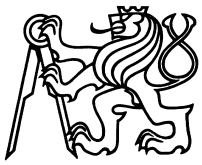
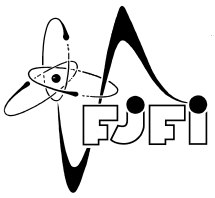


CZECH TECHNICAL UNIVERSITY IN PRAGUE
Faculty of Nuclear Sciences and Physical Engineering



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, November 29, 2018 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 [43] [37]. 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. [37]

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: [3]

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. [3]

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. [35]

Although EEG has significantly lower spatial resolution in comparison with other diagnostic techniques such as functional magnetic resonance sampling (fMRI) and magnetoencephalography (MEG) [51] and enables measuring only neural activity near the cortical surface, as a depression diagnostic tool, it has numerous benefits. Importantly, its significantly lower costs [56] [19], high portability, and ease of operation imply increased availability to the patients [49]. Moreover, it is perfectly noninvasive, which means less complications such as claustrophopia or anxiety [33].

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 ([41]). *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.*

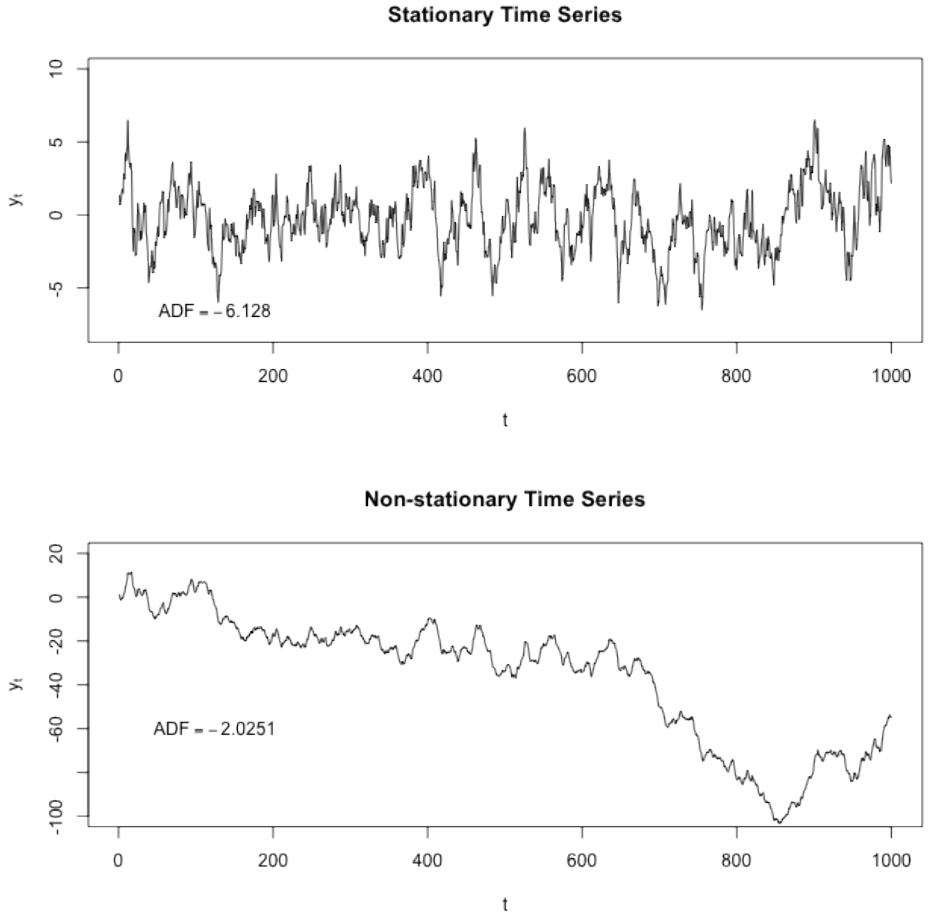


Figure 1.1: A comparison of stationary and non-stationary time series. (Courtesy: Protonk)

Definition 2 ([6]). *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* [24] [52]. 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 [8] [7] and diseases such as epilepsy [50], tremor [32], Parkinson's disease and depression [30]. 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. [55]

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. [24]

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. [16] [42]. 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 ([3]). *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 [29] and, since, in EEG signal analysis both M and d are unknown, we have no alternative but to work in Euclidean M.

Definition 4 ([3]). *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 differential equations*

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

such that there exists a mapping $f^t : M \to M$ satisfying

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

This mapping is called ***state evolution function***.

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 (if $\mathbf{F}$ satisfies certain reasonable properties). 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. [24]

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 $s : \mathbb{R}^d \to \mathbb{R}^{d'}$, where $d' \ll d$. 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.

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

1.4 Attractor

Depending on the properties of $\mathbf{F}$, there are several possibilities of how the system might evolve when as $t \to \inf$. In the following, we will focus on dissipative dynamical systems.

Definition 5 ([23]). *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. (TODO: How precise is this statement? Is it equivalent to the equation?)

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

[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.

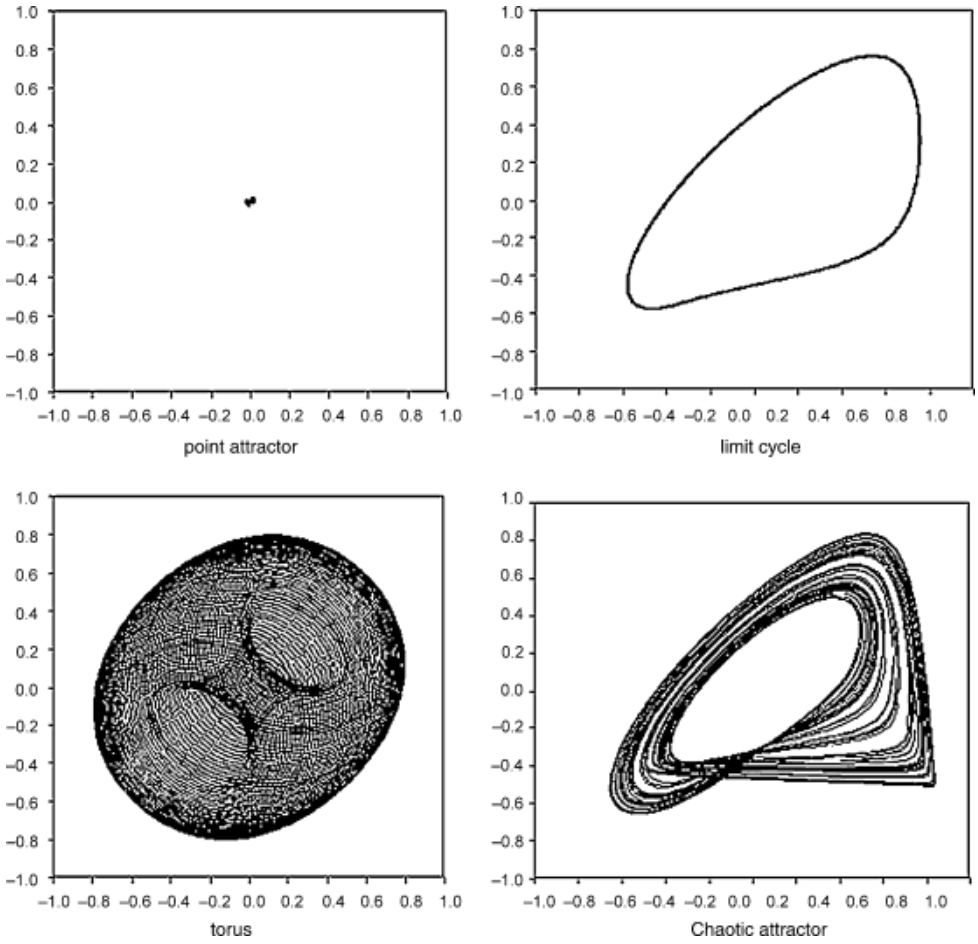


Figure 1.2: Visualization of four common attractor types (units are arbitrary). Left to right, top to bottom: *Point attractor* is the only type of attractor of linear deterministic dissipative systems. It consist of a single final state to which all points from the corresponding region of attraction 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. [3] ([52])

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 a self-similar attractor is shown on Figure 1.3). [3] For our purposes it is sufficient to say that this means that these attractors can be characterized as having (quantifiable) self-similarity.[2] However, the following definition will be useful:

Definition 6 ([11])**.** *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*

$$d_F = \lim_{\epsilon \to 0} \frac{\log N_\epsilon(F)}{\log \frac{1}{\epsilon}}, \tag{1.3}$$

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 $\epsilon \to 0$. [11]

1.4.1 Startionarity

(TODO: Define (non)stationarity.)

1.5 State space reconstruction

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

1. reconstruction of the dynamics of given system from recorded data,
2. characterization of the reconstructed attractor,
3. checking validity of the results with surrogate data testing. [52]

(TODO: Connect this to the content of this section. Expand on the steps.)

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*.

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

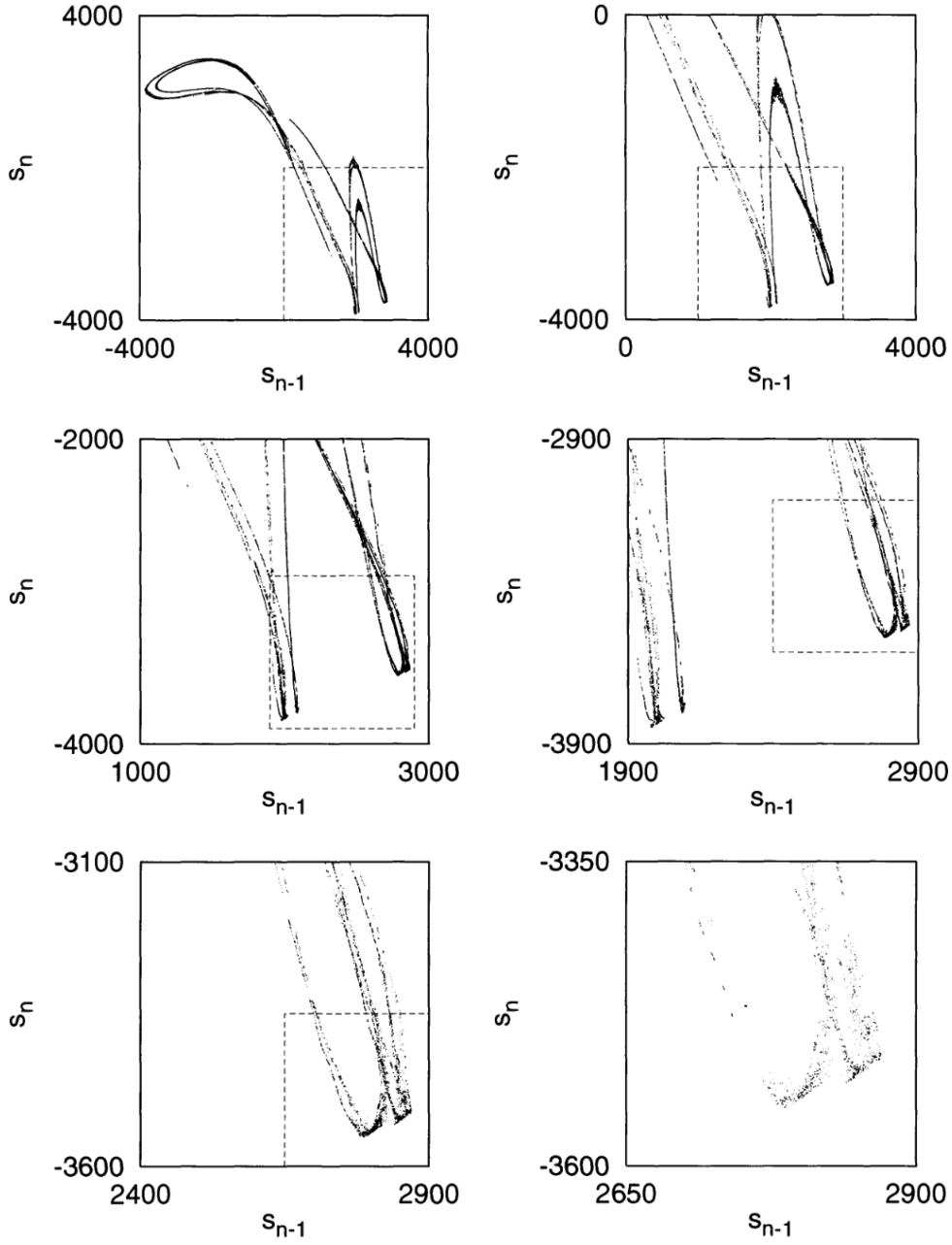


Figure 1.3: Noise-reduced visualization of successive enlargements of highly self-similar attractor. ([23])

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:

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

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 $\mathbf{s}_n = \Phi(\xi(n\Delta t))$. 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. [3] 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 [23].

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 M. This corresponds to the tangent space $D\Phi$ also being a one-to-one mapping.

(TODO: 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:

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.

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. [**?**] [3]) 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, this 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. Itwas 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 [3]). 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. [38] Indeed, since only a single sequence of scalar measurements is

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

available, our only option for constructing the vectors $\mathbf{s}_n$ is by using values recorded at multiple different times. Packard's technique, however, can be useful in general.

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[4]. Using the notation of 1.4, the time delay reconstruction is then formed by the following vectors:

$$\mathbf{s}_n = (s_{n-(m-1)\tau}, s_{n-(m-2)\tau}, \ldots, s_{n-\tau}, s_n), \tag{1.5}$$

for $n > (m-1)\tau$. [23]

A year after Packard's discovery, in [53], 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 ([53])**.** *Let M be a compact[5] smooth manifold specifying the state space of a deterministic dynamical system of dimension $d \in \mathbb{N}$, $s : M \to \mathbb{R}, s \in C^2$ 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. [48]

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 improvement came ten years later, when T. Sauer both generalized Takens' result as follows (in a simplified form):

Theorem 2 (Sauer, [47])**.** *Let A be a compact fractal with box-counting dimension d_A, and let A be a subset of a m-dimensional manifold. Then*

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

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 the vector $\mathbf{s}_n$ is a true embedding of the underlying attractor for almost any τ (note only sufficiency of the resulti - $\mathbf{s}_n$ 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". [54]

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

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

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 precisions, and absence of noise. Moreover, practical applications present further challanges, such as presence of noise.

(TODO: Expand.)

1.5.4 The choice of time delay

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. 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 $\tau_w = \tau(m-1)$ [26], 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. [3]

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. [9]

3. It was shown that increasing τ leads to rise in entropy. [22]

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 $s_t \approx s_{t+\tau}$, the reconstructed attractor will concentrate along the main diagonal (see Fig. 1.4, left hand side). Moreover, in this case, the effect of noise is amplified. [9]

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 Fig. 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).

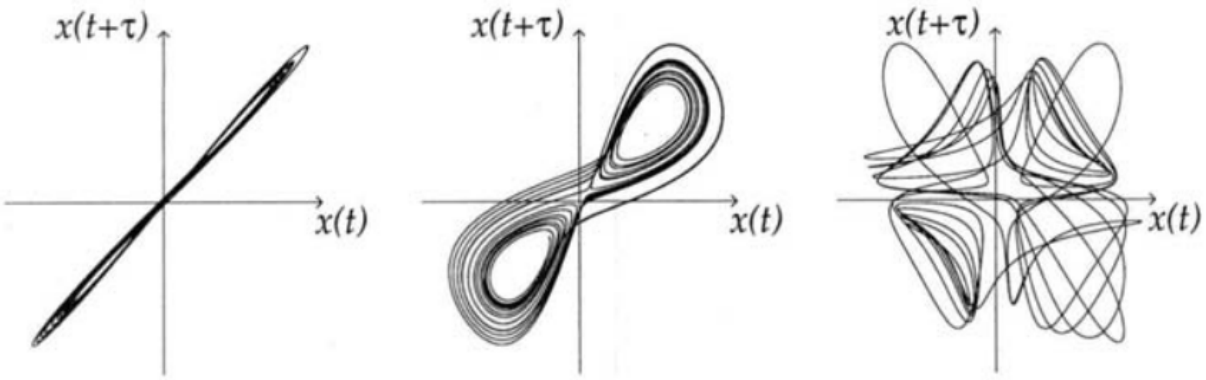


Figure 1.4: Time delay reconstructions of the Lorenz attractor for different values of τ. Figure on the left hand side shows choice of small τ and represents the case of redundance - the states concentrate along the main diagonal. Figure in the middle shows a successful reconstruction (although not an 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. ([3])

1.5.4.1 Autocorrelation

From the above, we know that for the optimal choice of time delay, we may want to study correlation of the elements of the reconstructed vectors $\mathbf{s}_n$. Thus, a natural method of estimating the time delay is studying the *autocorrelation function* A, and picking the first τ where $A(\tau)$ decays below a certain value (such as 0 or $A(0)/e$). [2]

Definition 7 ([23]). ***Autocorrelation*** *$A : \mathbb{R} \to \mathbb{R}$ for time delay τ is given by*

$$A(\tau) = \frac{1}{\sigma^2}\langle (s_n - \langle s_n \rangle)(s_{n-\tau} - \langle s_n \rangle) \rangle$$

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 $\mathbf{s}_n \in \mathbb{R}^m$ 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. s_t and $s_{t-\tau}$ being completely uncorrelated. [23] An obvious objection is that correlation between s_t and $s_{t-\tau}$ says nothing about correlation between s_t and $s_{t-2\tau}$, etc. Thus, this method is generally useful only for two-dimensional reconstructions.

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 $A(\tau) > A(0) - \sigma^2_{\text{noise}}/\sigma^2_{\text{signal}}$. [23]

1.5.4.2 Mutual information

Another commonly used method is to use the first minimum of the *time delayed mutual information*. [13]

Definition 8 ([23]). *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 s_t is in a bin i and $s_{t+\tau}$ is in a bin j. ***Delayed mutual information*** *$\mathcal{I}_\epsilon$ for time delay τ is defined as*

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

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 s_t and $s_{t-\tau}$ 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. [12]

Another criticism of delayed mutual information that some systems exhibit slowly decaying mutual information which has no minima. [31]

Moreover, both approaches described above 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 [6]: [26]

1. The reconstructed attractor must be expanded from the diagonal.
2. The components of the reconstructed vector $\mathbf{s}_n$ must be uncorrelated.

Those criteria are noticably similar, and bias towards larger estimates of τ. This leads many authors to suggest more advanced techniques.

(TODO: Maybe talk about techniques we didn't use, or we will yet try.)

1.5.5 The choice of embedding dimension

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. [25]

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 Fig. []). 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 will be true.

Let use denote by $y^{(r)}(n)$ the r-th nearest neighbor of $y(n)$. Then, let $R_m^2(n, r)$ denote the square Euclidean distance between $y(n)$ and its neighbor:

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

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

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:

$$\left[\frac{R_m^2(n,r) - R_{m+1}^2(n,r)}{R_m^2(n,r)}\right]^{1/2} = \frac{x(n+k\tau) - x^{(r)}(n+k\tau)}{R_m(n,r)} > R_{\text{tol}},$$

where $R_{\text{tol}} \in \mathbb{R}$ 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 that 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. [25]

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

$$\frac{R_{m+1}(n,r)}{R_A} > A_{\text{tol}}$$

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

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

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

Moreover, in practice, it has been found that the results of this method are sensitive not only to the tolerance parameters R_{tol} and A_{tol}, but also to the lag as well. [26] In this sense, they are somewhat arbitrary unless stable across different choices of τ.

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. [26]

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. [23]

1.5.5.2 Average nearest neighbors

(TODO: Average nearest neighbors.)

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.

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 Fig.).

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

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

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 [18]

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

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. It is commonly agreed that estimating Lyapunov exponents is even more difficult than esimating correlation dimension [3], although they have been successfully employed in EEG analysis. [45, 20, 52] 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. [18] 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. [44]

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

$$d(i) = d(0)e^{\lambda_1(i\Delta t)}, \quad d(i) \ll 1, \quad i \to \infty, \quad d(0) \to 0, \tag{1.7}$$

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. [23] Equivalently, $1/\lambda_1$ is (on average) an upper bound on predictability in the system. [3] 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. [44]

(TODO: Say what different values of λ_1 say about the system.)

In the following, we will describe *Rosenstein's algorithm* for computation of the largest Lyapunov exponent. [44] 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 each point on the trajectory $\mathbf{s}_n$, the algorithm locates the nearest neighbor $\mathbf{s}_{\hat{n}}$, such that their distance in the embedded space is minimized:

$$d_n(0) = \min_{n \neq \hat{n}} \|\mathbf{s}_n - \mathbf{s}_{\hat{n}}\|.$$

As an approximation, we want to assume $\mathbf{s}_n$ and $\mathbf{s}_{\hat{n}}$ 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:

$$\frac{1}{4} \text{ time series length} > |n - \hat{n}| > \text{mean period of the time series.}$$

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

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

By taking logarithm of both sides, we obtain

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

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

$$y(i) = \frac{1}{\Delta t}\langle \ln d_n(i) \rangle,$$

where the average $\langle \cdot \rangle$ is over n. 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.

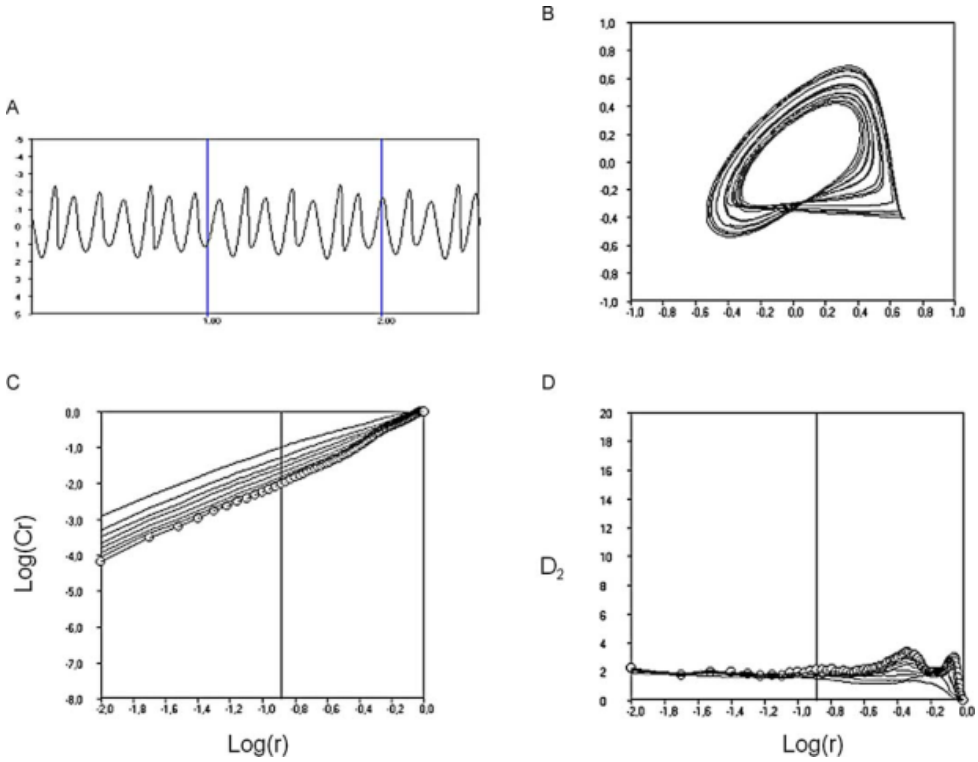


Figure 1.5: Computation of the correlation dimension [52]. TODO: Add description.

1.6.2 Correlation dimension

Correlation dimension is a characteristic measure which describes complexity of the geometry of chaotic attractors.

Correlation sum $C(r)$ is defined as the fraction of points in the phase space whose distance is smaller than r:

$$C(r) = \frac{2}{N(N-1)} \sum_{i<j} \Phi(r - \|\mathbf{s}_i - \mathbf{s}_j\|). \tag{1.8}$$

If $C(r)$ decreases according to the power law with $r \to 0$ such that $C(r) \approx r^D$, then D is called the correlation dimension, formally defined as

$$CD = \lim_{r \to 0} \frac{\log C(r)}{r}. \tag{1.9}$$

$$[[54]]\text{bulk} \approx \text{size}^{\text{dimension}}$$

$$[[54]]\text{dimension} = \lim_{\text{size} \to 0} \frac{\log \text{bulk}}{\log \text{size}}$$

1.7 Visual characterization of the dynamical system

1.7.1 Phase space plot

1.7.2 Poincare plot

1.7.3 Recurrence plot

When presented with a task of finding regularities in seemingly chaotic data, one possible approach is analysing at least approximate repetitions of simple patterns, which can be further used for reconstruction of more complicated rules. Recurrence plot is a method of visualizing obtained state-space trajectory segments in relation to each other to achieve this goal. Furthermore, it can be used to test necessary conditions for validity of dynamical parameters derivable from a non-linear time series such as the information dimension, entropy, Lyapunov exponents, dimension spectrum, etc. The information contained in recurrence plots is not easily obtainable by other known methods. [10]

Definition 9 ([10])**.** *Let N be the length of given time series, $\mathbf{s}_i$ for $i \in \{1, 2, \ldots, N\}$ be a i-th delay vector of any integer embedding dimension, $\|\cdot\|$ a norm, $\Theta(\cdot)$ a Heaviside step function, and $\epsilon \in \mathbb{R}_0^+$ a tolerance parameter. Then,* ***recurrence plot*** *is the matrix*

$$M_{ij} = \Theta(\epsilon - \|\boldsymbol{s}_i - \boldsymbol{s}_j\|)\,. \tag{1.10}$$

In other words, M_{ij} is a symmetric[7] binary NxN matrix, where $M_{ij} = 1$ when i-th and j-th points of the reconstructed trajectory enter each other's ϵ neighborhood.

The essential drawback of recurrence plot is their size - it is quadratic in the length of the time series. A simple way of reducing its dimension is to partition the time series into disjoined segments, and let M_{ij} represent the distance between those two segments. This is known as **meta-recurrence plot**. [23] (TODO: Find a justification for using them.)

(TODO: Cross-recurrence plots may be useful? Only between two series. Joint recurrence plots may be used to detect phase synchronization.)

1.8 Applications in disease diagnosis

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. [43]

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. [34] [39]

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. [45]

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. [1]

[7] Although this is true for our definition, it may not be true for an alternative definition using a more general topology instead of a norm.

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. [20]

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. [5]

Sleep disorder diagnosis may also relevant to this work for the very close connection of depression with disturbed sleep and insomnia [36]. The first study emplying techniques of non-linear analysis on human EEG was published in 1985 and dealt with sleep recordings. [4] 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 [52].

Chapter 2

Convolutional Neural Networks

2.1 Mathematical background

TODO: Do we really need this section???

Definition 10. *Let I be an image function, K a kernel. A (discrete)* ***convolution*** *of I and K is a functional defined as*

$$(I * K)(i, j) = \sum_m \sum_n I(m, n)K(i - m, j - n)\,. \tag{2.1}$$

Note that some machine learning libraries (such as Tensorflow) implement **cross-correlation** instead of convolution, but preserving the term convolution for the operation. Cross-correlation corresponds to convolution with kernel rotated by 90 degrees:

$$(I * K)(i, j) = \sum_m \sum_n I(m, n)K(i + m, j + n)\,. \tag{2.2}$$

Unlike convolution, cross-correlation is not commutative, but this property is not required for neural network applications.

Definition 11. *Let f be arbitrary function, and $\mathcal{D}$ its degradation operator. We say f is* ***invariant*** *under $\mathcal{D}$ if*

$$\mathcal{D}(f) \equiv f\,. \tag{2.3}$$

For the following, the reader needs to understand the term **equivariance**.

Definition 12 ([40])**.** *Let G be a group and X, Y its G-sets. Then $F : X \to Y$ is called an* ***equivariant function*** *if*

$$F(g(x)) = g(F(x)) \tag{2.4}$$

for all G actions g and $x \in X$.

For our purposes, we can view G as a group of transformations, and then equivariance as a commutative property of a function with regards to the transformations. In other words, computing the function and then applying the transformation has the same effect as applying the transformation and then computing the function.

Algorithm 1 Gradient descent algorithm.

1: Initialize random $x_0 \in D(f)$
2: $n \leftarrow 0$
3: step_size $\leftarrow 1$
4: **while** step_size < threshold and $n <$ iters_limit **do**
5: $\quad x_{n+1} = x_n - \epsilon \nabla_{x_n} f$
6: $\quad$ step_size $\leftarrow |x_{n+1} - x_n|$
7: $\quad n \leftarrow n + 1$
8: **end while**

TODO: This is probably too basic to be here

Gradient descent is a first order iterative method of finding an extremum a differentiable function $f : \mathbb{R} \to \mathbb{R}^n$, $f \in C^1$, based on continually moving a point in its domain in the direction of negative of its gradient at that point, until the absolute value of the gradient (or the step size) is below a certain threshold. See Algorithm 1.

Stochastic gradient descent...

2.2 History

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

preprocessing: supressing unwanted distortions and noise, enhancement beneficient for further processing,

object segmetation: separating disparate objects from the background,

feature extraction: gathering relevant information about the properties of the objects, removing irrelevant variations,

classification: categorizing segmented objects based on obtained features into classes.

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

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

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

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

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

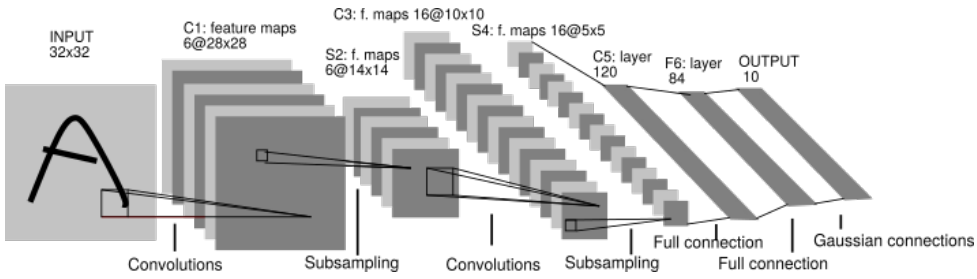


Figure 2.1: LeNet-5 architecture [28].

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

Although the main idea dates back 1980 with K. Fukushima's neocognitron [14], the back-propagation algorithm was not known at the time. The first convolutional architecture successfuly applied on an image pattern recognition problem by attempting to solve the aforementioned problems, dubbed LeNet-5, was proposed in 1998 by Y. LeCun, L. Bottou, Y. Bengio and P. Haffner [27].

2.3 Description

Bearing resemblence to visual processing in biological organisms [1], LeNet-5 proposed the following design principles to enforce *shift, scale and distortion invariance*: [28]

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

shared weights: each layer is composed of neurons organized in planes within which each neuron have the same weight vector (feature map),

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

2.3.1 Local receptive fields

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

For a visual explanation of the concept of receptive field, see Figure 2.2. The locality of of those receptive fields implies sparser connectivity, and hence more efficient computations in comparison with fully connected neural networks. A fully connected neural network with no hidden layers with m inputs and n outputs has $m \times n$ weight parameters, and the correspoding feed forward pass (matrix multiplication) is of $O(m \times n)$ time complexity per input. If the number of connections per output unit is limited to $k < m$,

[1] As early as in 1968, D. H. Hubel and T.N. Wiesel discovered that some cells (called simple cells) in cat's primary visual cortex (V1) with small receptive fields (shared by neighboring neurons) are sensitive to straight lines and edges of light of particular orientation, and other cells (called complex cells) with larger receptive fields further in the visual cortex also respond to straight lines and edges, but with invariance to translation [21].

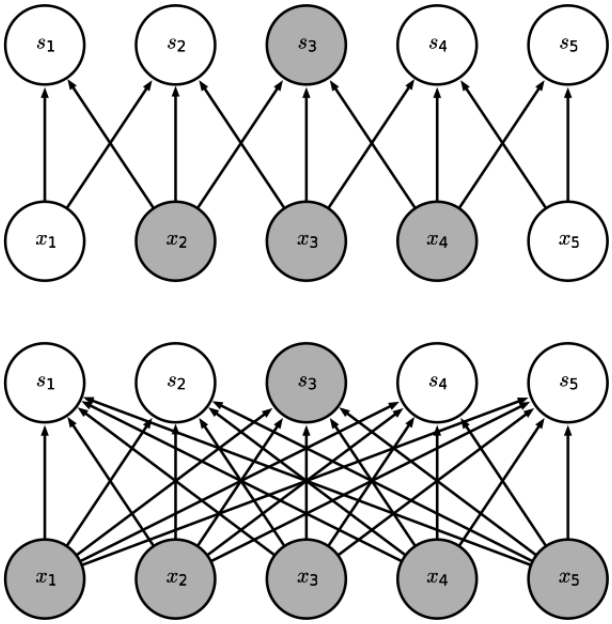


Figure 2.2: Receptive field. [15]

the achieved runtime is $O(k \times n)$, where k is usally in practice several orders of magintude smaller than m. [15]

In shallow neural networks, locality of receptive fields implies locality of "influence" of each input unit on the output. In deep neural networks, on the other hand, units in the deeper layers can be indirectly connected to some or all units of the input, thus enabling them to achieve aforementioned effect of combining more complex features from simpler ones.

2.3.2 Shared weights

With *shared weights*, neural units in a layer with differing receptive fields have the same feature map and the same feature detecting operation (convolution with feature map kernel followed by additive bias and a application of a non-linear function) is performed on differing parts of the image (see Figure 2.3). A single convolutional layer is composed of multiple feature detecting planes.

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

Although it does not reduce the time complexity of the feedforward pass, it does reduce the memory requirements. If the kernel size is k, m the number of inputs, n the number of outputs, the number of parameters per layer is k instead of $m \times n$ (per feature detecting plane) in a fully connected case. Since k is usually in practice several orders of magnitude smaller than m, and usually m and n are comparable in size, the memory savings are highly significant. [15]

One of the drawbacks of classical CNNs is that although convolution in combination with weight sharing causes layer output to be equivariant to translation of the input, this is not the case for scaling and rotation. Moreover, equivariance to input may not be always desirable. Consider a case of face detection, where all training and test images are centered. Then, the relative positions of individual

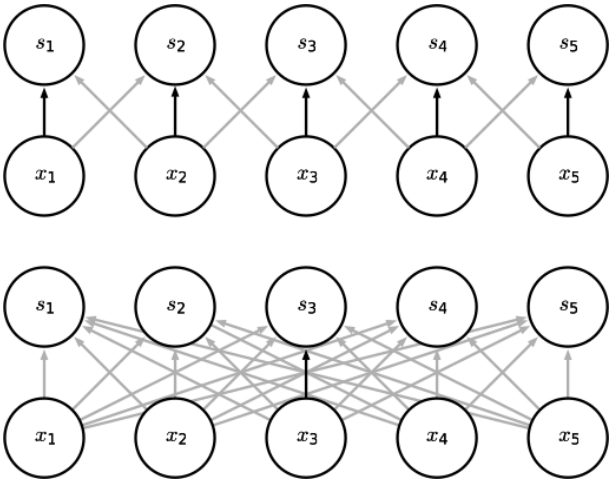


Figure 2.3: Shared weights. [15]

features are important, and it may be favorable to fix feature detectors (and thus weights) to certain locations in the image.

2.3.3 Pooling

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

1. linear unit activations are computed via the convolution operation,

2. a non-linear activation function is applied to the activations,

3. a spatial subsampling (pooling) operation is applied.

The rationale behind applying a non-linearity is it makes the network capable of modelling non-linear functions. Common activation functions include rectified linear $\max(0, x)$, sigmoid $\frac{1}{1+\exp(-x)}$, hyperbolic tangent tanh, and many others. They have varying properties making them useful in different situations. We will not explore them further here.

Pooling operation splits the neural units into sets of multiple adjacent activations and computes a summary statistic, such as the maximum element (max pooling) or the average (average pooling), per such set and outputs the result. If the stride between the sets is greater than one, the spatial dimension of output is decreased relative to input (subsampling).

The purpose of spatial subsampling is to ensure scale and distortion invariance[2] by reducing the precision at which a feature is encoded in a feature map by reducing its resolution - when scale and distortion invariance is assumed, the exact location of a feature becomes less important and is allowed to exhibit slight positional variance - roughly speaking, an "approximate" translation invariance.

Although the combination of convolution and pooling performs well in many practical situations, it has multiple drawbacks. For example, the learned representations are not rotation invariant and thus, to mitigate this, the capacity of the network has to be increased and the training dataset must be enhanced to contain examples of rotated features, often extending the amount of data necessary and training time. A

[2]Whether it achieves this goal has been famously doubted by Geoffrey Hinton: "The pooling operation used in convolutional neural networks is a big mistake and the fact that it works so well is a disaster." []

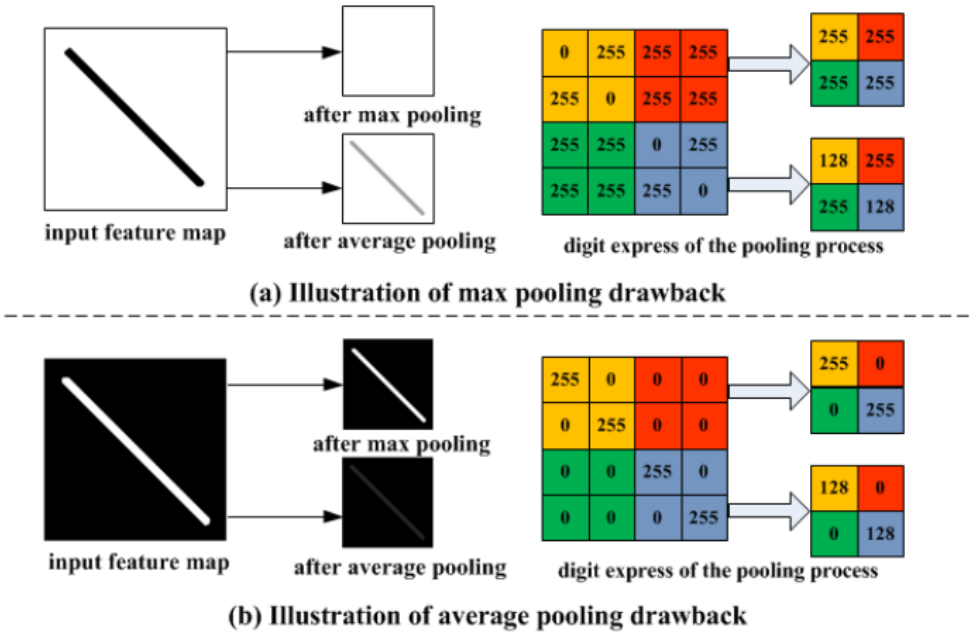


Figure 2.4: Examples of drawbacks of the pooling operation. Max pooling discards all except the maximum element, and valuable information may thus be lost. Average pooling considers all the values, and the information about their contrast is reduced. Moreover, extreme values may have undesired effects on the result. [57]

number of alternative approaches were suggested in the litarature.[3] For another example of a limitation, see Figure 2.4.

2.4 Applications

Maybe mention an example of how LeNet-5 was improved on subsequently (AlexNet, ResNet, etc.)? But this changes all the time...

2.5 CapsNets?

Does it make sense trying them? I found a only a few successful implementations. Maybe it would be better to try those after we have some results already, because it seems risky - we might end up with nothing.

[3] For instance, Hinton's *CapsNet*, described e.g. in [46], is an attempt to transform the manifold of images of similar shape (which is highly non-linear in the space of pixel intensities) to a space where it is globally linear by the way of using so called capsules instead of traditional convolutional layers.

Chapter 3

Experiments

3.1 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). 99 subjects EEG signal was measured 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.

3.2 Preprocessing

Recordings of insufficient duration were not used for further analysis. The threshold was selected to be 60 s, resulting in exclusion of 26 recordings from the total of 266.

All signals were highpass filtered with 0.5 Hz frequency and lowpass filtered with 70 Hz frequency. Our hypothesis is that such filter should not affect signals of our interest, since neural oscillations are known to lie inside the unmodified frequency band.

Lastly, recordings of $f_s = 1000$ Hz were downsampled to 250 Hz using the Fourier method.

3.3 Feature extraction

Based on multiple studies successfully applying non-linear dynamical analysis on the problem of psychological disorder diagnosis mentioned in Section 1.8, we decided to extract the following non-

linear features, quantifying the amount chaos and complexity of the signal.

In the first experiment, we used time-delay embedding (see Section 1.5.1), resulting in 5 x 19 features for each measurement.

3.3.1 Largest Lyapunov exponent

Largest Lyapunov exponent (LLE) was computed using the Rosenstein's algorithm [44]. This algorithm was found to be robust to noise and the choice of lag and embedding dimension.

First, it reconstructs phase space trajectory using the method of delays described in Section 1.5.1. The lag is selected as the value for which the autocorrelation function drops to below $1 - 1/e$ of its initial value. Then, for each point on the reconstructed trajectory $\mathbf{s}_n$, a nearest neighbor $\mathbf{s}_{\hat{n}}$ is found as

$$d_n(0) = \min_{\hat{n} \neq n} \|\mathbf{s}_n - \mathbf{s}_{\hat{n}}\|,$$

where, additionaly, the nearest neighbors have temporal separation greater then the reciprocal of the mean frequency of the power spectrum of the time series, so that they can be safely considered to be nearby initial conditions for different trajectories (this separation, however, is restricted to be at most $1/4$ of the time series length). The mean rate of separation between the nearest neighbors is then an unbiased estimator for the LLE (TODO: Find citation for this claim.)

From the definition of the LLE, the algorithm proceeds by assuming that each pair of nearest neighbors diverge approximately at the rate given by the LLE:

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

By taking the logarithm of both sides,

$$\ln d_n(i) \approx \ln d_n(0) + \lambda_1 n \Delta t,$$

we obtain a set of lines (one for each index n), whose slope is an approximation of the largest Lyapunov exponent. Hence, the value of λ_1 is approximated as the slope of least squares fitted line through the mean log divergence $\overline{d}$,

$$\overline{d}(i) = \langle \ln d_n(i) \rangle.$$

3.3.2 Correlation dimension

The simplest version of the Grassberger-Procaccia algorithm was used to compute the correlation dimension. [17] The value of the correlation integral $C(r)$ from the definition 1.8 is computed for multiple values of r in range from $0.1 * \sigma$ to $0.5 * \sigma$, where σ is the standard deviation of the time series, and a least squares straight line is plotted through the plot of $\ln C(r)$ against $\ln r$. Correlation dimension is approximated as the slope of the line.

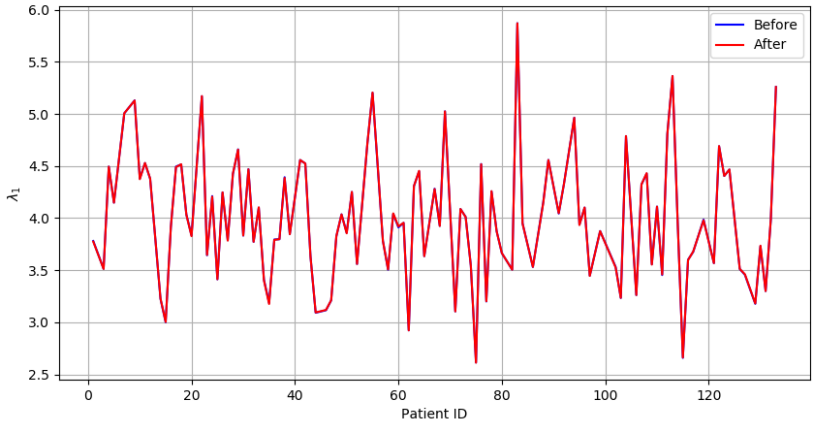


Figure 3.1

3.3.3 Sample entropy

3.3.4 Detrended fluctuation analysis

3.3.5 Hurst exponent

3.4 Unsupervised analysis of before / after treatment differences

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.

To this goal, we started by simply plotting each measure's mean value over channels for the recording before and after administration of drugs for all patients. Moreover, 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.

We found that for each measure except correlation dimension, its average over channels, although differring between subjects, is remarkably stable for all patients accross measurements. This means that except for correlation dimension, information about any change between measurements was not caputered by the computed non-linear measures (see Figure 3.1, 3.2).

Moreover, we found that for all measures, either for mean value or all channels, we cannot reject the hypothesis that the computed values are drawn from the same distribution. This is true even when patients responding and not responding to treatment are considered separately.

Apart from computing the mean, principal component analysis was used to reduce the number of dimensions of the 19 dimensional feature vectors. By visually inspecting projections into 2, 3 (2/3D plot) and 4 dimensions (a heatmap), we were unable to find any separation boundary between before and after group. [1] (see Figure 3.3, 3.4, 3.5).

Also, by looking at the subjects above 90th percentile of euclidean distance between before and after vectors in the projected space, we were unable to find any regularity. Subjects in those groups seem to be drawn randomly from the dataset.

[1] And neither for male / female, responding / non-responding, age < 40 / age > 50 groups, and in groups based on depression scores.

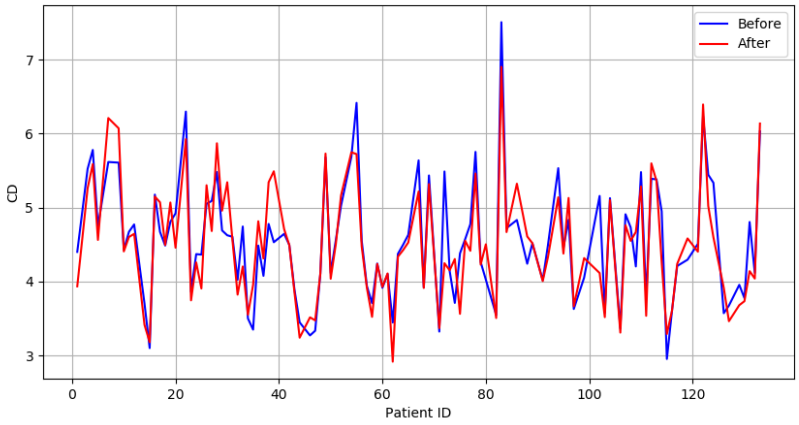


Figure 3.2

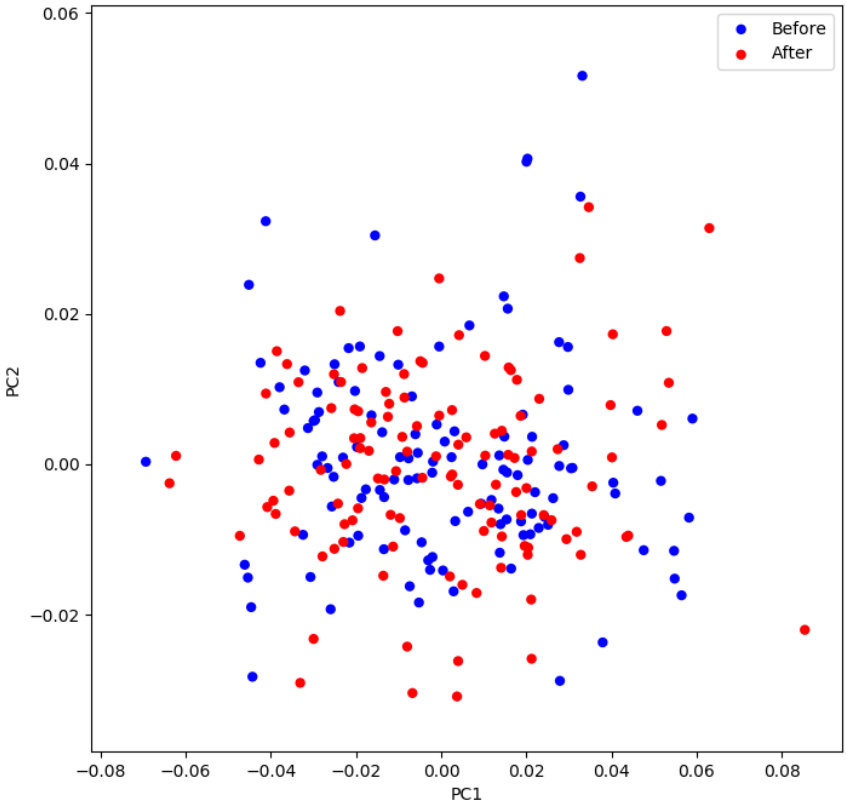


Figure 3.3

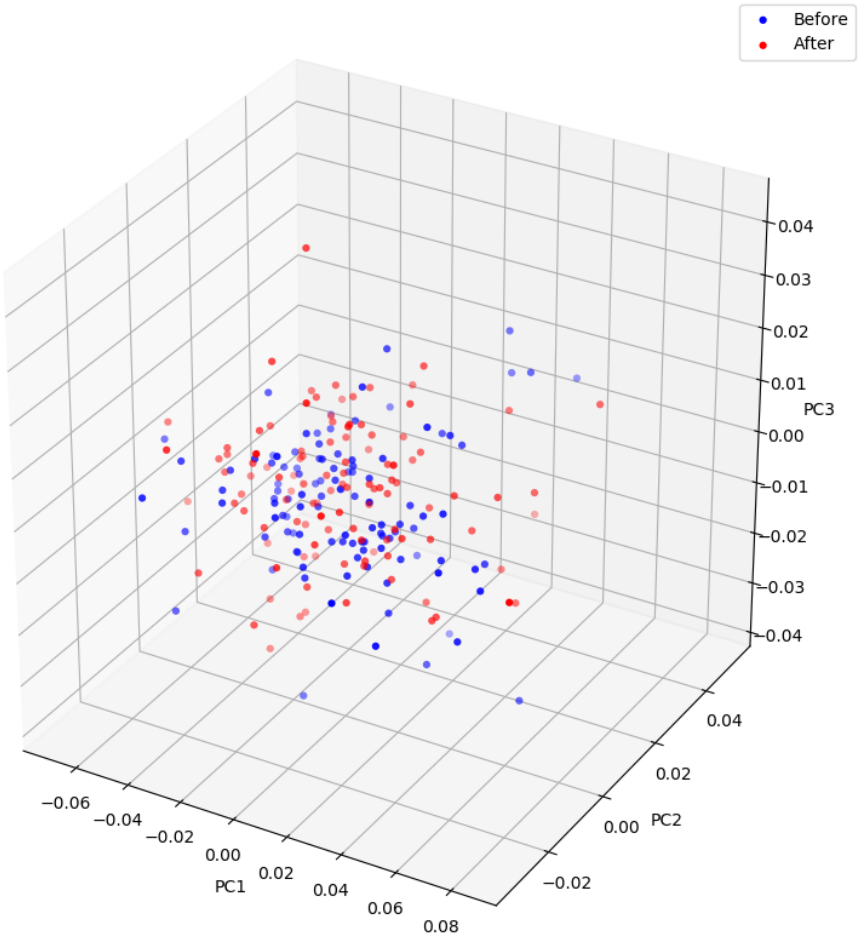

Before
After
PC1
PC2
PC3
−0.06
−0.04
−0.02
0.00
0.02
0.04
0.06
0.08
0.04
0.02
0.00
−0.02
0.04
0.03
0.02
0.01
0.00
−0.01
−0.02
−0.03
−0.04

Figure 3.4

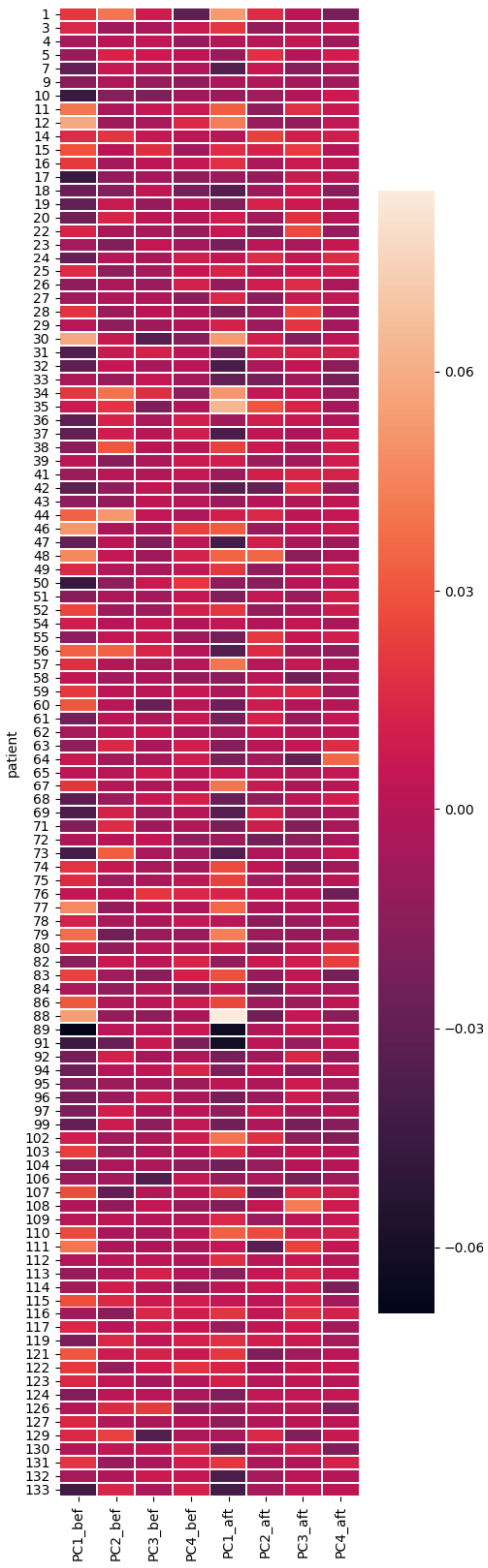

patient
PC1_bef
PC2_bef
PC3_bef
PC4_bef
PC1_aft
PC2_aft
PC3_aft
PC4_aft
0.06
0.03
0.00
−0.03
−0.06

Figure 3.5

3.5 Results

Conclusion

Bibliography

[1] 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.

[2] 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.

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

[4] 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.

[5] 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.

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

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

[8] 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.

[9] 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.

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

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

[12] 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.

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

[14] Kunihiko Fukushima and Sei Miyake. Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In *Competition and cooperation in neural nets*, pages 267–285. Springer, 1982.

[15] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

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

[17] Peter Grassberger and Itamar Procaccia. Characterization of strange attractors. *Physical review letters*, 50(5):346, 1983.

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

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

[20] 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.

[21] D. H. Hubel and T. N. Wiesel. Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology*, 1968.

[22] 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.

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

[24] 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.

[25] 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.

[26] 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.

[27] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2323, 1998.

[28] Yann LeCun, Patrick Haffner, Léon Bottou, and Yoshua Bengio. Object recognition with gradient-based learning. In *Shape, contour and grouping in computer vision*, pages 319–345. Springer, 1999.

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

[30] 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.

[31] 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.

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

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

[34] 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.

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

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

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

[38] 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.

[39] 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.

[40] Andrew M Pitts. *Nominal sets: Names and symmetry in computer science*. Cambridge University Press, 2013.

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

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

[43] 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.

[44] 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.

[45] 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.

[46] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. Dynamic Routing Between Capsules. (Nips), 2017.

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

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

[49] 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.

[50] 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.

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

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

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

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

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

[56] 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.

[57] Dingjun Yu, Hanli Wang, Peiqiu Chen, and Zhihua Wei. Mixed pooling for convolutional neural networks. In *International Conference on Rough Sets and Knowledge Technology*, pages 364–375. Springer, 2014.