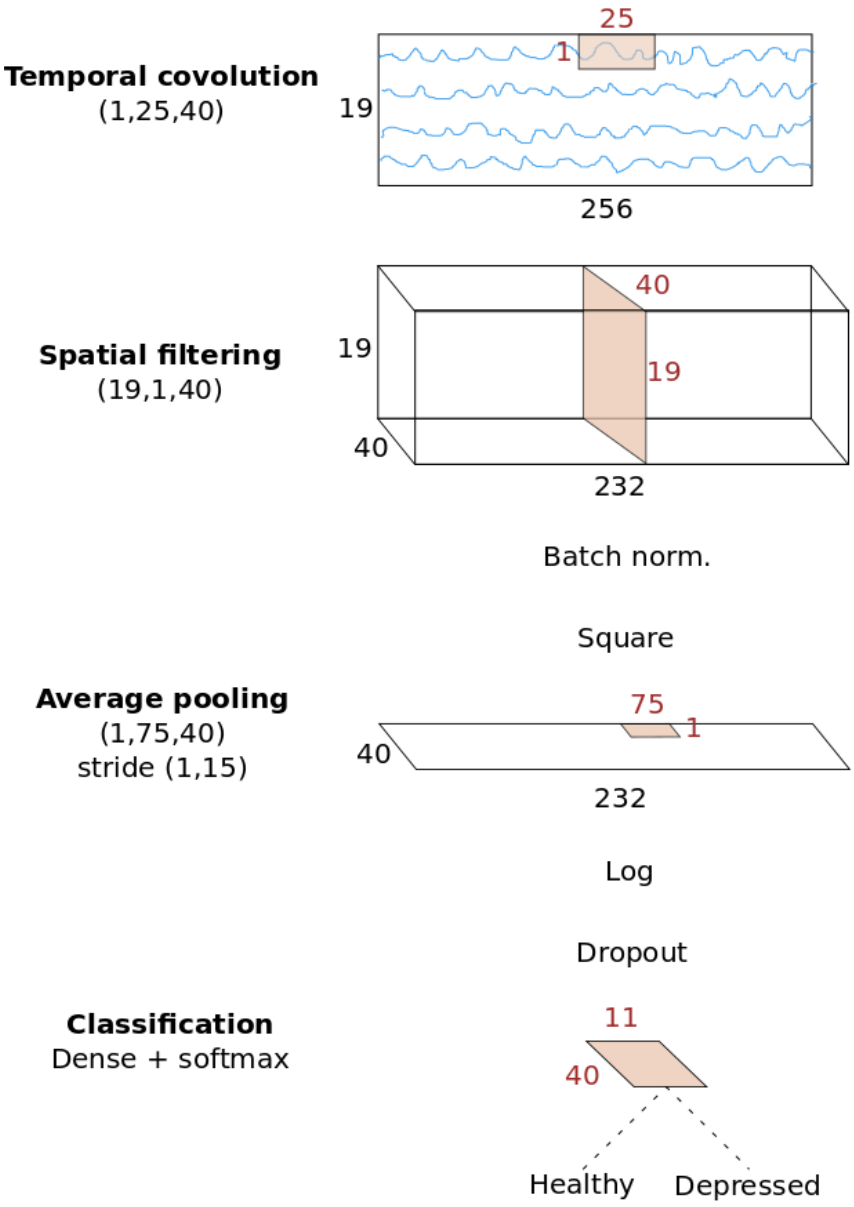


Figure 3.9: Shallow architecture, inspired by FBCSP algorithm, which achieved outstanding performance on the raw data.

Dataset	DEP		RES	
	Neg.	**Pos.**	**Neg.**	**Pos.**
Training	3278	3230	2684	2705
Validation	826	802	686	662
Test	1038	997	830	855
Overall	5142	5029	4200	4222

Table 3.1: Number of negative / positive samples in training, validation, test sets.

Label	Freq.	Arch.	Accuracy	
			Mean	**Std**
DEP	$0 - f_{\text{fin}}$	SHAL	0.85	0.13
	$4 - f_{\text{fin}}$	SHAL	0.84	0.11
	$0 - f_{\text{fin}}$	DEEP	**0.86**	0.01
	$4 - f_{\text{fin}}$	DEEP	0.85	0.02
RES	$0 - f_{\text{fin}}$	SHAL	**0.94**	0.02
	$4 - f_{\text{fin}}$	SHAL	0.94	0.03
	$0 - f_{\text{fin}}$	DEEP	0.88	0.01
	$4 - f_{\text{fin}}$	DEEP	0.86	0.02

(a) Raw data.

Label	Freq.	Meth.	Accuracy	
			Mean	**Std**
DEP	$0 - f_{\text{fin}}$	RP	**0.63**	0.02
	$4 - f_{\text{fin}}$	RP	0.61	0.01
	$0 - f_{\text{fin}}$	CS	0.59	0.02
	$4 - f_{\text{fin}}$	CS	0.58	0.01
RES	$0 - f_{\text{fin}}$	RP	0.61	0.03
	$4 - f_{\text{fin}}$	RP	**0.65**	0.02
	$0 - f_{\text{fin}}$	CS	0.55	0.02
	$4 - f_{\text{fin}}$	CS	0.63	0.01

(b) Image-encoded data.

Table 3.2: Evaluation of accuracies of the shallow (SHAL) and deep (DEEP) architectures on the raw and image-encoded data in classification of depression state (DEP) or prediction of response (RES).

3.7.2 Discussion

As evidenced in Section 2.4.1, nonlinear measures seem to be relatively stable over time for individual patients. Moreover, it has been observed that EEG signals contain other temporally stable patterns which can be used for individual identification [171]. Since the method we used for training and evaluation described in Section 3.7 involved cross validation over samples instead of patients, the test set includes distinct samples from the same patients. This raises the question whether the models indeed learn to identify the patterns contributing to the label, or whether instead they learn to identify the patterns associated with the patients, and thus they infer the label indirectly.

To investigate this question, we trained and evaluated the models as follows. The patients were split into two sets. The first set was again used for training and validation and containing 80% of the patients, and the second set, containing the remaing samples, was used for testing. Only either the recordings obtained during the first visit, or recordings obtained during the second visit, were used (this required changing the depression score thresholds to keep the classes balanced). This way, samples from patients used in the test set were never used for training. The first set was then split into training (80%) and validation (20%) sets of distinct patients, and the model was trained and evaluated on samples obtained from these distinct patients using the method described in Section 3.7. This entire procedure was repeated 5 times, resulting in 5-fold cross validation across patients.

Figure 3.11 shows loss and accuracy on the training and validation sets during one loop of the cross validation using the simpler, shallow model on the recordings obtained during the first session to predict the depression score level. In order to maximize the amount of training data, the thresholds for the classes were set to 26 as upper bound for healthy and 27 as lower bound for depressed, resulting in 63 and 62 patients in each group respectively. The figure shows that the model fits the training set of small number

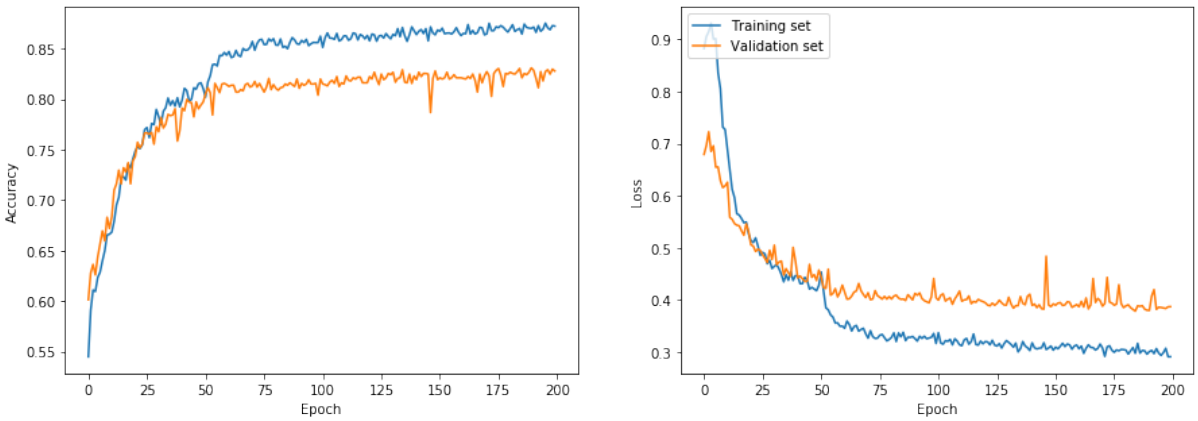


Figure 3.10: An example of accuracy and loss measured on the validation and test set during one cross validation cycle for the shallow network on the depression classification task.

Subrecording		Dataset sizes			Accuracy	
Length	**Overlap**(%)	**Train.**	**Val.**	**Test**	**Mean**	**Std**
256	0	5277	1453	1785	0.55	0.02
512	50	5202	1433	1759	0.57	0.03
1024	25	5052	1295	1707	0.55	0.02

Table 3.3: Accuracy of the shallow model on the depression classification task using raw data representation of recordings obtained before treatment. The dataset contained 63 patients labeled as healthy (depression score below 26) and 62 patients labeled as depressed (depression score above 27).

of patients, but does not generalize sufficiently to the patients the patients in the validation set. Thus, the number of training epochs before evaluation on test set was decreased. We addressed special effort in increasing the probability parameter of dropout. Nevertheless, adding regularization to the layers, increasing or shortening the recording length to either increase the amount of information in a sample or simplify the model, or increasing the overlap between samples and shifting more patients to the training set does not improve generalization. Table 3.3 shows accuracies obtained using this procedure with varying subrecording length and width.

We also changed the evaluation strategy on the test set to report percentage of successfully classified patients, where each patient was successfully classified if the majority of samples corresponding to the patient's recording was correctly classified. However, the obtained results were similar to the accuracy reached per sample. This means that the false predictions were evenly distributed across patients.

Another modification of the training procedure was evaluated. The split between training with validation and test was again performed across patients. However, the instead of splitting the training with validation set into training and validation sets across patients, we split it across samples. This way, we obtained good performance on the validation set. Nevertheless, the performance on the test set did not noticeably improve.

A possible explanation for the observed result is as follows. It seems that the relatively high performance obtained in the previous section was mostly due to small intrapersonal variance and high interpersonal variance of the patterns in the EEG signals resulting in the network being trained to map patient-characteristic patterns to labels, instead of mapping label-characteristic patterns to labels. Hence, the models overfit over a small number of epochs to the training set of small number of patients, and is

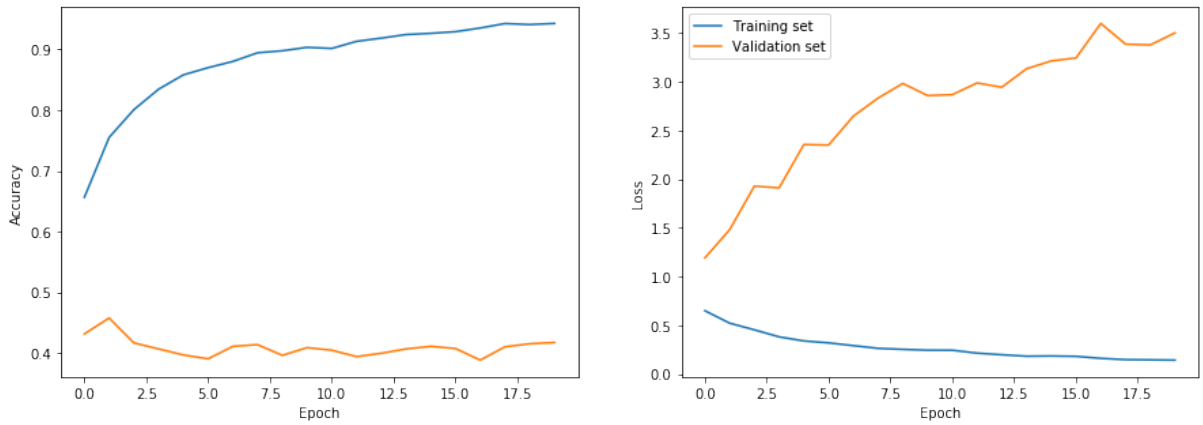


Figure 3.11: The accuracy and loss obtained for the depression classification task on training and validation sets split across patients. Only recordings acquired during the first session were used, and thus samples in each of training, validation and test sets were acquired from distinct patients. The subrecordings were non-overlapping, and the subrecording length was 256, equal to the subrecording length used for evaluation in Section 3.7. We can see that the model overfits the training set during the initial epochs, and fails to generalize to the validation set.

unable to generalize to another set of patients.

This hypothesis is supported by a considerable amount of research dedicated to finding subject specific traits in EEG [171]. There is ample evidence that certain aspects of EEG morphology are phenotypic. In [172], the authors concluded that EEG of monozygotic twins are identical, whereas EEG of dizygotic twins show lesser degree of similarity. This result has been replicated multiple times and using various methods. Furthermore, alpha peak frequency and peak frequency in occipital regions seems strongly heritable [147]. Since then, a number of EEG-based identity verification systems have been proposed [171]. If this hypothesis holds, then the observed lack of generalization may be improved by increasing the number of patients included in the dataset. Assuming that the dataset would then contain large enough aggregate of patients with small variance in patient-charecterstic patterns and, at the same time, high enough variance in depression scores and treatment response, large training accuracy could be achieved by fitting the label-characteristic patterns. Conceivably, models trained on such dataset would generalize to patients not included in the training set.

3.8 Implementation

For the implementation in this chapter, we used the same software as in Chapter 2, which was listed in Section 2.6. In addition, to design and evaluate the models, we used the deep learning library Keras.

Conclusion

In this thesis, we studied nonlinear measures and machine learning in relation to EEG signal analysis. We incorporate both thorough review of the theory of nonlinear dynamical systems in relation to EEG analysis along with overview of the relevant algorithms and techniques for nonlinear measures extraction and input parameter estimation, comprehensive literature review, and practical application of those techniques on an original dataset. Furthermore, we develop and evaluate an approach to the same task based on multilayer convolutional neural networks and objective function optimization using iterative optimization. In this chapter, we conclude with a summary of the contributions to the scientific knowledge and suggestions for future research.

Contributions

Analysis of Correlation between Depression Score and Nonlinear Measures

Investigating the significance of correlations between nonlinear measures and depression scores, we found that LLE, DFA, HE, HD, and SE computed from EEG signals recorded from EEG channels near certain cortical areas are significantly correlated with depression score. The most statistically significant correlations were found for LLE, which, was found as least discriminative in a similar experimental setting [10]. Although CD showed no significant correlations with depression score per channel, using it in combination with other nonlinear measures or by combining a number of channels resulted in one of the most accurate depression classifiers evaluated. Out of the feature selection algorithms used for selection of the most discriminative measures and channels, genetic algorithm was shown to be the most effective one.

All the studies reviewed in Section 1.6.1 included a control group of symptomless, clinically healthy patients, whereas our dataset consists exclusively of patients suffering from MDD of various symptom severity. Nevertheless, the accuracies achieved by the classifiers are substantial, indicating possible relationship between depression severity and values of nonlinear measures recorded in EEG signals obtained from various cortical regions, and opening a possibility for developing a regression model predicting depression score making use of nonlinear measures.

Largest Lyapunov Exponents are Predictive of Positive Treatment Response

As we outlined in Section 1.6.2, most of the EEG-based treatment outcome prediction has been focused on quantifying frontal θ band. This approach leans on strong neuroscientific support. In the analysis performed in Section 2.4.3, we found significant differences in distributions of LLE estimated from signals recorded in channels across multitude of cortical areas, especially near frontal (F3, F4) and left temporal (T6) areas. Almost no significant differences were found using other nonlinear measures. Interestingly, although LLE was positively correlated depression score, it is negatively correlated with

treatment response. In other words, depressed patients exhibit slightly higher chaoticity than healthy patients (see Section 2.4.2), but out of those depressed patients, those with lower chaoticity at the beginning of the treatment may be more likely to respond positively to that treatment (see Section 2.4.3). However, note that [11] observed higher predictability, and thus lower complexity in EEG signals of depressed patients using different methods.

Indeed, in Section 2.5.3, LLE was shown to be potent feature for treatment outcome prediction classifiers. These results may encourage further exploration of the potential use of nonlinear dynamical measures for treatment outcome prediction, which, to our knowledge, is nonexistent at present.

Analysis of Spatial Distribution of Nonlinear Measures across Brain Regions in Depression

Although some neuroscientific theories suggest that medical depression is caused by malfuction in the right hemisphere [173, 174], in Sections 2.4.2 and 2.5, we found nonlinear measures computed in both hemispheres contribute significantly to high depression score. The distributions of correlation significance is notably distinct between measures, possibly indicating that each measure captures different dynamical and spatial aspects of the brain activity correlating with depression severity. For example, DFA and HE show most significant correlations with depression score in right parietal regions, LLE and SE in right temporal regions. Nevertheless, the most successful depression classifiers made use especially of channels in temporal areas (T6 in particular).

Analysis of Nonlinear Measure and Input Parameter Estimation Algorithms and Procedures for EEG Analysis

Since computation of nonlinear measures requires laborious selection of input parameters and knowledge of the algorithms in order to interpret the results, these methods often remain inaccessible to practitioners who has not yet become experts in the field of nonlinear dynamical analysis [137]. Therefore, there is a need for automated procedures to ease application of these methods and increase their use and availability. For the purpose of computing LLE and CD, in Section 2.3, we evaluated a range of algorithms for estimation of embedding dimension and time delay parameters, including implementations of automated pipelines for computation of said measures. Although these input parameter estimation algorithms remain widely used (see Sections 2.3.3.3 and 2.3.4.3), we found that, for our dataset, their outputs varied widely for each method, and provided only a crude approximation for parameters which produced more discriminative measures. The most consistent and informative estimation algorithm was ILD of our own implementation. However, the procedure suffers from relatively high computational costs and potentially could be optimized.

Evaluation of FBCSP-inspired Neural Network Architectures for Depression Diagnosis and Prognosis

In Section 3.7, we evaluated two CNN architectures which were previously shown to be successfully applied to decoding task-related information [169, 175]. The models achieved relatively high accuracy when trained and evaluated using cross validation across samples. However, when trained and evaluated using cross validation across recordings obtained before treatment from individual, distinct patients, the models achieved only *modest* accuracy, and did not generalize to the test set (see Section 3.7.2). This observation may be explained by low intrapersonal, and high intrapersonal variance of the EEG signals, which results in high correlation of samples corresponding to one patient, and thus in the presence of high number of samples from small number of patients, the model is trained to associate labels to samples

based on the patient-charecterstic patterns, instead of label-characteristic patterns. Indeed, a number of studies suggest presence of phenotypic traits in EEG, and personal identification using EEG is a flourishing field [171] (see Section 3.7.2). If this hypothesis holds, then performance of these models may be improved by extending the dataset with enough patients with similar patient-characteritic patterns and different label-characteristic patterns. Such dataset may be obtained by tracking patients' depression score across extended time span.

Recommendations for Future Work

The nature of our dataset restricts us to analysis of patients suffering from depressive symptoms to varying degrees. Although obtaining accuracy lower in comparison with studies employing healthy, symptom-free group as well as depressed patients is justifiable, we believe that this accuracy can still be significantly improved with more advanced noise-reduction preprocessing techniques, and recording environment insulated from power-line, electronic devices, and other sources of electromagnetic interference. Moreover, the patients should be advised to keep their eyes closed, an ECG may recorded simulatenously and later used to remove interference from the heart muscle. Potentially, machine learning techniques may be used improve the signal quality by continuously estimating it, and even detecting events such as lost electrode contact [169].

Due to the fact that computing nonlinear dynamical measures on a nonstationary signal is not theoretically justified (see Section 1.2.3), another possibility for improving the estimation is by acquiring longer recordings, and computing the nonlinear measures on short, overlapping time windows [15]. In spite of the reduction in estimation accuracy (see Sections 1.4.1.2 and 1.4.2.2), it may be beneficial in case of subtantial nonstationarity (which is common in EEG recordings), since over short enough time interval, even severely nonstationary time series may be regarded as stationary [15]. Although out of the main scope of study, we already initiated this analysis.

Nevertheless, the approach to nonlinear dynamical analysis we employed in this study is not conceptually ideal in itself. It has been shown repeatedly that the initial hypothesis of low dimensional attractor explaining braing dynamics is incorrect [37, 117, 15].[5] Although this does not mean that these measures are meaningless [60] (see Section 1.6.3), development of novel nonlinear measures reflecting the advancements in understanding of brain dynamics is called for [5]. For example, it has been suggested that measures of nonlinear coupling between time series, reflecting the discovery of generalized synchronization, may be considerably more relevant than local measures studying properties of a low-dimensional attractor [5].

We hypothesized that generating a dataset containing multiple recordings obtained from the same individuals with different depression scores, e.g. by tracking long-term changes in depression score while recording EEG, may improve classification of deep learning models. At the same time, tracking long-term reactions to treatment using carefully recorded EEG signals and depression scores may facilitate development of tools which advise on the optimal patient-specific treatment method. Moreover, a common problem in treatment of depressive patients is frequent disease relapse. Hence, this may provide further motivation for continuous tracking of depression score using EEG for the purpose of early relapse detection, resulting in positive feedback loop and progressive improvement in accuracy.

Another potential avenue for improvement of the deep learning approach may lie in deep learning architectures using modalities from nonlinear measures and other biomarkers, or in architectures inspired by the estimation algorithms of those measures.

[5] With possible exeption of brain dynamics in epileptic seizures [5].

Finally, we found significant differences between LLE values for patients responding and nonresponding to treatment, as well as developed simple, yet relatively accurate response prediction models based on nonlinear measures. Our hope is that these findings may spark interest in nonlinear measures for prediction of treatment outcome.

Bibliography

[1] Germán Rodríguez-Bermúdez and Pedro J García-Laencina. Analysis of EEG Signals using Nonlinear Dynamics and Chaos : A review. *Applied Mathematics & Information Sciences*, 9(5):2309–2321, 2015.

[2] World Heatlh Organization. Depression. `http://www.who.int/en/news-room/fact-sheets/detail/depression`, 2018. [Online; accessed 18-August-2018].

[3] Katie M Smith, Perry F Renshaw, and John Bilello. The diagnosis of depression: current and emerging methods. *Comprehensive psychiatry*, 54(1):1–6, 2013.

[4] Sebastian Olbrich and Martijn Arns. Eeg biomarkers in major depressive disorder: discriminative power and prediction of treatment response. *International Review of Psychiatry*, 25(5):604–618, 2013.

[5] C. J. Stam. Nonlinear dynamical analysis of EEG and MEG: Review of an emerging field. *Clinical Neurophysiology*, 116(10):2266–2301, 2005.

[6] Oliver Faust, Peng Chuan Alvin Ang, Subha D Puthankattil, and Paul K Joseph. Depression diagnosis support system based on eeg signal entropies. *Journal of mechanics in medicine and biology*, 14(03):1450035, 2014.

[7] U Rajendra Acharya, Vidya K Sudarshan, Hojjat Adeli, Jayasree Santhosh, Joel EW Koh, Subha D Puthankatti, and Amir Adeli. A novel depression diagnosis index using nonlinear features in eeg signals. *European neurology*, 74(1-2):79–83, 2015.

[8] Mehran Ahmadlou, Hojjat Adeli, and Amir Adeli. Fractality analysis of frontal brain in major depressive disorder. *International Journal of Psychophysiology*, 85(2):206–211, 2012.

[9] Maie Bachmann, Jaanus Lass, Anna Suhhova, and Hiie Hinrikus. Spectral asymmetry and higuchi's fractal dimension measures of depression electroencephalogram. *Computational and mathematical methods in medicine*, 2013, 2013.

[10] Behshad Hosseinifard, Mohammad Hassan Moradi, and Reza Rostami. Classifying depression patients and normal subjects using machine learning techniques and nonlinear features from eeg signal. *Computer methods and programs in biomedicine*, 109(3):339–345, 2013.

[11] Jean Louis Nandrino, Laurent Pezard, Jacques Martinerie, Farid El Massioui, Bernard Renault, Roland Jouvent, Jean François Allilaire, and Daniel Widlöcher. Decrease of complexity in EEG as a symptom of depression. *NeuroReport*, 5(4):528–530, 1994.

[12] J Röschke, Juergen Fell, and P Beckmann. Nonlinear analysis of sleep eeg in depression: calculation of the largest lyapunov exponent. *European archives of psychiatry and clinical neuroscience*, 245(1):27–35, 1995.

[13] Jun-Seok Lee, Byung-Hwan Yang, Jang-Han Lee, Jun-Ho Choi, Ihn-Geun Choi, and Sae-Byul Kim. Detrended fluctuation analysis of resting eeg in depressed outpatients and healthy controls. *Clinical Neurophysiology*, 118(11):2489–2496, 2007.

[14] James Meiss. Dynamical systems. `http://www.scholarpedia.org/article/dynamical_systems`, 2007. [online; accessed 6-March-2019].

[15] Galka Andreas. *Topics in nonlinear time series analysis, with implications for EEG analysis*, volume 14. World Scientific, 2000.

[16] Paul L Nunez, Ramesh Srinivasan, et al. *Electric fields of the brain: the neurophysics of EEG*. Oxford University Press, USA, 2006.

[17] Ramesh Srinivasan. Methods to improve the spatial resolution of eeg. *International Journal of Bioelectromagnetism*, 1(1):102–111, 1999.

[18] Paul M Vespa, Val Nenov, and Marc R Nuwer. Continuous eeg monitoring in the intensive care unit: early findings and clinical efficacy. *Journal of Clinical Neurophysiology*, 16(1):1–13, 1999.

[19] Matti Hämäläinen, Riitta Hari, Risto J Ilmoniemi, Jukka Knuutila, and Olli V Lounasmaa. Magnetoencephalography—theory, instrumentation, and applications to noninvasive studies of the working human brain. *Reviews of modern Physics*, 65(2):413, 1993.

[20] Teal L Schultz. Technical tips: Mri compatible eeg electrodes: advantages, disadvantages, and financial feasibility in a clinical setting. *The Neurodiagnostic Journal*, 52(1):69–81, 2012.

[21] Kieran J Murphy and James A Brunberg. Adult claustrophobia, anxiety and sedation in mri. *Magnetic resonance imaging*, 15(1):51–54, 1997.

[22] Ryan T Canolty, Erik Edwards, Sarang S Dalal, Maryam Soltani, Srikantan S Nagarajan, Heidi E Kirsch, Mitchel S Berger, Nicholas M Barbaro, and Robert T Knight. High gamma power is phase-locked to theta oscillations in human neocortex. *science*, 313(5793):1626–1628, 2006.

[23] György Buzsáki and Andreas Draguhn. Neuronal oscillations in cortical networks. *science*, 304(5679):1926–1929, 2004.

[24] Vladimir Shusterman and William C Troy. From baseline to epileptiform activity: a path to synchronized rhythmicity in large-scale neural networks. *Physical Review E*, 77(6):061911, 2008.

[25] JH McAuley and CD Marsden. Physiological and pathological tremors and rhythmic central motor control. *Brain*, 123(8):1545–1567, 2000.

[26] Rodolfo R Llinás, Urs Ribary, Daniel Jeanmonod, Eugene Kronberg, and Partha P Mitra. Thalamocortical dysrhythmia: a neurological and neuropsychiatric syndrome characterized by magnetoencephalography. *Proceedings of the National Academy of Sciences*, 96(26):15222–15227, 1999.

[27] Sven Vanneste, Jae-Jin Song, and Dirk De Ridder. Thalamocortical dysrhythmia detected by machine learning. *Nature communications*, 9(1):1103, 2018.

[28] Maurice Bertram Priestley. Non-linear and non-stationary time series analysis. 1988.

[29] Jianqing Fan and Qiwei Yao. *Nonlinear time series: nonparametric and parametric methods*. Springer Science & Business Media, 2008.

[30] Alexander Ya Kaplan, Andrew A Fingelkurts, Alexander A Fingelkurts, Sergei V Borisov, and Boris S Darkhovsky. Nonstationary nature of the brain activity as revealed by eeg/meg: methodological, practical and conceptual challenges. *Signal processing*, 85(11):2190–2212, 2005.

[31] John Lee. *Introduction to topological manifolds*, volume 202. Springer Science & Business Media, 2010.

[32] Jürgen Schmidhuber. Don't forget randomness is still just a hypothesis. `http://people.idsia.ch/~juergen/randomness.html`, 2006. [online; accessed 6-March-2019].

[33] Steven H Strogatz and Donald E Herbert. Nonlinear dynamics and chaos. *Medical Physics-New York-Institute of Physics*, 23(6):993–995, 1996.

[34] Holger Kantz and Thomas Schreiber. *Nonlinear time series analysis*, volume 7. Cambridge university press, 2004.

[35] Peter Grassberger and Itamar Procaccia. Measuring the strangeness of strange attractors. *Physica D: Nonlinear Phenomena*, 9(1-2):189–208, 1983.

[36] Michael C Mackey and Leon Glass. Oscillation and chaos in physiological control systems. *Science*, 197(4300):287–289, 1977.

[37] James E Skinner. Low-dimensional chaos in biological systems. *Bio/technology*, 12(6):596, 1994.

[38] David Ruelle and Floris Takens. On the nature of turbulence. *Les rencontres physiciens-mathématiciens de Strasbourg-RCP25*, 12:1–44, 1971.

[39] Kenneth Falconer. *Fractal geometry: mathematical foundations and applications*. John Wiley & Sons, 2004.

[40] Heinz Isliker and Juergen Kurths. A test for stationarity: finding parts in time series apt for correlation dimension estimates. *International Journal of Bifurcation and Chaos*, 3(06):1573–1579, 1993.

[41] J-P Eckmann and David Ruelle. Ergodic theory of chaos and strange attractors. In *The Theory of Chaotic Attractors*, pages 273–312. Springer, 1985.

[42] J.-P Eckmann, S. Oliffson Kamphorst, and D Ruelle. Recurrence Plots of Dynamical Systems. *Europhysics Letters (EPL)*, 4(9):973–977, 1987.

[43] Radu Manuca and Robert Savit. Stationarity and nonstationarity in time series analysis. *Physica D: Nonlinear Phenomena*, 99(2-3):134–161, 1996.

[44] Norbert Marwan, M Carmen Romano, Marco Thiel, and Jürgen Kurths. Recurrence plots for the analysis of complex systems. *Physics reports*, 438(5-6):237–329, 2007.

[45] Norbert Marwan. How to avoid potential pitfalls in recurrence plot based data analysis. *International Journal of Bifurcation and Chaos*, 21(04):1003–1017, 2011.

[46] Dimitris Kugiumtzis. State space reconstruction parameters in the analysis of chaotic time series-the role of the time window length. *arXiv preprint comp-gas/9602002*, 1996.

[47] Andrew M Fraser. Reconstructing attractors from scalar time series: A comparison of singular system and redundancy criteria. *Physica D: Nonlinear Phenomena*, 34(3):391–404, 1989.

[48] Walter S Pritchard, Kelly K Krieble, and Dennis W Duke. On the validity of estimating eeg correlation dimension from a spatial embedding. *Psychophysiology*, 33(4):362–368, 1996.

[49] Laurent Pezard, Jean-Philippe Lachaux, Nitza Thomasson, and Jacques Martinerie. Why bother to spatially embed eeg? comments on pritchard et al., psychophysiology, 33, 362–368, 1996. *Psychophysiology*, 36(4):527–531, 1999.

[50] Jean-Philippe Lachaux, Laurent Pezard, Line Garnero, Christophe Pelte, Bernard Renault, Francisco J Varela, and Jacques Martinerie. Spatial extension of brain activity fools the single-channel reconstruction of eeg dynamics. *Human brain mapping*, 5(1):26–47, 1997.

[51] Holger Kantz and Eckehard Olbrich. Scalar observations from a class of high-dimensional chaotic systems: Limitations of the time delay embedding. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 7(3):423–429, 1997.

[52] Hassler Whitney. Differentiable manifolds. *Annals of Mathematics*, pages 645–680, 1936.

[53] Norman H Packard, James P Crutchfield, J Doyne Farmer, and Robert S Shaw. Geometry from a time series. *Physical review letters*, 45(9):712, 1980.

[54] Floris Takens. Detecting strange attractors in turbulence. In *Dynamical systems and turbulence, Warwick 1980*, pages 366–381. Springer, 1981.

[55] Timothy D. Sauer. Attractor reconstruction. `http://www.scholarpedia.org/article/Attractor_reconstruction`, 2006. [Online; accessed 28-November-2018].

[56] Tim Sauer, James A Yorke, and Martin Casdagli. Embedology. *Journal of statistical Physics*, 65(3-4):579–616, 1991.

[57] James Theiler. Estimating fractal dimension. *JOSA A*, 7(6):1055–1073, 1990.

[58] Martin Casdagli, Stephen Eubank, J Doyne Farmer, and John Gibson. State space reconstruction in the presence of noise. *Physica D: Nonlinear Phenomena*, 51(1-3):52–98, 1991.

[59] Holger Kantz and Eckehard Olbrich. Scalar observations from a class of high-dimensional chaotic systems: Limitations of the time delay embedding. *Chaos*, 7(3):423–429, 1997.

[60] AM Albano and PE Rapp. On the reliability of dynamical measures of eeg signals. In *The 2nd Annual Conference on Nonlinear Dynamics Analysis of the EEG, World Scientific, Singapore*, pages 117–139, 1993.

[61] Henry Abarbanel. *Analysis of observed chaotic data*. Springer Science & Business Media, 2012.

[62] Anna Krakovská, Kristína Mezeiová, and Hana Budáčová. Use of false nearest neighbours for selecting variables and embedding parameters for state space reconstruction. *Journal of Complex Systems*, 2015, 2015.

[63] Andrew M Fraser and Harry L Swinney. Independent coordinates for strange attractors from mutual information. *Physical review A*, 33(2):1134, 1986.

[64] JM Martinerie, Alfonso M Albano, AI Mees, and PE Rapp. Mutual information, strange attractors, and the optimal estimation of dimension. *Physical Review A*, 45(10):7058, 1992.

[65] Michael T Rosenstein, James J Collins, and Carlo J De Luca. Reconstruction expansion as a geometry-based framework for choosing proper delay times. *Physica D: Nonlinear Phenomena*, 73(1-2):82–98, 1994.

[66] AI Mees, PE Rapp, and LS Jennings. Singular-value decomposition and embedding dimension. *Physical Review A*, 36(1):340, 1987.

[67] Th Buzug and G Pfister. Optimal delay time and embedding dimension for delay-time coordinates by analysis of the global static and local dynamical behavior of strange attractors. *Physical review A*, 45(10):7073, 1992.

[68] Matthew B Kennel, Reggie Brown, and Henry DI Abarbanel. Determining embedding dimension for phase-space reconstruction using a geometrical construction. *Physical review A*, 45(6):3403, 1992.

[69] Liangyue Cao. Practical method for determining the minimum embedding dimension of a scalar time series. *Physica D: Nonlinear Phenomena*, 110(1-2):43–50, 1997.

[70] Peter Grassberger, Thomas Schreiber, and Carsten Schaffrath. Nonlinear time sequence analysis. *International journal of bifurcation and chaos*, 1(03):521–547, 1991.

[71] J-P Eckmann, S Oliffson Kamphorst, David Ruelle, and S Ciliberto. Liapunov exponents from time series. *Physical Review A*, 34(6):4971, 1986.

[72] Joachim Röschke, Jürgen Fell, and Peter Beckmann. Nonlinear analysis of sleep eeg data in schizophrenia: calculation of the principal lyapunov exponent. *Psychiatry research*, 56(3):257–269, 1995.

[73] Michael T. Rosenstein, James J. Collins, and Carlo J. De Luca. A practical method for calculating largest Lyapunov exponents from small data sets. *Physica D: Nonlinear Phenomena*, 65(1-2):117–134, 1993.

[74] Jürgen Fell and Peter E Beckmann. Resonance-like phenomena in lyapunov calculations from data reconstructed by the time-delay method. *Physics Letters A*, 190(2):172–176, 1994.

[75] Holger Kantz. A robust method to estimate the maximal lyapunov exponent of a time series. *Physics letters A*, 185(1):77–87, 1994.

[76] Alan Wolf, Jack B Swift, Harry L Swinney, and John A Vastano. Determining lyapunov exponents from a time series. *Physica D: Nonlinear Phenomena*, 16(3):285–317, 1985.

[77] J-P Eckmann and David Ruelle. Fundamental limitations for estimating dimensions and lyapunov exponents in dynamical systems. *Physica D: Nonlinear Phenomena*, 56(2-3):185–187, 1992.

[78] Peter Grassberger. Grassberger-Procaccia algorithm. `http://www.scholarpedia.org/article/Grassberger-Procaccia_algorithm`, 2007. [Online; accessed 20-December-2018].

[79] Richard Hardstone, Simon-Shlomo Poil, Giuseppina Schiavone, Rick Jansen, Vadim V Nikulin, Huibert D Mansvelder, and Klaus Linkenkaer-Hansen. Detrended fluctuation analysis: a scale-free view on neuronal oscillations. *Frontiers in physiology*, 3:450, 2012.

[80] Jan Beran. *Statistics for long-memory processes*. Routledge, 2017.

[81] Klaus Linkenkaer-Hansen, Simo Monto, Heikki Rytsälä, Kirsi Suominen, Erkki Isometsä, and Seppo Kähkönen. Breakdown of long-range temporal correlations in theta oscillations in patients with major depressive disorder. *Journal of Neuroscience*, 25(44):10131–10137, 2005.

[82] Harold E Hurst. The problem of long-term storage in reservoirs. *Hydrological Sciences Journal*, 1(3):13–27, 1956.

[83] Harold Edwin Hurst. A suggested statistical model of some time series which occur in nature. *Nature*, 180(4584):494, 1957.

[84] Rafał Weron. Estimating long-range dependence: finite sample properties and confidence intervals. *Physica A: Statistical Mechanics and its Applications*, 312(1-2):285–299, 2002.

[85] Benoit B Mandelbrot. Limit theorems on the self-normalized range for weakly and strongly dependent processes. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 31(4):271–285, 1975.

[86] Jens Feder. *Fractals*. Springer Science & Business Media, 2013.

[87] Tomoyuki Higuchi. Approach to an irregular time series on the basis of the fractal theory. *Physica D: Nonlinear Phenomena*, 31(2):277–283, 1988.

[88] Joshua S Richman and J Randall Moorman. Physiological time-series analysis using approximate entropy and sample entropy. *American Journal of Physiology-Heart and Circulatory Physiology*, 278(6):H2039–H2049, 2000.

[89] Kyongsik Yun, Hee-Kwon Park, Do-Hoon Kwon, Yang-Tae Kim, Sung-Nam Cho, Hyun-Jin Cho, Bradley S Peterson, and Jaeseung Jeong. Decreased cortical complexity in methamphetamine abusers. *Psychiatry Research: Neuroimaging*, 201(3):226–232, 2012.

[90] Erik R Hauge, Jan Øystein Berle, Ketil J Oedegaard, Fred Holsten, and Ole Bernt Fasmer. Nonlinear analysis of motor activity shows differences between schizophrenia and depression: a study using fourier analysis and sample entropy. *PloS one*, 6(1):e16291, 2011.

[91] Eugene N Bruce, Margaret C Bruce, and Swetha Vennelaganti. Sample entropy tracks changes in eeg power spectrum with sleep state and aging. *Journal of clinical neurophysiology: official publication of the American Electroencephalographic Society*, 26(4):257, 2009.

[92] James Theiler, Stephen Eubank, André Longtin, Bryan Galdrikian, and J Doyne Farmer. Testing for nonlinearity in time series: the method of surrogate data. *Physica D: Nonlinear Phenomena*, 58(1-4):77–94, 1992.

[93] James Theiler. On the evidence for low-dimensional chaos in an epileptic electroencephalogram. *Physics Letters A*, 196(1-2):335–341, 1994.

[94] Laurent Pezard, Jean Louis Nandrino, Bernard Renault, Farid El Massioui, Jean François Allilaire, Johannes Müller, Francisco J. Varela, and Jacques Martinerie. Depression as a dynamical disease. *Biological Psychiatry*, 39(12):991–999, 1996.

[95] Martijn Arns, Alexander Cerquera, Rafael M Gutiérrez, Fred Hasselman, and Jan A Freund. Non-linear eeg analyses predict non-response to rtms treatment in major depressive disorder. *Clinical Neurophysiology*, 125(7):1392–1399, 2014.

[96] Verner Knott, Colleen Mahoney, Sidney Kennedy, and Kenneth Evans. Eeg power, frequency, asymmetry and coherence in male depression. *Psychiatry Research: Neuroimaging*, 106(2):123–140, 2001.

[97] Stefan Debener, André Beauducel, Doreen Nessler, Burkhard Brocke, Hubert Heilemann, and Jürgen Kayser. Is resting anterior eeg alpha asymmetry a trait marker for depression? *Neuropsychobiology*, 41(1):31–37, 2000.

[98] VP Omel'Chenko and VG Zaika. Changes in the eeg-rhythms in endogenous depressive disorders and the effect of pharmacotherapy. *Human Physiology*, 28(3):275–281, 2002.

[99] Dan V Iosifescu, Scott Greenwald, Philip Devlin, David Mischoulon, John W Denninger, Jonathan E Alpert, and Maurizio Fava. Frontal eeg predictors of treatment outcome in major depressive disorder. *European Neuropsychopharmacology*, 19(11):772–777, 2009.

[100] Hiie Hinrikus, Anna Suhhova, Maie Bachmann, Kaire Aadamsoo, Ülle Võhma, Jaanus Lass, and Viiu Tuulik. Electroencephalographic spectral asymmetry index for detection of depression. *Medical & biological engineering & computing*, 47(12):1291, 2009.

[101] John JB Allen, Heather L Urry, Sabrina K Hitt, and James A Coan. The stability of resting frontal electroencephalographic asymmetry in depression. *Psychophysiology*, 41(2):269–280, 2004.

[102] Mehran Ahmadlou, Hojjat Adeli, and Amir Adeli. Spatiotemporal analysis of relative convergence of eegs reveals differences between brain dynamics of depressive women and men. *Clinical EEG and neuroscience*, 44(3):175–181, 2013.

[103] Christian Gold, Jörg Fachner, and JAAKKO ERKKILÄ. Validity and reliability of electroencephalographic frontal alpha asymmetry and frontal midline theta as biomarkers for depression. *Scandinavian Journal of Psychology*, 54(2):118–126, 2013.

[104] Martijn Arns, Wilhelmus H Drinkenburg, Paul B Fitzgerald, and J Leon Kenemans. Neurophysiological predictors of non-response to rtms in depression. *Brain Stimulation*, 5(4):569–576, 2012.

[105] David Nutt, Sue Wilson, and Louise Paterson. Sleep disorders as core symptoms of depression. *Dialogues in clinical neuroscience*, 10(3):329, 2008.

[106] A Babloyantz. Strange attractors in the dynamics of brain activity. In *Complex systems—Operational approaches in neurobiology, physics, and computers*, pages 116–122. Springer, 1985.

[107] Jan Pieter M Pijn, Demetrios N Velis, Marcel J van der Heyden, Jaap DeGoede, Cees WM van Veelen, and Fernando H Lopes da Silva. Nonlinear dynamics of epileptic seizures on basis of intracranial eeg recordings. *Brain topography*, 9(4):249–270, 1997.

[108] Fatemeh Bahari and Amin Janghorbani. Eeg-based emotion recognition using recurrence plot analysis and k nearest neighbor classifier. In *Biomedical Engineering (ICBME), 2013 20th Iranian Conference on*, pages 228–233. IEEE, 2013.

[109] Enrique Garcia-Ceja, Md Zia Uddin, and Jim Torresen. Classification of recurrence plots' distance matrices with a convolutional neural network for activity recognition. *Procedia computer science*, 130(C):157–163, 2018.

[110] U Rajendra Acharya, Vidya K Sudarshan, Hojjat Adeli, Jayasree Santhosh, Joel EW Koh, and Amir Adeli. Computer-aided diagnosis of depression using eeg signals. *European neurology*, 73(5-6):329–336, 2015.

[111] Subha D Puthankattil and Paul K Joseph. Classification of eeg signals in normal and depression conditions by ann using rwe and signal entropy. *Journal of Mechanics in Medicine and Biology*, 12(04):1240019, 2012.

[112] Aimee M Hunter, Ian A Cook, and Andrew F Leuchter. The promise of the quantitative electroencephalogram as a predictor of antidepressant treatment outcomes in major depressive disorder. *Psychiatric Clinics of North America*, 30(1):105–124, 2007.

[113] Craig E Tenke, Jürgen Kayser, Carlye G Manna, Shiva Fekri, Christopher J Kroppmann, Jennifer D Schaller, Daniel M Alschuler, Jonathan W Stewart, Patrick J McGrath, and Gerard E Bruder. Current source density measures of electroencephalographic alpha predict antidepressant treatment response. *Biological psychiatry*, 70(4):388–394, 2011.

[114] Ahmad Khodayari-Rostamabad, James P Reilly, Gary M Hasey, Hubert de Bruin, and Duncan J MacCrimmon. A machine learning approach using eeg data to predict response to ssri treatment for major depressive disorder. *Clinical Neurophysiology*, 124(10):1975–1985, 2013.

[115] Peter Grassberger. Do climatic attractors exist? *Nature*, 323(6089):609, 1986.

[116] Itamar Procaccia. Complex or just complicated? *Nature*, 333:498–499, 1988.

[117] Paul E Rapp. Is there evidence for chaos in the human central nervous system. *Chaos theory in psychology and the life sciences*, 89:100, 1995.

[118] Janet BW Williams and Kenneth A Kobak. Development and reliability of a structured interview guide for the montgomery-åsberg depression rating scale (sigma). *The British Journal of Psychiatry*, 192(1):52–58, 2008.

[119] N Herrmann, SE Black, J Lawrence, C Szekely, and JP Szalai. The sunnybrook stroke study: a prospective study of depressive symptoms and functional outcome. *Stroke*, 29(3):618–624, 1998.

[120] Asangi. Electrode locations of International 10-20 system for EEG (electroencephalography) recording . `https://commons.wikimedia.org/wiki/File:21_electrodes_of_International_10-20_system_for_EEG.svg`, 2010. [Online; accessed 18-March-2019].

[121] Ahmad Diab, Mahmoud Hassan, Brynjar Karlsson, and Catherine Marque. Effect of decimation on the classification rate of non-linear analysis methods applied to uterine emg signals. *IRBM*, 34(4-5):326–329, 2013.

[122] Gabriele Gratton. Dealing with artifacts: The eog contamination of the event-related brain potential. *Behavior Research Methods, Instruments, & Computers*, 30(1):44–53, 1998.

[123] J Mateo, Eva M Sánchez-Morla, and JL Santos. A new method for removal of powerline interference in ecg and eeg recordings. *Computers & Electrical Engineering*, 45:235–248, 2015.

[124] Nicholas Rohrbacker. Analysis of Electroencephologram Data Using Time-Delay Embeddings to Reconstruct Phase Space. *Dynamics at the Horsetooth*, 1:1–11, 2009.

[125] Mainak Jas, Eric Larson, Denis-Alexander Engemann, Jaakko Leppakangas, Samu Taulu, Matti Hamalainen, and Alexandre Gramfort. MEG/EEG group study with MNE: recommendations, quality assessments and best practices. *bioRxiv*, page 240044, 2017.

[126] Atin Das, Pritha Das, and AB Roy. Applicability of lyapunov exponent in eeg data analysis. *Complexity International*, 9(das01):1–8, 2002.

[127] Agnessa Babloyantz and Alain Destexhe. Low-dimensional chaos in an instance of epilepsy. *Proceedings of the National Academy of Sciences*, 83(10):3513–3517, 1986.

[128] S Blanco, H Garcia, R Quian Quiroga, L Romanelli, and OA Rosso. Stationarity of the eeg series. *IEEE Engineering in medicine and biology Magazine*, 14(4):395–399, 1995.

[129] Dominique Gallez and Agnessa Babloyantz. Predictability of human eeg: a dynamical approach. *Biological Cybernetics*, 64(5):381–391, 1991.

[130] Andrew D Krystal, Craig Zaidman, Henry S Greenside, Richard D Weiner, and C Edward Coffey. The largest lyapunov exponent of the eeg during ect seizures as a measure of ect seizure adequacy. *Electroencephalography and clinical neurophysiology*, 103(6):599–606, 1997.

[131] Jürgen Fell, Joachim Röschke, and P Beckmann. Deterministic chaos and the first positive lyapunov exponent: a nonlinear analysis of the human electroencephalogram during sleep. *Biological cybernetics*, 69(2):139–146, 1993.

[132] J Röschke, J Fell, and P Beckmann. The calculation of the first positive lyapunov exponent in sleep eeg data. *Electroencephalography and clinical neurophysiology*, 86(5):348–352, 1993.

[133] Ljubomir I Aftanas, Natalia V Lotova, Vladimir I Koshkarov, Vera L Pokrovskaja, Serguei A Popov, and Victor P Makhnev. Non-linear analysis of emotion eeg: calculation of kolmogorov entropy and the principal lyapunov exponent. *Neuroscience letters*, 226(1):13–16, 1997.

[134] Ernesto Pereda, Antoni Gamundi, Rubén Rial, and Julián González. Non-linear behaviour of human eeg: fractal exponent versus correlation dimension in awake and sleep stages. *Neuroscience letters*, 250(2):91–94, 1998.

[135] Ivan Dvorak and Jaromir Siska. On some problems encountered in the estimation of the correlation dimension of the eeg. *Physics Letters A*, 118(2):63–66, 1986.

[136] Martha Koukkou, D Lehmann, J Wackermann, I Dvorak, and B Henggeler. Dimensional complexity of eeg brain mechanisms in untreated schizophrenia. *Biological psychiatry*, 33(6):397–407, 1993.

[137] Gary Bruno Schmid and Rudolf M Dünki. Indications of nonlinearity, intraindividual specificity and stability of human eeg: The unfolding dimension. *Physica D: Nonlinear Phenomena*, 93(3-4):165–190, 1996.

[138] Carlos Gómez, Ángela Mediavilla, Roberto Hornero, Daniel Abásolo, and Alberto Fernández. Use of the higuchi's fractal dimension for the analysis of meg recordings from alzheimer's disease patients. *Medical engineering & physics*, 31(3):306–313, 2009.

[139] Xiang Jie, Rui Cao, and Li Li. Emotion recognition based on the sample entropy of eeg. *Biomedical materials and engineering*, 24(1):1185–1192, 2014.

[140] Yuedong Song, Jon Crowcroft, and Jiaxiang Zhang. Automatic epileptic seizure detection in eegs based on optimized sample entropy and extreme learning machine. *Journal of neuroscience methods*, 210(2):132–146, 2012.

[141] U Rajendra Acharya, Filippo Molinari, S Vinitha Sree, Subhagata Chattopadhyay, Kwan-Hoong Ng, and Jasjit S Suri. Automated diagnosis of epileptic eeg using entropies. *Biomedical Signal Processing and Control*, 7(4):401–408, 2012.

[142] Ruhi Mahajan and Bashir I Morshed. Unsupervised eye blink artifact denoising of eeg data with modified multiscale sample entropy, kurtosis, and wavelet-ica. *IEEE journal of Biomedical and Health Informatics*, 19(1):158–165, 2015.

[143] M Poulos, M Rangoussi, N Alexandris, and A Evangelou. Person identification from the eeg using nonlinear signal classification. *Methods of information in Medicine*, 41(01):64–75, 2002.

[144] RM Dünki, GB Schmid, and HH Stassen. Intraindividual specificity and stability of human eeg: comparing a linear vs a nonlinear approach. *Methods of information in medicine*, 39(01):78–82, 2000.

[145] John M Kane. Factors which can make patients difficult to treat. *The British Journal of Psychiatry*, 169(S31):10–14, 1996.

[146]

[147] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

[148] Andrew M Pitts. *Nominal sets: Names and symmetry in computer science*. Cambridge University Press, 2013.

[149] Paul Werbos. Beyond regression: New tools for prediction and analysis in the behavioral sciences. *Ph. D. dissertation, Harvard University*, 1974.

[150] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[151] David E Rumelhart, Geoffrey E Hinton, Ronald J Williams, et al. Learning representations by back-propagating errors. *Cognitive modeling*, 5(3):1, 1988.

[152] Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26–31, 2012.

[153] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[154] Yann LeCun, Patrick Haffner, Léon Bottou, and Yoshua Bengio. Object recognition with gradient-based learning. In *Shape, contour and grouping in computer vision*, pages 319–345. Springer, 1999.

[155] Kunihiko Fukushima and Sei Miyake. Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In *Competition and cooperation in neural nets*, pages 267–285. Springer, 1982.

[156] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2323, 1998.

[157] Yann LeCun, Patrick Haffner, Léon Bottou, and Yoshua Bengio. Object recognition with gradient-based learning. In *Shape, contour and grouping in computer vision*, pages 319–345. Springer, 1999.

[158] D. H. Hubel and T. N. Wiesel. Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology*, 1968.

[159] Shiyu Liang and R Srikant. Why deep neural networks for function approximation? *arXiv preprint arXiv:1610.04161*, 2016.

[160] George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

[161] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*, 2015.

[162] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. Dynamic Routing Between Capsules. (Nips), 2017.

[163] Dingjun Yu, Hanli Wang, Peiqiu Chen, and Zhihua Wei. Mixed pooling for convolutional neural networks. In *International Conference on Rough Sets and Knowledge Technology*, pages 364–375. Springer, 2014.

[164] Benjamin Blankertz, Ryota Tomioka, Steven Lemm, Motoaki Kawanabe, and K-R Muller. Optimizing spatial filters for robust eeg single-trial analysis. *IEEE Signal processing magazine*, 25(1):41–56, 2008.

[165] Zoltan J Koles, Michael S Lazar, and Steven Z Zhou. Spatial patterns underlying population differences in the background eeg. *Brain topography*, 2(4):275–284, 1990.

[166] Kai Keng Ang, Zheng Yang Chin, Haihong Zhang, and Cuntai Guan. Filter bank common spatial pattern (fbcsp) in brain-computer interface. In *Neural Networks, 2008. IJCNN 2008.(IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on*, pages 2390–2397. IEEE, 2008.

[167] Lan Ma, James W Minett, Thierry Blu, and William SY Wang. Resting state eeg-based biometrics for individual identification using convolutional neural networks. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 2848–2851. IEEE, 2015.

[168] Zhiguang Wang and Tim Oates. Imaging time-series to improve classification and imputation. *arXiv preprint arXiv:1506.00327*, 2015.

[169] Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. Deep learning with convolutional neural networks for eeg decoding and visualization. *Human brain mapping*, 38(11):5391–5420, 2017.

[170] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[171] Marcos Del Pozo-Banos, Jesús B Alonso, Jaime R Ticay-Rivas, and Carlos M Travieso. Electroencephalogram subject identification: A review. *Expert Systems with Applications*, 41(15):6537–6554, 2014.

[172] Hallowell Davis and Pauline A Davis. Action potentials of the brain: In normal persons and in normal states of cerebral activity. *Archives of Neurology & Psychiatry*, 36(6):1214–1224, 1936.

[173] David Hecht. Depression and the hyperactive right-hemisphere. *Neuroscience research*, 68(2):77–87, 2010.

[174] Jun Soo Kwon, Tak Youn, and Hee Yeon Jung. Right hemisphere abnormalities in major depression: quantitative electroencephalographic findings before and after treatment. *Journal of affective disorders*, 40(3):169–173, 1996.

[175] R Schirrmeister, Lukas Gemein, Katharina Eggensperger, Frank Hutter, and Tonio Ball. Deep learning with convolutional neural networks for decoding and visualization of eeg pathology. In *2017 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, pages 1–7. IEEE, 2017.

Appendix

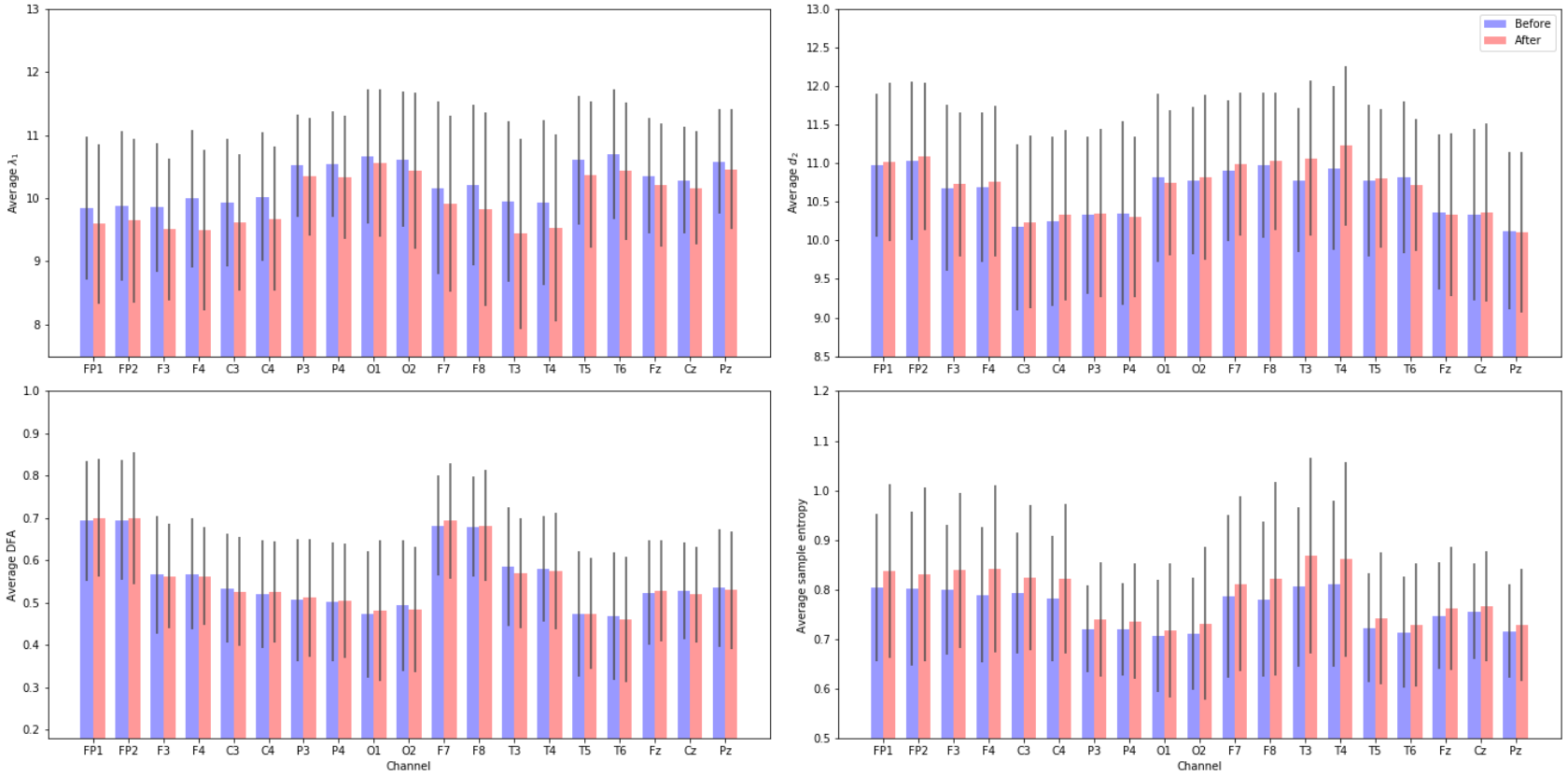


Figure 12: Values of individual measures computed before and after treatment.

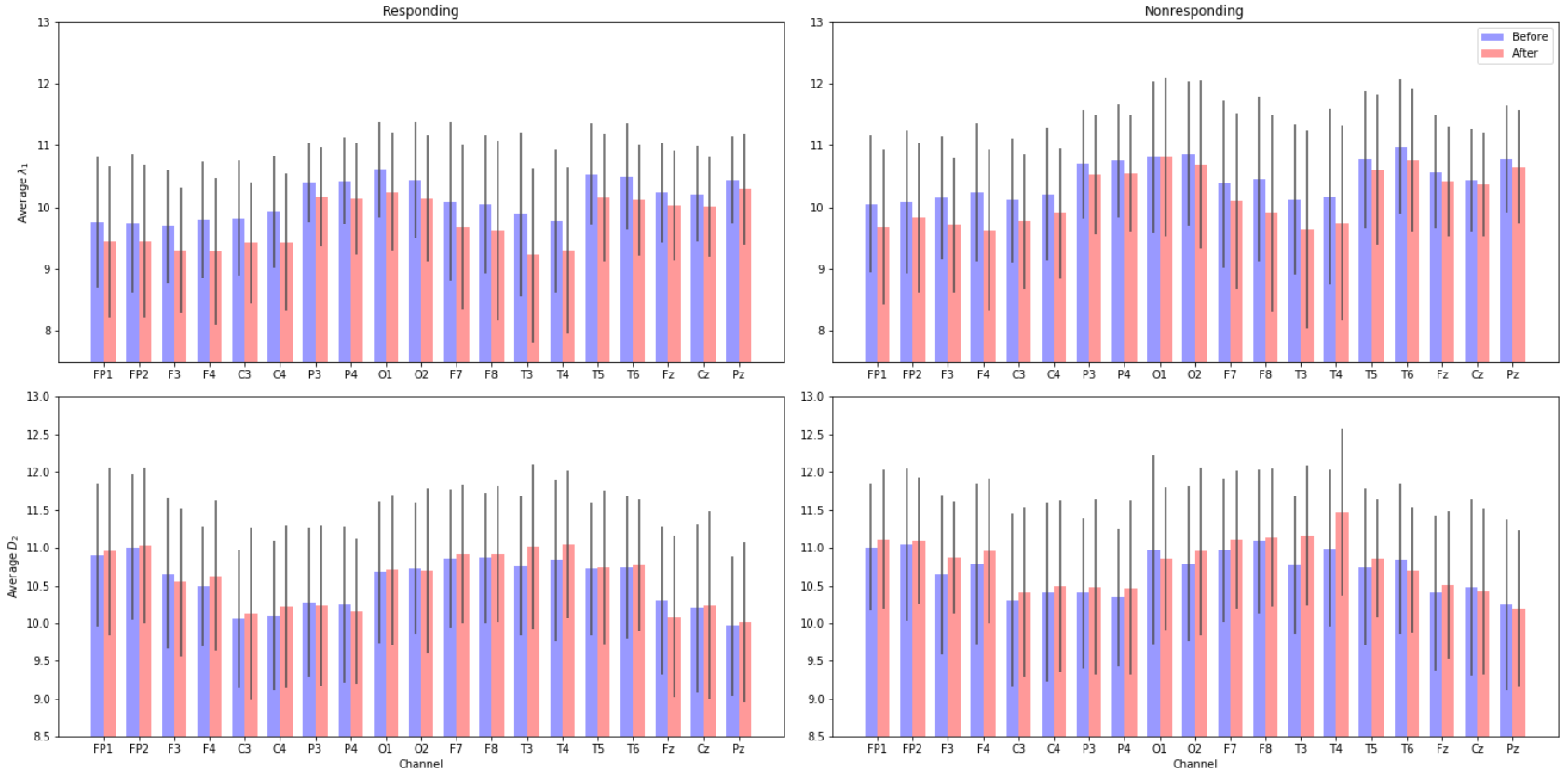


Figure 13: Comparison of mean values of largest Lyapunov exponent and correlation dimension between responders and nonresponders.

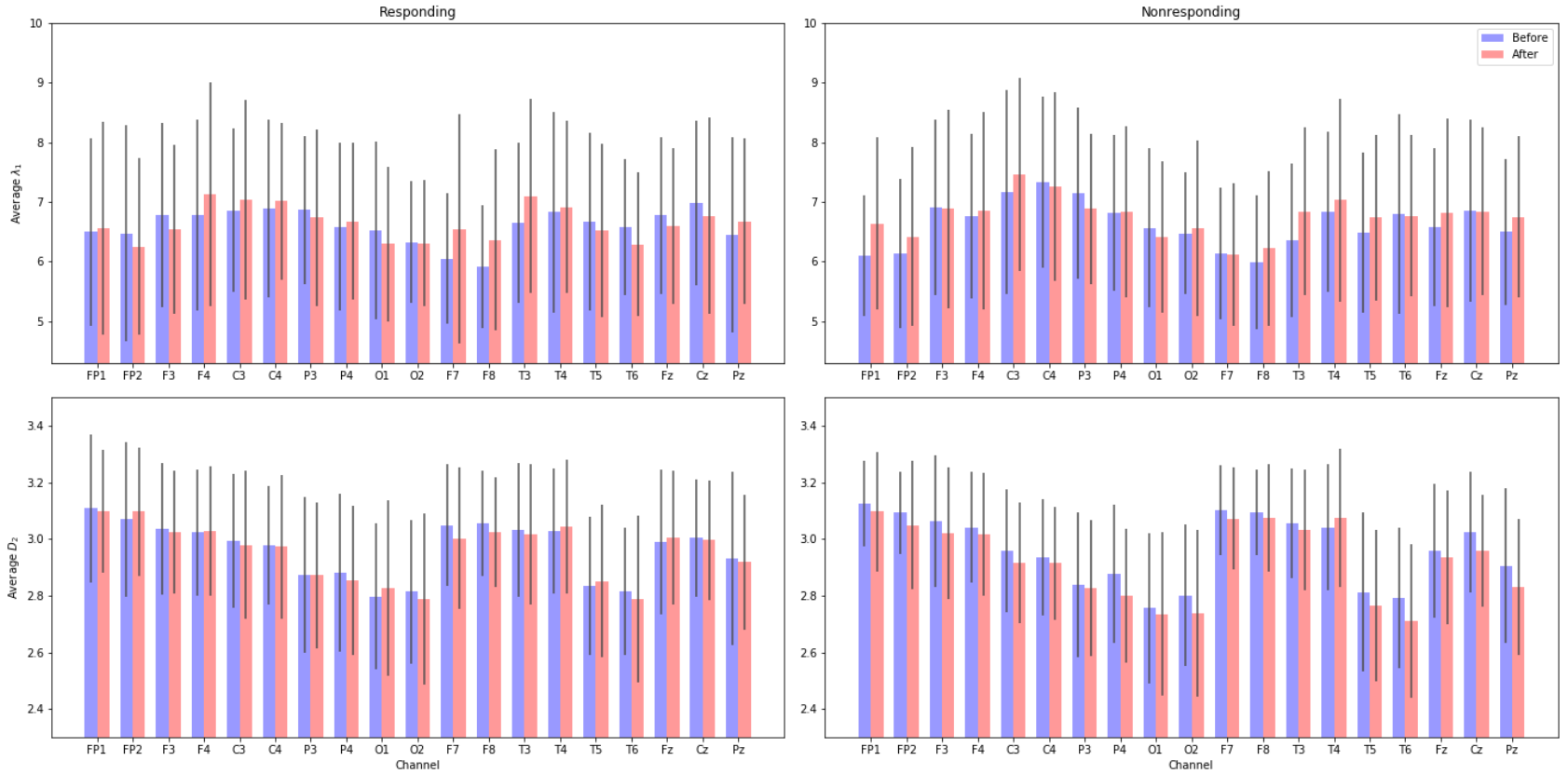


Figure 14: Comparison of mean values of largest Lyapunov exponent and correlation dimension between responders and nonresponders computed using automatic procedure described in Section 2.3.3.2.

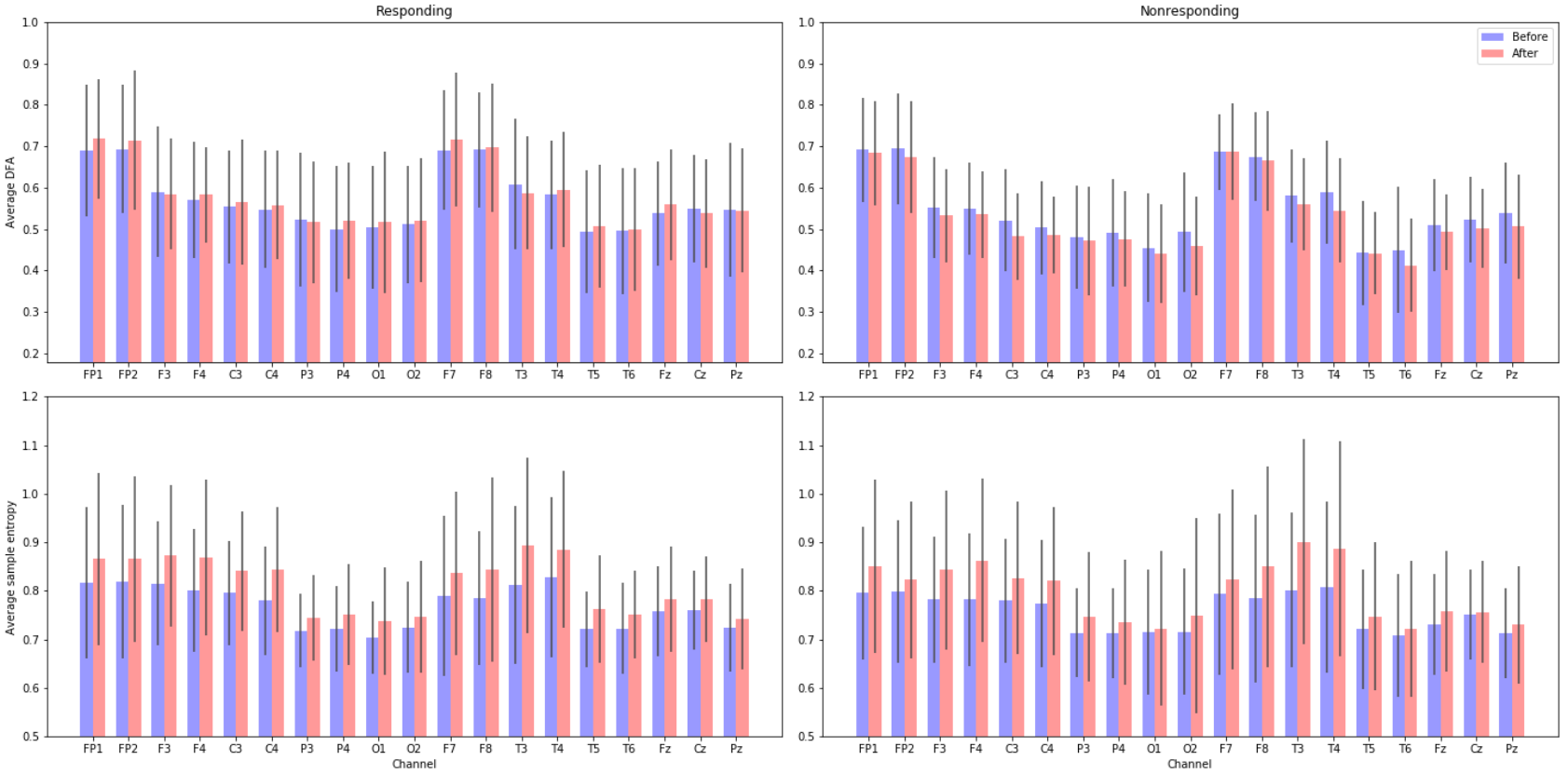


Figure 15: Comparison of mean values of computed detrended fluctuation analysis and sample entropy between responders and nonresponders.

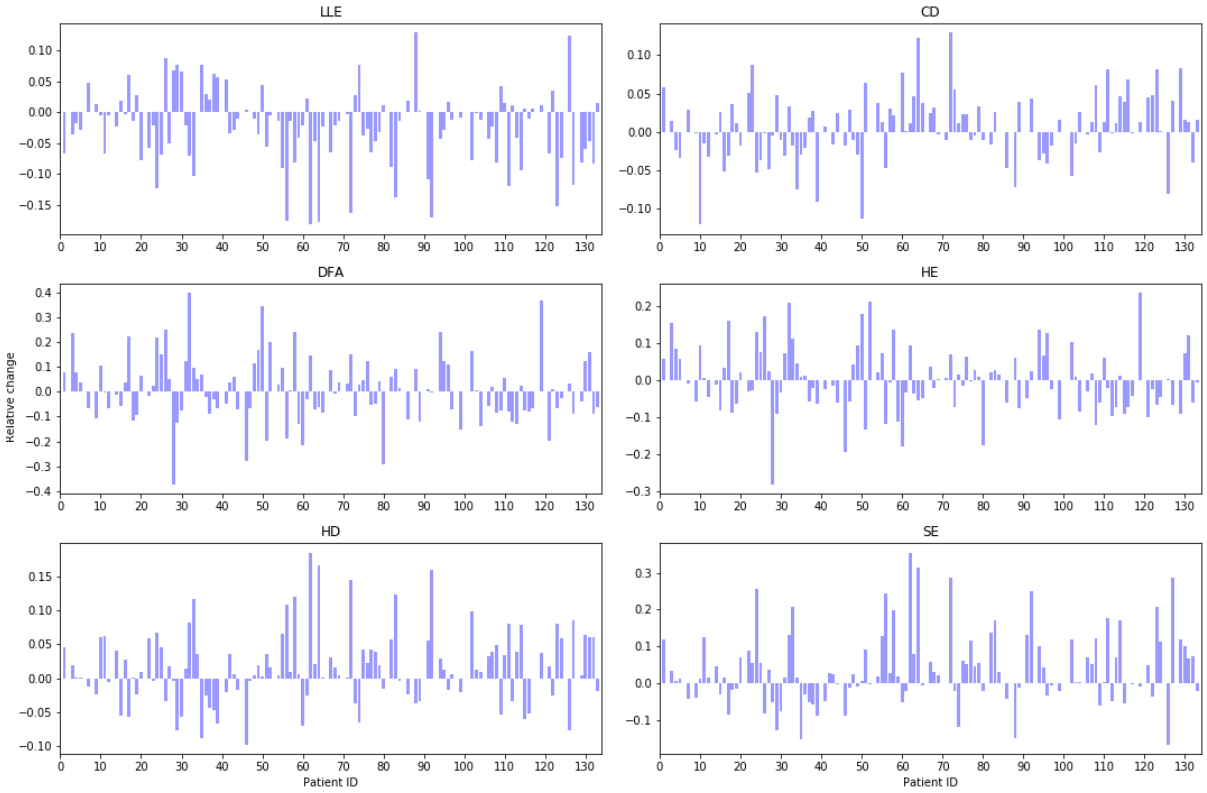


Figure 16: Overview of relative change in each measure for each patient. The relative change is computed for each patient and measure as (mean after − mean before)/|mean before|, where mean before is the mean of the nonlinear measure values across channels in the recording obtained before treatment, and mean after is the analogous mean for the recording obtained after treatment.

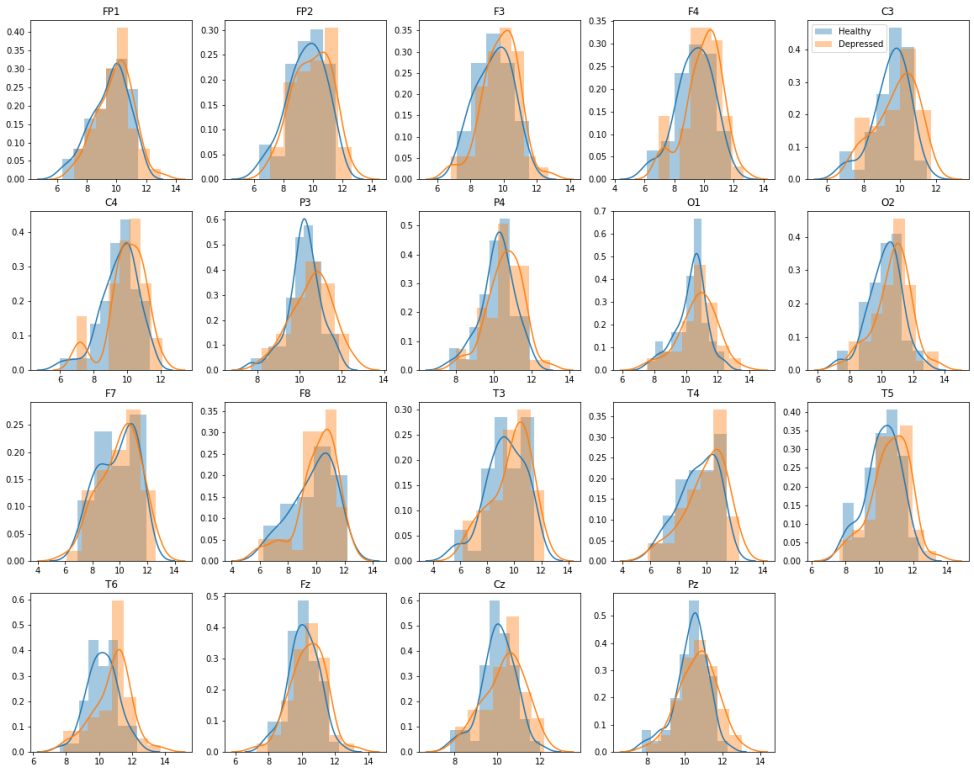


Figure 17: Distributions of the largest Lyapunov exponents between healthy and depressed patients. Most notable differences can be observed in the left and right temporal areas, T3 and T6. The distributions seem generally normal (however, this is not true for all measures).

Channel	Healthy	Depressed	p-value	Sig.
mean	0.559 ± 0.112	0.558 ± 0.111	0.790	
std	0.106 ± 0.029	0.108 ± 0.028	0.712	
FP1	0.693 ± 0.140	0.700 ± 0.139	0.986	
FP2	0.696 ± 0.142	0.699 ± 0.156	0.993	
F3	0.566 ± 0.140	0.563 ± 0.124	0.932	
F4	0.568 ± 0.131	0.563 ± 0.116	0.882	
C3	0.534 ± 0.129	0.527 ± 0.129	0.447	
C4	0.521 ± 0.127	0.525 ± 0.120	0.860	
P3	0.507 ± 0.145	0.512 ± 0.139	0.701	
P4	0.502 ± 0.141	0.506 ± 0.135	0.817	
O1	0.473 ± 0.149	0.482 ± 0.166	0.752	
O2	0.494 ± 0.154	0.484 ± 0.147	0.714	
F7	0.682 ± 0.118	0.694 ± 0.136	0.560	
F8	0.680 ± 0.118	0.682 ± 0.132	0.929	
T3	0.586 ± 0.140	0.570 ± 0.129	0.498	
T4	0.580 ± 0.125	0.574 ± 0.138	0.788	
T5	0.473 ± 0.148	0.475 ± 0.132	0.855	
T6	0.469 ± 0.151	0.462 ± 0.148	0.678	
Fz	0.524 ± 0.123	0.529 ± 0.119	0.901	
Cz	0.529 ± 0.114	0.520 ± 0.114	0.447	
Pz	0.536 ± 0.139	0.530 ± 0.138	0.539	

(a) DFA

Channel	Healthy	Depressed	p-value	Sig.
mean	0.597 ± 0.085	0.594 ± 0.080	0.595	
std	0.071 ± 0.022	0.071 ± 0.021	0.980	
FP1	0.671 ± 0.090	0.674 ± 0.082	0.882	
FP2	0.677 ± 0.089	0.670 ± 0.098	0.579	
F3	0.603 ± 0.097	0.600 ± 0.087	0.725	
F4	0.607 ± 0.095	0.600 ± 0.085	0.743	
C3	0.588 ± 0.096	0.577 ± 0.094	0.222	
C4	0.579 ± 0.099	0.577 ± 0.083	0.692	
P3	0.559 ± 0.112	0.562 ± 0.108	0.877	
P4	0.558 ± 0.112	0.560 ± 0.101	0.944	
O1	0.536 ± 0.117	0.539 ± 0.126	0.917	
O2	0.551 ± 0.120	0.544 ± 0.114	0.652	
F7	0.683 ± 0.079	0.685 ± 0.085	0.992	
F8	0.680 ± 0.077	0.678 ± 0.087	0.997	
T3	0.620 ± 0.094	0.604 ± 0.089	0.183	
T4	0.619 ± 0.085	0.609 ± 0.091	0.543	
T5	0.536 ± 0.115	0.538 ± 0.099	0.909	
T6	0.531 ± 0.119	0.525 ± 0.108	0.703	
Fz	0.580 ± 0.094	0.581 ± 0.084	0.936	
Cz	0.588 ± 0.089	0.585 ± 0.084	0.498	
Pz	0.584 ± 0.106	0.579 ± 0.102	0.546	

(b) Hurst exponent

Channel	Healthy	Depressed	p-value	Sig.
mean	10.245 ± 0.906	9.976 ± 1.015	0.064	
std	0.617 ± 0.239	0.703 ± 0.303	0.047	*
FP1	9.845 ± 1.129	9.593 ± 1.258	0.181	
FP2	9.877 ± 1.184	9.645 ± 1.291	0.223	
F3	9.853 ± 1.021	9.509 ± 1.124	0.042	*
F4	9.994 ± 1.082	9.491 ± 1.272	0.004	***
C3	9.931 ± 1.010	9.612 ± 1.080	0.032	*
C4	10.027 ± 1.017	9.675 ± 1.142	0.019	**
P3	10.516 ± 0.816	10.343 ± 0.931	0.262	
P4	10.536 ± 0.835	10.336 ± 0.975	0.257	
O1	10.664 ± 1.067	10.557 ± 1.164	0.576	
O2	10.617 ± 1.076	10.438 ± 1.231	0.266	
F7	10.164 ± 1.361	9.908 ± 1.395	0.157	
F8	10.209 ± 1.266	9.826 ± 1.528	0.109	
T3	9.955 ± 1.271	9.437 ± 1.508	0.007	***
T4	9.928 ± 1.310	9.528 ± 1.484	0.060	
T5	10.606 ± 1.018	10.375 ± 1.160	0.220	
T6	10.704 ± 1.028	10.432 ± 1.092	0.080	
Fz	10.356 ± 0.917	10.211 ± 0.972	0.315	
Cz	10.287 ± 0.849	10.164 ± 0.898	0.384	
Pz	10.584 ± 0.828	10.458 ± 0.949	0.518	

(c) Largest Lyapunov exponent

Channel	Healthy	Depressed	p-value	Sig.
mean	0.761 ± 0.108	0.790 ± 0.130	0.155	
std	0.071 ± 0.040	0.086 ± 0.048	0.029	*
FP1	0.804 ± 0.149	0.837 ± 0.176	0.296	
FP2	0.802 ± 0.156	0.830 ± 0.175	0.313	
F3	0.800 ± 0.132	0.839 ± 0.156	0.125	
F4	0.790 ± 0.137	0.842 ± 0.168	0.024	*
C3	0.793 ± 0.122	0.825 ± 0.147	0.133	
C4	0.781 ± 0.126	0.821 ± 0.151	0.031	*
P3	0.720 ± 0.087	0.740 ± 0.115	0.414	
P4	0.720 ± 0.093	0.736 ± 0.116	0.634	
O1	0.707 ± 0.113	0.718 ± 0.134	0.993	
O2	0.712 ± 0.113	0.732 ± 0.154	0.607	
F7	0.786 ± 0.163	0.811 ± 0.176	0.405	
F8	0.781 ± 0.156	0.821 ± 0.195	0.204	
T3	0.806 ± 0.160	0.867 ± 0.197	0.023	*
T4	0.812 ± 0.167	0.861 ± 0.197	0.079	
T5	0.723 ± 0.110	0.743 ± 0.133	0.557	
T6	0.714 ± 0.112	0.729 ± 0.123	0.403	
Fz	0.747 ± 0.107	0.762 ± 0.124	0.583	
Cz	0.756 ± 0.096	0.767 ± 0.110	0.672	
Pz	0.716 ± 0.093	0.728 ± 0.113	0.806	

(d) Sample entropy

Channel	Healthy	Depressed	p-value	Sig.
mean	1.369 ± 0.118	1.391 ± 0.129	0.249	
std	0.078 ± 0.037	0.093 ± 0.047	0.024	*
FP1	1.429 ± 0.157	1.458 ± 0.175	0.324	
FP2	1.430 ± 0.163	1.450 ± 0.180	0.461	
F3	1.413 ± 0.141	1.442 ± 0.157	0.195	
F4	1.405 ± 0.149	1.450 ± 0.171	0.053	
C3	1.406 ± 0.124	1.431 ± 0.140	0.179	
C4	1.396 ± 0.129	1.427 ± 0.143	0.096	
P3	1.323 ± 0.108	1.335 ± 0.115	0.552	
P4	1.319 ± 0.105	1.329 ± 0.120	0.722	
O1	1.295 ± 0.132	1.304 ± 0.144	0.859	
O2	1.297 ± 0.128	1.312 ± 0.165	0.794	
F7	1.411 ± 0.161	1.430 ± 0.172	0.540	
F8	1.407 ± 0.155	1.446 ± 0.185	0.144	
T3	1.413 ± 0.159	1.461 ± 0.187	0.050	*
T4	1.416 ± 0.165	1.455 ± 0.182	0.119	
T5	1.307 ± 0.117	1.319 ± 0.127	0.700	
T6	1.294 ± 0.122	1.304 ± 0.124	0.509	
Fz	1.356 ± 0.130	1.363 ± 0.136	0.743	
Cz	1.377 ± 0.113	1.387 ± 0.122	0.656	
Pz	1.319 ± 0.112	1.323 ± 0.120	0.971	

(e) Higuchi's fractal dimension

Channel	Healthy	Depressed	p-value	Sig.
mean	10.635 ± 0.825	10.683 ± 0.806	0.700	
std	0.658 ± 0.198	0.680 ± 0.203	0.402	
FP1	10.978 ± 0.923	11.012 ± 1.026	0.710	
FP2	11.031 ± 1.023	11.092 ± 0.956	0.664	
F3	10.681 ± 1.081	10.725 ± 0.930	0.603	
F4	10.683 ± 0.968	10.767 ± 0.975	0.516	
C3	10.172 ± 1.079	10.234 ± 1.120	0.838	
C4	10.243 ± 1.094	10.328 ± 1.107	0.447	
P3	10.326 ± 1.015	10.347 ± 1.089	0.823	
P4	10.352 ± 1.191	10.299 ± 1.042	0.622	
O1	10.814 ± 1.092	10.744 ± 0.943	0.862	
O2	10.771 ± 0.954	10.813 ± 1.069	0.945	
F7	10.907 ± 0.915	10.983 ± 0.926	0.589	
F8	10.974 ± 0.945	11.026 ± 0.894	0.634	
T3	10.778 ± 0.934	11.065 ± 1.004	0.026	*
T4	10.935 ± 1.061	11.226 ± 1.032	0.024	*
T5	10.773 ± 0.984	10.801 ± 0.901	0.728	
T6	10.812 ± 0.985	10.722 ± 0.856	0.481	
Fz	10.366 ± 1.000	10.331 ± 1.052	0.658	
Cz	10.335 ± 1.108	10.361 ± 1.152	0.918	
Pz	10.124 ± 1.022	10.102 ± 1.038	0.812	

(f) Correlation dimension

Table 4: Comparison of mean values of measures computed from recordings obtained before and after treatment. Both before and after groups contain 110 samples.

Channel	Healthy	Depressed	p-value	Sig.
mean	0.576 ± 0.127	0.524 ± 0.104	0.024	*
std	0.105 ± 0.027	0.110 ± 0.030	0.469	
FP1	0.712 ± 0.150	0.665 ± 0.152	0.195	
FP2	0.710 ± 0.168	0.663 ± 0.145	0.202	
F3	0.583 ± 0.141	0.538 ± 0.126	0.074	
F4	0.573 ± 0.126	0.538 ± 0.120	0.075	
C3	0.561 ± 0.146	0.507 ± 0.135	0.048	*
C4	0.548 ± 0.135	0.498 ± 0.128	0.042	*
P3	0.522 ± 0.148	0.473 ± 0.146	0.048	*
P4	0.526 ± 0.145	0.454 ± 0.124	0.009	***
O1	0.502 ± 0.178	0.431 ± 0.141	0.031	*
O2	0.509 ± 0.162	0.450 ± 0.141	0.045	*
F7	0.709 ± 0.162	0.652 ± 0.115	0.035	*
F8	0.696 ± 0.154	0.643 ± 0.117	0.026	*
T3	0.583 ± 0.134	0.548 ± 0.132	0.085	
T4	0.596 ± 0.141	0.544 ± 0.123	0.044	*
T5	0.496 ± 0.152	0.437 ± 0.137	0.041	*
T6	0.489 ± 0.160	0.433 ± 0.136	0.091	
Fz	0.554 ± 0.138	0.487 ± 0.113	0.013	**
Cz	0.534 ± 0.127	0.504 ± 0.116	0.116	
Pz	0.547 ± 0.145	0.490 ± 0.146	0.030	*

(a) DFA

Channel	Healthy	Depressed	p-value	Sig.
mean	0.604 ± 0.091	0.574 ± 0.080	0.060	
std	0.068 ± 0.022	0.076 ± 0.025	0.081	
FP1	0.679 ± 0.088	0.654 ± 0.102	0.138	
FP2	0.674 ± 0.104	0.662 ± 0.101	0.730	
F3	0.609 ± 0.098	0.582 ± 0.084	0.104	
F4	0.604 ± 0.094	0.586 ± 0.089	0.197	
C3	0.596 ± 0.105	0.568 ± 0.093	0.113	
C4	0.587 ± 0.097	0.563 ± 0.093	0.164	
P3	0.567 ± 0.119	0.536 ± 0.115	0.160	
P4	0.573 ± 0.108	0.527 ± 0.106	0.031	*
O1	0.549 ± 0.133	0.505 ± 0.116	0.072	
O2	0.561 ± 0.128	0.519 ± 0.112	0.059	
F7	0.695 ± 0.105	0.667 ± 0.083	0.108	
F8	0.677 ± 0.103	0.658 ± 0.082	0.094	
T3	0.611 ± 0.092	0.597 ± 0.087	0.366	
T4	0.625 ± 0.091	0.594 ± 0.087	0.061	
T5	0.555 ± 0.114	0.509 ± 0.106	0.048	*
T6	0.544 ± 0.115	0.502 ± 0.111	0.071	
Fz	0.594 ± 0.096	0.553 ± 0.088	0.015	**
Cz	0.594 ± 0.093	0.571 ± 0.084	0.138	
Pz	0.588 ± 0.114	0.554 ± 0.117	0.073	

(b) Hurst exponent

Channel	Healthy	Depressed	p-value	Sig.
mean	9.848 ± 0.947	10.236 ± 1.043	0.030	*
std	0.681 ± 0.264	0.670 ± 0.291	0.730	
FP1	9.538 ± 1.257	9.888 ± 1.202	0.200	
FP2	9.551 ± 1.274	9.946 ± 1.275	0.168	
F3	9.413 ± 1.081	9.812 ± 1.097	0.062	
F4	9.363 ± 1.234	9.963 ± 1.254	0.012	**
C3	9.518 ± 1.010	9.803 ± 1.155	0.093	
C4	9.513 ± 1.131	9.956 ± 1.172	0.027	*
P3	10.226 ± 0.806	10.569 ± 0.961	0.041	*
P4	10.200 ± 0.898	10.593 ± 0.941	0.034	*
O1	10.321 ± 0.976	10.736 ± 1.201	0.026	*
O2	10.240 ± 1.059	10.700 ± 1.205	0.017	**
F7	9.814 ± 1.401	9.988 ± 1.466	0.572	
F8	9.772 ± 1.528	10.061 ± 1.471	0.341	
T3	9.336 ± 1.454	9.731 ± 1.498	0.138	
T4	9.408 ± 1.378	9.922 ± 1.520	0.054	
T5	10.216 ± 1.039	10.649 ± 1.142	0.033	*
T6	10.200 ± 0.929	10.787 ± 1.157	0.002	***
Fz	10.097 ± 0.901	10.358 ± 1.040	0.158	
Cz	10.048 ± 0.816	10.339 ± 0.963	0.082	
Pz	10.343 ± 0.888	10.680 ± 1.016	0.081	

(c) Largest Lyapunov exponent

Channel	Healthy	Depressed	p-value	Sig.
mean	0.797 ± 0.123	0.760 ± 0.125	0.041	*
std	0.087 ± 0.046	0.076 ± 0.044	0.238	
FP1	0.843 ± 0.180	0.788 ± 0.154	0.101	
FP2	0.841 ± 0.179	0.790 ± 0.157	0.111	
F3	0.848 ± 0.149	0.799 ± 0.145	0.065	
F4	0.852 ± 0.171	0.792 ± 0.160	0.033	*
C3	0.833 ± 0.141	0.801 ± 0.146	0.101	
C4	0.832 ± 0.151	0.791 ± 0.155	0.043	*
P3	0.740 ± 0.101	0.714 ± 0.102	0.065	
P4	0.742 ± 0.109	0.705 ± 0.105	0.039	*
O1	0.729 ± 0.123	0.704 ± 0.128	0.104	
O2	0.739 ± 0.130	0.705 ± 0.126	0.040	*
F7	0.817 ± 0.174	0.799 ± 0.179	0.620	
F8	0.824 ± 0.199	0.796 ± 0.189	0.440	
T3	0.873 ± 0.195	0.834 ± 0.188	0.220	
T4	0.864 ± 0.179	0.816 ± 0.194	0.060	
T5	0.750 ± 0.120	0.715 ± 0.125	0.046	*
T6	0.743 ± 0.109	0.704 ± 0.115	0.017	**
Fz	0.771 ± 0.116	0.739 ± 0.119	0.111	
Cz	0.775 ± 0.101	0.745 ± 0.108	0.055	
Pz	0.734 ± 0.109	0.701 ± 0.105	0.062	

(d) Sample entropy

Channel	Healthy	Depressed	p-value	Sig.
mean	1.408 ± 0.129	1.357 ± 0.131	0.024	*
std	0.093 ± 0.042	0.086 ± 0.042	0.338	
FP1	1.474 ± 0.186	1.406 ± 0.166	0.050	
FP2	1.468 ± 0.182	1.405 ± 0.176	0.074	
F3	1.465 ± 0.160	1.405 ± 0.158	0.039	*
F4	1.466 ± 0.180	1.398 ± 0.167	0.049	*
C3	1.450 ± 0.135	1.406 ± 0.140	0.057	
C4	1.446 ± 0.149	1.393 ± 0.139	0.045	*
P3	1.346 ± 0.109	1.305 ± 0.111	0.044	*
P4	1.345 ± 0.119	1.295 ± 0.106	0.021	*
O1	1.322 ± 0.138	1.279 ± 0.136	0.068	
O2	1.330 ± 0.136	1.283 ± 0.141	0.038	*
F7	1.450 ± 0.178	1.412 ± 0.181	0.212	
F8	1.462 ± 0.190	1.405 ± 0.183	0.065	
T3	1.474 ± 0.191	1.431 ± 0.184	0.174	
T4	1.468 ± 0.168	1.414 ± 0.192	0.059	
T5	1.333 ± 0.117	1.290 ± 0.127	0.037	*
T6	1.327 ± 0.116	1.274 ± 0.121	0.011	**
Fz	1.381 ± 0.137	1.340 ± 0.147	0.086	
Cz	1.404 ± 0.115	1.356 ± 0.125	0.026	*
Pz	1.337 ± 0.118	1.295 ± 0.118	0.064	

(e) Higuchi's fractal dimension

Channel	Healthy	Depressed	p-value	Sig.
mean	10.591 ± 0.879	10.816 ± 0.716	0.177	
std	0.664 ± 0.197	0.651 ± 0.154	0.874	
FP1	10.939 ± 1.125	11.222 ± 0.915	0.128	
FP2	11.015 ± 1.038	11.278 ± 0.978	0.231	
F3	10.616 ± 1.053	10.986 ± 0.911	0.053	
F4	10.614 ± 0.975	10.892 ± 0.944	0.149	
C3	10.147 ± 1.136	10.425 ± 0.864	0.119	
C4	10.218 ± 1.120	10.504 ± 0.981	0.180	
P3	10.226 ± 1.053	10.596 ± 0.928	0.090	
P4	10.169 ± 0.973	10.420 ± 0.825	0.143	
O1	10.690 ± 1.007	11.014 ± 1.144	0.242	
O2	10.725 ± 1.088	10.863 ± 0.983	0.484	
F7	10.913 ± 0.962	10.948 ± 0.818	0.740	
F8	10.923 ± 0.892	11.087 ± 0.953	0.419	
T3	11.015 ± 1.113	10.995 ± 0.812	0.829	
T4	11.059 ± 1.002	11.109 ± 0.958	0.812	
T5	10.787 ± 1.048	10.831 ± 0.913	0.840	
T6	10.740 ± 0.946	10.998 ± 0.764	0.257	
Fz	10.151 ± 1.118	10.585 ± 0.906	0.038	*
Cz	10.297 ± 1.258	10.451 ± 1.038	0.403	
Pz	9.991 ± 1.073	10.295 ± 0.891	0.155	

(f) Correlation dimension

Table 5: Comparison of mean values of measures computed for 50 healthy (depression score $\leq$ 16) and 50 depressed (depression score $\geq$ 28) patients.

Channel	Responding	Nonresponding	p-value	Sig.
mean	0.566 ± 0.122	0.554 ± 0.098	0.533	
std	0.106 ± 0.034	0.108 ± 0.026	0.568	
FP1	0.689 ± 0.154	0.697 ± 0.121	0.820	
FP2	0.690 ± 0.150	0.704 ± 0.127	0.714	
F3	0.579 ± 0.152	0.566 ± 0.129	0.764	
F4	0.567 ± 0.141	0.569 ± 0.112	0.895	
C3	0.548 ± 0.134	0.526 ± 0.124	0.330	
C4	0.541 ± 0.136	0.503 ± 0.110	0.113	
P3	0.515 ± 0.162	0.493 ± 0.126	0.302	
P4	0.499 ± 0.157	0.499 ± 0.123	0.866	
O1	0.492 ± 0.161	0.462 ± 0.136	0.356	
O2	0.497 ± 0.154	0.502 ± 0.156	0.936	
F7	0.687 ± 0.144	0.685 ± 0.092	0.563	
F8	0.688 ± 0.134	0.677 ± 0.103	0.510	
T3	0.600 ± 0.157	0.576 ± 0.115	0.302	
T4	0.576 ± 0.127	0.591 ± 0.127	0.468	
T5	0.486 ± 0.154	0.454 ± 0.130	0.253	
T6	0.485 ± 0.154	0.460 ± 0.148	0.371	
Fz	0.537 ± 0.129	0.514 ± 0.110	0.288	
Cz	0.544 ± 0.124	0.522 ± 0.100	0.185	
Pz	0.536 ± 0.165	0.534 ± 0.119	0.618	

(a) DFA

Channel	Responding	Nonresponding	p-value	Sig.
mean	0.603 ± 0.093	0.593 ± 0.072	0.330	
std	0.071 ± 0.027	0.072 ± 0.018	0.644	
FP1	0.672 ± 0.103	0.669 ± 0.071	0.371	
FP2	0.680 ± 0.098	0.675 ± 0.076	0.468	
F3	0.610 ± 0.103	0.602 ± 0.086	0.412	
F4	0.606 ± 0.100	0.609 ± 0.080	0.982	
C3	0.597 ± 0.100	0.580 ± 0.091	0.302	
C4	0.594 ± 0.104	0.563 ± 0.087	0.050	
P3	0.564 ± 0.129	0.550 ± 0.098	0.285	
P4	0.558 ± 0.127	0.555 ± 0.098	0.687	
O1	0.549 ± 0.126	0.527 ± 0.105	0.308	
O2	0.556 ± 0.128	0.555 ± 0.112	0.764	
F7	0.684 ± 0.097	0.685 ± 0.059	0.618	
F8	0.684 ± 0.086	0.678 ± 0.070	0.450	
T3	0.626 ± 0.104	0.617 ± 0.077	0.360	
T4	0.622 ± 0.088	0.623 ± 0.080	0.947	
T5	0.546 ± 0.123	0.522 ± 0.097	0.147	
T6	0.547 ± 0.122	0.519 ± 0.116	0.227	
Fz	0.587 ± 0.099	0.577 ± 0.080	0.269	
Cz	0.600 ± 0.093	0.579 ± 0.080	0.085	
Pz	0.582 ± 0.130	0.583 ± 0.085	0.618	

(b) Hurst exponent

Channel	Responding	Nonresponding	p-value	Sig.
mean	10.123 ± 0.766	10.458 ± 0.952	0.030	*
std	0.628 ± 0.217	0.604 ± 0.252	0.404	
FP1	9.762 ± 1.057	10.052 ± 1.110	0.200	
FP2	9.740 ± 1.132	10.088 ± 1.150	0.133	
F3	9.688 ± 0.917	10.156 ± 1.000	0.008	***
F4	9.802 ± 0.941	10.241 ± 1.117	0.020	**
C3	9.821 ± 0.931	10.115 ± 1.004	0.061	
C4	9.921 ± 0.909	10.217 ± 1.076	0.035	*
P3	10.410 ± 0.638	10.700 ± 0.884	0.043	*
P4	10.423 ± 0.700	10.754 ± 0.918	0.027	*
O1	10.609 ± 0.769	10.815 ± 1.218	0.098	
O2	10.443 ± 0.943	10.863 ± 1.167	0.019	**
F7	10.092 ± 1.289	10.386 ± 1.358	0.200	
F8	10.045 ± 1.116	10.459 ± 1.329	0.030	*
T3	9.885 ± 1.324	10.126 ± 1.221	0.356	
T4	9.777 ± 1.164	10.176 ± 1.418	0.045	*
T5	10.531 ± 0.824	10.773 ± 1.114	0.079	
T6	10.500 ± 0.860	10.981 ± 1.086	0.007	***
Fz	10.238 ± 0.801	10.570 ± 0.913	0.050	
Cz	10.213 ± 0.770	10.447 ± 0.835	0.155	
Pz	10.447 ± 0.702	10.775 ± 0.876	0.032	*

(c) Largest Lyapunov exponent

Channel	Responding	Nonresponding	p-value	Sig.
mean	0.764 ± 0.102	0.749 ± 0.106	0.213	
std	0.075 ± 0.038	0.064 ± 0.041	0.069	
FP1	0.806 ± 0.154	0.789 ± 0.129	0.553	
FP2	0.810 ± 0.162	0.787 ± 0.136	0.500	
F3	0.809 ± 0.128	0.776 ± 0.122	0.089	
F4	0.799 ± 0.134	0.772 ± 0.130	0.168	
C3	0.797 ± 0.116	0.781 ± 0.122	0.178	
C4	0.785 ± 0.122	0.767 ± 0.126	0.164	
P3	0.718 ± 0.084	0.711 ± 0.086	0.275	
P4	0.720 ± 0.096	0.709 ± 0.089	0.238	
O1	0.699 ± 0.080	0.705 ± 0.119	0.412	
O2	0.718 ± 0.102	0.702 ± 0.119	0.187	
F7	0.781 ± 0.163	0.775 ± 0.156	0.553	
F8	0.783 ± 0.146	0.766 ± 0.159	0.241	
T3	0.810 ± 0.169	0.792 ± 0.151	0.687	
T4	0.818 ± 0.164	0.795 ± 0.168	0.216	
T5	0.719 ± 0.085	0.715 ± 0.116	0.256	
T6	0.722 ± 0.104	0.702 ± 0.117	0.089	
Fz	0.753 ± 0.100	0.729 ± 0.097	0.105	
Cz	0.758 ± 0.090	0.744 ± 0.090	0.348	
Pz	0.720 ± 0.098	0.705 ± 0.087	0.241	

(d) Sample entropy

Channel	Responding	Nonresponding	p-value	Sig.
mean	1.378 ± 0.113	1.348 ± 0.113	0.113	
std	0.080 ± 0.036	0.074 ± 0.037	0.262	
FP1	1.442 ± 0.165	1.400 ± 0.138	0.145	
FP2	1.445 ± 0.169	1.405 ± 0.143	0.173	
F3	1.431 ± 0.138	1.381 ± 0.133	0.045	*
F4	1.420 ± 0.143	1.380 ± 0.139	0.096	
C3	1.413 ± 0.115	1.385 ± 0.120	0.147	
C4	1.403 ± 0.124	1.376 ± 0.126	0.112	
P3	1.328 ± 0.101	1.304 ± 0.104	0.095	
P4	1.322 ± 0.103	1.302 ± 0.103	0.119	
O1	1.298 ± 0.106	1.284 ± 0.139	0.178	
O2	1.311 ± 0.125	1.281 ± 0.128	0.159	
F7	1.418 ± 0.159	1.385 ± 0.152	0.173	
F8	1.419 ± 0.144	1.387 ± 0.160	0.102	
T3	1.421 ± 0.169	1.388 ± 0.143	0.315	
T4	1.422 ± 0.166	1.398 ± 0.165	0.371	
T5	1.312 ± 0.099	1.290 ± 0.118	0.143	
T6	1.308 ± 0.110	1.278 ± 0.130	0.076	
Fz	1.367 ± 0.123	1.330 ± 0.119	0.031	*
Cz	1.383 ± 0.106	1.359 ± 0.103	0.171	
Pz	1.328 ± 0.111	1.301 ± 0.106	0.089	

(e) Higuchi's fractal dimension

Channel	Responding	Nonresponding	p-value	Sig.
mean	10.546 ± 0.755	10.696 ± 0.848	0.346	
std	0.649 ± 0.191	0.653 ± 0.162	0.765	
FP1	10.900 ± 0.951	11.006 ± 0.834	0.438	
FP2	11.005 ± 0.967	11.043 ± 1.010	0.812	
F3	10.659 ± 1.000	10.648 ± 1.049	0.913	
F4	10.486 ± 0.796	10.785 ± 1.056	0.268	
C3	10.058 ± 0.918	10.308 ± 1.145	0.182	
C4	10.100 ± 0.990	10.412 ± 1.180	0.219	
P3	10.272 ± 0.987	10.402 ± 0.993	0.513	
P4	10.245 ± 1.036	10.341 ± 0.903	0.859	
O1	10.677 ± 0.933	10.968 ± 1.251	0.568	
O2	10.728 ± 0.875	10.789 ± 1.021	0.859	
F7	10.859 ± 0.915	10.968 ± 0.948	0.584	
F8	10.864 ± 0.869	11.088 ± 0.952	0.258	
T3	10.760 ± 0.920	10.769 ± 0.914	0.925	
T4	10.837 ± 1.070	10.992 ± 1.040	0.285	
T5	10.719 ± 0.884	10.743 ± 1.041	0.563	
T6	10.739 ± 0.939	10.847 ± 0.993	0.859	
Fz	10.301 ± 0.979	10.402 ± 1.023	0.686	
Cz	10.197 ± 1.114	10.472 ± 1.167	0.172	
Pz	9.964 ± 0.917	10.249 ± 1.132	0.295	

(f) Correlation dimension

Table 6: Comparison of mean values of measures computed from recordings obtained during the first session (i.e. before drug administration) for 50 patients responding positively to the treatment and 50 nonresponding patients, i.e. patients retaining their symptoms and patients worsening their symptoms.

Channel	Responding	Nonresponding	p-value	Sig.
mean	0.571 ± 0.122	0.546 ± 0.094	0.081	
std	0.105 ± 0.031	0.109 ± 0.027	0.342	
FP1	0.700 ± 0.149	0.692 ± 0.121	0.377	
FP2	0.701 ± 0.157	0.695 ± 0.130	0.689	
F3	0.580 ± 0.143	0.553 ± 0.117	0.170	
F4	0.571 ± 0.131	0.559 ± 0.108	0.394	
C3	0.555 ± 0.139	0.510 ± 0.116	0.013	**
C4	0.549 ± 0.131	0.496 ± 0.104	0.002	***
P3	0.516 ± 0.155	0.490 ± 0.127	0.105	
P4	0.511 ± 0.152	0.493 ± 0.119	0.225	
O1	0.498 ± 0.167	0.462 ± 0.139	0.076	
O2	0.503 ± 0.156	0.485 ± 0.140	0.271	
F7	0.698 ± 0.152	0.684 ± 0.102	0.191	
F8	0.691 ± 0.143	0.673 ± 0.108	0.171	
T3	0.590 ± 0.145	0.569 ± 0.117	0.300	
T4	0.584 ± 0.131	0.574 ± 0.135	0.649	
T5	0.495 ± 0.150	0.451 ± 0.118	0.020	*
T6	0.489 ± 0.154	0.443 ± 0.137	0.026	*
Fz	0.546 ± 0.132	0.507 ± 0.100	0.021	*
Cz	0.541 ± 0.123	0.513 ± 0.098	0.044	*
Pz	0.539 ± 0.158	0.523 ± 0.122	0.151	

(a) DFA

Channel	Responding	Nonresponding	p-value	Sig.
mean	0.605 ± 0.091	0.587 ± 0.068	0.056	
std	0.069 ± 0.025	0.072 ± 0.018	0.163	
FP1	0.676 ± 0.095	0.668 ± 0.072	0.140	
FP2	0.678 ± 0.100	0.670 ± 0.079	0.340	
F3	0.611 ± 0.099	0.595 ± 0.081	0.117	
F4	0.607 ± 0.095	0.600 ± 0.077	0.400	
C3	0.597 ± 0.101	0.567 ± 0.086	0.012	**
C4	0.594 ± 0.097	0.561 ± 0.077	0.002	***
P3	0.563 ± 0.124	0.549 ± 0.096	0.147	
P4	0.565 ± 0.120	0.551 ± 0.091	0.150	
O1	0.551 ± 0.129	0.529 ± 0.108	0.079	
O2	0.559 ± 0.127	0.545 ± 0.101	0.193	
F7	0.690 ± 0.101	0.680 ± 0.060	0.128	
F8	0.681 ± 0.094	0.677 ± 0.072	0.292	
T3	0.619 ± 0.098	0.608 ± 0.079	0.300	
T4	0.623 ± 0.088	0.609 ± 0.088	0.250	
T5	0.555 ± 0.115	0.520 ± 0.089	0.008	***
T6	0.548 ± 0.117	0.508 ± 0.106	0.011	**
Fz	0.591 ± 0.096	0.571 ± 0.075	0.034	*
Cz	0.599 ± 0.091	0.576 ± 0.076	0.016	**
Pz	0.583 ± 0.124	0.575 ± 0.082	0.129	

(b) Hurst exponent

Channel	Responding	Nonresponding	p-value	Sig.
mean	9.943 ± 0.856	10.316 ± 0.985	0.001	***
std	0.658 ± 0.242	0.674 ± 0.316	0.740	
FP1	9.603 ± 1.147	9.867 ± 1.192	0.109	
FP2	9.596 ± 1.185	9.959 ± 1.181	0.031	*
F3	9.495 ± 0.983	9.933 ± 1.066	0.001	***
F4	9.544 ± 1.101	9.934 ± 1.245	0.006	***
C3	9.624 ± 0.966	9.947 ± 1.055	0.009	***
C4	9.678 ± 1.039	10.060 ± 1.073	0.003	***
P3	10.290 ± 0.732	10.614 ± 0.925	0.002	***
P4	10.284 ± 0.817	10.653 ± 0.931	0.001	***
O1	10.431 ± 0.882	10.818 ± 1.242	0.000	***
O2	10.295 ± 0.990	10.778 ± 1.262	0.000	***
F7	9.885 ± 1.320	10.243 ± 1.393	0.053	
F8	9.834 ± 1.305	10.182 ± 1.483	0.022	*
T3	9.557 ± 1.401	9.881 ± 1.436	0.082	
T4	9.544 ± 1.275	9.965 ± 1.508	0.013	**
T5	10.343 ± 0.946	10.689 ± 1.161	0.004	***
T6	10.308 ± 0.896	10.867 ± 1.119	0.000	***
Fz	10.133 ± 0.846	10.497 ± 0.901	0.003	***
Cz	10.110 ± 0.795	10.409 ± 0.828	0.012	**
Pz	10.370 ± 0.804	10.718 ± 0.892	0.003	***

(c) Largest Lyapunov exponent

Channel	Responding	Nonresponding	p-value	Sig.
mean	0.785 ± 0.114	0.762 ± 0.118	0.029	*
std	0.082 ± 0.042	0.075 ± 0.049	0.091	
FP1	0.831 ± 0.167	0.808 ± 0.149	0.307	
FP2	0.832 ± 0.170	0.796 ± 0.143	0.141	
F3	0.835 ± 0.139	0.798 ± 0.138	0.015	**
F4	0.830 ± 0.156	0.801 ± 0.150	0.080	
C3	0.820 ± 0.131	0.794 ± 0.135	0.028	*
C4	0.814 ± 0.139	0.783 ± 0.135	0.023	*
P3	0.732 ± 0.096	0.721 ± 0.106	0.044	*
P4	0.734 ± 0.105	0.715 ± 0.104	0.021	*
O1	0.717 ± 0.107	0.705 ± 0.132	0.035	*
O2	0.733 ± 0.118	0.714 ± 0.154	0.011	**
F7	0.805 ± 0.168	0.786 ± 0.166	0.287	
F8	0.811 ± 0.174	0.792 ± 0.178	0.204	
T3	0.848 ± 0.184	0.824 ± 0.182	0.249	
T4	0.846 ± 0.173	0.822 ± 0.190	0.083	
T5	0.739 ± 0.106	0.721 ± 0.128	0.016	**
T6	0.736 ± 0.108	0.704 ± 0.122	0.001	***
Fz	0.766 ± 0.110	0.735 ± 0.105	0.017	**
Cz	0.769 ± 0.097	0.744 ± 0.093	0.025	*
Pz	0.729 ± 0.105	0.710 ± 0.099	0.038	*

(d) Sample entropy

Channel	Responding	Nonresponding	p-value	Sig.
mean	1.397 ± 0.124	1.357 ± 0.112	0.006	***
std	0.087 ± 0.040	0.084 ± 0.046	0.279	
FP1	1.464 ± 0.177	1.419 ± 0.144	0.029	*
FP2	1.463 ± 0.176	1.410 ± 0.142	0.013	**
F3	1.453 ± 0.151	1.396 ± 0.130	0.002	***
F4	1.447 ± 0.165	1.406 ± 0.145	0.050	
C3	1.435 ± 0.129	1.394 ± 0.124	0.006	***
C4	1.429 ± 0.139	1.385 ± 0.121	0.008	***
P3	1.339 ± 0.108	1.309 ± 0.105	0.012	**
P4	1.336 ± 0.115	1.303 ± 0.104	0.010	***
O1	1.314 ± 0.126	1.282 ± 0.142	0.013	**
O2	1.324 ± 0.132	1.289 ± 0.164	0.006	***
F7	1.440 ± 0.169	1.394 ± 0.153	0.022	*
F8	1.446 ± 0.170	1.409 ± 0.170	0.037	*
T3	1.452 ± 0.183	1.414 ± 0.163	0.113	
T4	1.448 ± 0.169	1.417 ± 0.177	0.110	
T5	1.327 ± 0.109	1.294 ± 0.121	0.008	***
T6	1.321 ± 0.115	1.276 ± 0.124	0.001	***
Fz	1.379 ± 0.134	1.331 ± 0.112	0.002	***
Cz	1.396 ± 0.114	1.356 ± 0.099	0.006	***
Pz	1.335 ± 0.117	1.299 ± 0.104	0.011	**

(e) Higuchi's fractal dimension

Channel	Responding	Nonresponding	p-value	Sig.
mean	10.563 ± 0.803	10.749 ± 0.799	0.118	
std	0.657 ± 0.193	0.681 ± 0.191	0.401	
FP1	10.927 ± 1.030	11.057 ± 0.875	0.236	
FP2	11.019 ± 0.993	11.069 ± 0.924	0.672	
F3	10.602 ± 0.986	10.760 ± 0.910	0.172	
F4	10.557 ± 0.901	10.869 ± 1.006	0.044	*
C3	10.092 ± 1.031	10.359 ± 1.130	0.073	
C4	10.160 ± 1.029	10.452 ± 1.151	0.066	
P3	10.251 ± 1.017	10.441 ± 1.076	0.307	
P4	10.201 ± 0.994	10.404 ± 1.033	0.355	
O1	10.691 ± 0.962	10.913 ± 1.105	0.352	
O2	10.711 ± 0.982	10.872 ± 1.063	0.373	
F7	10.886 ± 0.912	11.035 ± 0.927	0.285	
F8	10.891 ± 0.882	11.110 ± 0.927	0.111	
T3	10.887 ± 1.010	10.965 ± 0.939	0.570	
T4	10.942 ± 1.024	11.230 ± 1.093	0.053	
T5	10.730 ± 0.946	10.801 ± 0.916	0.910	
T6	10.752 ± 0.901	10.775 ± 0.913	0.727	
Fz	10.197 ± 1.024	10.456 ± 0.994	0.089	
Cz	10.217 ± 1.175	10.447 ± 1.131	0.099	
Pz	9.987 ± 0.987	10.220 ± 1.082	0.215	

(f) Correlation dimension

Table 7: Comparison of mean values of measures computed from recordings obtained during both sessions (i.e. both before and after drug administration) for 50 patients responding positively to the treatment and 50 nonresponding patients.

Channel	Before	After	p-value	Sig.
mean	10.123 ± 0.766	9.763 ± 0.910	0.039	*
std	0.628 ± 0.217	0.689 ± 0.264	0.256	
FP1	9.762 ± 1.057	9.445 ± 1.220	0.171	
FP2	9.740 ± 1.132	9.452 ± 1.231	0.259	
F3	9.688 ± 0.917	9.302 ± 1.019	0.079	
F4	9.802 ± 0.941	9.286 ± 1.196	0.028	*
C3	9.821 ± 0.931	9.428 ± 0.970	0.074	
C4	9.921 ± 0.909	9.435 ± 1.110	0.018	**
P3	10.410 ± 0.638	10.170 ± 0.803	0.101	
P4	10.423 ± 0.700	10.145 ± 0.904	0.086	
O1	10.609 ± 0.769	10.252 ± 0.956	0.069	
O2	10.443 ± 0.943	10.146 ± 1.023	0.129	
F7	10.092 ± 1.289	9.679 ± 1.331	0.147	
F8	10.045 ± 1.116	9.623 ± 1.452	0.308	
T3	9.885 ± 1.324	9.230 ± 1.413	0.013	**
T4	9.777 ± 1.164	9.311 ± 1.349	0.089	
T5	10.531 ± 0.824	10.155 ± 1.028	0.096	
T6	10.500 ± 0.860	10.117 ± 0.898	0.023	*
Fz	10.238 ± 0.801	10.029 ± 0.885	0.253	
Cz	10.213 ± 0.770	10.007 ± 0.814	0.281	
Pz	10.447 ± 0.702	10.293 ± 0.895	0.514	

Channel	Before	After	p-value	Sig.
mean	10.458 ± 0.952	10.175 ± 1.008	0.202	
std	0.604 ± 0.252	0.744 ± 0.359	0.052	
FP1	10.052 ± 1.110	9.682 ± 1.252	0.216	
FP2	10.088 ± 1.150	9.829 ± 1.210	0.312	
F3	10.156 ± 1.000	9.710 ± 1.093	0.054	
F4	10.241 ± 1.117	9.628 ± 1.302	0.030	*
C3	10.115 ± 1.004	9.778 ± 1.089	0.139	
C4	10.217 ± 1.076	9.903 ± 1.058	0.101	
P3	10.700 ± 0.884	10.527 ± 0.966	0.529	
P4	10.754 ± 0.918	10.551 ± 0.943	0.391	
O1	10.815 ± 1.218	10.821 ± 1.278	0.826	
O2	10.863 ± 1.167	10.692 ± 1.357	0.613	
F7	10.386 ± 1.358	10.099 ± 1.426	0.250	
F8	10.459 ± 1.329	9.905 ± 1.587	0.085	
T3	10.126 ± 1.221	9.636 ± 1.598	0.173	
T4	10.176 ± 1.418	9.754 ± 1.579	0.173	
T5	10.773 ± 1.114	10.605 ± 1.211	0.608	
T6	10.981 ± 1.086	10.753 ± 1.151	0.387	
Fz	10.570 ± 0.913	10.425 ± 0.893	0.505	
Cz	10.447 ± 0.835	10.370 ± 0.828	0.603	
Pz	10.775 ± 0.876	10.661 ± 0.914	0.709	

Table 8: Mean values of λ_1 of responding / nonresponding patients before and after treatment.

Channel	Before	After	p-value	Sig.
mean	10.546 ± 0.755	10.581 ± 0.856	0.985	
std	0.649 ± 0.191	0.666 ± 0.196	0.708	
FP1	10.900 ± 0.951	10.954 ± 1.114	0.783	
FP2	11.005 ± 0.967	11.032 ± 1.028	0.961	
F3	10.659 ± 1.000	10.546 ± 0.979	0.610	
F4	10.486 ± 0.796	10.627 ± 0.998	0.508	
C3	10.058 ± 0.918	10.126 ± 1.143	0.794	
C4	10.100 ± 0.990	10.220 ± 1.074	0.578	
P3	10.272 ± 0.987	10.231 ± 1.056	0.806	
P4	10.245 ± 1.036	10.157 ± 0.959	0.443	
O1	10.677 ± 0.933	10.705 ± 1.000	0.985	
O2	10.728 ± 0.875	10.694 ± 1.088	0.664	
F7	10.859 ± 0.915	10.913 ± 0.918	0.889	
F8	10.864 ± 0.869	10.917 ± 0.903	0.823	
T3	10.760 ± 0.920	11.014 ± 1.087	0.192	
T4	10.837 ± 1.070	11.048 ± 0.976	0.175	
T5	10.719 ± 0.884	10.741 ± 1.015	0.800	
T6	10.739 ± 0.939	10.765 ± 0.872	0.961	
Fz	10.301 ± 0.979	10.093 ± 1.068	0.246	
Cz	10.197 ± 1.114	10.237 ± 1.244	0.985	
Pz	9.964 ± 0.917	10.009 ± 1.062	0.943	

Channel	Before	After	p-value	Sig.
mean	10.696 ± 0.848	10.802 ± 0.751	0.523	
std	0.653 ± 0.162	0.710 ± 0.214	0.240	
FP1	11.006 ± 0.834	11.107 ± 0.920	0.412	
FP2	11.043 ± 1.010	11.094 ± 0.840	0.931	
F3	10.648 ± 1.049	10.873 ± 0.741	0.175	
F4	10.785 ± 1.056	10.953 ± 0.958	0.281	
C3	10.308 ± 1.145	10.410 ± 1.126	0.877	
C4	10.412 ± 1.180	10.493 ± 1.133	0.573	
P3	10.402 ± 0.993	10.480 ± 1.163	0.877	
P4	10.341 ± 0.903	10.468 ± 1.155	0.783	
O1	10.968 ± 1.251	10.858 ± 0.947	0.865	
O2	10.789 ± 1.021	10.955 ± 1.108	0.513	
F7	10.968 ± 0.948	11.102 ± 0.910	0.470	
F8	11.088 ± 0.952	11.133 ± 0.912	0.937	
T3	10.769 ± 0.914	11.161 ± 0.932	0.027	*
T4	10.992 ± 1.040	11.468 ± 1.104	0.015	**
T5	10.743 ± 1.041	10.858 ± 0.779	0.203	
T6	10.847 ± 0.993	10.703 ± 0.831	0.421	
Fz	10.402 ± 1.023	10.510 ± 0.973	0.642	
Cz	10.472 ± 1.167	10.422 ± 1.106	0.754	
Pz	10.249 ± 1.132	10.190 ± 1.040	0.719	

Table 9: Mean values of D_2 of responding / non-respoding patients before and after treatment.

Channel	Before	After	p-value	Sig.
mean	0.764 ± 0.102	0.806 ± 0.122	0.064	
std	0.075 ± 0.038	0.089 ± 0.046	0.173	
FP1	0.806 ± 0.154	0.855 ± 0.178	0.187	
FP2	0.810 ± 0.162	0.854 ± 0.176	0.192	
F3	0.809 ± 0.128	0.860 ± 0.147	0.093	
F4	0.799 ± 0.134	0.861 ± 0.171	0.070	
C3	0.797 ± 0.116	0.842 ± 0.141	0.092	
C4	0.785 ± 0.122	0.843 ± 0.149	0.021	*
P3	0.718 ± 0.084	0.745 ± 0.105	0.202	
P4	0.720 ± 0.096	0.748 ± 0.113	0.262	
O1	0.699 ± 0.080	0.735 ± 0.127	0.323	
O2	0.718 ± 0.102	0.747 ± 0.132	0.275	
F7	0.781 ± 0.163	0.828 ± 0.171	0.210	
F8	0.783 ± 0.146	0.838 ± 0.196	0.259	
T3	0.810 ± 0.169	0.885 ± 0.192	0.037	*
T4	0.818 ± 0.164	0.874 ± 0.178	0.095	
T5	0.719 ± 0.085	0.759 ± 0.120	0.173	
T6	0.722 ± 0.104	0.750 ± 0.111	0.113	
Fz	0.753 ± 0.100	0.779 ± 0.119	0.295	
Cz	0.758 ± 0.090	0.780 ± 0.103	0.323	
Pz	0.720 ± 0.098	0.738 ± 0.113	0.533	

Channel	Before	After	p-value	Sig.
mean	0.749 ± 0.106	0.775 ± 0.129	0.319	
std	0.064 ± 0.041	0.086 ± 0.053	0.046	*
FP1	0.789 ± 0.129	0.827 ± 0.166	0.371	
FP2	0.787 ± 0.136	0.805 ± 0.150	0.692	
F3	0.776 ± 0.122	0.819 ± 0.152	0.175	
F4	0.772 ± 0.130	0.830 ± 0.164	0.082	
C3	0.781 ± 0.122	0.806 ± 0.146	0.395	
C4	0.767 ± 0.126	0.799 ± 0.143	0.187	
P3	0.711 ± 0.086	0.731 ± 0.122	0.676	
P4	0.709 ± 0.089	0.721 ± 0.117	0.855	
O1	0.705 ± 0.119	0.705 ± 0.144	0.618	
O2	0.702 ± 0.119	0.726 ± 0.183	0.942	
F7	0.775 ± 0.156	0.797 ± 0.177	0.450	
F8	0.766 ± 0.159	0.818 ± 0.194	0.129	
T3	0.792 ± 0.151	0.857 ± 0.205	0.180	
T4	0.795 ± 0.168	0.849 ± 0.209	0.229	
T5	0.715 ± 0.116	0.726 ± 0.140	0.994	
T6	0.702 ± 0.117	0.706 ± 0.128	0.907	
Fz	0.729 ± 0.097	0.742 ± 0.114	0.895	
Cz	0.744 ± 0.090	0.745 ± 0.097	0.843	
Pz	0.705 ± 0.087	0.715 ± 0.111	0.965	

Table 10: Mean values of SE, responding and nonresponding patients before and after treatment.