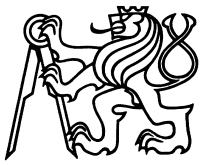
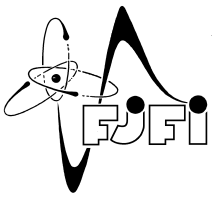


CZECH TECHNICAL UNIVERSITY IN PRAGUE
Faculty of Nuclear Sciences and Physical Engineering



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 13, 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 [30] [25]. 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. [25]

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

1.1 EEG signal

Nature, processes in the brain, way of measuring, limitations, complications

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. [23]

Although EEG has significantly lower spatial resolution in comparison with other diagnostic techniques such as functional magnetic resonance sampling (fMRI) and magnetoencephalography (MEG) [36] and enables measuring only neural activity near the cortical surface, as a depression diagnostic tool, it has numerous benefits. Importantly, its significantly lower costs [40] [12], high portability, and ease of operation imply increased availability to the patients [34]. Moreover, it is perfectly noninvasive, which means less complications such as claustrophopia or anxiety [21].

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 ([29]). *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.*

Definition 2 ([4]). *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* [16] [37]. 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 [6] [5] and diseases such as epilepsy [35], tremor [20], Parkinson's disease and depression [19]. 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. [39]

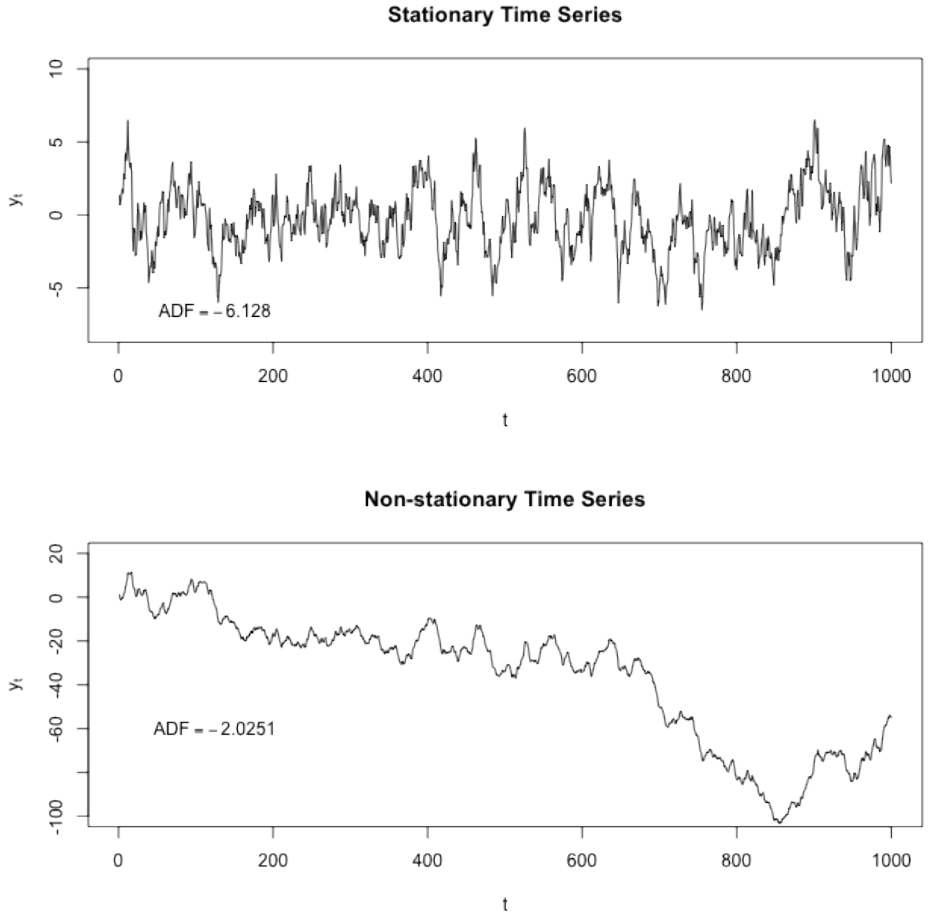


Figure 1.1: A comparison of stationary and non-stationary time series. (Courtesy: Protonk)

1.2 Theory

Modern neuroscientific work suggests that the 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, it is common to apply principles derived from nonlinear dynamics to characterize these neuronal systems.[16] In this section, we will describe some of those principles.

Definition 3 ([15])**.** *Let* $\mathbf{x} \in \mathbb{R}^m$ *be an* $m \in \mathbb{N}$ *dimensional state space vector dependent on time. A deterministic* ***dynamical system*** *is described by a set of m first-order differential equations*

$$\frac{d}{dt}\mathbf{x}(t) = \mathbf{f}(\mathbf{x}(t)), \qquad t \in \mathbb{R} \tag{1.2}$$

In other words, a dynamical system is a model that determines the evolution of a system only given by the initial state, and current state is a function of the previous state. Hence, a total description of a dynamical system is given by the initial state and a set of equations. 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. [16]

1.2.1 Attractor

Definition 4 ([15])**.** *A dynamical system is called dissipative, when it is the case that*

$$E[\mathrm{div}\mathbf{f}] < 1 \tag{1.3}$$

In other words, average space space volume (containing the initial state) is contracted as the system evolves.

Dissipative dynamical system have the property that a set of initial states (of positive measure) will converge to a subspace of the overall state space. This subspace 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 complex, and exhibit what is known as *fractal geometry* (an example of a self-similar attractor is shown on Figure 1.3). . For our purposes, this means they can be characterized as having quantifiable self-similarity.[1] However, the following definition will be useful:

Definition 5 ([8])**.** *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*

$$d_F = \lim_{\epsilon \to 0} \frac{\log N_\epsilon(F)}{\log \frac{1}{\epsilon}}, \tag{1.4}$$

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 $\epsilon \to 0$. [8]

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

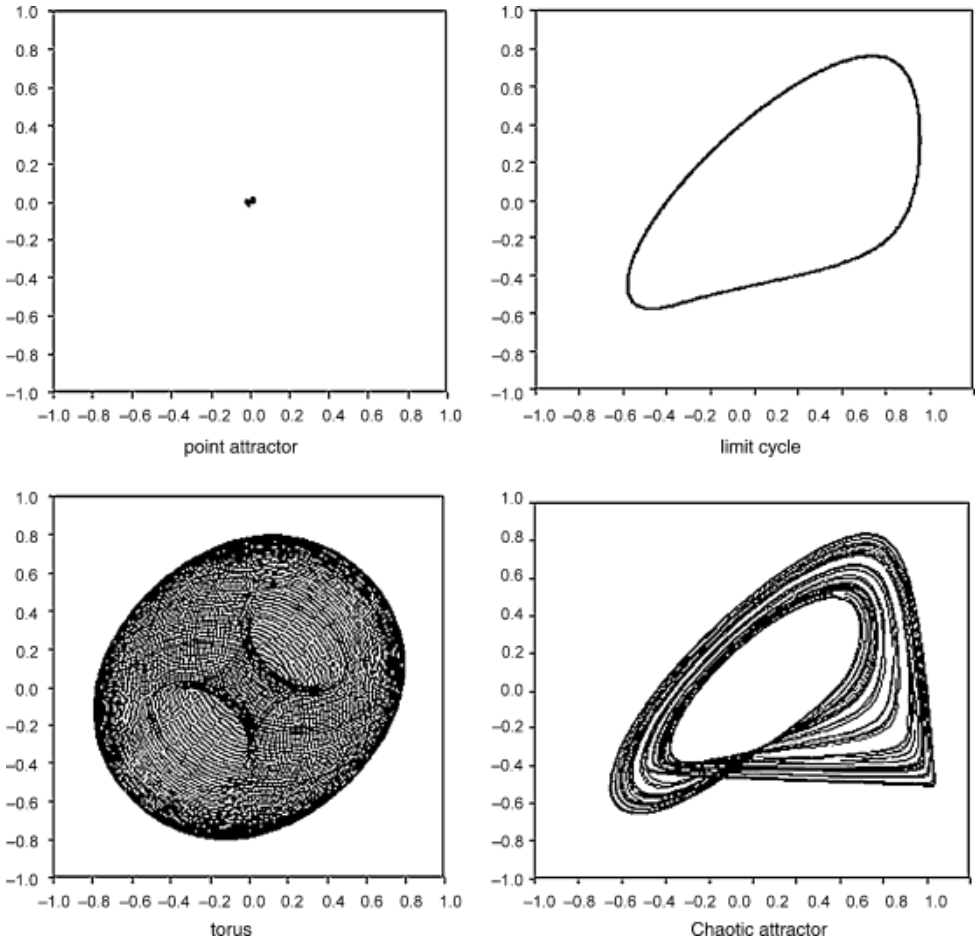


Figure 1.2: Visualization of four common attractor types (units are arbitrary). Left to right, top to bottom: *Point attractor* is the only type of attractor of linear deterministic dissipative systems. *Limit cycle* corresponds to a peridic system. *Torus attractor* corresponds to a quasi-periodic system, resulting (in this example) from a superposition of two periodic oscillations. *Chaotic* (strange) *attractor*, characterized by strong dependence on the original conditions. ([37])

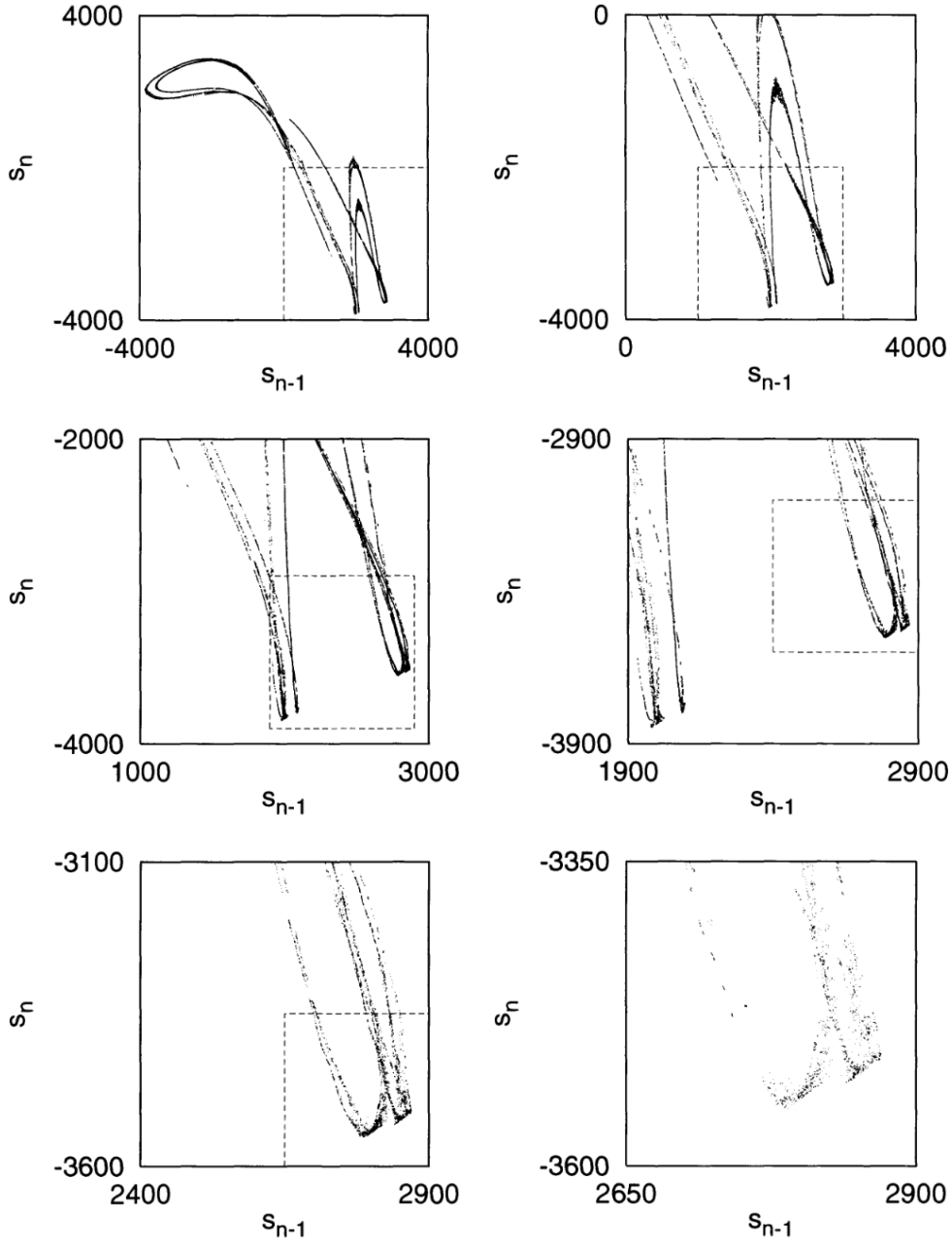


Figure 1.3: Noise-reduced visualization of successive enlargements of highly self-similar attractor. ([15])

1.2.2 Embedding

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. [37]

In the following text, we will describe multiple techniques of characterizing the properties of attractors (and hence ways of describing the dynamics of the dynamical system) using these steps.

In order to reconstruct the dynamics of given system from recorded data, 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). The process known as *state* (phase) *space reconstruction via the method of time delays* is a way of reconstructing the attractor of given dynamical system from single (or multiple) sequences of scalar measurements.

The following result from 1936 was important for theoretical understanding of time delay embeddings:

Theorem 1 (Whitney embedding theorem, [41])**.** *Every n-dimensional differentiable manifold can be embedded in* $\mathbb{R}^{2n}$.

N. H. Packard has proposed in [26] a process by which a time series of measurements sampled at regular intervals can be embededded into a n-dimensional space in such a way that the series of reconstructed vectors is representative of the dynamics of the system. A year later, in [38], F. Takens has proved theoretically, using Whitney's result, that the reconstructed attractor has the same dynamical properties (entropy, dimension, Lyapunov spectrum) as attractor of the original system.

Takens delay embedding theorem is an important result of non-linear time series analysis and can be stated as follows:

Theorem 2 ([38])**.** *Let M be a compact*[2] *manifold specifying the state space of a deterministic dynamical system of dimension $n \in \mathbb{N}$, $s : M \to \mathbb{R}, s \in C^2$ a smooth measurement function, $f : M \to M, f \in C^2$ a smooth diffeomorphic state evolution function with a strange attractor of box-counting dimension d_A, and let $k > 2d_A$. Then a map $\phi : M \to \mathbb{R}^k$, defined by*

$$\phi(x) = (s(\mathbf{x}), s(f(\mathbf{x})), \ldots, s(f^{k-1}(\mathbf{x}))) \tag{1.5}$$

is an embedding.

In what follows, we will describe how these results can be applied in practice. Let us have a time series of scalar measurements of a quantity depending on the current state of the system:

$$s_n = s(\mathbf{x}(n\Delta t)) + \eta_n\,, \tag{1.6}$$

where $\mathbf{x}$ is a state space vector, $s(\cdot)$ is a measurement function and η_n is a measurement noise.

Then, the time delay reconstruction is then simply formed by the following vectors:

$$\mathbf{s}_n = (s_{n-(m-1)\tau}, s_{n-(m-2)\tau}, \ldots, s_{n-\tau}, s_n)\,, \tag{1.7}$$

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

for $n > (m-1)\tau$. [15]

The purpose of this embedding process is to unfold the observed scalar projection of the origin system back to a multivariate state space which is representative of the original system. Theorem 2 ensures that when m is chosen such that $m > 2d_A$, then the vector $\mathbf{s}_n$ is a true embedding of the underlying attractor (note only sufficiency of the result).

Of course, this method requires proper choice of parameters m and τ. From Theorem 2 follows that vectors $\{\mathbf{s}_n\}$ are difeomorphic to the attractor A of the system M if m is sufficiently large, specifically $m > 2d_A$ (note that d_A may be much smaller than dim M). On the other hand, m adds redundancy, makes successive alogorithms less efficient. Moreover, the total time delay $m\tau$ is limited by the recording time.

There are multiple methods of choosing the parameters. We will start with a description of an algorithm for determining the minimal sufficient embedding dimension m, known as *false nearest neighbors* algorithm.

(TODO: Mention spatial embedding. Is it worth using for EEG analysis? Pro: it does feature extraction for us. Con: we may lose valuable information.)

1.2.3 False nearest neighbors

The false nearest neighbors algorithm is based on simple topological insight. If the embedding dimension m is too small, some points that are close to each other in dimension $m + 1$ (see Fig. []). Neighboring points satisfying this property are called *false neighbors*.

(TODO: Mutual information algorithm for τ)

1.3 Non-linear measures

1.3.1 Lyapunov exponents

Lyapunov exponents can be used to quantify sensitivity of a dynamical system to initial conditions. Consider a small sphere of similar initial conditions in the phase space of the attractor. As the the system evolves, this sphere expands at exponentially in all the m directions at rates given by the spectrum of Lyapunov exponents. Hence, m dimensional system has exactly m Lyapunov exponents and the average rate of separation between two points in the phase space with similar initial conditions can be characterized by the largest Lyapunov exponent. A sigle positive exponents is a sufficient indication of chaos.

In order to compute entire Lyapunov spectrum, it is necessary to measure the rate of separation along the appropriate Lyapunov directions. This requirement, however, is unnecessary for computation of only the Largest lyapunov exponent, as its direction dominates growth. Hence, the largest Lyapunov exponent λ_1 can be defined as

$$d(t) = Ce^{\lambda_1 t}, \tag{1.8}$$

where d is a function of the average divergence, and C is a factor normalizing for the initial separation.

1.3.2 Correlation dimension

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

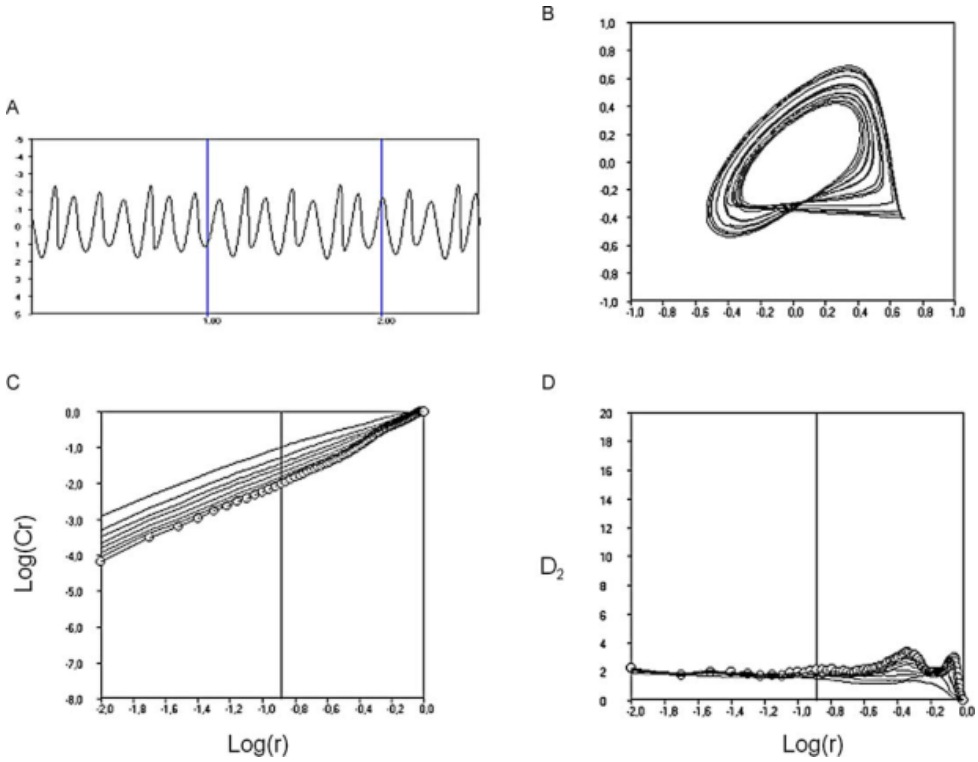


Figure 1.4: Computation of the correlation dimension [37]. TODO: Add description.

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

$$C(r) = \frac{2}{N(N-1)} \sum_{i<j} \Phi(r - \|\mathbf{s}_i - \mathbf{s}_j\|). \tag{1.9}$$

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

$$CD = \lim_{r \to 0} \frac{\log C(r)}{r}. \tag{1.10}$$

1.4 Visual characterization of the dynamical system

1.4.1 Phase space plot

1.4.2 Poincare plot

1.4.3 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. [7]

Definition 6 ([7]). *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*

$$M_{ij} = \Theta(\epsilon - \|\boldsymbol{s}_i - \boldsymbol{s}_j\|)\,. \tag{1.11}$$

In other words, M_{ij} is a symmetric[3] 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**. [15] (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.)

1.5 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. [30]

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. [22] [27]

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. [32]

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]

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. [13]

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. [3]

Sleep disorder diagnosis may also relevant to this work for the very close connection of depression with disturbed sleep and insomnia [24]. The first study emplying techniques of non-linear analysis on human EEG was published in 1985 and dealt with sleep recordings. [2] This early success sparked

[3] 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.

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 [37].

Chapter 2

Convolutional Neural Networks

2.1 Mathematical background

TODO: Do we really need this section???

Definition 7. *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 8. *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 9 ([28])**.** *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,

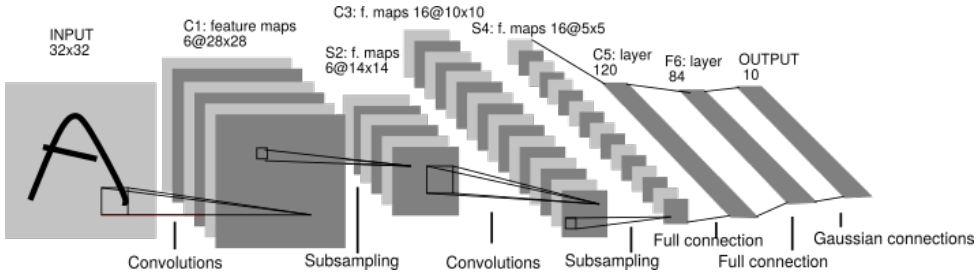


Figure 2.1: LeNet-5 architecture [18].

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 [9], 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 [17].

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*: [18]

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 [14].

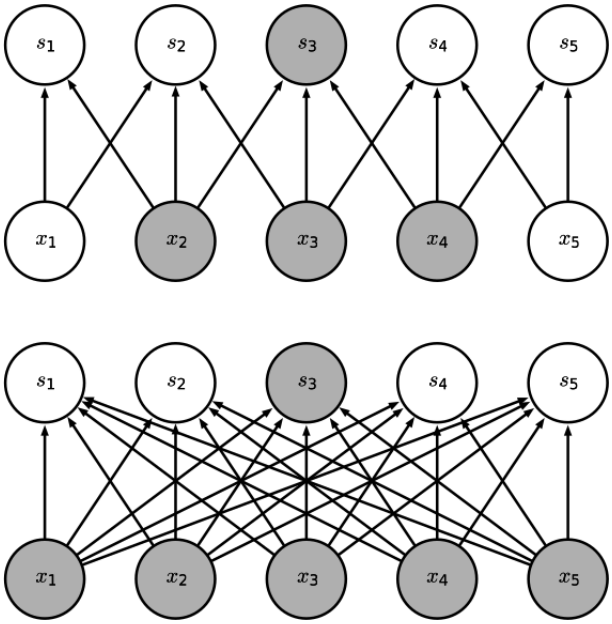


Figure 2.2: Receptive field. [10]

the achieved runtime is $O(k \times n)$, where k is usally in practice several orders of magintude smaller than m. [10]

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. [10]

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

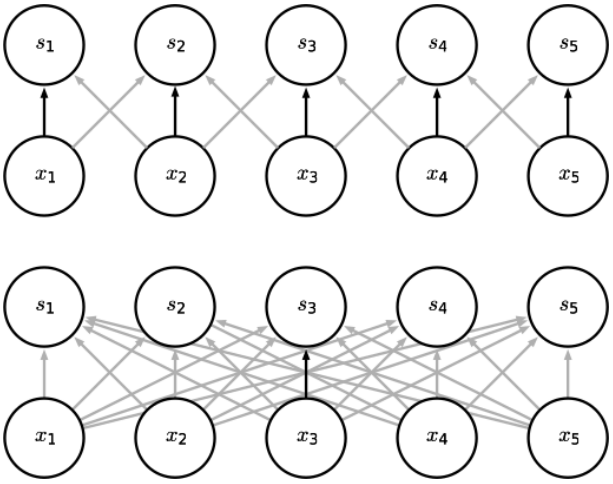


Figure 2.3: Shared weights. [10]

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." []

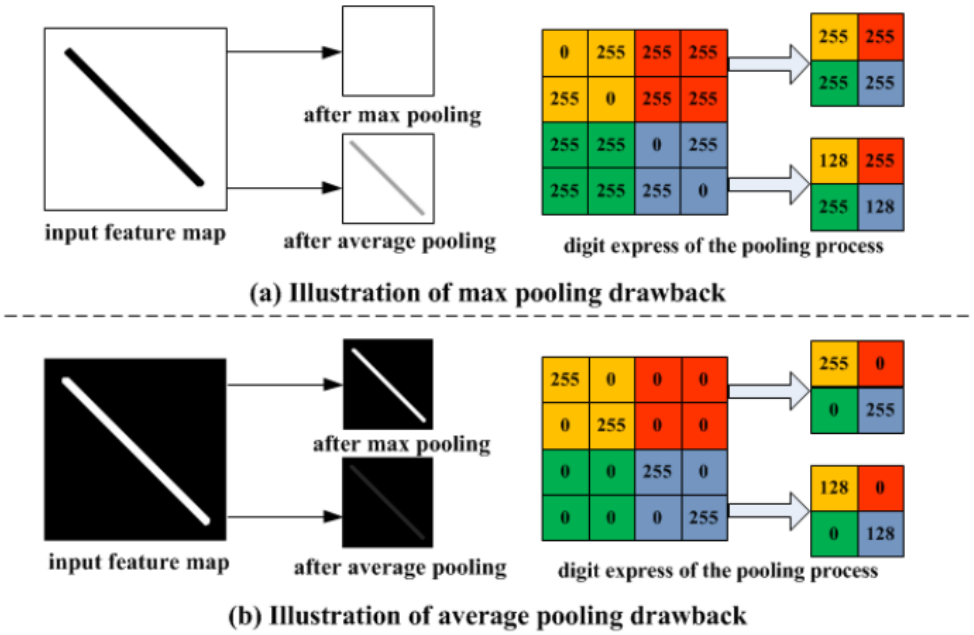


Figure 2.4: Examples of drawbacks of the pooling operation. Max pooling discards all except the maximum element, and valuable information may thus be lost. Average pooling considers all the values, and the information about their contrast is reduced. Moreover, extreme values may have undesired effects on the result. [42]

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 [33], 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 1.5, 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 1.2.2), 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 [31]. 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 1.2.2. 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 1.8, 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. [11] The value of the correlation integral $C(r)$ from the definition 1.9 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.

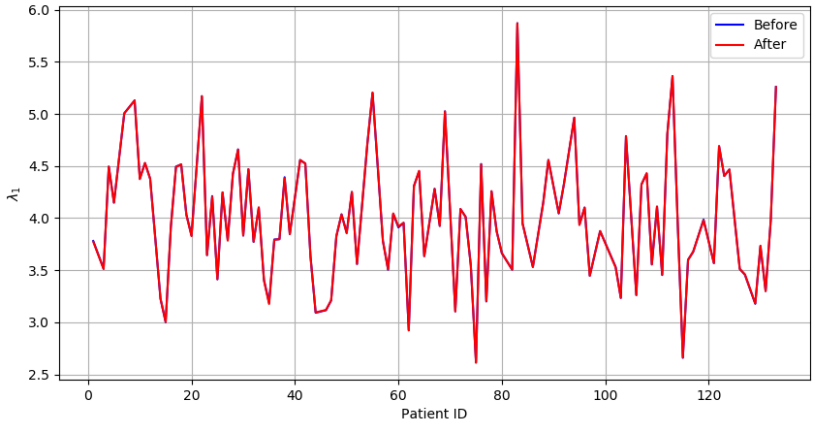


Figure 3.1

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.

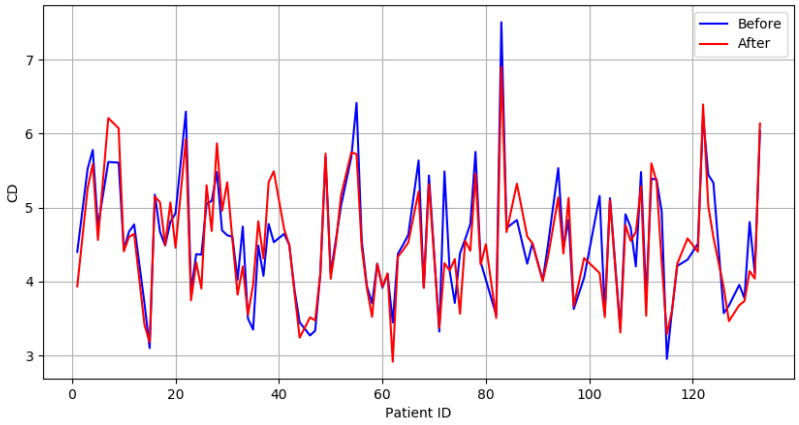


Figure 3.2

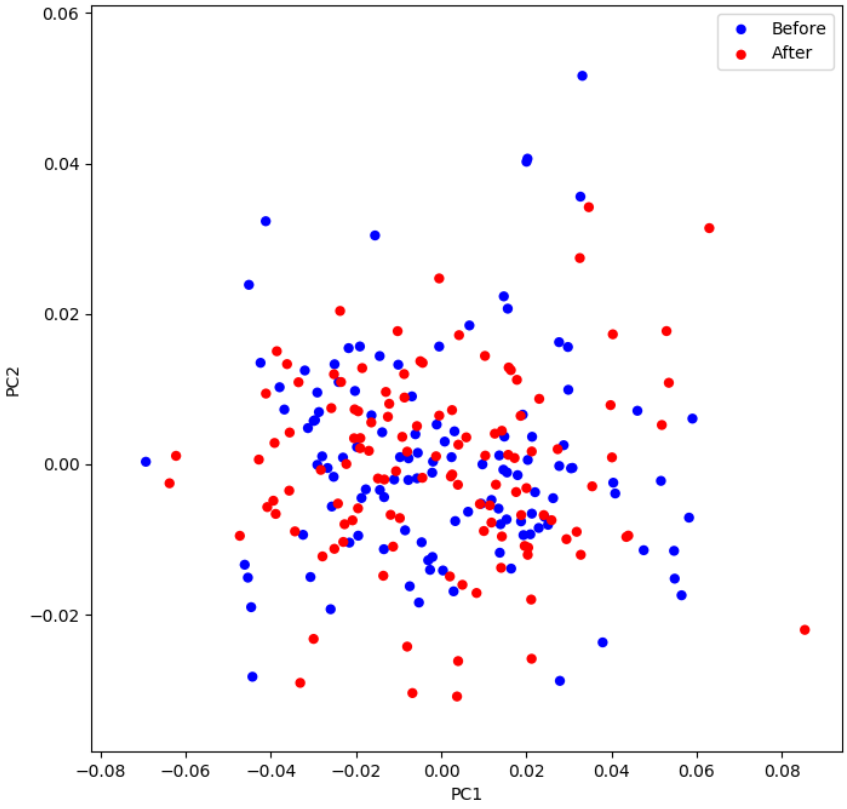


Figure 3.3

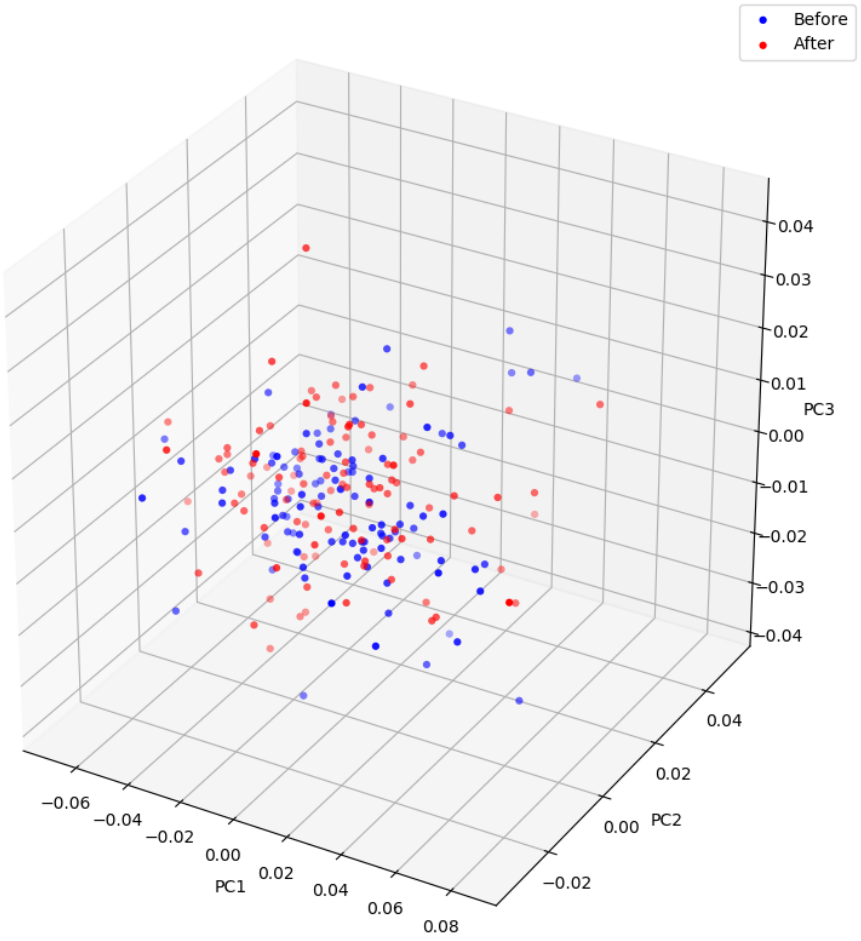

Before
After
PC1
PC2
PC3
−0.06
−0.04
−0.02
0.00
0.02
0.04
0.06
0.08
0.04
0.02
0.00
−0.02
0.04
0.03
0.02
0.01
0.00
−0.01
−0.02
−0.03
−0.04

Figure 3.4

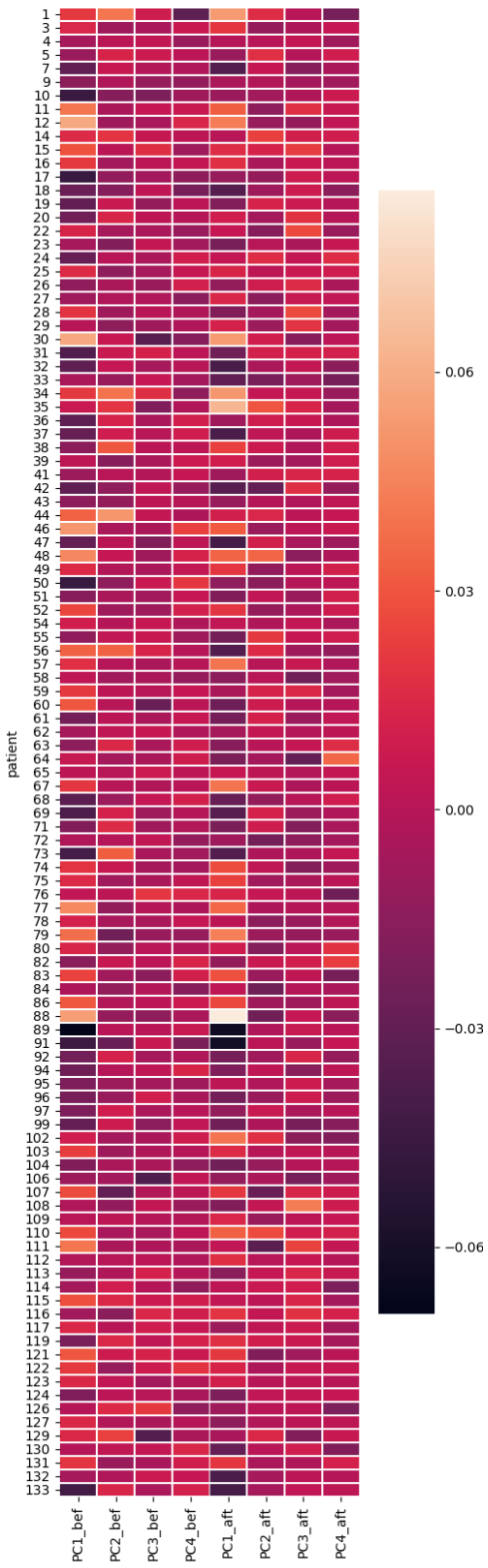

patient
PC1_bef
PC2_bef
PC3_bef
PC4_bef
PC1_aft
PC2_aft
PC3_aft
PC4_aft
0.06
0.03
0.00
−0.03
−0.06

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] 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.

[3] 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.

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

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

[6] 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.

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

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

[9] 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.

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

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

[12] 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.

[13] 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.

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

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

[16] 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.

[17] 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.

[18] 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.

[19] 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.

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

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

[22] 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.

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

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

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

[26] 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.

[27] 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.

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

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

[30] 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.

[31] 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.

[32] 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.

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

[34] 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.

[35] 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.

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

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

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

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

[40] 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.

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

[42] 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.