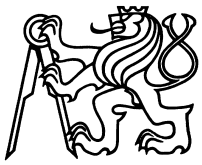
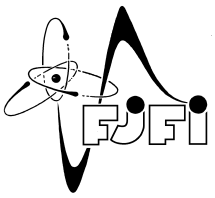


CZECH TECHNICAL UNIVERSITY IN PRAGUE
Faculty of Nuclear Sciences and Physical Engineering



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: **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, January 7, 2019 Miroslav Kovář

Název práce:

Analýza biomarkerů psychiatrických pacientů pomocí analýzy EEG signálu a strojového učení

Autor: 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: Miroslav Kovář

Abstract:

Key words:

Contents

Introduction

Depression is one of the most common brain disorders - it affacts 121-300 million people worldwide, and this number is expected to increase in the future [47] [41]. Although effective treatments are known, World Health Organization estimates that fewer than half of those affected receive those treatments. Major barriers include insufficient resources, lack of properly trained practitioners, inaccurate assessment and misdiagnosis. [41]

For these reasons, it is important that affordable, fast, accurate, and easy to use methods for its diagnosis are developed. Although electroencephalography (EEG) [1] may be one such method thanks to its comparatively low-cost and easy recording process, comparatively little research has been focused on this area. Non-linear dynamical analysis in particular has been proven very effective at diagnosing mental disorders, and this work is aimed at contributing to this important and relatively new topic.

In **Chapter 1**, we present some of the classical theory and methods of non-linear dynamical analysis and chaos theory, with focus on the terms used in the following text.

In **Chapter 2**, we introduce the basic concepts and terminology used in design and evaluation of convolutional neural networks.

In **Chapter 3**, we describe the methods proposed, experiments performed, and results obtained.

[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.

Chapter 1

Non-linear 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 most simple 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 non-linear dynamics: [4]

Deterministic process : periodic (or quasi-periodic), fully describable by its Fourier spectrum.

Stochasitc process : influenced by forces unpredictable under all circumstances.

Non-linear dynamical analysis studies a third class of processes, which are irregular, non-periodic, yet still deterministic. Every non-periodic, deterministic process is non-linear (bot 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 to be developed only with the rising feasibility of numerical simulations, peaking in 1980s. [4]

1.1 EEG signal

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. [39]

Although EEG has significantly lower spatial resolution in comparison with other diagnostic techniques such as functional magnetic resonance sampling (fMRI) and magnetoencephalography (MEG) [57] and enables measuring only neural activity near the cortical surface, as a depression diagnostic tool, it has numerous benefits. Importantly, its significantly lower costs [65] [23], high portability, and ease of operation imply increased availability to the patients [55]. Moreover, it is perfectly noninvasive, which means less complications such as claustrophopia or anxiety [37].

Although the science of EEG signal analysis as a diagnostic tool brings compelling clinical promise as a result of the aforementioned benefits, it also presents multiple technical and conceptual challenges.

Definition 1 ([44]). *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 $k \in \mathrm{N}$, $P[X_{t_1}, X_{t_2}, \ldots, X_{t_n}] = P[X_{t_1+k}, X_{t_2+k}, \ldots, X_{t_n+k}]$, i.e. the joint probability distribution of $\{X_t\}_{t\in\mathbb{Z}}$ is not a function of time. It is called **non-stationary**, if it is not stationary.*

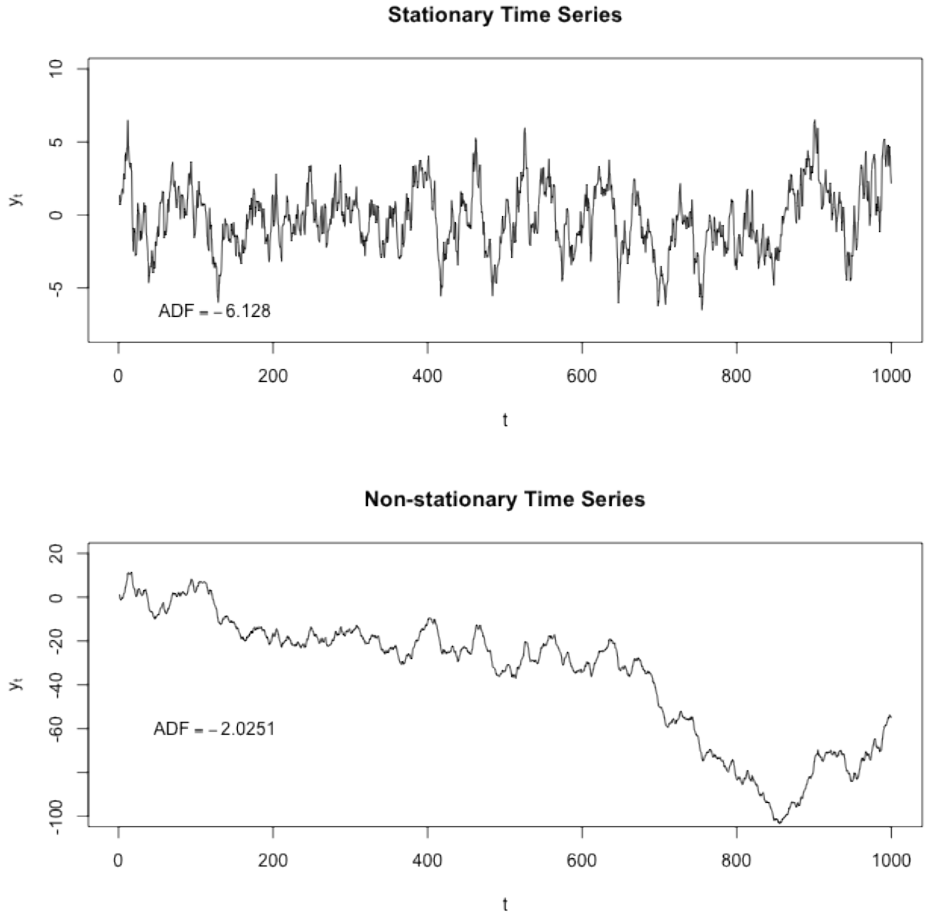


Figure 1.1: A comparison of stationary and non-stationary time series. (Courtesy: Protonk)

Definition 2 ([7]). *A series $\{X_t\}_{t\in\mathbb{Z}}$ is called (noisy chaotic)* ***non-linear****, if it satisfies the relation*

$$X_t = f(X_{t-1}) + \epsilon_t \tag{1.1}$$

for a general $f : \mathbb{R} \to \mathbb{R}$.

EEG signals are prone to be infected with *noise* due to imperfect isolation from surrounding environment. They are known to be *transient, non-Gaussian, non-stationary and nonlinear* [29] [58]. 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. And finally, EEG records are relatively high dimensional - 16 electrodes sampling at 256 Hz result 4096 data points par second.

Moreover, 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 band have been linked to various mental states [10] [8] and diseases such as epilepsy [56], tremor [35], Parkinson's disease and depression [33]. 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. [64]

1.2 Limitations in application to EEG

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 non-linear differential equations, principles derived from nonlinear dynamics are applicable to characterization of these neuronal systems. [29]

The approach of estimating a finite embedding dimension, however, has been doubted by some of the most prominent figures in the field of non-linear dynamical analysis, such as the originators of Grassberger-Procaccia algorithm. 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. Presumably, the the observed estimates of low dimension are due to artifacts or limited data size. [20] [45]. However, as we will see in Section 1.8, 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 non-linear dynamical analysis still manages to capture some of its important aspects.

1.3 Dynamical systems

Definition 3 ([4]). *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 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 [32] and, since, in EEG signal analysis both M and d are unknown, we have no alternative but to work in Euclidean M.

Definition 4 ([4]). *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 mapping $f^t : M \to M$ satisfying [2]

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

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.*

In late 1800s, H. Poincare 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. [59]

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 non-linear dynamical system is a system where the differential equations describing its dynamics are non-linear. Unlike in a linear system, changes in the initial state of a non-dynamical system are allowed to have a non-linear relationship to the state space trajectory of the system. [29]

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, $s(\xi(t))$ for some (generally non-invertible) measurement function $s : \mathbb{R}^d \to \mathbb{R}^n$, where $n \ll d$. Morover, the time between subsequent measurements is limited by a sampling frequency and the values of the variables themselves are taken and stored with a limited precision.

Add a few examples (Lorenz, Rossler, Mackey-Glass). Create my own plots instead of reusing.

1.3.1 Nonstationarity

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. [25] 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 numerical applications. [4] However, a relevant subset of nonstationary systems can be defined more explicitly:

Definition 5 ([4]). *A dynamical system is called* ***nonautonomous*** *if its dynamics* $\mathbf{F}$ *are explicitly dependent on time:*

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

[1] In this work, we are going to assume that brain is a deterministic dynamical system, and that any stochastic component is small and does not change non-linear properties of the system. Thus, by the term dynamical system, we will always mean a deterministic dynamical system.

[2] This condition is equivalent to satisfaction of the assumptions of the uniqueness theorem of differential equations.

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. [25]

This weaker definition employs only linear quantites, and is therefore not strictly suitable for nonlinear time series analysis. On the other hand, statistical tests of this property exist. In this text, we use the following test discussed by H. Isliker and J. Kurths in [25].

This technique attempts to approximate a projection of so called *physical invariant measure* ρ defined as [15]

$$\rho := \lim_{T\to\infty} \frac{1}{T} \int_0^t \delta_{\mathbf{x}(t)}\, \mathrm{d}t$$

into one coordinate of the state space (given by the time series). Loosely 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".

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

Since this measure is ergodic [3], the ergodic theorem basically states that the space and time averages are equal almost everywhere, i.e.

$$\int_{\text{statespace}} f(\mathbf{x})\rho(\mathrm{d}\mathbf{x}) = \lim_{T\to\infty} \frac{1}{T} \int_0^T f(\mathbf{x}(t))\, \mathrm{d}t$$

for any $f \in C$ defined on the state space.

Let x_1 represent the measured quantity, and N be the length of the time series. The range of the time series is divided into K 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:

Is this confusing? Should I just say that they intervals are "equiprobable"?

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

where $\xi_{[a,b]}$ 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

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

If the system is stationary, then the distribution for the first half of the time the same. 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 ξ^2-test:

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

where $Z = \lceil N/2 \rceil = \sum_k n_k^{\text{half}}$. [25]

[3] This means, loosely, that it is "decomposable" into several different pieces, each again invariant.

1.4 Attractor

Depending on the properties of **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.

Definition 6 ([28]). *A dynamical system is called dissipative, when it is the case that*

$$E[\mathrm{div}\mathbf{F}] < 0, \tag{1.2}$$

where 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 can be seen on Figure 1.2.

Since most physiogenerated signals are chaotic, their analysis is concerned primarily with *chaotic* (strange) *attractors*. These attractors are relatively complex, characteristic of dynamical systems with extending volumes in some directions. This property results fast divergence of two initial states, one of which has nonzero component in the direction of growth, i.e. sensitive dependence on the initial conditions. However, since atractors are bounded, the divergence eventually stops and the two trajectories fold together. This continuous expansion and folding may create an attractor with a *fractal structure* (an example of such an attractor is shown on Figure 1.3). [4] For our purposes it is sufficient to say that this means that these attractors can be characterized as having (quantifiable) self-similarity.[4] However, the following definition related to fractals will be useful in Section 1.5.1:

Definition 7 ([17]). *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 which can cover F. Then, the **box-counting dimension** (also known as Minkowski–Bouligand dimension) is defined as*

$$d_0(F) = \lim_{\epsilon \to 0} -\frac{\log N_\epsilon(F)}{\log \epsilon}, \tag{1.3}$$

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 $\epsilon \to 0$. [17]

1.5 State space reconstruction

Broadly, one possible approach to non-linear time series analysis consists of the following steps:

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. [58]

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

Connect this to the content of this section. Expand on the steps.

[4] Cantor set being a canonical example of self-similarity.

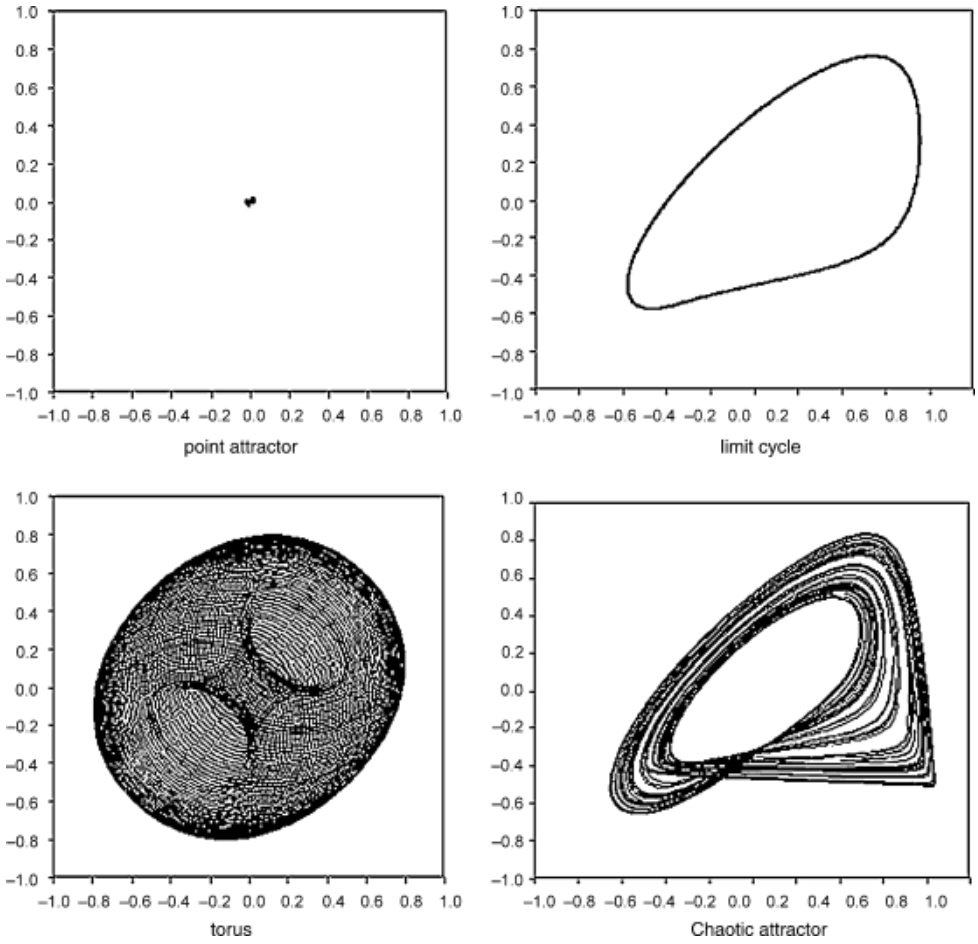


Figure 1.2: Visualization of four common attractor types (units are arbitrary). Left to right, top to bottom: **Point attractor** is the only type of attractor of linear deterministic dissipative systems. It consist of a single final state to which all points from the corresponding region of attraction 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 is completely deterministic. [4] ([58])

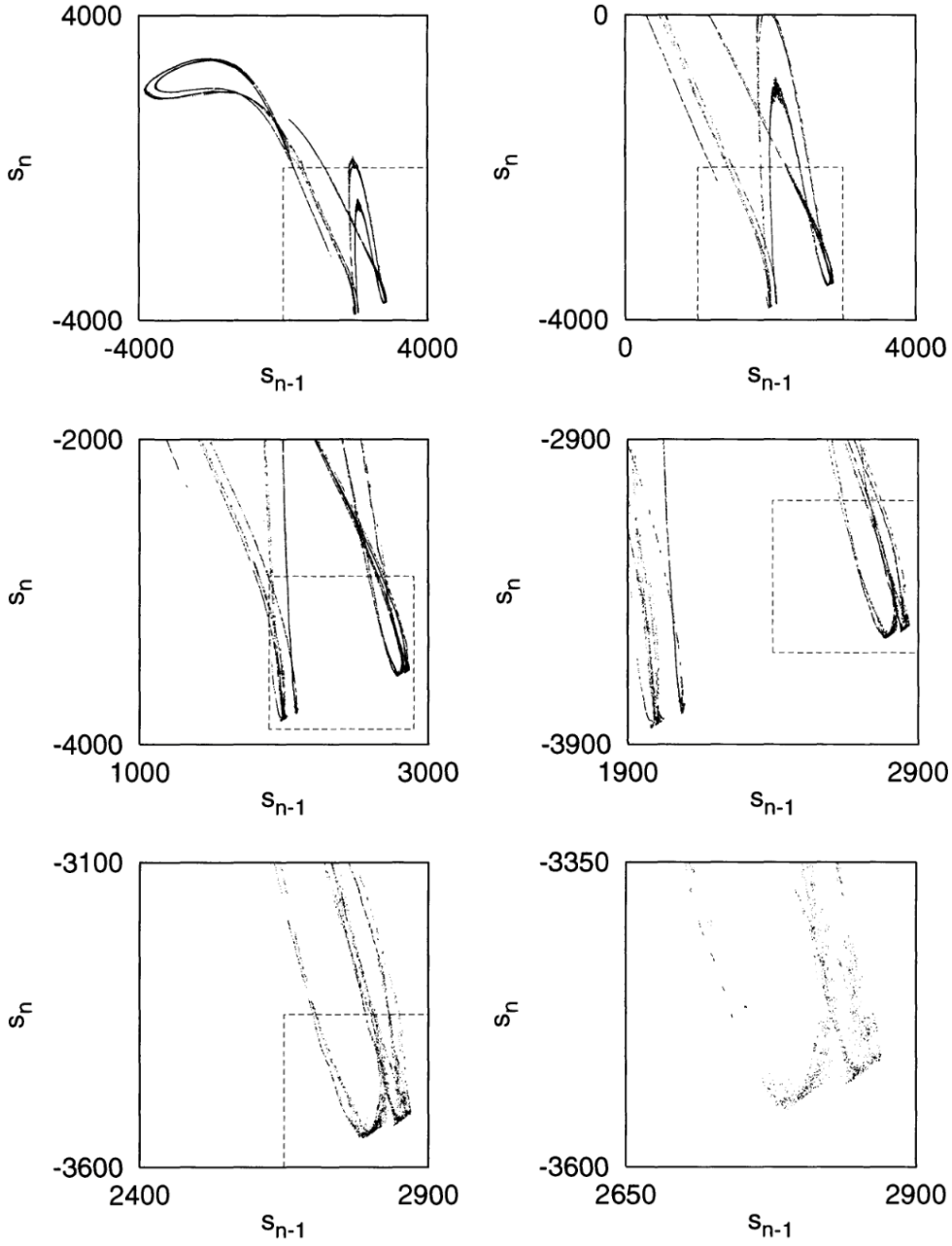


Figure 1.3: Noise-reduced visualization of successive enlargements of highly self-similar attractor. ([28])

1.5.1 Embedding

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 (a sequence of scalar measurements), 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*.

To this goal, let $\mathbf{s}_n$ be the reconstructed vector we are trying to find, and let us have a time series of scalar measurements of a quantity depending on the current state of the system:

$$x_n = s(\xi(n\Delta t)) + \eta_n(n\Delta t)\,, \tag{1.4}$$

where ξ is a state space vector, $s(\cdot)$ is a measurement function and η_n is a measurement noise. Furthermore, let us consider a function $\Phi : M \to \mathbb{R}^m$, such that $\mathbf{x}_n = \Phi(\xi(n\Delta t))$. Such function is called an **embedding**. In the following, we will discuss what properties does Φ have to satisfy so that it provides useful information about the true state space trajectories.

Before we do that, let us mention the following. As we have stated in Section 1.3, our observations are formed by application of non-invertible 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. Therefore, 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. [4] 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 [28].

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.

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, in order to preserve these properties in the embedded space, we may require Φ to preserve the differential structure of the state space M. This corresponds to the tangent space $D\Phi$ also being a one-to-one mapping.

Add images illustrating these two conditions.

1.5.2 Method of time delays

There are two common 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 time series corresponds to a coordinate of the state space vector.

Int the following text, we will focus on the first one, because we are not using the second one in this thesis, and it has been widely criticised.

Add some citations.

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

maps, which is known as Whitney's theorem. [66] [5]) 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 non-linear dynamical system analysis by N. H. Packard in 1980 (although it was already being used in different fields in 1950s [4]). 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. [42] 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 $t - \tau_w \leq t' \leq t$. 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[6]. Using the notation of 1.4, the time delay reconstruction is then formed by the following vectors:

$$\mathbf{x}_n = (x_{n-(m-1)\tau}, x_{n-(m-2)\tau}, \ldots, x_{n-\tau}, x_n)\,, \tag{1.5}$$

for $n > (m-1)\tau = \tau_w$. [28]

A year after Packard's discovery, in [60], 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 non-linear time series analysis and can be stated as follows:

Theorem 1 ([60])**.** *Let M be a compact[7] 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*

$$\phi_{(s,f^t)}(x) = (s(\xi), s(f^{-\tau}(\xi)), \ldots, s(f^{-2d\tau}(\xi))), \tag{1.6}$$

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. [54]

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, [53])**.** *Let A be a compact fractal with box-counting dimension d_A, and let A be a subset of a m-dimensional manifold. Then*

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

[5]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 $d_1 + d_2 \geq m$.

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

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

In conclusion, Theorem 1 and Theorem 2 together ensure that when m is chosen such that $m > d_A$i (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". [61]

1.5.3 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: [12]

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 (1.4). [4]

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 [12], 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.

1.5.4 Time delay selection

A careful reader might have noted that the results of theorems in Section 1.5.2 do not depend on the value of the delay τ.[8]. 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 $\tau_w = \tau(m-1)$ [31], as we will see, the choice of time delay has effects independent of the choice of embedding dimension, and vice versa.

For example:

1. The embedding may fail to be a one-to-one map due to finite precision, or presence of noise in the data. [4]

[8] This is because of the fact that the measurements are infinitely precise. [12]

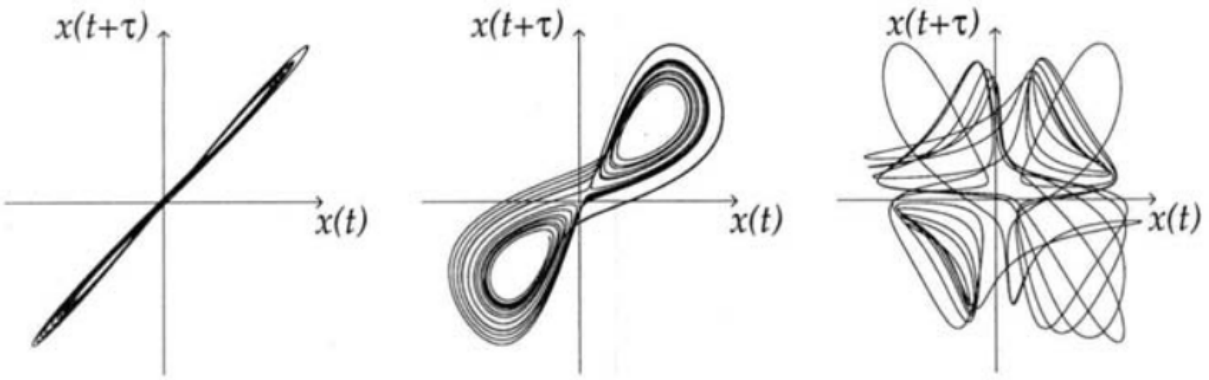


Figure 1.4: Time delay reconstructions of the Lorenz attractor for different values of τ. Figure on the left hand side shows choice of small τ and represents the case of redundance - the states concentrate along the main diagonal. Figure in the middle shows a successful reconstruction (although not an 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. ([4])

2. Highly chaotic systems with large Lyapunov exponents (see Section 1.6.1) and large dimension, projection to a low dimensional time series causes explosion in the noise amplification. 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. [12]

3. It was shown that increasing τ leads to rise in entropy. [27]

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.

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 1.4, left hand side). Moreover, in this case, the effect of noise is amplified. [12]

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 1.4, 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).

1.5.4.1 Autocorrelation

From the above, we understand that statistical non-correlation between values of coordinates of the reconstructed vectors $\mathbf{x}_n$ are desirable property of a time delay embedding. 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 $A(0)/e$ [58], $1 - A(0)/e$ [28], or even the first local minimum [3, 1] or the first 0 crossing [28].

Definition 8 ([28]). ***Autocorrelation*** *$A : \mathbb{R} \to \mathbb{R}$ for time delay τ is given by*

$$A(\tau) = \frac{E[(x_i - \overline{x})(x_{i-\tau} - \overline{x})]}{\sigma^2},$$

where $\overline{x}$ is the mean of the time series, and σ^2 is its standard deviation.

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. [28] 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. Thus, this method, since it computes correlations only between two successive coordinates, is generally useful only for low dimensional systems.

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 the the variation of noise. This means that τ should be selected such that $A(\tau) > A(0) - \sigma^2_{\text{noise}}/\sigma^2_{\text{signal}}$. [28]

1.5.4.2 Delayed mutual information

Another commonly used method is to use the first minimum of the *time delayed mutual information*. [19]

Definition 9 ([28]). *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 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. [18]

Another criticism of delayed mutual information that some systems exhibit slowly decaying mutual information which has no minima. [34]

1.5.4.3 Average displacement from diagonal

Average displacement from diaognal 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 main diagonal, and found ADFD to be the most computationally efficient, robust to noise, and accurate. [51] They also experimentily identified optimal τ as the one for which the slope of ADFD drops below 40% of its initial value.

1.5.4.4 Singular values analysis

All the approaches described so far address the issue of irrelevance, but not that of redundance. In fact, based mostly on empirical, rather than the most time delay estimation techniques optimize for the following criteria [9]: [31]

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 noticeably 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. [4]

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

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

then

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

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.

[9] However, additional criteria may arise depending on the particular application.

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. [4]

One drawback of this method is that its evaluation is largely subjective. 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. [36]

1.5.4.5 Integral local deformation

The uniqueness theorem of differential equations requires that any no trajectories in the state space intersect. 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. [9] 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 $\mathbf{x}_i^{com}(0)$ 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:

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

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:

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

1.5.5 Embedding dimension selection

1.5.5.1 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. [30]

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 (see Figure []). 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 use denote by $y^{(r)}(n)$ the r-th nearest neighbor of $y(n)$. Then, let $R_m(n,r)$ denote the Euclidean distance between $y(n)$ and its neighbor:

$$R_m(n,r) = \sqrt{\sum_{k=0}^{m-1} [x_{n+k\tau} - x_{n+k\tau}^{(r)}]^2}$$

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:

$$\left[\frac{R_m^2(n,r) - R_{m+1}^2(n,r)}{R_m^2(n,r)}\right]^{1/2} = \frac{x_{n+k\tau} - x_{n+k\tau}^{(r)}}{R_m(n,r)} > R, \tag{1.7}$$

where $R \in \mathbb{R}$ is some threashold. The m for which the relative proportion of false neigbors to all neigbors 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 $R_m(n,r) \approx R_A$, 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. [30]

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

$$\frac{R_{m+1}(n,r)}{R_A} > A \tag{1.8}$$

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

$$R_A = \frac{1}{N}\sum_{n=1}^{N} [x_n - \overline{x}]^2 .$$

Although this technique is commonly used, it is not without its drawbacks. An obvious point is that altough it is true that distance between neigbors 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. [31]

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. [31]

Lastly, in presence of measurement noise, the proportion of false neigbors may increase after transition to a higher dimension, since even identical vectors will diverge. [28]

1.5.5.2 Average false neighbors

This technique by Cao [11] 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

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

where $n(i, m)$ denotes the nearest neighbor of vector $\mathbf{x}_i$ in dimension m, and $\|\cdot\|_\infty$ denotes the Chebyshev norm [10]. 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.

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

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

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.

1.6 Non-linear measures

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

[10] This norm suggested by the author, but presumably, another norm can be used.

1.6.1 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 (see Figure).

Lyapunov exponents can be used to quantify this sensitivity. 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, for simplicity, 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 $T^{(n)}$, where

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

Isn't there a theorem for that?

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

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

Then, the Lyapunov exponents can be defined as [22]

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

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 [14]). It is commonly agreed that estimating Lyapunov exponents is even more difficult than esimating correlation dimension [4], although they have been successfully employed in EEG analysis. [52, 24, 58] 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. [22] 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. [50]

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

$$d(i) = d(0)e^{\lambda_1(i\Delta t)}, \quad d(i) \ll 1, \quad i \to \infty, \quad d(0) \to 0, \tag{1.9}$$

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. [28] Equivalently, $1/\lambda_1$ is (on average) an upper bound on predictability in the system. [4] 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. [50]

Say what different values of λ_1 say about the system.

1.6.1.1 Rosenstein's algorithm

In the following, we will describe *Rosenstein's algorithm* for computation of the largest Lyapunov exponent. [50] 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 1.5.1). The suggested method of time delay selection is the autocorrelation method (see Section 1.5.4.1).

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 will impose a condition on their temporal separation:

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

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

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

Note that the sampling period Δt plays no role - one can decide to set $\Delta t = 1$ and work with units of time series indeces instead of seconds interchangeably. Relatedly, we can even rescale or shift the data, since Lyapunov exponents are invariant under any smooth invertible map.

Another prominent and widely used algorithm for estimation of the largest Lyapunov exponent is Wolf's algorithm [67], but due to its instability and the impossibility of distinguishing exponential divergence, it cannot be recommended. [28]

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. [28] 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.

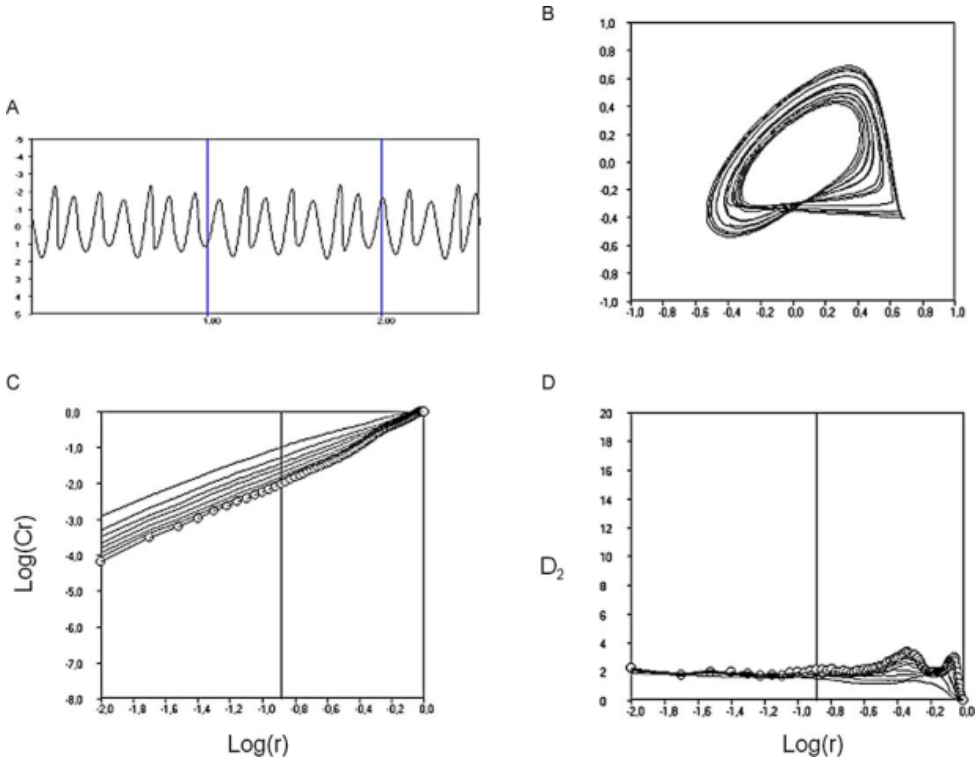


Figure 1.5: Computation of the correlation dimension [58]. TODO: Add description.

1.6.1.2 Dataset size requirements

The minimum dataset requirements was estimated by Eckmann and Ruelle in [16] 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.

1.6.2 Correlation dimension

The world of mathematics offers numerous definitions of dimension (box-counting dimension (1.3), Hausdorff dimension, information dimension, etc.) and similar quantities, but many of them can be regarded as variations of the following, simple and intuitive analogy:

$$[[61]]\text{bulk} \approx \text{size}^{\text{dimension}} \implies \text{dimension} = \lim_{\text{size}\to 0} \frac{\log \text{bulk}}{\log \text{size}}.$$

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. [4]

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 7), 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

$$d_2(A) = \lim_{r\to 0} \frac{\log C(r, 2)}{\log r}.$$

1.6.2.1 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.

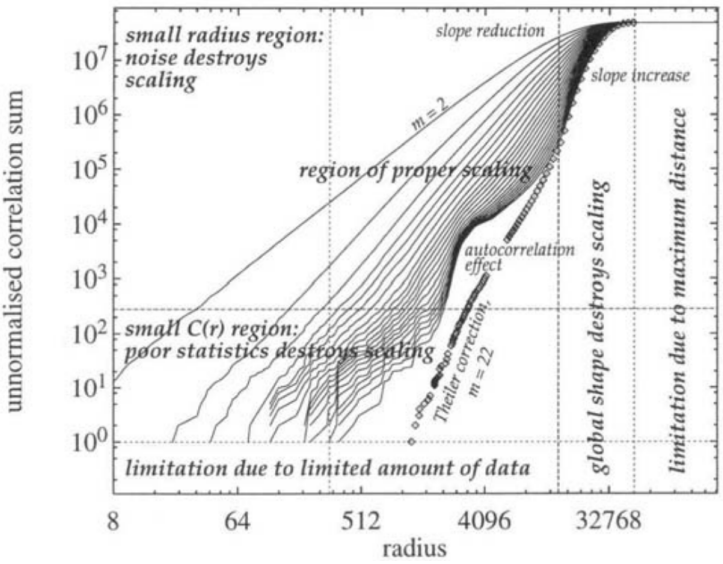


Figure 1.6: Log-log plot of typical behavior of $C(r)$.

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

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

which corresponds to the fraction of 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. [21]

Typical behavior of the correlation sum is shown in Figure 1.6. 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. [4]

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

$$d_2(r) = \frac{\partial \log C(r)}{\partial \log r}.$$

In our implementation, we perform a least squares fit of values $(\log r, \log C(r))$ for a window of 6 neighboring points for each sampled r. Expected behavior of the resulting function in a favorable case can be seen in Figure 1.7.

1.6.2.2 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-

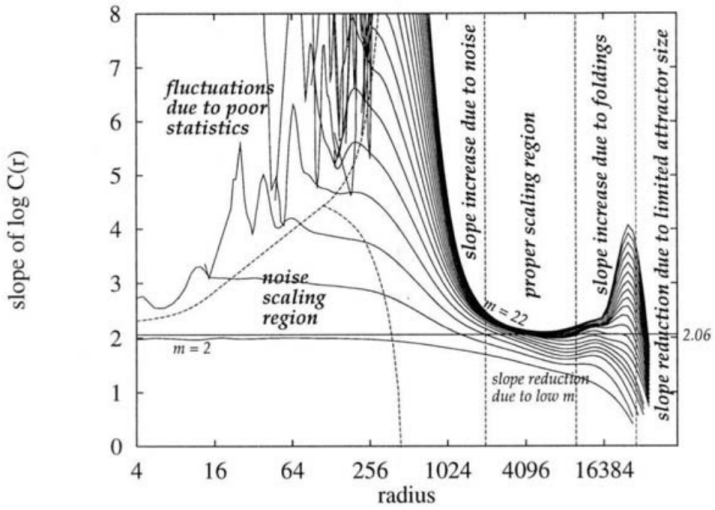


Figure 1.7: Log plot of a typical local dimension estimates in favorable case.

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, 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 $\rho = 5\%$. [4]

1.7 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. [46] Hence, interpretation of results obtained by non-linear analysis require judgement. For example. obtaining finite-dimensional estimates for d_2 is not evidence of non-linearity, but may indicate lack of data, measurement error, on numerical inaccuracies. The result of these algorithms does not include any error estimate, and sometimes even non-linearity of the underlying process is uncertain. So, one may ask: would we obtain the same estimate for data with the same (non-measured) linear properties as the original, but differs in the (measured) non-linear property? In the following, we will describe a method for answering this question.

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

To this end, we construct a Monte Carlo hypothesis test of non-linearity. 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

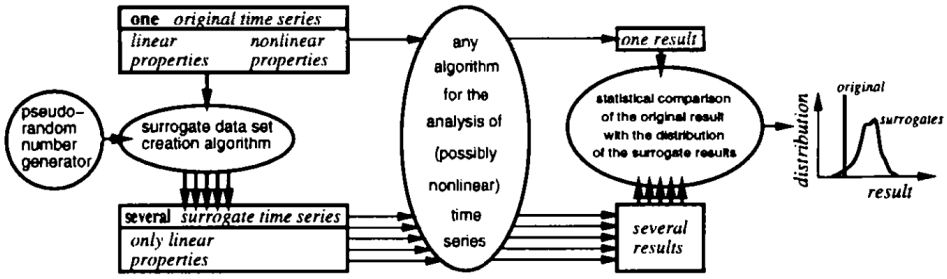


Figure 1.8: [4]

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 1.8.

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

$$\mathcal{S} \equiv \frac{|Q_{\text{orig}} - \mu_{\text{surr}}|}{\sigma_{\text{surr}}},$$

where Q_{orig} is the statistic computed for the original time series, and μ_{surr}, σ_{surr} are the mean and variance of the statistic computed for the surrogate time series. [63] 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. [28]

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 $n_s = 38$. [4]

1.7.0.1 Generating improved amplitude adjusted surrogates

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

$$x_{t+1} = \mu + \sum_{j=0}^{p-1} a_j x_{t-j} + \sigma e_t, \tag{1.12}$$

with unknown parameters $a_j, e_t, \mu, \sigma \in \mathbb{R}$. [63]

If the computed non-linear statistic depends on the free parameters in AR(p) (1.12) (which is not true for certain statistics, such as 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. [63][11] 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).

[11] 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.

However, there is a drawback of FTPR. It has been shown that if the amplitudes of AR(p) are not Gaussian (as in (1.12)), e.g. non-linear,then the surrogates created using this method show non-linear behavior. [28] Rarely do the amplitudes of an experimental process follow a Gaussian distribution. Hence, we change our model to correspond a non-linear, time independent filter applied to the output of AR(p). Surrogate creation algorithm for this model was described by Theiler in [63]: 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. [62]

e.g. or i.e. here?

Even this method is not without its drawbacks: due to the final reordering, the original power spectrum is slightly distorted in the surrogate. In [62], 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: [4]

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:

   Create a set of random numbers with Gaussian distribution.

   Rank order the original time series, and reorder the random numbers created in the previous step such that they achieve the same ordering as the original time series.

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

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

1.8 Applications in disease diagnosis

This section is probably not sufficiently exhaustive.

Although non-linear 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 non-linear time series analysis for clinical depression diagnosis is relatively limited. [47]

It has been found that the EEG dynamics of depressed patients exhibit more predictability than those of non-depressed ones, with this indicator receding after treatment. [38] [43]

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 relative to control. [52]

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. [2]

In the same year, Hosseinifard et al. extracted Higuchi's correlation dimension, Lyapunov exponents and Higuchi's fractal dimension from 4 EEG channels of 90 patients split evenly between depressed and non-depressed subjects, achieving 90% accuracy using a logistic regression classifier. [24]

In 2013, Bachmann et al. compared two non-linear analysis methods, spectral assymetry index (SASI) and Higuchi's fractal dimension (HFD), 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 HFD provided true detection rate of 94% in the depressives and 76% in the controls. [6]

Sleep disorder diagnosis may also relevant to this work for the very close connection of depression with disturbed sleep and insomnia [40]. The first study emplying techniques of non-linear analysis on human EEG was published in 1985 and dealt with sleep recordings. [5] This early success sparked intensive research focus on applying non-linear analysis to sleep data, thus generating relatively large amount of results.

Many studies focused of 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 dimensionality lower values of the largest Lyapunov exponent [58].

Chapter 2

Non-linear analysis approach

2.1 Dataset

The EEG recordings were performed by and obtained from the Czech National Institute of Mental Health. The dataset comprises total of 133 subjects, 104 women and 29 men, ranging in age from 30 to 65 (47.7 ± 9.58). Geriatric Depession Scale 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).

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 drugs 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.

A typical recoding can be seen on Figure 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. [13]

In further analysis, unless otherwise specified, recordings were shortened to a fixed length. To balance the data requirements (see Sections 1.6.1.2 and 1.6.2.2), decrease in dataset size due to removal of too short recordings, and stationarity (see Section 1.3.1), the threshold was selected to be 60 s (15 000 datapoints per time series), resulting in exclusion of 26 recordings from the total of 266.

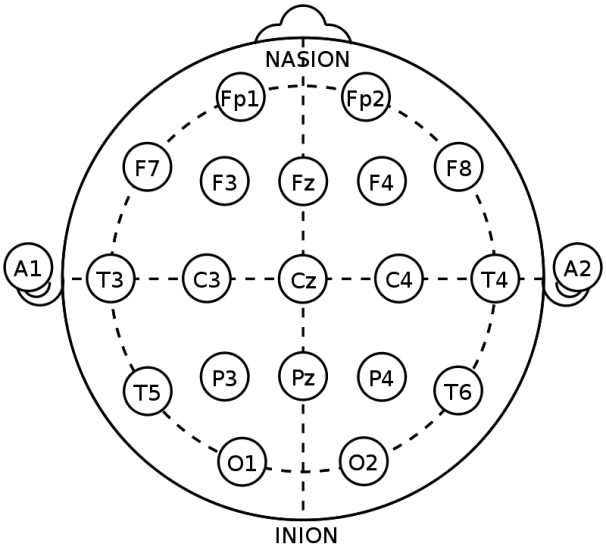


Figure 2.1

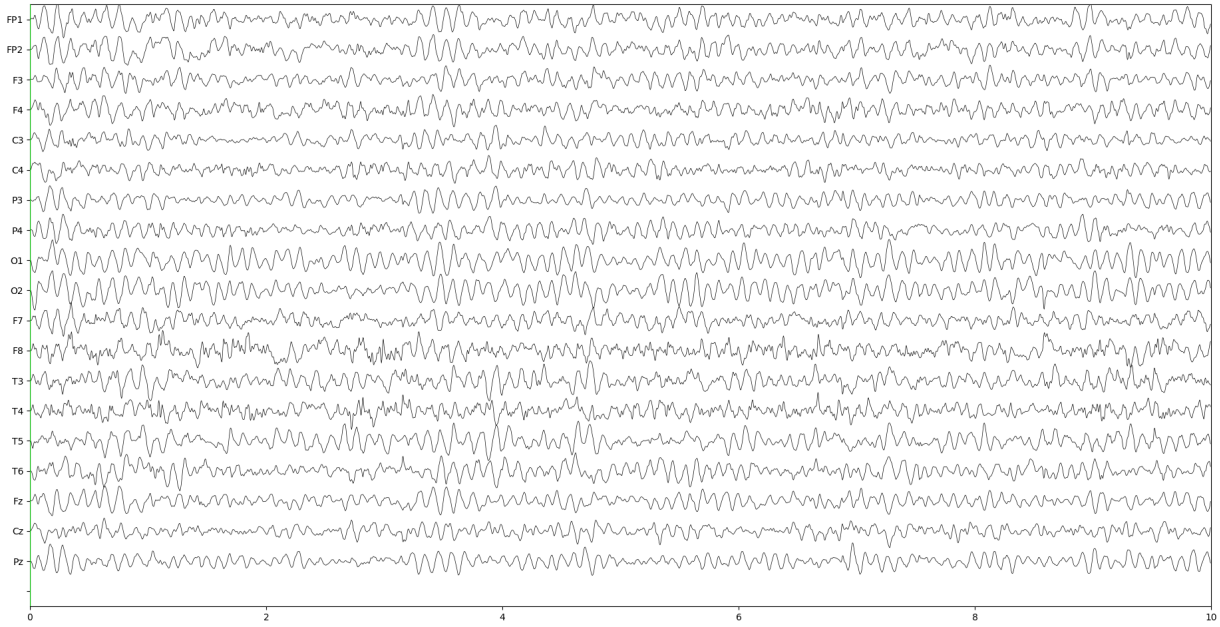


Figure 2.2

Duration\Av. p-value	Original	Surrogate
10 s		
30 s		
60 s		

Table 2.1: Results of stationarity tests.

In some studies, 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). [24] Sometimes, it is suggested to notch filter at power line frequencies (40 Hz or 50 Hz). However, some authors suggest that linear filtering may adversely affect the results of non-linear analysis. [4] Others, on the other hand, observed that simple linear filtering does not influence the reconstruction of embedding space considerably. [48] If quality of the data is sufficient, filtering is not necessary. [26] By visual inspection, we found our data to be of sufficient quality and therefore decided to not risk influencing the results by filtering.

In fact, we tried both, but for filtered the results looked slightly more uniform.

2.3 Stationarity

Stationarity was evaluated on multiple time scales using the stationarity test described in Section 1.3.1. For results, see Table 2.1.

2.4 State space reconstruction

2.4.1 Surrogate data

Describe process of generating surrogate data.

See Figure 2.3.

2.4.2 Time delay

In order to estimate the time delay, we used the following techniques:

1. Reconstruction plots
2. Autocorrelation $A(\tau)$ (see Section 1.5.4.1)
3. Delayed mutual information $\mathcal{I}(\tau)$ (see Section 1.5.4.2)
4. Average displacement from diagonal (ADFD) (see Section)
5. PCA reconstructions comparison (see Section)
6. Integral local deformation (ILD) (see Section)

In this section, we will analyze the results of these techniques for time series obtained from FP1 electrode of patient 75, second session, shown in Figure 2.3. The time series was clipped to 60 s (15000 data points). In the following sections , we will explain how these techniques were used to obtain estimates of individual non-linear measures.

Add concrete references.

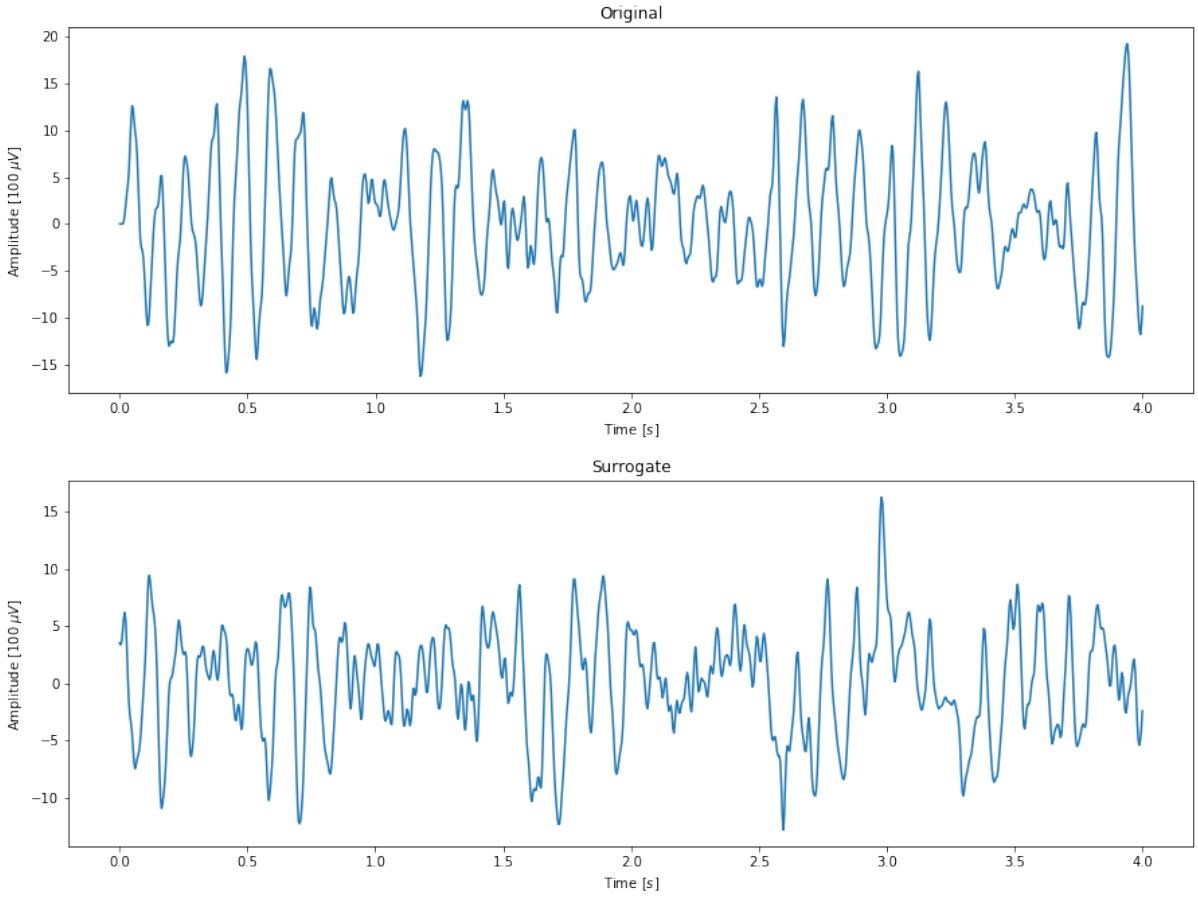


Figure 2.3: The first 4 *s* of a time series and its surrogate.

Figure 2.4 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 on Figure 2.5. First local minima of DMI and first τ for which $A(\tau) \leq 1/e$ are marked by yellow dots. For this channel, these are $\tau_{\text{DMI}} = 7$ and $\tau_A = 6$. These values were computed for all channels of this recording, and their distribution for both DMI and autocorrelation can be seen in Figure 2.6. For this patient, autocorrelation shows less variance and lower suggested time delays. This behavior was observed across patients.

I should provide mean τ reported by $A(\tau)$ and DMI for all recordings.

Figure 2.7 shows singular values of the PCA reconstruction as functions of τ. The two prominent singular values correspond to the main axes of the attractor. We can see several collapses: smaller ones at $\tau = 5$ and $\tau = 7$, and larger one at $\tau = 14$, corresponding to the sharp peak of ILD on Figure 2.9. For $\tau = 3$, the three largest singular values show convergence - however, the small singular values suggests that the attractor has not fully unfolded in their corresponding directions. Overall, this behavior suggests $\tau_{\text{svd}} = 6$ as optimal. Note that results of this technique are difficult to evaluate by an automatic procedure.

Is this correct?

Not sure if this wording is precise and understandable.

The results obtained by ADFD can be seen in Figure 2.8. The average displacement tends to increase with m, is not monotonically increasing on its domain, and saturates for relavely small values of τ - thus, the estimated time delays are (consistently)lower than those obtained by other techniques. Moreover, this technique requires prior selection of m. However, the algorithms for selection of m, require estimation τ, making this technique impractical.

We need to smartly separate general observations with this analysis.

The result of ILD, the most powerful algorithm for estimation of the embedding parameters we used, can be seen on Figure 2.9. There is a clear minimum at $\tau_{\text{ILD}} = 4$, and the ILD curves become very similar for approximately all $m \geq 10$, except near the minimum, where they converge slower. Almost identical behavior was observed across all channels in this recording. This algorithm is computationally expensive (it takes around an hour to generate a single plot), and so is impractical for large datasets.

There are interesting patterns in these plots across patients and channels.

As explained in Section 1.5.4, 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 non-linear 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 [30]).

Find some studies doing this also. Is there a way to justify this theoretically?

Table 2.2 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 wil ILD.

Add description of Theiler window to 1.6.1.1

In the Section 2.5.1, we will show the effects of increasing τ on the average divergence.

Should I implement some better methods? MI and acorr are not practical for EEG and other high dimensional systems.

2.4.3 Embedding dimension

For estimating the embedding dimension, we used combination of *false nearest neighbors* (FNN) algorithm described in Section 1.5.5.1 and average false neighbors (AFN) described in Section 1.5.5.2. The convergence of ILD curves and saturation of correlation dimension also provides insight into optimal choice of embedding dimension.

	Optimal time delay estimate
Reconstruction plot	-
Autocorrelation	6, 4
Delayed mutual information	7
Singular values analysis	6
Average displacement	2, 3
Integral local deformation	4

Table 2.2: Optimal time delay estimates of individual techniques for patient 75, second session.

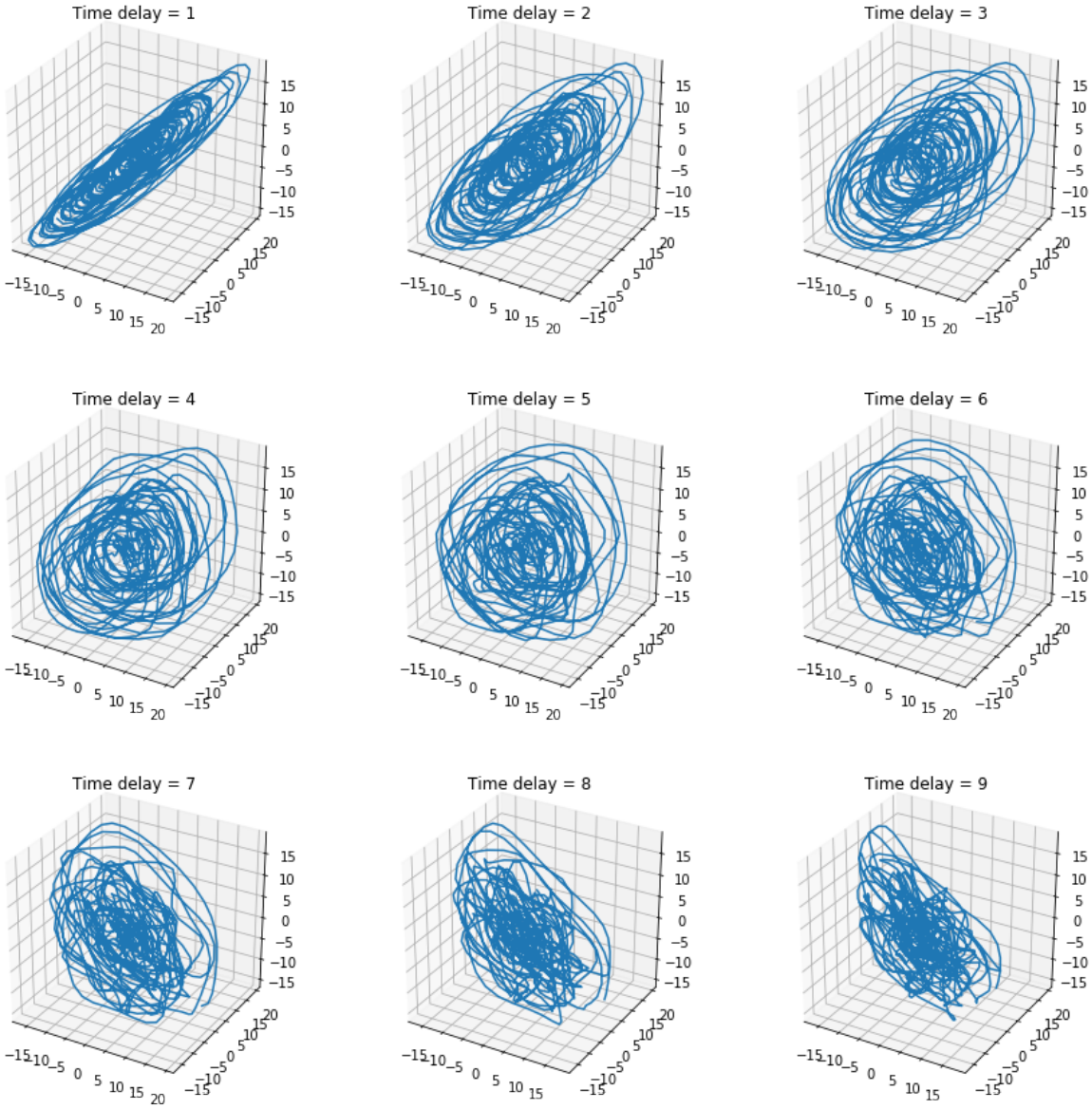


Figure 2.4: 3D time delay reconstructions for various values of τ.

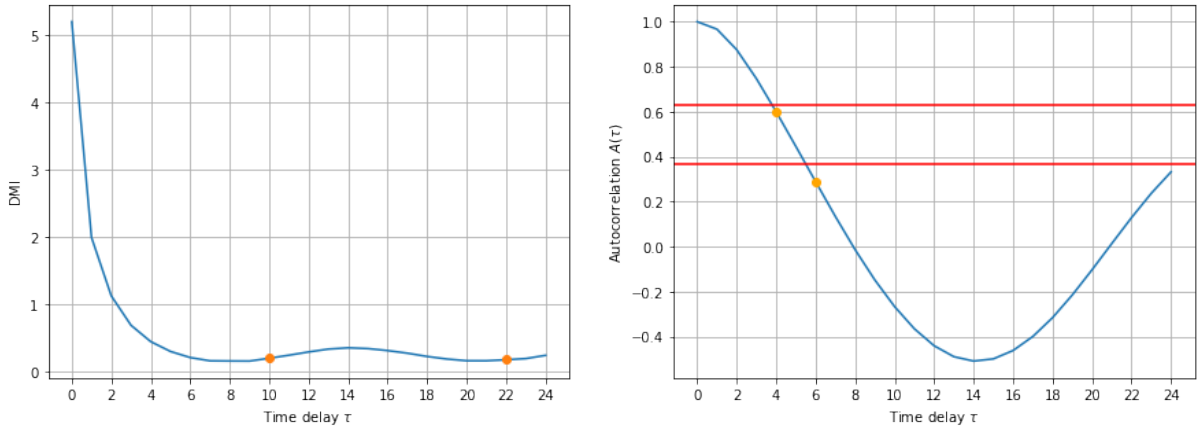


Figure 2.5: Delayed mutual information and autocorrelation as functions of τ. The red line shows threshold values $1 - 1/e$ and $1/e$ respectively. The plots of surrogate data are equivalent.

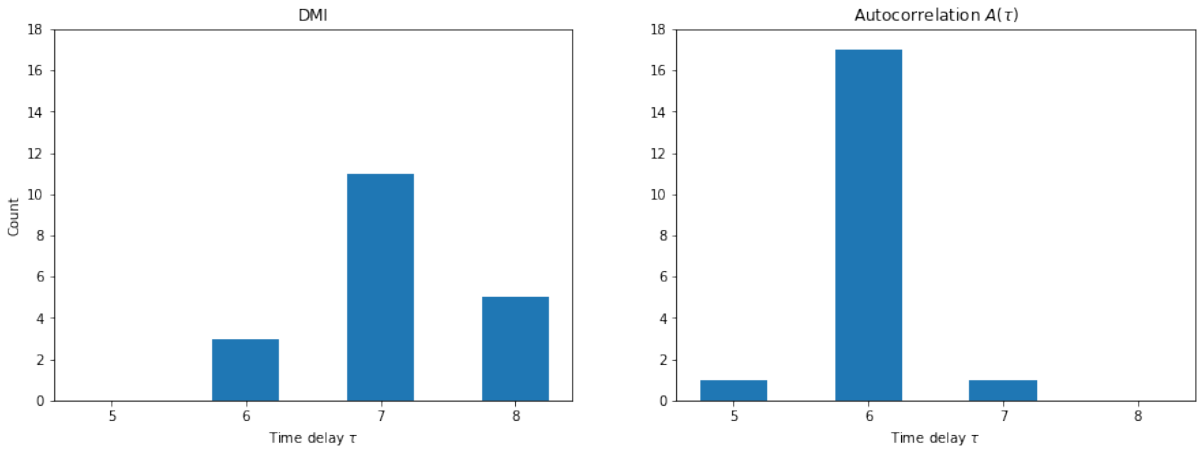


Figure 2.6: Distributions of time delays computed using delayed mutual information and autocorrelation for threshold $1/e$.

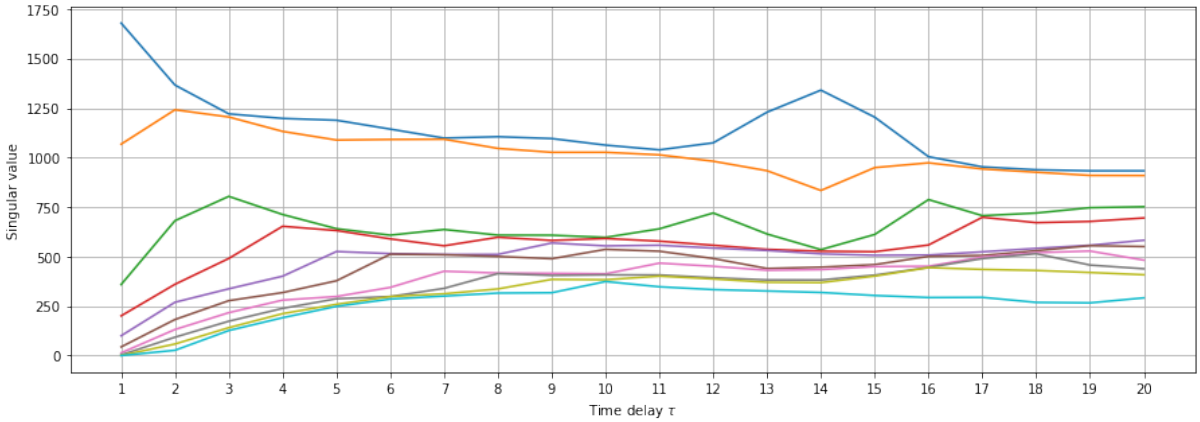


Figure 2.7: Plot of singular values as functions of τ for $m = 10$.

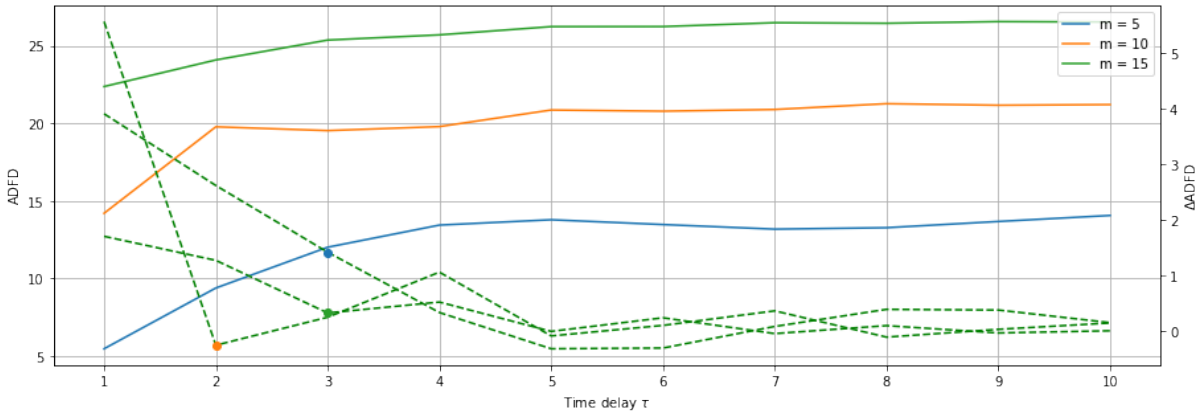


Figure 2.8: Plot of average displacement from diagonal for $m = 10$.

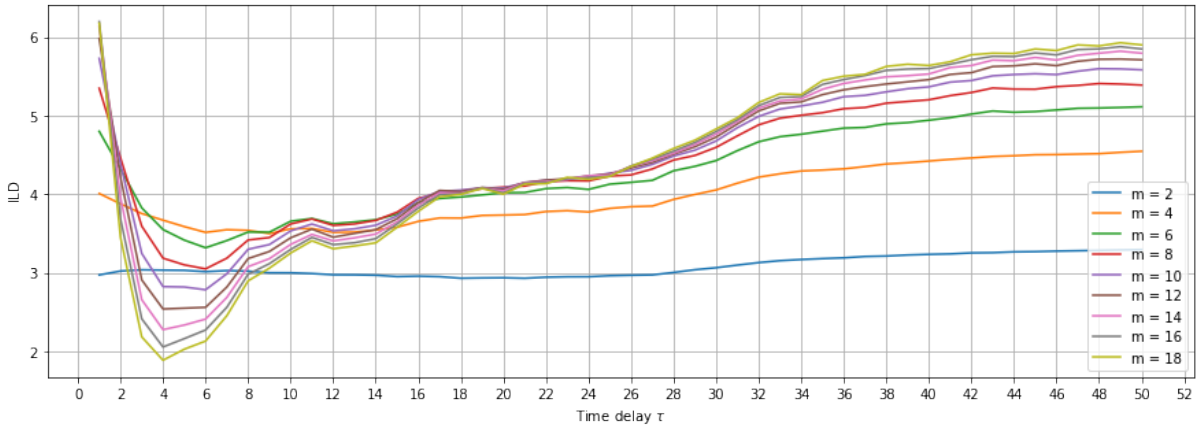


Figure 2.9: Plot of integral local deformation. The parameters used for this computation are $q_{\max} = 10$, $t_e = 3$, $N_{\text{ref}} = N_v$, $k = 20$ and $w_t = 10$.

The percentage of reported false neighbors depends strongly on the selected values of R and A from equations (1.7) and (1.8). This is illustrated on Figure 2.10, showing the percentage of false 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 is indicates that, regardless of τ, the same percentage of near neighbors changes their distance proportionally with increase in m. As explained in Section 1.5.5.1, 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.

Actually explain it there - nearest ≠ close, etc. . . , [30]

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 and AFN.

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 2.11. The E_2 statistic, tends to oscillate in small neighborhood of value 1, which is an indecation of nondeterminism [11].

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.

We plotted the value of correlation dimension against m for various values of τ (see Figure 2.16) - for details about the computation, see Section 2.5.2.

Add statistics of m computed this way.

2.5 Estimation of non-linear features

2.5.1 Largest Lyapunov exponents

For all computations of the largest Lyapunov exponent, we used the Rosenstein's algorithm [50] described in Section 1.6.1.1, 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.

Why? This is unexpected.

Figure 2.12 shows divergence plots for different values of the embedding dimension m and time delay τ. Let us remind the reader that longer scaling regions correspond to higher certainty of the estimate. The short scaling regions and high slopes for small embedding dimension may 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. Therefore, selecting proper embedding dimension based on divergence plots is a balancing act 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 computed using Kantz' algorithm also exhibits this behavior. The explanation is as follows: let $x_1, x_2, \ldots, x_N$ represent equidistantly sampled time series, and $y_i \in \mathbb{R}^m$ an embedded point in the reconstructed orbit. Then

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

and so if $x_i \approx x_{i+\tau}$ for enough i, then $y_i \approx y_{i+\tau} \approx y_{i+2\tau} \approx$, and this oscillation with period τ gradually vanishes over $m - 1$ periods.

Why doesn't this happen always?

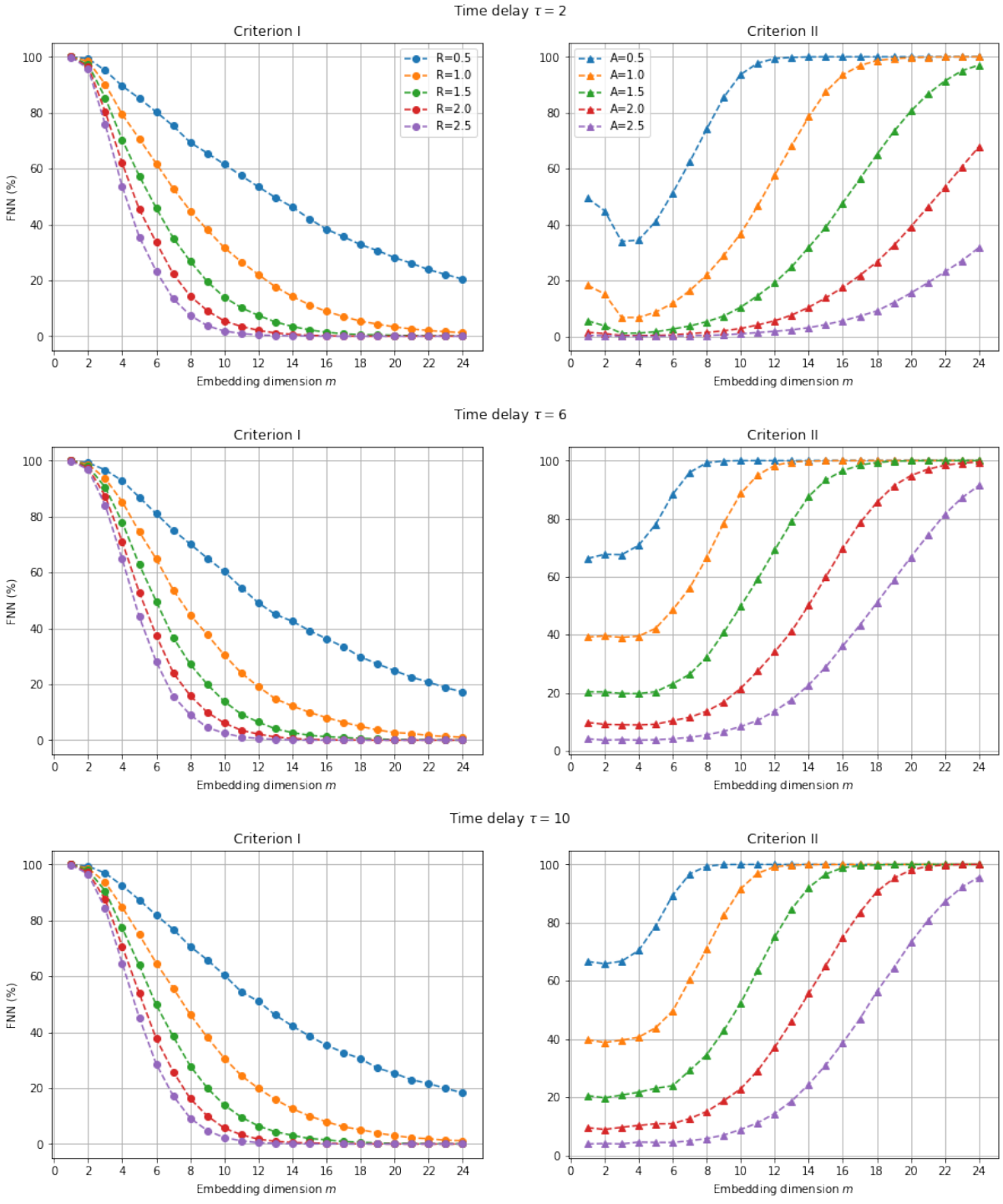


Figure 2.10: The effect of values of the tolerance parameters on the percentage of false neighbors reported by I. criterion (1.7) and II. criterion (1.8), Theiler window $w_t = 50$.

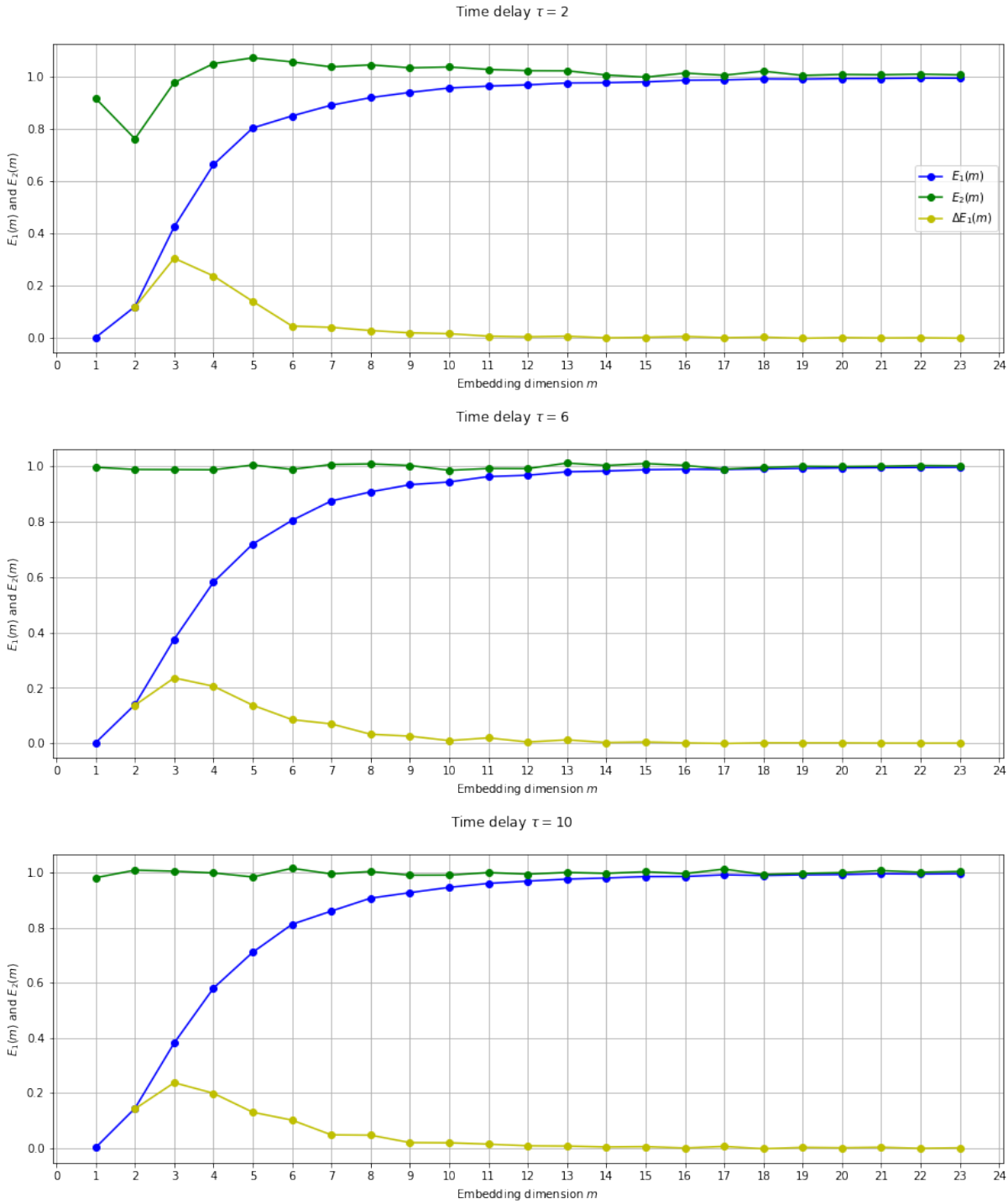


Figure 2.11: The results of AFN for varying values of time delay τ, Theiler window $w_t = 50$.

This effect can be aleviated by choosing smaller τ, but we are unaware of any way of eliminating it completely for Rosenstein's algorithm. [1] Another metric to optimize after length of the scaling region is, therefore, reduction of the degree of deformation of the scaling region by the periodic oscillations.

Are there others? Somehow introduce randomness?

We computed LLE for fixed m and τ, and then by selecting them automatically. Describe the reasoning for selecting the fixed ones, and the process of automatic selection via the method in the previous section.

Oscillation-like behavior was observed for white noise data in [50], and for periodic data with period equal to the dominant period of the system in [28].

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 1.4) 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 [28].

We observe very similar behavior of the average divergence for the surrogate data. This, together with the observations made in previous sections, gives rise to the hypothesis of lack of chaos in the data. We tested the hypothesis of linear Gaussian process in Section 2.5.7.

Good thing is that Eckmann's algorithm gives similar results with very different approach. Maybe we shoud incorporate this somehow?

To compute the LLE estimates with automatic selection of proper embedding parameters, we proceeded as follows. First, we found the 60 s subsection of the time series with the lowest p-value of the χ^2 stationarity test using moving window of length 15000 and slide 100. Selection of time delay was done using autocorrelation function with threshold $1 - 1/e$. 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. $m_1 = \arg\min_{m' \in \{1,\dots,20\}} \mathrm{FNN}(m')$. The tolerance parameters wer $R = 2.5$, $A = 2.0$ and $w_t = 50$. Moreover, we found the first embedding dimension m_2 for which $E_1(m_2) - E_1(m_2 - 1) < 0.008$. The selected embedding dimension was $m = \lceil m_1 + m_2)/2 \rceil$. The length of the scaling region $t_{\max} = m\tau$ and the Theiler window $t_w = 50$.

Add the average m's computed this way.

2.5.2 Correlation dimension

For corr dim, we also used two ways to compute it - automatic and fixed. Explain here why have we chosen $m = 10$ and $\tau = 3$.

To compute the correlation sum $C(r)$, we used classical Grassberger-Procaccia algorithm described in Section 1.6.2 using Chebyshev metric, $w_t = 50$, for values of r either in geometrical progression of 100 values from 0.05 to 10 or by an automatic procedure described further. Then, these $(r, C(r))$ pairs were used to compute local least square fits of the equation $C(r) = r^{d_2}$ inside windows of length 7 for each pair.

This paragraph can be much improved (wording, etc.).

Figure 2.13 shows log-log plots of normalized correlations sum $C(r)$ against radius r for varying values of time delay τ. There are clear straight lines indicating expected relationship $C(r) \propto r^{d_2}$. We can see that the lines shift to the right, increasing their slopes with m. The correlation sum is almost independent of time delay.

Figure 2.14 shows the log plot of local slope of of $\log C(r)$ as a function of r. There are no apparent scaling regions at all. Morever, by comparing with the same plot for iAAFT surrogate of the same time series (see Figure 2.15), we cannot even reject the hypothesis of a linear stochastic process.

We decided to compute the correlation dimension as follows. First, as with computation of the Lyapunov exponent, we use a moving window of length 15000 datapoints and shift 100 to locate 60 s section of the time series with the lowest p-value for the χ^2 stationarity test described in Section . Then, we create embeddings for 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 $[r_{\mathrm{lower}}, r_{\mathrm{upper}}]$, where r_{lower} corresponds to the average nearest neighbor distance on the reconstructed attractor, r_{upper} is given by

Add description of the test.

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

[1] There are algorithms, such as modifications of Wolf's algorithm [49], whose results are almost independent on τ (as is theoretically expected).

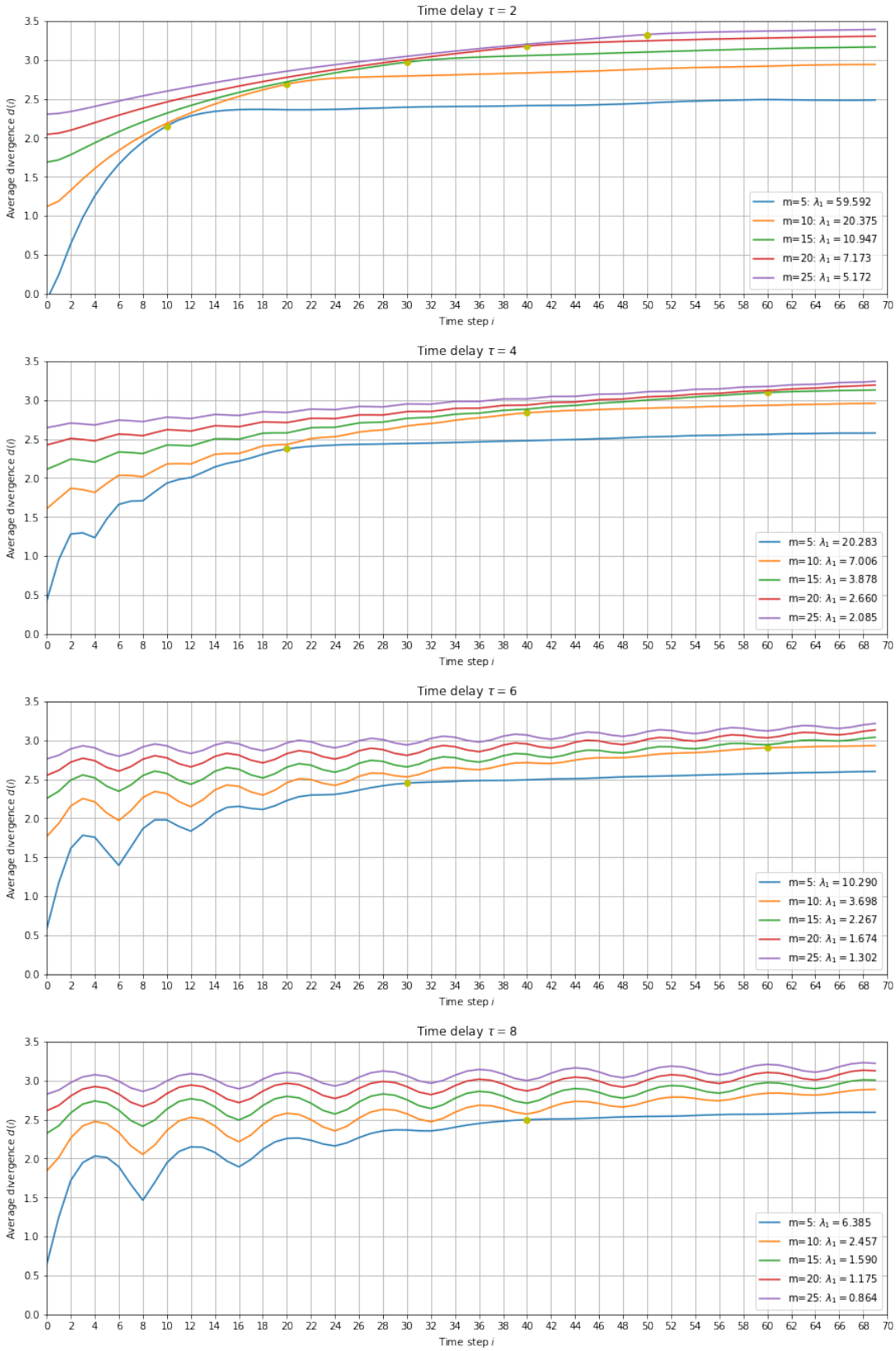


Figure 2.12: Average divergence plots for varying values of m and τ.

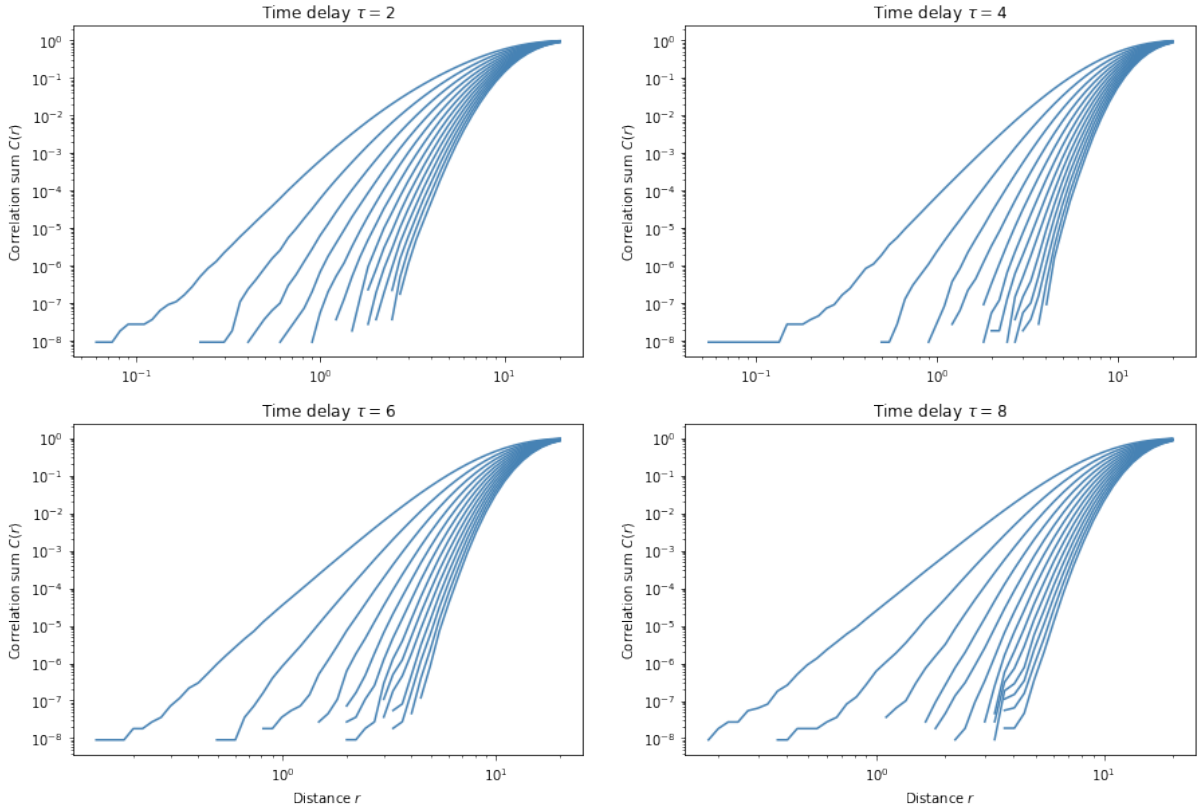


Figure 2.13: Normalized correlation sum as a function of radius r for dimensions in range from 5 to 30 (from left to right).

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 [4].

Figure 2.16 shows d_2 computed this way as a function of the embedding dimension m for varying values of the embedding dimension τ. There d_2 are no signs of saturation, correlation dimension reaches a global maximum and then starts to decrease.

Conclusion? No finite value, or no chaos?

2.5.3 Detrended fluctuation analysis

2.5.4 Hurst exponent

2.5.5 Higuchi fractal dimension

2.5.6 Sample entropy

2.5.7 Surrogate analysis

2.6 Analysis of measure distributions between groups

2.6.1 Before and after treatment

As the first step of our analysis, we conducted an investigation of the differences in the non-linear 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

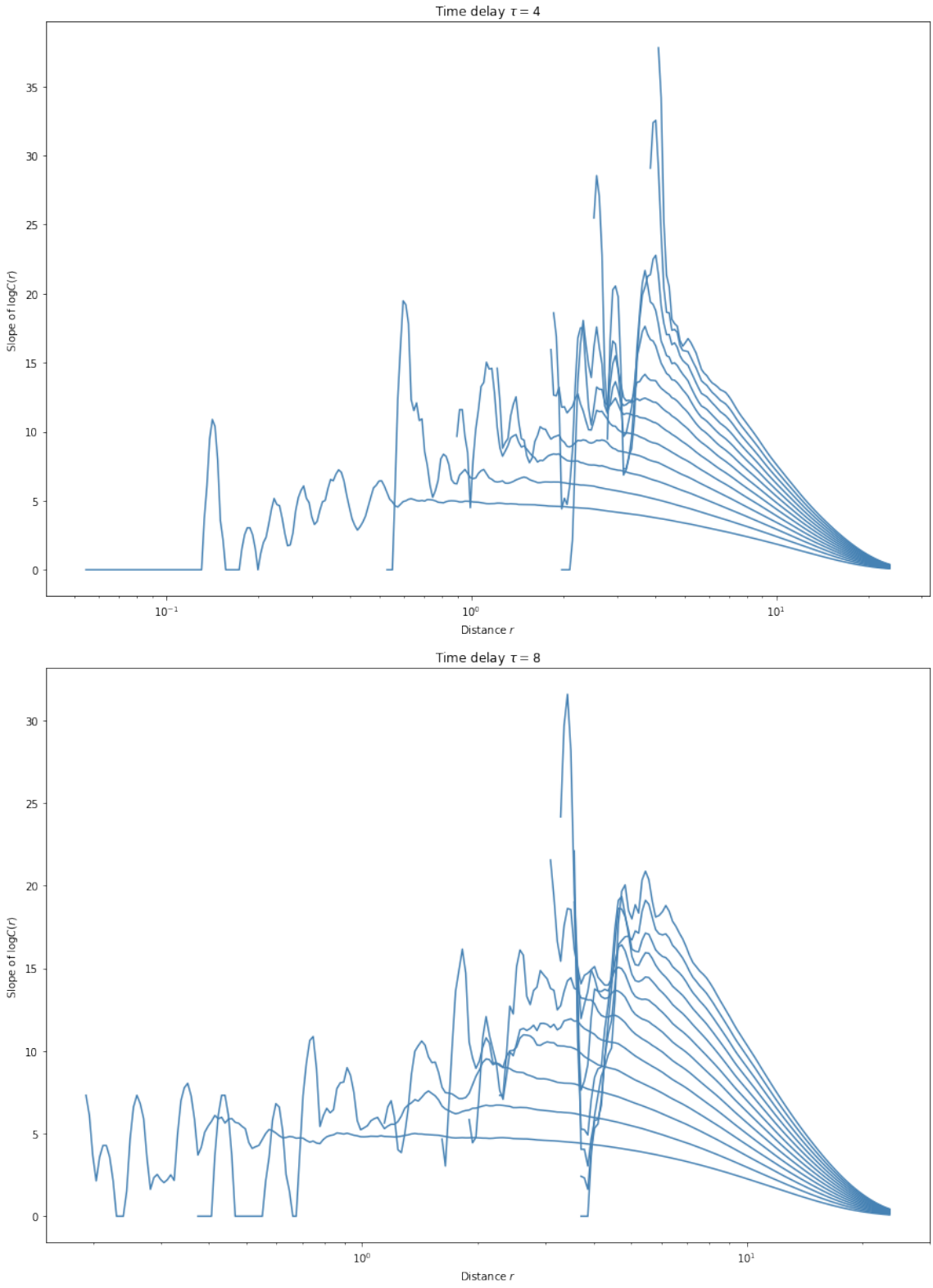


Figure 2.14: Local correlation dimension d_2 as a function of radius r for dimensions in range from 5 to 30 (from bottom to top) and time delays $\tau = 4$ and $\tau = 8$.

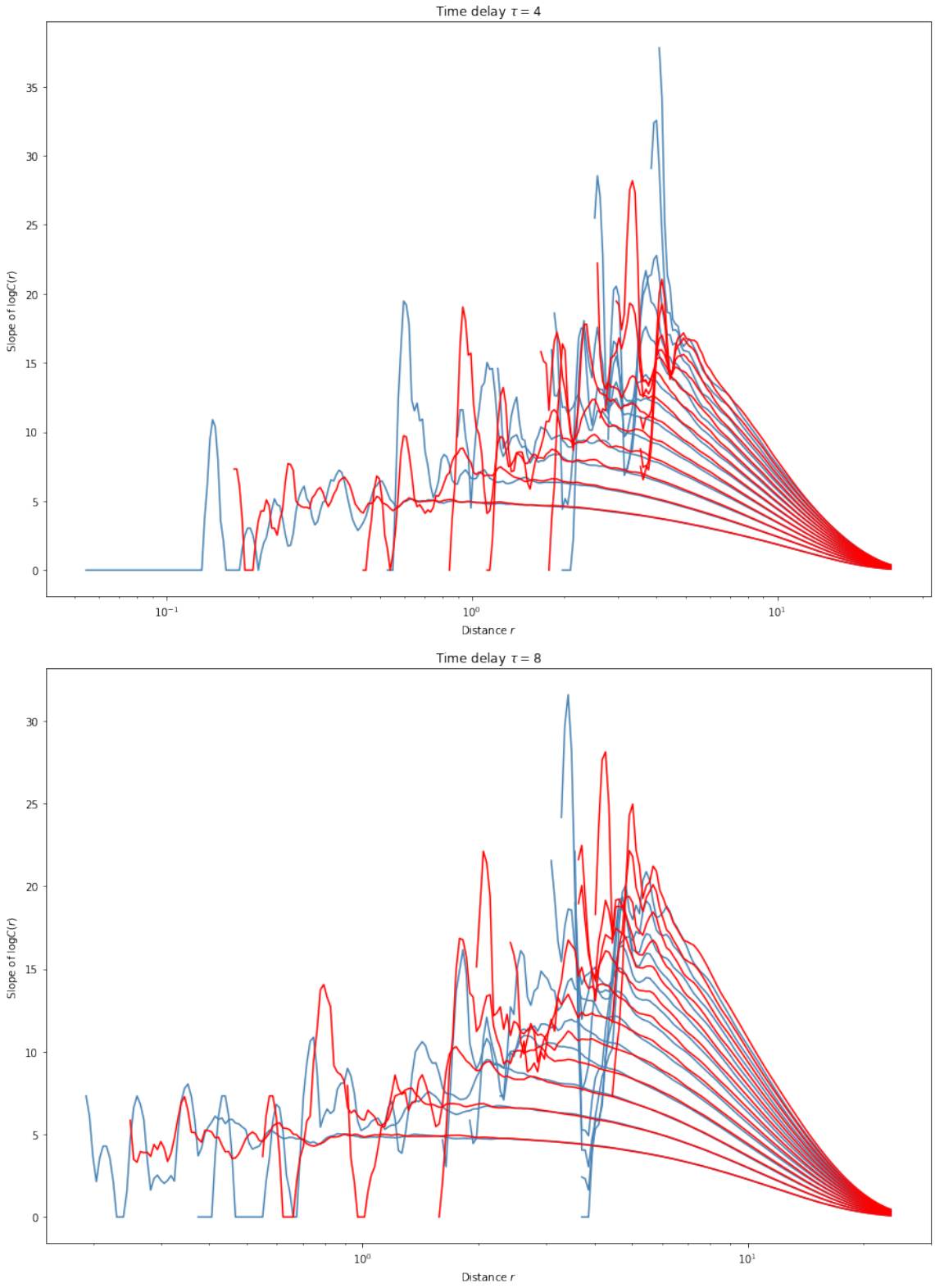


Figure 2.15: Local correlation dimension d_2 as a function of radius r for dimensions in range from 5 to 30 (from bottom to top) and time delays $\tau = 4$ and $\tau = 8$ for the original series (blue) and its surrogate series computed using iAAFT.

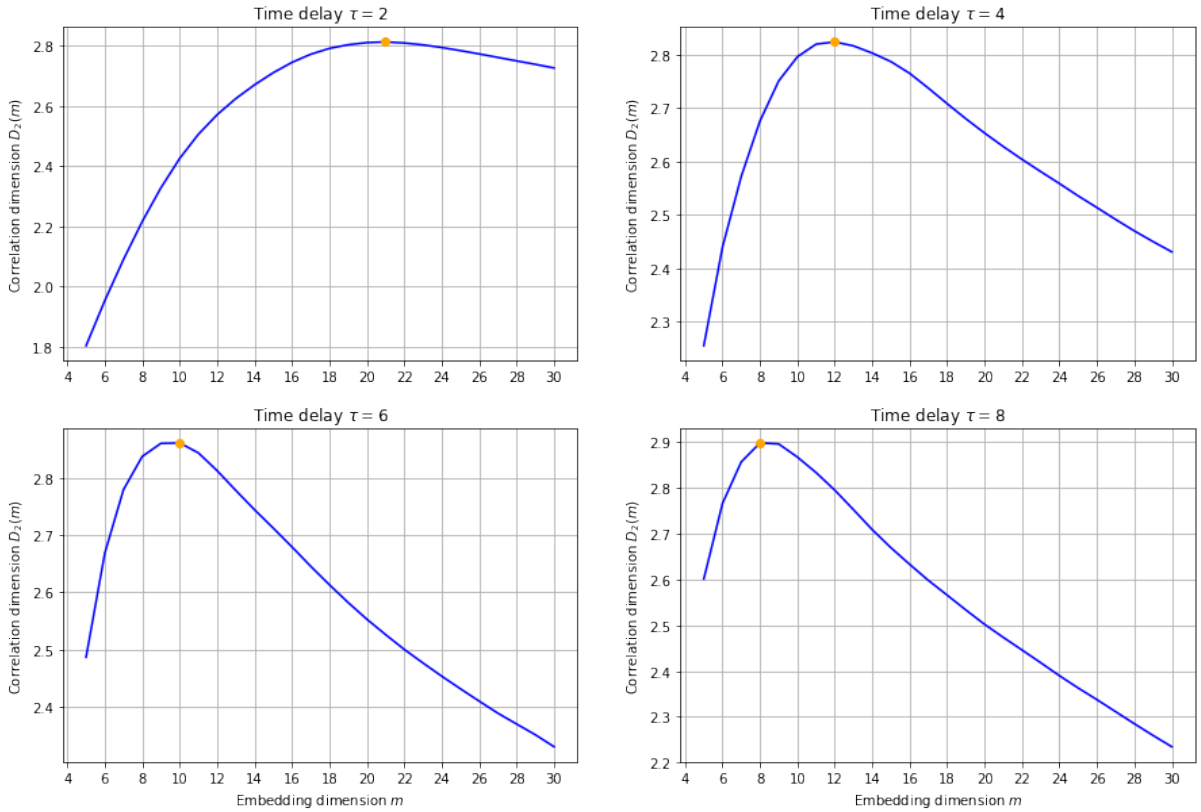


Figure 2.16: Correlation dimension as function of the embedding dimension m.

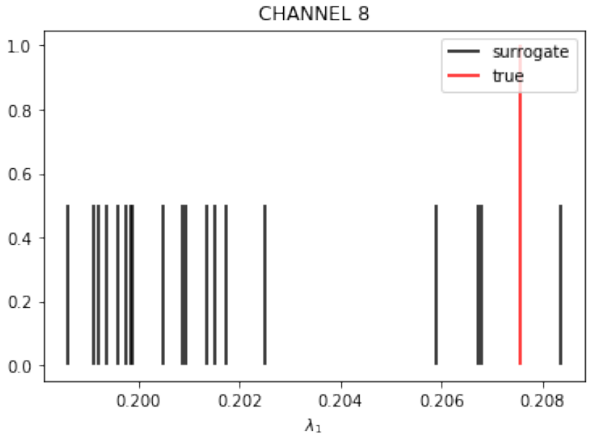


Figure 2.17: Example distribution of the largest Lyapunov exponent for surrogate data and the original.

This should be cited!

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.

We separated the patients into terciles according to the ratio of the depression scores before and after treatment. Out of these three populations, we selected the first and last, containing 46 and 44 samples respectively, to obtain population we call *responding* (responders) to treatment and *non-responding* (non-responders) to treatment. The second tercile was not considered in this analysis to minimize the effect of inaccuracy of the self-reported depression score. Comparison of mean values of individual measures between the two populations can be seen in Figures 2.19, 2.20 and 2.21. Length of an error bar corresponds to one standard deviation.

Kruskal is better, redirect to distributions.

For each group, we performed two-sided Kolmogorov-Smirnov test for the null hypothesis that the distributions of values computed for measurements before and after treatment are the same. No significant differences in distributions were found for d_2 computed using automatic selection of embedding parameters so in this section, we used λ_1 and d_2 computed for $m = 10$, $\tau = 3$, $w_t = 50$. Moreover, we found no significant differences in DFA, so we decided to leave it out of this analysis. The results can be seen in Tables 2.3, 2.4 and 2.5.

However, we found that responders had significantly lower λ_1 computed using these methods in C3 and C4 electrodes (6.899 ± 1.278 vs. 7.342 ± 1.838 for C3, 6.731 ± 1.116 vs. 7.365 ± 1.475 for C4) on recording performed before treatment.

For all measures computed this way, we found significantly differences in temporal areas, especially T3. The distributions of Largest Lyapunov exponents were also significantly different in the frontal and "central" areas, whereas d_2 differed mainly in prefrontal areas.Sample entropy mimics the pattern seen in λ_1, differing mainly in frontal and "central" areas.

We also performed unsupervised analysis of before / after groupds 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.

Which are associated with depression, but we want to leave that out in this section.

Change the tables and text to using Kruskal test instead of KS. Justify by saying the distributions are not generally normal.

Channel	Before	After	p-value	Significance
mean	10.151 ± 0.950	9.919 ± 1.074	0.121	
std	0.628 ± 0.239	0.724 ± 0.295	0.089	
FP1	9.770 ± 1.130	9.545 ± 1.287	0.432	
FP2	9.764 ± 1.186	9.565 ± 1.281	0.432	
F3	9.794 ± 1.082	9.493 ± 1.177	0.065	*
F4	9.862 ± 1.090	9.413 ± 1.330	0.010	***
C3	9.846 ± 1.068	9.579 ± 1.117	0.089	
C4	9.922 ± 1.046	9.598 ± 1.196	0.033	**
P3	10.447 ± 0.865	10.291 ± 1.055	0.212	
P4	10.437 ± 0.883	10.266 ± 1.046	0.832	
O1	10.539 ± 1.174	10.485 ± 1.271	0.965	
O2	10.518 ± 1.198	10.409 ± 1.312	0.273	
F7	10.096 ± 1.351	9.886 ± 1.402	0.432	
F8	10.118 ± 1.297	9.785 ± 1.545	0.273	
T3	9.872 ± 1.308	9.387 ± 1.544	0.000	***
T4	9.842 ± 1.317	9.449 ± 1.534	0.065	*
T5	10.506 ± 1.092	10.329 ± 1.262	0.273	
T6	10.584 ± 1.087	10.380 ± 1.189	0.347	
Fz	10.257 ± 1.004	10.117 ± 1.096	0.161	
Cz	10.204 ± 0.906	10.075 ± 0.998	0.273	
Pz	10.490 ± 0.897	10.408 ± 1.032	0.735	

Table 2.3: Mean values of λ_1 of all patients before and after treatment.

Channel	Before	After	p-value	Significance
mean	7.522 ± 0.441	7.593 ± 0.433	0.481	
std	0.383 ± 0.125	0.414 ± 0.165	0.071	*
FP1	7.812 ± 0.611	7.880 ± 0.704	0.387	
FP2	7.826 ± 0.650	7.935 ± 0.790	0.035	**
F3	7.594 ± 0.592	7.681 ± 0.586	0.179	
F4	7.639 ± 0.602	7.726 ± 0.582	0.387	
C3	7.342 ± 0.592	7.395 ± 0.591	0.585	
C4	7.334 ± 0.550	7.412 ± 0.574	0.387	
P3	7.274 ± 0.515	7.319 ± 0.522	0.305	
P4	7.325 ± 0.573	7.349 ± 0.506	0.888	
O1	7.539 ± 0.566	7.543 ± 0.524	0.987	
O2	7.516 ± 0.518	7.569 ± 0.547	0.387	
F7	7.680 ± 0.530	7.812 ± 0.550	0.305	
F8	7.702 ± 0.534	7.822 ± 0.565	0.179	
T3	7.669 ± 0.585	7.877 ± 0.624	0.011	***
T4	7.684 ± 0.588	7.840 ± 0.556	0.024	**
T5	7.556 ± 0.523	7.593 ± 0.481	0.585	
T6	7.536 ± 0.518	7.593 ± 0.483	0.585	
Fz	7.339 ± 0.535	7.350 ± 0.525	0.987	
Cz	7.359 ± 0.566	7.354 ± 0.533	0.998	
Pz	7.199 ± 0.494	7.210 ± 0.543	0.888	

Table 2.4: Mean values of d_2 of all patients before and after treatment.

Channel	Before	After	p-value	Significance
mean	0.761 ± 0.108	0.790 ± 0.130	0.240	
std	0.071 ± 0.040	0.086 ± 0.048	0.094	
FP1	0.804 ± 0.149	0.837 ± 0.176	0.403	
FP2	0.802 ± 0.156	0.830 ± 0.175	0.403	
F3	0.800 ± 0.132	0.839 ± 0.156	0.179	
F4	0.790 ± 0.137	0.842 ± 0.168	0.046	**
C3	0.793 ± 0.122	0.825 ± 0.147	0.314	
C4	0.781 ± 0.126	0.821 ± 0.151	0.046	**
P3	0.720 ± 0.087	0.740 ± 0.115	0.619	
P4	0.720 ± 0.093	0.736 ± 0.116	0.975	
O1	0.707 ± 0.113	0.718 ± 0.134	0.734	
O2	0.712 ± 0.113	0.732 ± 0.154	0.314	
F7	0.786 ± 0.163	0.811 ± 0.176	0.619	
F8	0.781 ± 0.156	0.821 ± 0.195	0.403	
T3	0.806 ± 0.160	0.867 ± 0.197	0.006	***
T4	0.812 ± 0.167	0.861 ± 0.197	0.131	
T5	0.723 ± 0.110	0.743 ± 0.133	0.403	
T6	0.714 ± 0.112	0.729 ± 0.123	0.506	
Fz	0.747 ± 0.107	0.762 ± 0.124	0.506	
Cz	0.756 ± 0.096	0.767 ± 0.110	0.840	
Pz	0.716 ± 0.093	0.728 ± 0.113	0.996	

Table 2.5: Mean values of sample entropy of all patients before and after treatment.

Channel	Before	After	p-value	Significance
mean	9.994 ± 0.890	9.655 ± 1.064	0.022	**
std	0.639 ± 0.229	0.701 ± 0.267	0.452	
FP1	9.599 ± 1.086	9.335 ± 1.360	0.625	
FP2	9.590 ± 1.090	9.281 ± 1.293	0.308	
F3	9.588 ± 1.119	9.190 ± 1.190	0.123	
F4	9.682 ± 0.999	9.199 ± 1.339	0.072	*
C3	9.690 ± 1.065	9.349 ± 1.111	0.072	*
C4	9.827 ± 1.052	9.407 ± 1.221	0.041	**
P3	10.294 ± 0.797	10.032 ± 1.050	0.072	*
P4	10.265 ± 0.873	10.004 ± 1.104	0.308	
O1	10.343 ± 1.081	10.117 ± 1.176	0.452	
O2	10.261 ± 1.160	9.961 ± 1.212	0.123	
F7	9.998 ± 1.324	9.682 ± 1.419	0.199	
F8	9.991 ± 1.170	9.659 ± 1.492	0.308	
T3	9.789 ± 1.387	9.172 ± 1.492	0.005	***
T4	9.703 ± 1.261	9.164 ± 1.403	0.022	**
T5	10.370 ± 1.091	10.073 ± 1.214	0.011	***
T6	10.335 ± 0.954	10.021 ± 1.166	0.123	
Fz	10.096 ± 0.970	9.849 ± 1.126	0.072	*
Cz	10.150 ± 0.886	9.847 ± 1.006	0.123	
Pz	10.318 ± 0.805	10.113 ± 1.062	0.801	

Channel	Before	After	p-value	Significance
mean	10.400 ± 0.969	10.015 ± 1.088	0.423	
std	0.623 ± 0.260	0.813 ± 0.370	0.018	***
FP1	10.034 ± 1.166	9.510 ± 1.325	0.108	
FP2	10.045 ± 1.196	9.752 ± 1.269	0.778	
F3	10.116 ± 1.034	9.619 ± 1.200	0.108	
F4	10.098 ± 1.146	9.343 ± 1.364	0.034	**
C3	10.160 ± 1.010	9.585 ± 1.147	0.018	***
C4	10.162 ± 1.060	9.681 ± 1.170	0.062	*
P3	10.711 ± 0.874	10.468 ± 1.065	0.423	
P4	10.720 ± 0.897	10.453 ± 1.005	0.595	
O1	10.765 ± 1.292	10.772 ± 1.352	0.924	
O2	10.823 ± 1.242	10.604 ± 1.434	0.595	
F7	10.234 ± 1.340	9.875 ± 1.459	0.282	
F8	10.307 ± 1.405	9.556 ± 1.715	0.108	
T3	10.073 ± 1.207	9.292 ± 1.602	0.004	***
T4	10.018 ± 1.431	9.394 ± 1.726	0.179	
T5	10.709 ± 1.140	10.490 ± 1.301	0.778	
T6	10.933 ± 1.072	10.649 ± 1.219	0.778	
Fz	10.568 ± 0.951	10.327 ± 1.024	0.423	
Cz	10.383 ± 0.867	10.291 ± 0.939	0.924	
Pz	10.744 ± 0.894	10.630 ± 1.023	0.423	

Table 2.6: Mean values of λ_1 of responding / non-responding patients before and after treatment.

Channel	Before	After	p-value	Significance
mean	7.536 ± 0.394	7.585 ± 0.465	0.765	
std	0.401 ± 0.121	0.400 ± 0.134	0.917	
FP1	7.851 ± 0.588	7.841 ± 0.751	0.580	
FP2	7.921 ± 0.647	7.903 ± 0.553	0.580	
F3	7.614 ± 0.579	7.714 ± 0.634	0.765	
F4	7.640 ± 0.575	7.696 ± 0.591	0.408	
C3	7.399 ± 0.575	7.416 ± 0.659	0.989	
C4	7.303 ± 0.481	7.378 ± 0.615	0.765	
P3	7.247 ± 0.488	7.288 ± 0.552	0.580	
P4	7.338 ± 0.510	7.337 ± 0.543	0.765	
O1	7.554 ± 0.479	7.593 ± 0.571	0.917	
O2	7.539 ± 0.464	7.599 ± 0.560	0.765	
F7	7.662 ± 0.585	7.797 ± 0.601	0.269	
F8	7.717 ± 0.469	7.762 ± 0.574	0.408	
T3	7.694 ± 0.524	7.902 ± 0.636	0.269	
T4	7.682 ± 0.563	7.826 ± 0.522	0.100	
T5	7.606 ± 0.532	7.589 ± 0.477	0.765	
T6	7.578 ± 0.460	7.625 ± 0.485	0.765	
Fz	7.335 ± 0.522	7.340 ± 0.571	0.989	
Cz	7.321 ± 0.574	7.354 ± 0.532	0.917	
Pz	7.188 ± 0.451	7.162 ± 0.538	0.765	

Channel	Before	After	p-value	Significance
mean	7.483 ± 0.432	7.648 ± 0.369	0.548	
std	0.366 ± 0.116	0.450 ± 0.162	0.048	**
FP1	7.709 ± 0.552	7.984 ± 0.658	0.149	
FP2	7.749 ± 0.582	7.909 ± 0.544	0.244	
F3	7.517 ± 0.576	7.717 ± 0.548	0.244	
F4	7.585 ± 0.621	7.886 ± 0.603	0.086	
C3	7.271 ± 0.579	7.427 ± 0.583	0.377	
C4	7.335 ± 0.561	7.435 ± 0.527	0.738	
P3	7.272 ± 0.474	7.392 ± 0.488	0.548	
P4	7.268 ± 0.457	7.417 ± 0.518	0.548	
O1	7.510 ± 0.693	7.530 ± 0.484	0.548	
O2	7.494 ± 0.533	7.623 ± 0.563	0.377	
F7	7.643 ± 0.419	7.874 ± 0.486	0.048	**
F8	7.651 ± 0.557	7.914 ± 0.554	0.012	***
T3	7.616 ± 0.582	8.032 ± 0.646	0.012	***
T4	7.687 ± 0.604	7.984 ± 0.648	0.086	
T5	7.517 ± 0.518	7.625 ± 0.441	0.548	
T6	7.488 ± 0.494	7.613 ± 0.493	0.377	
Fz	7.287 ± 0.509	7.359 ± 0.459	0.548	
Cz	7.380 ± 0.536	7.326 ± 0.490	0.902	
Pz	7.195 ± 0.498	7.262 ± 0.499	0.548	

Table 2.7: Mean values of d_2 of responding / non-respoding patients before and after treatment.

Channel	Before	After	p-value	Significance
mean	0.768 ± 0.093	0.811 ± 0.107	0.086	
std	0.078 ± 0.039	0.093 ± 0.047	0.377	
FP1	0.816 ± 0.156	0.866 ± 0.178	0.548	
FP2	0.819 ± 0.158	0.866 ± 0.171	0.377	
F3	0.815 ± 0.128	0.872 ± 0.145	0.086	
F4	0.801 ± 0.126	0.868 ± 0.160	0.048	**
C3	0.796 ± 0.107	0.841 ± 0.123	0.149	
C4	0.780 ± 0.112	0.844 ± 0.129	0.012	***
P3	0.718 ± 0.075	0.744 ± 0.089	0.548	
P4	0.722 ± 0.089	0.751 ± 0.103	0.548	
O1	0.704 ± 0.074	0.738 ± 0.111	0.244	
O2	0.725 ± 0.094	0.746 ± 0.116	0.149	
F7	0.790 ± 0.166	0.837 ± 0.168	0.149	
F8	0.786 ± 0.138	0.845 ± 0.190	0.548	
T3	0.813 ± 0.163	0.894 ± 0.181	0.012	***
T4	0.828 ± 0.165	0.886 ± 0.162	0.048	**
T5	0.722 ± 0.078	0.762 ± 0.110	0.086	
T6	0.723 ± 0.094	0.752 ± 0.090	0.086	
Fz	0.758 ± 0.094	0.782 ± 0.109	0.244	
Cz	0.760 ± 0.082	0.783 ± 0.089	0.377	
Pz	0.724 ± 0.090	0.742 ± 0.103	0.902	

Channel	Before	After	p-value	Significance
mean	0.757 ± 0.113	0.798 ± 0.137	0.676	
std	0.067 ± 0.044	0.095 ± 0.057	0.061	*
FP1	0.796 ± 0.137	0.850 ± 0.178	0.479	
FP2	0.799 ± 0.146	0.823 ± 0.162	0.975	
F3	0.783 ± 0.130	0.843 ± 0.163	0.479	
F4	0.782 ± 0.137	0.863 ± 0.169	0.111	
C3	0.780 ± 0.128	0.826 ± 0.157	0.193	
C4	0.774 ± 0.131	0.820 ± 0.153	0.193	
P3	0.713 ± 0.092	0.747 ± 0.134	0.676	
P4	0.713 ± 0.092	0.735 ± 0.130	0.975	
O1	0.715 ± 0.130	0.723 ± 0.159	0.975	
O2	0.716 ± 0.130	0.750 ± 0.201	0.975	
F7	0.794 ± 0.167	0.824 ± 0.186	0.314	
F8	0.784 ± 0.172	0.850 ± 0.207	0.314	
T3	0.802 ± 0.159	0.901 ± 0.211	0.031	**
T4	0.807 ± 0.176	0.887 ± 0.222	0.193	
T5	0.722 ± 0.123	0.747 ± 0.152	0.863	
T6	0.709 ± 0.127	0.721 ± 0.140	0.975	
Fz	0.731 ± 0.104	0.757 ± 0.124	0.314	
Cz	0.751 ± 0.093	0.757 ± 0.105	0.975	
Pz	0.713 ± 0.092	0.730 ± 0.121	0.863	

Table 2.8: Mean values of sample entropy of responding / non-responding patients before and after treatment.

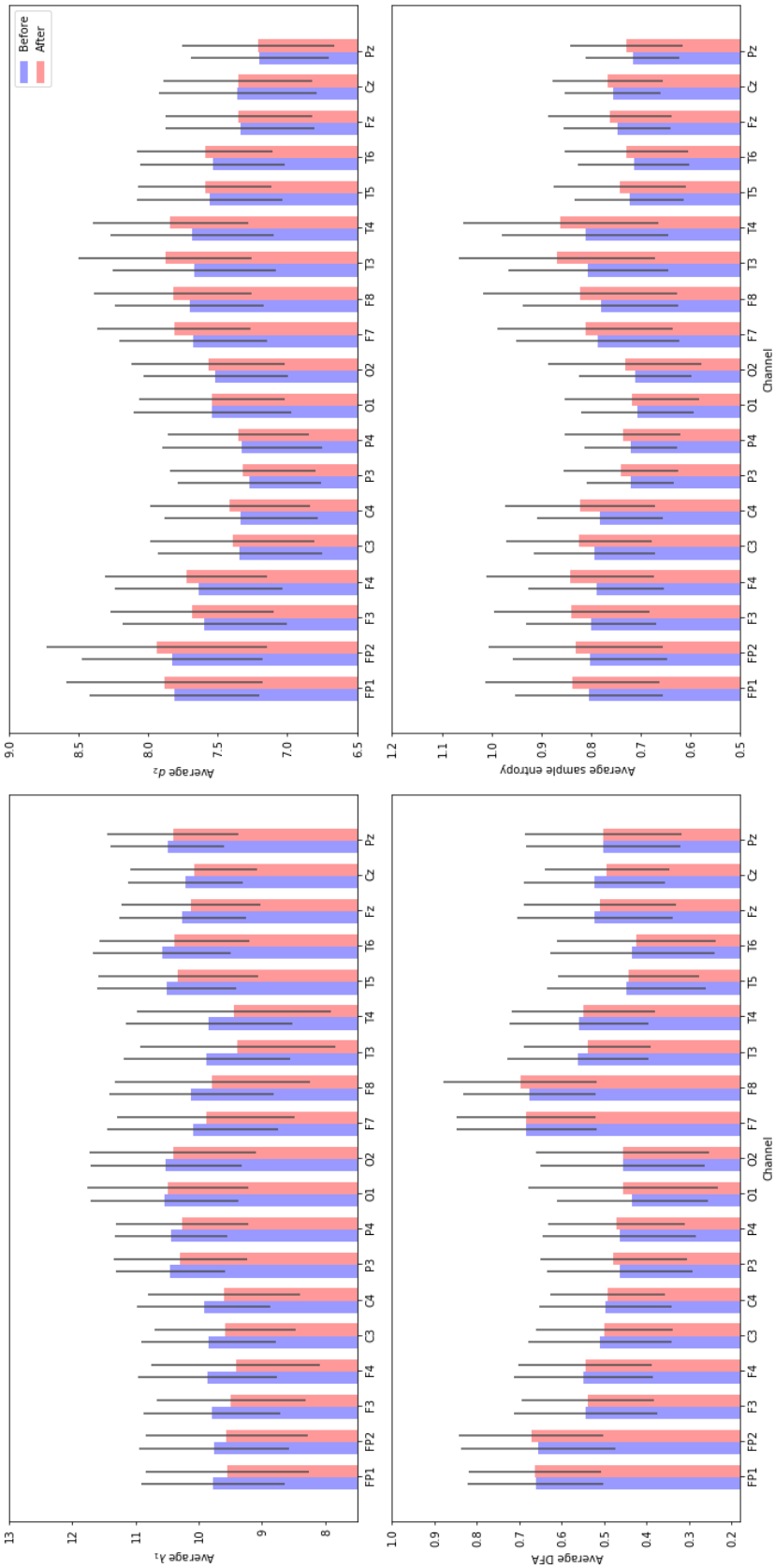


Figure 2.18: Values of individual measures computed before and after treatment.

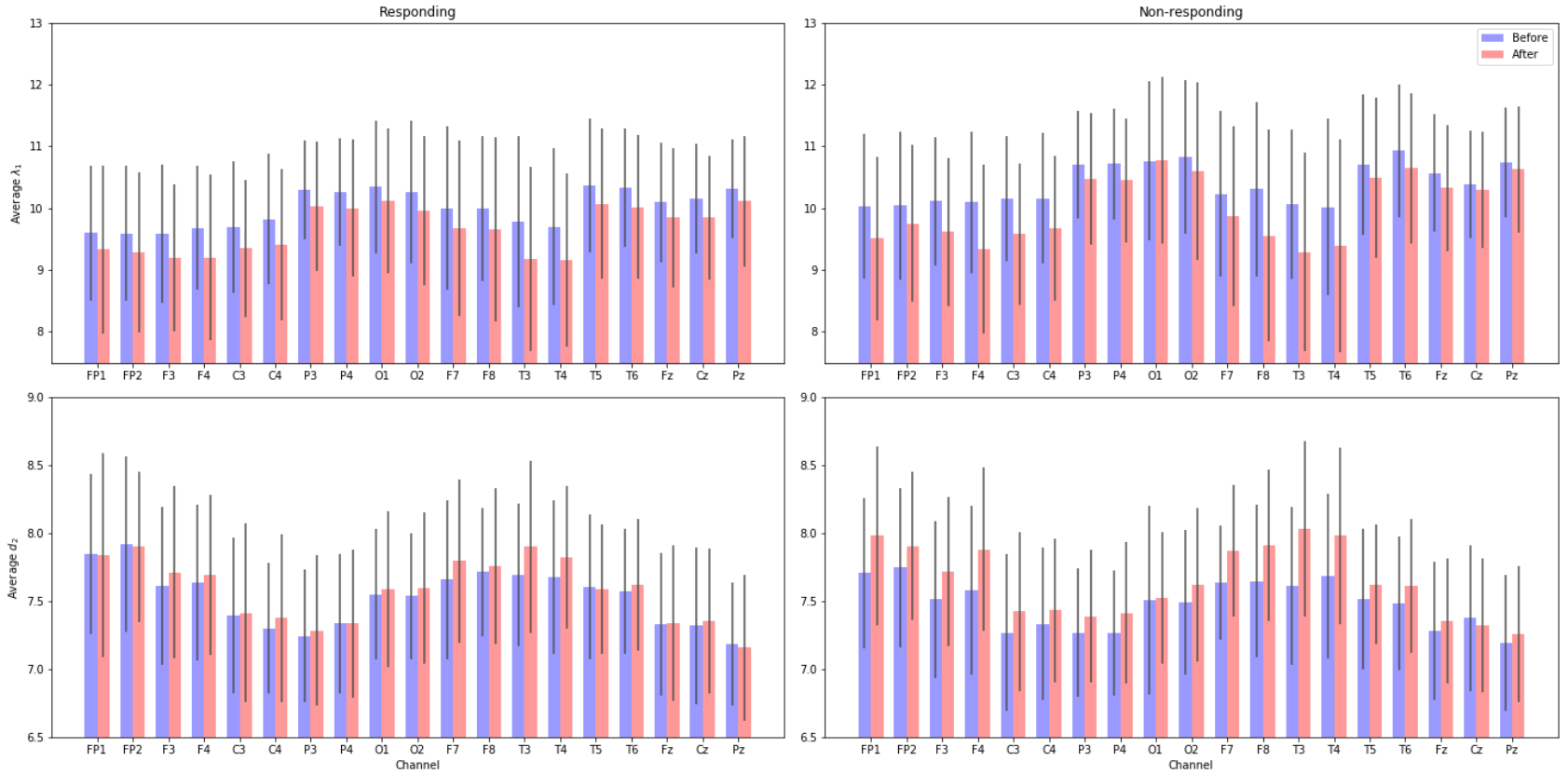


Figure 2.19: Comparison of mean values of largest Lyapunov exponent and correlation dimension between responders and non-responders computed using embedding dimension $m = 10$ and time delay $\tau = 3$.

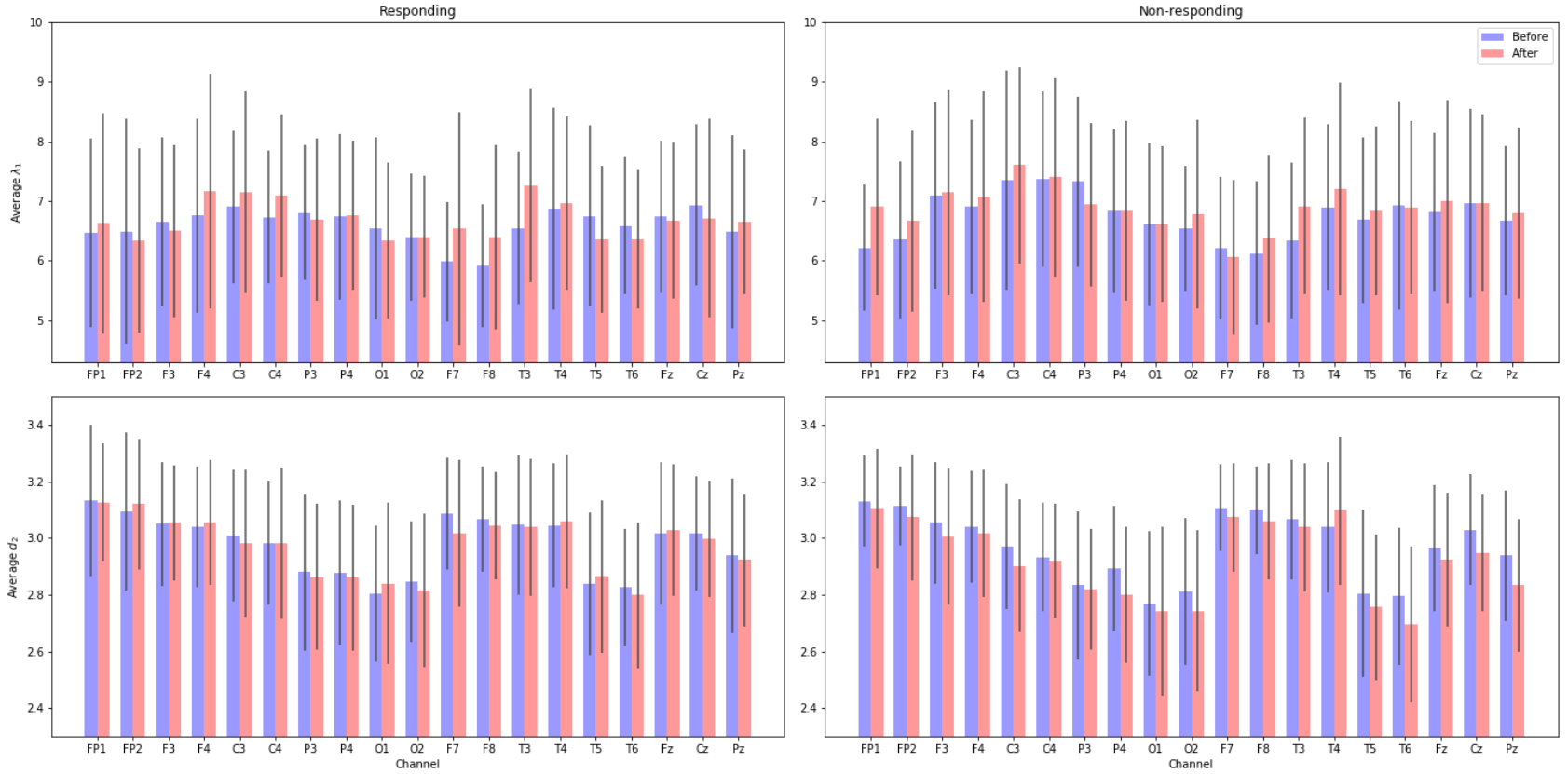


Figure 2.20: Comparison of mean values of largest Lyapunov exponent and correlation dimension between responders and non-responders computed using automatic procedure described in Section .

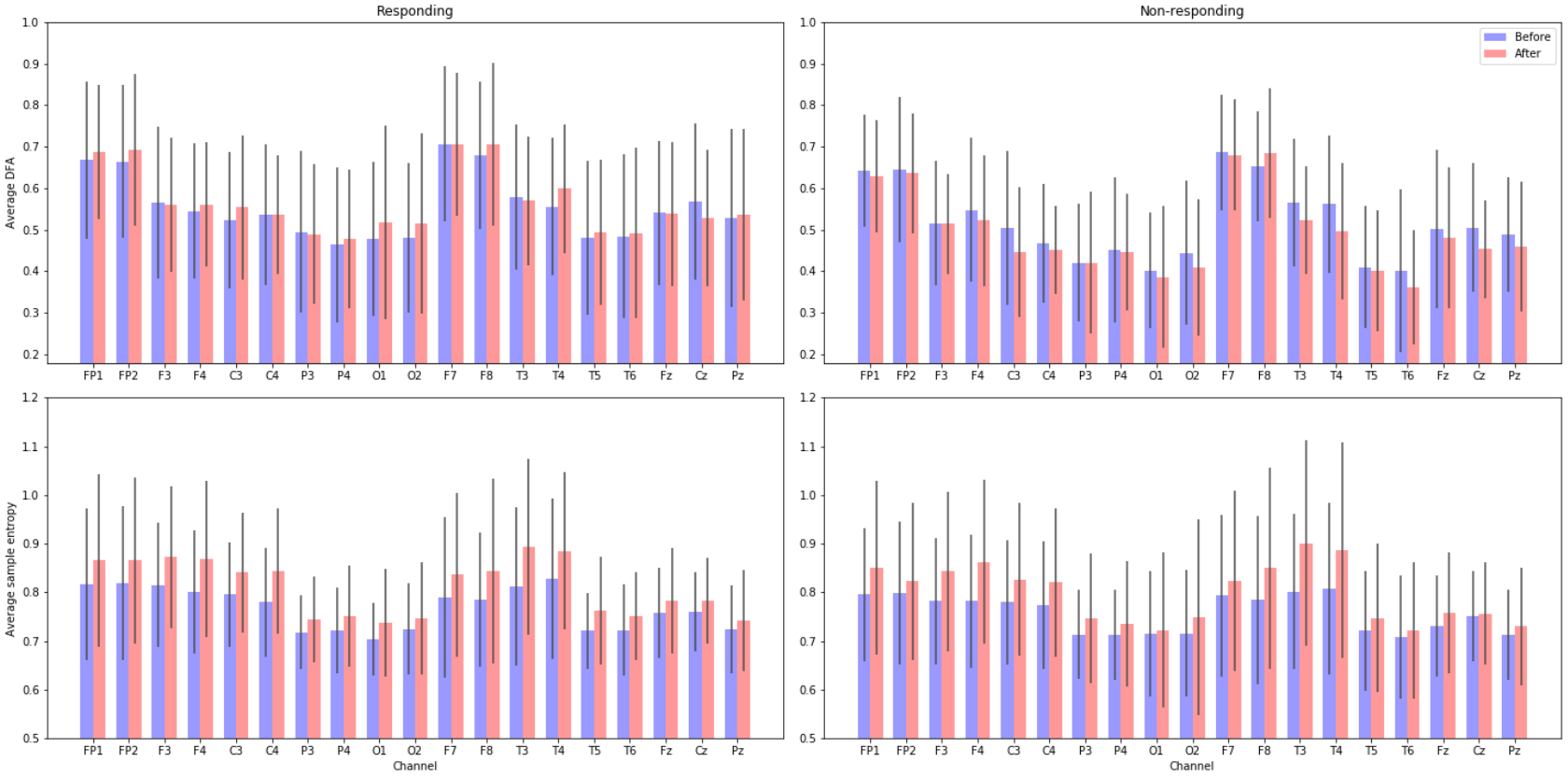


Figure 2.21: Comparison of mean values of computed detrended fluctuation analysis and sample entropy between responders and non-responders.

2.6.2 Low and high depression score

Here, we have to explain limitations of training classifiers on this test. There is no healthy group. The effect of depression score is meddled with the effect of drugs in after treatment measurements. But still, there are correlations. The goal of this is to inform and maybe interpret the results of classification, and which measures show correlations and how they are distributed between groups, which is inherently important and relevant for eventual interpretation.

2.6.3 Low and high remission

There was no control group given no drugs. Each patient assigned different treatment. Each patient started on different level of depression. We recognize this is inherently bad study design.

2.7 Classification

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. Feature selection was performed on LR with regularization strength 1 and SVM with regularization strenth 1 and linear kernel (i.e. $k(\boldsymbol{x_1}, \boldsymbol{x_2}) = \boldsymbol{x_1} \cdot \boldsymbol{x_2}$) using

- 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 3-fold cross validation (scoring models based on ROC AUC, population size 50, 50 generations, crossover probability 0.8, mutation probability 0.2, and tournament size 5).

Am I justified in using χ^2?

Then, a brute force grid search with 3-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?

These classifiers were evaluated on the test set. The best performing classifiers (based on accuracy, precision, recall, f-score, ROC AUC and confusion matrices) for each measure were selected, and then manually optimized by adding, removing, or replacing features based on differences in distributions described in Section 2.6, performance of similar classifiers on similar measures, and results of other studies. In some cases, hyperparameters were modified. The results are described in the following text.

2.7.1 Depression

2.7.2 Remission

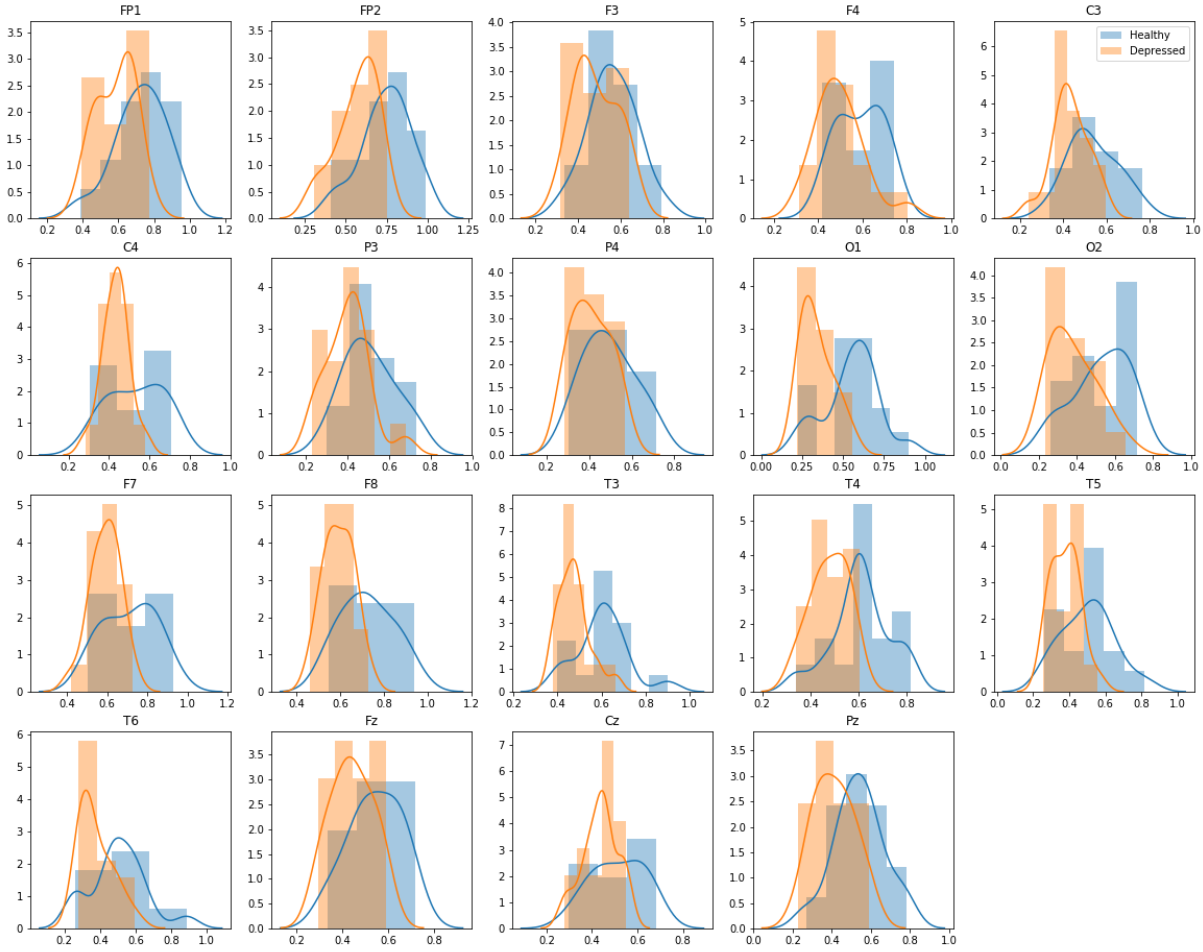


Figure 2.22: Distributions of DFA between healthy and depressed patients. Healthy patients seem to have, on average, lower values of DFA, and the distributions are not generally normal.

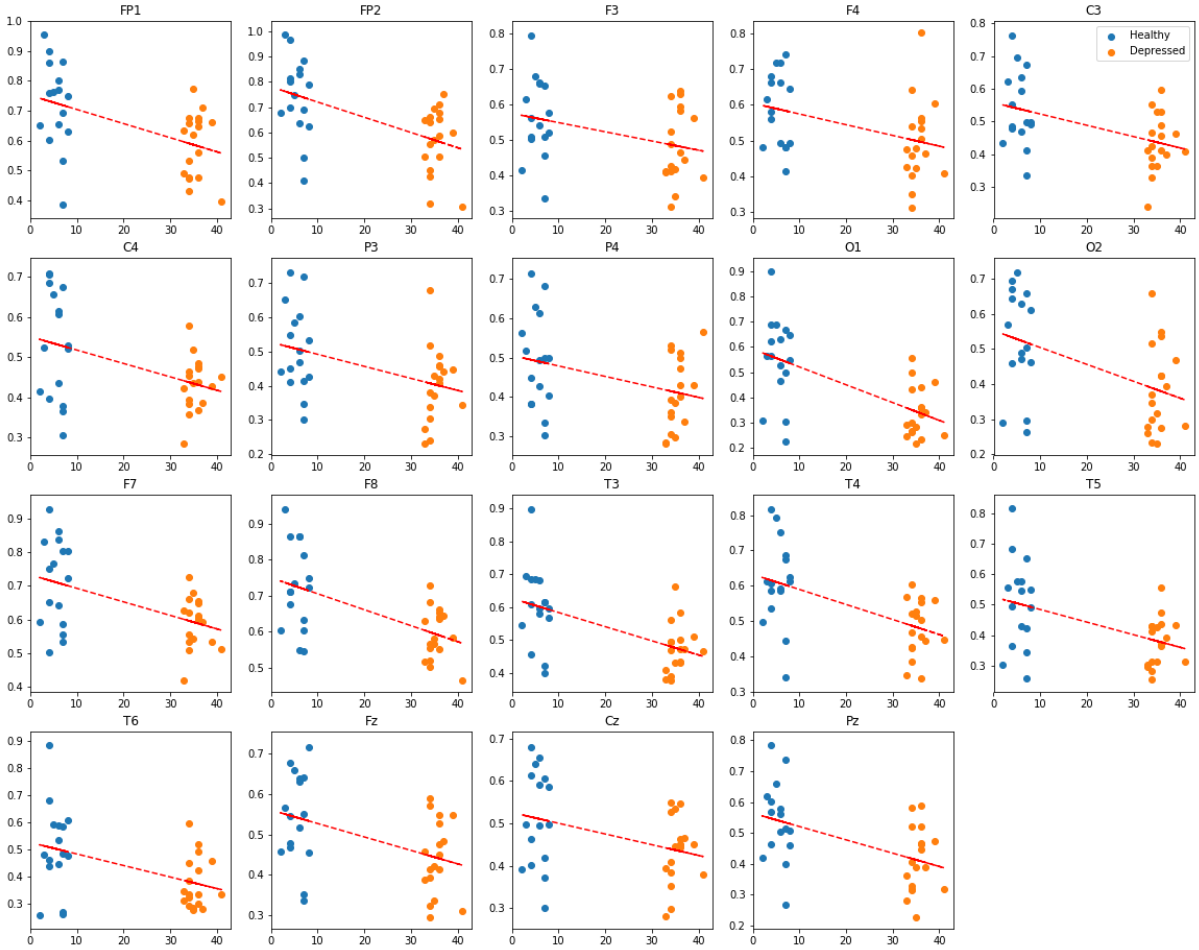


Figure 2.23: Trend of values of DFA as a function of depression score. For all channels, correlation is significantly ($p < 0.05$) negative.

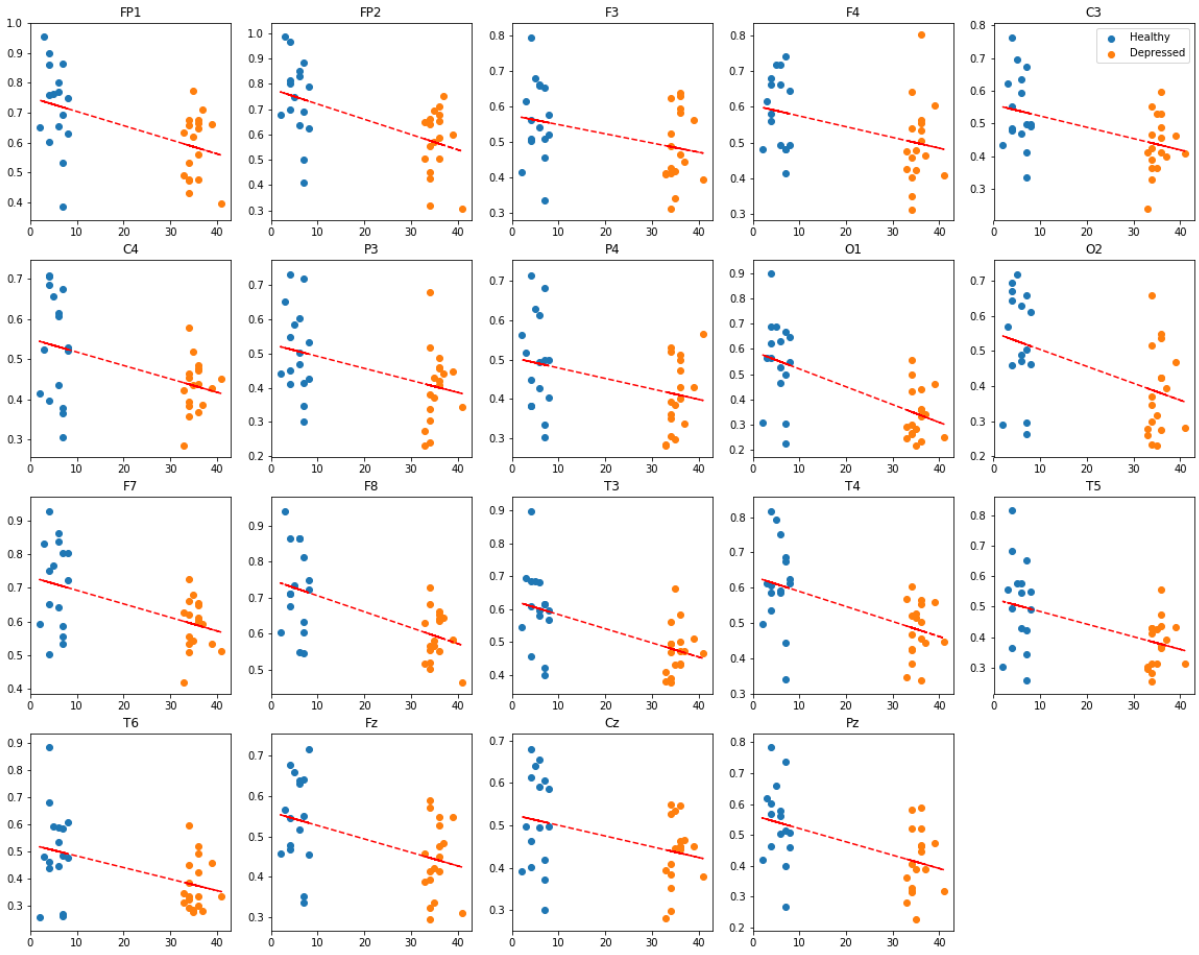


Figure 2.24: Trend of values of largest Lyapunov exponent as a function of depression score. The correlation is significantly ($p < 0.05$) positive for all channels with exception of FP1, FP2, F3, C3, C4, F7, F8, T3.

Channel	Depressed	Healthy	p-value	Sig.
mean	0.503 ± 0.104	0.561 ± 0.125	0.027	*
std	0.109 ± 0.032	0.105 ± 0.025	0.681	
FP1	0.637 ± 0.144	0.694 ± 0.154	0.168	
FP2	0.641 ± 0.166	0.710 ± 0.169	0.131	
F3	0.520 ± 0.131	0.553 ± 0.134	0.153	
F4	0.527 ± 0.127	0.566 ± 0.125	0.104	
C3	0.500 ± 0.128	0.528 ± 0.126	0.183	
C4	0.474 ± 0.105	0.535 ± 0.132	0.049	*
P3	0.442 ± 0.148	0.487 ± 0.149	0.100	
P4	0.435 ± 0.127	0.479 ± 0.141	0.135	
O1	0.401 ± 0.144	0.507 ± 0.178	0.011	*
O2	0.437 ± 0.155	0.503 ± 0.155	0.076	
F7	0.627 ± 0.131	0.695 ± 0.159	0.035	*
F8	0.624 ± 0.129	0.694 ± 0.155	0.014	*
T3	0.527 ± 0.119	0.580 ± 0.146	0.057	
T4	0.516 ± 0.119	0.595 ± 0.130	0.007	**
T5	0.436 ± 0.151	0.493 ± 0.141	0.088	
T6	0.421 ± 0.140	0.482 ± 0.138	0.066	
Fz	0.471 ± 0.111	0.528 ± 0.130	0.053	
Cz	0.469 ± 0.094	0.518 ± 0.119	0.040	*
Pz	0.453 ± 0.140	0.521 ± 0.150	0.051	

(a) DFA

Channel	Depressed	Healthy	p-value	Sig.
mean	0.557 ± 0.079	0.592 ± 0.095	0.051	
std	0.080 ± 0.027	0.068 ± 0.022	0.038	*
FP1	0.638 ± 0.102	0.669 ± 0.096	0.139	
FP2	0.649 ± 0.117	0.668 ± 0.113	0.554	
F3	0.565 ± 0.086	0.585 ± 0.103	0.250	
F4	0.576 ± 0.091	0.603 ± 0.098	0.189	
C3	0.562 ± 0.090	0.576 ± 0.107	0.293	
C4	0.548 ± 0.074	0.574 ± 0.102	0.163	
P3	0.513 ± 0.121	0.538 ± 0.121	0.278	
P4	0.513 ± 0.110	0.542 ± 0.113	0.212	
O1	0.475 ± 0.119	0.554 ± 0.138	0.018	*
O2	0.499 ± 0.121	0.552 ± 0.123	0.079	
F7	0.650 ± 0.091	0.674 ± 0.103	0.168	
F8	0.643 ± 0.092	0.674 ± 0.118	0.061	
T3	0.587 ± 0.082	0.604 ± 0.103	0.332	
T4	0.571 ± 0.084	0.621 ± 0.091	0.012	*
T5	0.498 ± 0.112	0.555 ± 0.107	0.043	*
T6	0.489 ± 0.113	0.540 ± 0.102	0.038	*
Fz	0.537 ± 0.085	0.576 ± 0.100	0.043	*
Cz	0.548 ± 0.080	0.579 ± 0.094	0.063	
Pz	0.525 ± 0.117	0.567 ± 0.120	0.068	

(b) Hurst exponent

Channel	Depressed	Healthy	p-value	Sig.
mean	10.144 ± 1.166	9.830 ± 0.884	0.194	
std	0.761 ± 0.304	0.662 ± 0.258	0.173	
FP1	9.650 ± 1.363	9.574 ± 1.122	0.831	
FP2	9.874 ± 1.387	9.543 ± 1.126	0.324	
F3	9.627 ± 1.260	9.387 ± 1.094	0.421	
F4	9.876 ± 1.421	9.383 ± 1.137	0.119	
C3	9.500 ± 1.234	9.483 ± 0.948	0.895	
C4	9.810 ± 1.302	9.550 ± 0.955	0.212	
P3	10.508 ± 1.060	10.223 ± 0.763	0.301	
P4	10.588 ± 1.051	10.178 ± 0.914	0.131	
O1	10.744 ± 1.380	10.317 ± 0.893	0.063	
O2	10.724 ± 1.394	10.070 ± 0.985	0.030	*
F7	9.941 ± 1.666	9.757 ± 1.452	0.657	
F8	9.953 ± 1.655	9.794 ± 1.403	0.543	
T3	9.442 ± 1.541	9.367 ± 1.382	0.742	
T4	9.710 ± 1.663	9.422 ± 1.357	0.375	
T5	10.667 ± 1.211	10.240 ± 1.027	0.107	
T6	10.877 ± 1.226	10.165 ± 0.948	0.017	*
Fz	10.296 ± 1.171	10.098 ± 0.851	0.349	
Cz	10.284 ± 1.038	9.972 ± 0.754	0.243	
Pz	10.663 ± 1.140	10.254 ± 0.902	0.200	

(c) Largest Lyapunov exponent

Channel	Depressed	Healthy	p-value	Sig.
mean	0.768 ± 0.141	0.796 ± 0.103	0.119	
std	0.087 ± 0.045	0.085 ± 0.045	0.870	
FP1	0.819 ± 0.172	0.834 ± 0.150	0.611	
FP2	0.788 ± 0.169	0.839 ± 0.138	0.135	
F3	0.819 ± 0.170	0.848 ± 0.141	0.264	
F4	0.801 ± 0.182	0.851 ± 0.145	0.115	
C3	0.835 ± 0.158	0.830 ± 0.118	0.793	
C4	0.807 ± 0.172	0.823 ± 0.106	0.119	
P3	0.715 ± 0.111	0.733 ± 0.089	0.257	
P4	0.702 ± 0.115	0.736 ± 0.108	0.148	
O1	0.701 ± 0.146	0.729 ± 0.111	0.066	
O2	0.705 ± 0.146	0.753 ± 0.127	0.016	*
F7	0.803 ± 0.202	0.815 ± 0.178	0.681	
F8	0.802 ± 0.216	0.815 ± 0.174	0.421	
T3	0.862 ± 0.208	0.869 ± 0.177	0.634	
T4	0.835 ± 0.218	0.860 ± 0.169	0.230	
T5	0.714 ± 0.134	0.748 ± 0.097	0.030	*
T6	0.694 ± 0.115	0.747 ± 0.092	0.008	**
Fz	0.744 ± 0.132	0.768 ± 0.098	0.218	
Cz	0.749 ± 0.113	0.782 ± 0.082	0.079	
Pz	0.696 ± 0.117	0.736 ± 0.117	0.131	

(d) Sample entropy

Channel	Depressed	Healthy	p-value	Sig.
mean	1.357 ± 0.139	1.402 ± 0.113	0.082	
std	0.095 ± 0.043	0.091 ± 0.042	0.793	
FP1	1.423 ± 0.186	1.456 ± 0.177	0.450	
FP2	1.396 ± 0.188	1.459 ± 0.163	0.123	
F3	1.413 ± 0.177	1.459 ± 0.157	0.173	
F4	1.396 ± 0.181	1.462 ± 0.153	0.094	
C3	1.435 ± 0.149	1.448 ± 0.125	0.693	
C4	1.401 ± 0.147	1.436 ± 0.111	0.097	
P3	1.299 ± 0.115	1.336 ± 0.104	0.194	
P4	1.291 ± 0.108	1.336 ± 0.120	0.111	
O1	1.265 ± 0.141	1.322 ± 0.123	0.035	*
O2	1.273 ± 0.146	1.338 ± 0.129	0.031	*
F7	1.406 ± 0.190	1.437 ± 0.173	0.301	
F8	1.402 ± 0.198	1.447 ± 0.173	0.153	
T3	1.448 ± 0.188	1.468 ± 0.174	0.532	
T4	1.427 ± 0.214	1.452 ± 0.147	0.243	
T5	1.281 ± 0.126	1.334 ± 0.106	0.035	*
T6	1.258 ± 0.113	1.330 ± 0.093	0.004	***
Fz	1.335 ± 0.152	1.375 ± 0.124	0.148	
Cz	1.352 ± 0.123	1.411 ± 0.098	0.040	*
Pz	1.287 ± 0.126	1.334 ± 0.123	0.131	

(e) Higuchi fractal dimension

Channel	Depressed	Healthy	p-value	Sig.
mean	10.892 ± 0.681	10.734 ± 0.723	0.332	
std	0.667 ± 0.166	0.629 ± 0.161	0.393	
FP1	11.439 ± 0.814	11.067 ± 1.058	0.076	
FP2	11.405 ± 0.921	11.141 ± 0.998	0.522	
F3	11.100 ± 0.881	10.673 ± 0.809	0.076	
F4	10.888 ± 0.884	10.844 ± 0.947	0.870	
C3	10.481 ± 0.872	10.289 ± 0.895	0.375	
C4	10.495 ± 0.880	10.402 ± 0.978	0.706	
P3	10.665 ± 0.854	10.419 ± 0.861	0.301	
P4	10.465 ± 0.766	10.313 ± 0.774	0.460	
O1	11.002 ± 1.225	11.023 ± 0.944	0.646	
O2	10.990 ± 1.022	10.894 ± 1.020	0.657	
F7	10.979 ± 0.799	10.842 ± 0.837	0.402	
F8	11.116 ± 1.020	10.858 ± 0.835	0.358	
T3	11.133 ± 0.763	11.085 ± 0.946	0.974	
T4	11.256 ± 0.924	11.090 ± 0.965	0.565	
T5	10.988 ± 0.927	10.970 ± 0.995	0.961	
T6	11.038 ± 0.729	10.911 ± 0.743	0.588	
Fz	10.614 ± 0.792	10.322 ± 0.877	0.206	
Cz	10.444 ± 1.047	10.545 ± 0.962	0.767	
Pz	10.454 ± 0.959	10.254 ± 0.886	0.384	

(f) Correlation dimension

Table 2.9: Comparison of mean values of measures computed for depressed and healthy patients.

Measure	Class.	Acc.	F-sc.	ROC AUC	CM	Channels
DFA	LR	0.88	0.88	0.91	$\begin{pmatrix} 9 & 2 \\ 0 & 6 \end{pmatrix}$	C3, C4, O1, T4
HE	LR	0.76	0.77	0.78	$\begin{pmatrix} 8 & 3 \\ 1 & 5 \end{pmatrix}$	O1, T4
CD	SVM (poly.)	0.76	0.76	0.74	$\begin{pmatrix} 9 & 2 \\ 2 & 4 \end{pmatrix}$	F3, F4, P3, O1, T4, T5, Cz, Pz
LLE	SVM (lin.)	0.71	0.71	0.77	$\begin{pmatrix} 6 & 5 \\ 0 & 6 \end{pmatrix}$	FP1, FP2, F4, C3, O1, T3, T6
SE	LR	0.71	0.71	0.73	$\begin{pmatrix} 7 & 4 \\ 1 & 5 \end{pmatrix}$	P3, P4, T5, T6
DFA	SVM (rbf)	0.70	0.70	0.72	$\begin{pmatrix} 8 & 7 \\ 1 & 11 \end{pmatrix}$	O1, T5, T6
CD	SVM (rbf)	0.70	0.70	0.71	$\begin{pmatrix} 10 & 5 \\ 3 & 9 \end{pmatrix}$	C4, P4, T4, T6, Cz
HE	LR	0.70	0.70	0.70	$\begin{pmatrix} 11 & 4 \\ 4 & 8 \end{pmatrix}$	F3, F4, T3, T5, Fz, Cz

Table 2.10: Evaluation of depression classification. The first part corresponds to classifiers trained on 15 healthy (DS $\leq$ 10) and 24 depressed (DS $\geq$ 30) patients. The second part corresponds to classifiers trained on 24 healthy (DS $\leq$ 12) and 36 depressed (DS $\geq$ 32). Both before and after treatment sessions are included.

Measure	Class.	Acc.	F-sc.	ROC AUC	CM	Channels
CD	SVM (rbf)	1.00	1.00	1.00	$\begin{pmatrix} 10 & 0 \\ 0 & 10 \end{pmatrix}$	F4, C4, O1, F7, F8, T5, T6
LLE	LR	0.80	0.80	0.80	$\begin{pmatrix} 9 & 1 \\ 3 & 7 \end{pmatrix}$	FP2, F3, F4, P3, P4, F7, F8, T6, Fz, Cz
DFA	LR	0.80	0.80	0.80	$\begin{pmatrix} 8 & 2 \\ 2 & 8 \end{pmatrix}$	FP2, F3, O1, T5, T6, Cz
HE	LR	0.70	0.70	0.70	$\begin{pmatrix} 7 & 3 \\ 3 & 7 \end{pmatrix}$	FP2, F3, O1, T5, T6, Cz
LLE	SVM (rbf)	0.76	0.76	0.77	$\begin{pmatrix} 16 & 7 \\ 3 & 16 \end{pmatrix}$	F3, F4, O1, O2, T6, Fz
HE	LR	0.71	0.71	0.73	$\begin{pmatrix} 14 & 9 \\ 3 & 16 \end{pmatrix}$	P3, F7, T4, T6, Cz
CD	LR	0.69	0.69	0.70	$\begin{pmatrix} 14 & 9 \\ 4 & 15 \end{pmatrix}$	F4, C4, T5
HD	LR	0.69	0.69	0.70	$\begin{pmatrix} 13 & 10 \\ 3 & 16 \end{pmatrix}$	FP2
DFA	LR	0.64	0.64	0.64	$\begin{pmatrix} 15 & 8 \\ 7 & 12 \end{pmatrix}$	C3, P4, F7, T6

Table 2.11: Evaluation of remission. The first part corresponds to split of the whole dataset according to the first and last septiles of m_2/m_1 (considered non-remitting and remitting patients respectively). The second part corresponds to split according to first and last terciles of m_2/m_1. Both before and after treatment sessions are included.

Chapter 3

Machine learning approach

Conclusion

Bibliography

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

[2] 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.

[3] 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.

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

[5] 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.

[6] 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.

[7] Peter J Bickel and Peter Bühlmann. What is a linear process? *Proceedings of the National Academy of Sciences*, 93(22):12128–12131, 1996.

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

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

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

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

[16] 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.

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

[18] 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.

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

[20] Peter Grassberger. Do climatic attractors exist? *Nature*, 323(6089):609, 1986.

[21] Peter Grassberger. Grassberger-Procaccia algorithm. `http://www.scholarpedia.org/article/Grassberger-Procaccia_algorithm`, 2007. [Online; accessed 20-December-2018].

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

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

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

[29] 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.

[30] 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.

[31] 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.

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

[33] 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.

[34] 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.

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

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

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

[38] 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.

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

[40] David Nutt, Sue Wilson, and Louise Paterson. Sleep disorders as core symptoms of depression. *Dialogues in clinical neuroscience*, 10(3):329, 2008.

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

[42] 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.

[43] 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.

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

[45] Itamar Procaccia. Complex or just complicated? *Nature*, 333:498–499, 1988.

[46] Paul E Rapp, Alfonso M Albano, TI Schmah, and LA Farwell. Filtered noise can mimic low-dimensional chaotic attractors. *Physical review E*, 47(4):2289, 1993.

[47] 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.

[48] Nicholas Rohrbacker. Analysis of Electroencephologram Data Using Time-Delay Embeddings to Reconstruct Phase Space. *Dynamics at the Horsetooth*, 1:1–11, 2009.

[49] 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.

[50] 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.

[51] 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.

[52] J. R??schke, J. 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.

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

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

[55] 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.

[56] 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.

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

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

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

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

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

[62] James Theiler. On the evidence for low-dimensional chaos in an epileptic electroencephalogram. *Physics Letters A*, 196(1-2):335–341, 1994.

[63] 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.

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

[65] 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.

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

[67] 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.