Chapter 1

Introduction

If we want to analyze experimental or simulated data we might encounter the following tasks:

- Characterization of the source of the signal and diagnosis
- Studying dependencies
- Prediction
- Modeling

These tasks are not independent. In fact, they are interrelated, but not identical. Modeling is the most general, but also most challenging task: If you have a good model for your data, you can use it to predict future data, you can use the model parameters to characterize the data and in particular you can use the parameters representing coupling constants between different observables to characterize dependencies between these observables.
The main focus of this lecture is the analysis of time series, i.e. the analysis of possibly vector valued measurements $\boldsymbol{x}_i$, that are characterized by an one dimensional index, which is usually the time, but could be also a spatial direction.
Classical examples from the statistics literature are the sun spot time series or the Canadian lynx population data. Other areas with time series data are geophysics, astrophysics, physiological time series such as ECG and EEG. In economy there is a whole special area called econometrics dealing with time series data. Moreover, also DNA and RNA sequences might be considered as time series. The latter are series of observables with discrete states, we will, however, in this lecture consider mainly continuous valued time series.

The traditional models in mathematical statistics were and still are linear models, i.e. models in which the next value $\boldsymbol{x}_{n+1}$ is a linear function of the past values plus a stochastic noise term, the residuals. The most general stationary model of this form can be written in the from of a autoregressive (AR) moving average (MA) model, short ARMA-model.

$$X_n = \sum_{k=1}^{p} a_k X_{n-k} + \epsilon_n + \sum_{l=1}^{q} b_l \epsilon_{n-l} \; . \tag{1.1}$$

The residuals ϵ_n are uncorrelated in time, i.e. $\langle \epsilon_k \epsilon_l \rangle = \delta_{kl}$. If they were not, than the model would be not the best linear model, because this dependency should then be also included in the model. The residuals might be, however, not independent. But modeling these dependencies would require nonlinear functions. This would lead to nonlinear stochastic models.
But there is a second approach to time series analysis mainly developed by physicists. It is based on the discovery the phenomenon of deterministic chaos, i.e. the fact that low-dimensional deterministic systems can also produce aperiodic seemingly random behavior, and not only constant, periodic or quasi-periodic motion as had been thought before. Thus the model class of nonlinear deterministic systems was added as an alternative:

$$x_n = f(x_{n-1}, \ldots, x_{n-p}) \tag{1.2}$$

In many cases the original hope that these phenomena can be described by such low dimensional deterministic systems had to be abandoned. Examples are the sleep EEG, the "climate attractor" or the stock market.
During this lecture we will look at some of these examples in more detail. If we assume that the degree of non-linearity and the degree of stochasticity could be quantified, the the linear stochastic and the nonlinear deterministic models are at the two axes of the diagram. The actual scientific challenge is to fill the large area in the middle — to develop methods for nonlinear stochastic systems. After a short introduction we will start with linear models and the related methods such as correlation functions and spectral analysis. At the end of this part we will deal with the Kalman filter, which is important also beyond the area of linear time series analysis.
After an intermezzo devoted to wavelet analysis we will proceed in the second part of the lecture to nonlinear deterministic systems and the corresponding methods, e.g. the estimate of fractal dimensions, dynamical entropies and Ljapunov exponents. Finally we will consider some first approaches to deal with nonlinear stochastic systems: Fitting Langevin equations or Fokker-Planck equations, respectively, from data.

1.1 Simple Characterizations

The starting point is a generally vector valued time series $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ representing k observables. In the following we will at first restrict ourselves to the case of a scalar time series, i.e. k=1. In order to proceed we have to assume stationarity, i.e. that the data were generated by a process/system, which remained constant during the time of observation. If cannot assume that then we have either to shorten the observation time or to extend our model in order to include also the slow temporal change of the system. Mathematically we can distinguish between weak an strong stationarity. Weak stationarity means that the mean and the variance of the process do not change with time. Strong stationarity means that all probability distributions characterizing the process are time independent. To describe this in a more formal way we have to introduce the concept of a random variable:

1.1.1 Random variable

At first we need a Probability space $(\Omega, \mathcal{A}, P)$ containing of a

Set of possible events Ω: Set of outcomes of an random experiment — in the case of a coin toss $\Omega = (heads, tails)$. Elements denoted by $\omega \in \Omega$.

σ-algebra of subsets $\mathcal{A}$: Set of subsets of Ω.

Probability measure P: Each set of events $A \subseteq \mathcal{A}$ has a probability $0 \leq P(A) \leq 1$. $P(\Omega) = 1$.

A **random variable** X is then a measurable function $X : (\Omega, \mathcal{A}) \to S$ to a measurable space S (frequently taken to be the real numbers with the standard measure). The probability measure $PX^{-1} : S \to \mathbb{R}$ associated to the random variable is defined by $PX^{-1}(s) = P(X^{-1}(s))$. A random variable has either an associated probability distribution (discrete random variable) or probability density function (continuous random variable).
This was the mathematical definition. For physicists one could simply say that a random variable is an observable equipped with a probability for each of its possible outcomes. In the following we will denote random variables by capital letters and there values by lower case letters. A random variable X is said to be *discrete* if the set $\{X(\omega) : \omega \in \Omega\}$ (i.e. the range of X) is finite or countable.

Alphabet: Set $\mathcal{X}$ of values of the random variable X.

Probability: $p(x) = P(X = x), x \in \mathcal{X}$.

Normalization:

$$\sum_{x \in \mathcal{X}} p(x) = 1$$

Expectation value of X:

$$E_P[X] = \sum_{x \in \mathcal{X}} x p(x)$$

If the states of our observable are continuous we have a continuous random variable and we can consider the cumulative distribution function:

Cumulative distribution

$$F(x) = P_{\leq}(x) = P(X \leq x) = \int_{-\infty}^{x} f(y) dy$$

A distribution has a density function if and only if its cumulative distribution function F(x) is absolutely continuous. In this case: F is almost everywhere differentiable, and its derivative can be used as probability density:

$$f(x) = \frac{dF}{dx}$$

Probability density $f(x)$: The density itself is not a probability (it can be > 1), it is related to a probability by

$$P(a \leq x \leq b) = \int_a^b f(x) dx \ .$$

Normalization

$$\int_{x_{min}}^{x_{max}} f(x) dx = 1 \ .$$

Expectation value, mean:

$$E[X] = \mu = \mu_1 = \int_{-\infty}^{\infty} x f(x) dx$$

Moments:

$$E[X^m] = \mu_m = \int_{-\infty}^{\infty} x^m f(x) dx$$

Median $x_{1/2}$

$$F(x_{1/2}) = \frac{1}{2}$$

Variance, standard deviation: Variance:

$$(E[X - E[X]])^2 = E[X^2] - (E[X])^2 = \sigma^2(X) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx$$

σ is called the standard deviation.

Covariance: For two random variables, the covariance is defined as

$$\mathrm{Cov}(X,Y) = E[(X - E[X]) \cdot (Y - E[Y])] = E[XY] - E[X]E[Y] \; .$$

The correlation coefficient is the normalized covariance

$$\rho_{X,Y} = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Cov}(X,X)\mathrm{Cov}(Y,Y)}}$$

1.1.2 Stochastic process

If we have measured a time series and describe any single measurement by a random variable $X_t \quad t \in T$, then the family of all random variables $X = (X_t)_{t \in T}$ is called a **stochastic process**. The distributions $F(X_{t_1}, \ldots, X_{t_m})$ are called the finite dimensional marginal distributions of the process X. If all finite dimensional marginal distributions are invariant with respect to a shift in time, i.e.

$$F(X_{t_1}, \ldots, X_{t_m}) = F(X_{t_1+\tau}, \ldots, X_{t_m+\tau})$$

the process is called **stationary**. This condition, however, cannot be tested in most cases. Therefore there is the weaker condition of weak stationarity, which is also related to linear systems. To define it we need the notion of the auto-covariance or autocorrelation function, respectively. The auto-covariance function is the covariance between X_t at different times t_1 and t_2: $\mathrm{Cov}[X_{t_1}, X_{t_2}]$. The autocorrelation function is the normalized auto-covariance

$$\rho(t_1, t_2) = \frac{\mathrm{Cov}[X_{t_1}, X_{t_2}]}{\sqrt{\mathrm{Cov}[X_{t_1}, X_{t_1}]\mathrm{Cov}[X_{t_2}, X_{t_2}]}}$$

i.e. the correlation coefficient between the values of X at different times.
If the mean of X_t does not depend on time and the auto-covariance does only depend on the time lag between the two arguments the process is called **weakly stationary**.

1.1.3 Independent random variables

Given a time series $\{x_1, x_2, \ldots, x_N\}$, the simplest model for it is to assume that the values of X at different points in time are independent, i.e. $p(x_i, x_j) = p(x_i)p(x_j)$ with the same distribution or density function $p(\cdot)$. The only thing we can know and the only thing we have to know for an optimal prediction is this distribution or density function $p(\cdot)$. For continuous random variables the most common density functions are

Gaussian distribution:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} ,$$

which is called normal distribution for $\mu = 0$ and $\sigma = 1$. This distribution is ubiquitous, because the sum of random variables with finite mean and variance is Gaussian distributed (central limit theorem).

Exponential distribution:

$$f(x) = \lambda e^{-\lambda x} \qquad F(x) = 1 - e^{-\lambda x}$$

It describes the inter-event interval distribution of a Poisson process, i.e. events occurring randomly with the rate λ.

Log-normal distribution: How is a product X of positive random numbers asymptotically distributed? The logarithm of the product is the sum of the logarithms and therefore the logarithm of X is normal distributed, the product itself is log-normal distributed:

$$g(\ln x) = \frac{1}{\sqrt{2\pi\sigma^2}} \left(-\frac{(\ln x - \mu)^2}{2\sigma^2} \right) \tag{1.3}$$

$$f(x) = g(\ln x)\frac{d\ln x}{dx} = \frac{1}{x\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2} \right) \tag{1.4}$$

The log-normal distribution is not a power law, but it can look like a power law in the log-log plot

$$\ln p(x) = -\ln x - \frac{(\ln x - \mu)^2}{2\sigma^2} = -\frac{(\ln x)^2}{2\sigma^2} + \left(\frac{\mu}{\sigma^2} - 1\right)\ln x - \frac{\mu^2}{2\sigma^2} \tag{1.5}$$

All these distributions depend on parameters. Then a description of a sample of data by such a distribution would be a parametric model and modeling would then mean to estimate these parameters from the data.
Let us consider the case of the Gaussian distribution: If we know (or assume) that the data were drawn from a Gaussian distribution, we have to estimate two parameters, the mean and the variance of the data.

1.1.4 Mean

The estimator of the mean is well known - the sample mean is estimated by

$$\hat{\mu} = \frac{1}{N}\sum_{n=1}^{N} x_n \; . \tag{1.6}$$

It is *unbiased* and consistent. What does it mean?
Let $\hat{f}_n = f(x_1, \ldots, x_n)$ be the estimate of the parameter λ for a given sample $\{x_1, \ldots, x_n\}$. f is called *unbiased* (erwartungstreu oder unverzerrt), if

$$E[f(x_1, \ldots, x_n)] = \lambda \tag{1.7}$$

for any n, i.e. if there is no systematic error.
A consistent estimator is an estimator that converges in probability to the quantity being estimated as the sample size grows without bound. An estimator $\hat{f}_n$ (where n is the sample size) is a consistent estimator for parameter λ if and only if, for all $\epsilon > 0$, no matter how small, we have

$$\lim_{n\to\infty} P\{|\hat{f}_n - \lambda| < \epsilon\} = 1$$

In our case this is equivalent to a asymptotically vanishing variance, i.e.

$$\lim_{n\to\infty} \sigma\left(f(x_1, \ldots, x_n)\right) = 0 \tag{1.8}$$

with $\sigma^2(f) := E((f - E(f))^2)$. How can we see that the estimator of the mean (1.6) is unbiased and consistent?
Unbiased:

$$\begin{aligned} E\left(\hat{\mu}\right) &= E\left(\frac{1}{n}\sum_{n=1}^{n} x_i\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} E\left(x_i\right) \\ &= \mu \end{aligned}$$

Consistent:

$$\begin{aligned} \sigma^2(\hat{\mu}) &= E\left(\hat{\mu} - E(\hat{\mu})\right)^2 \\ &= E\left(\frac{1}{n}\sum_{n=1}^{n}(x_i - \mu)\right)^2 \\ &= \frac{1}{n}\sigma^2(x) \; . \end{aligned}$$

If we consider the mean square error (MSE) of our estimator

$$MSE(f) = E[(f - \lambda)^2] \ ,$$

it can be decomposed into the variance of the estimator and a contribution of the bias

$$MSE(f) = E[(f - E[f])^2] + (E[f] - \lambda)^2 \ . \tag{1.9}$$

1.1.5 Variance

The variance of a sample could be estimated by

$$\hat{\sigma'}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2 \ .$$

Now, what about the bias of this estimator?

$$\begin{aligned} E(\hat{\sigma'}^2) &= \frac{1}{n}E\left(\sum_{i=1}^{n}(x_i - \hat{\mu})^2\right) \\ &= \\ &= \frac{n-1}{n}\sigma^2(x) \end{aligned}$$

Thus, this estimator is biased. An unbiased estimator of the variance is therefore

$$s_n^2 = \hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu})^2 \ .$$

1.2 Hypothesis testing

If we have only the data, however, we can only calculate a value of the parameter, but we cannot calculate the bias and the variance of the estimator. How reliable is our estimate? There are several possibilities to deal with this situation: A possibility often encountered is the use of confidence intervals. If we have an estimates $\hat{x}$, lying in the confidence interval $[\hat{x} - \Delta x, \hat{x} + \Delta x]$ with a confidence level 0.95, this means that if we could repeat the experiment infinitely often in 95% of the cases the true value would lie in the interval. It does NOT say that the true value is in the interval with probability 0.95.
A second possibility is to estimate the likelihood of a certain observation,

i.e. how likely was the observation of the given data under the assumption that the observed parameter is the true one. This is not so informative for a single estimation, but it is useful to compare different models for the same data (testing two specific hypothesis against each other). Moreover, it is used to derive estimators for model parameters, which are then called maximum likelihood estimators. We will come back to that.
A directly related question is the problem of hypothesis testing. Usually we estimate these parameters in order to test some hypothesis. One example which we already encountered is stationarity. We could ask, whether the mean and the variance of our data are constant in time, i.e. whether our data are (weakly) stationary. Thus we can estimate the mean and the variances for different subsets of our data and we have then to decide whether they agree with the assumption of stationarity or not. That is, we have to test against the hypothesis of stationarity, which is called the null hypothesis in this case. Another simple example is the following: Let us assume we have two samples of data recorded under different conditions and we want to know, if this condition influences our observable. Thus our hypothesis would be that the two distributions are different. The simplest thing one can ask then, is, whether the mean of the two samples is different or not. This is done in the following way: First we need a so called **test statistic** T, which is a function of the measured sample. First a null hypothesis is formulated - this is the negative result we want to test against. In our case this would be that the condition has no influence and the two means are equal and therefore the expectation value of our test statistic is zero. Then we characterize our estimate of the test statistic (the difference of the two means) by the probability that this difference (or a larger one) would have been produced simply by chance supposed the null hypothesis is true, i.e. the mean values of the underlying distributions are equal. This probability is given by

$$p = P(abs(T) \geq \hat{T} | \mu_1 = \mu_2) \ .$$

This probability is often called the "p-value". The difference between out two sample means is significant if its p-value is smaller than some threshold - 0.05 or 0.01 are typical significance thresholds. This p-value measures the probability of an error of first kind or false positive and the corresponding threshold is often denoted by α in a test setting and called the size of the test. There is, however, the second possibility that despite that the null hypothesis is false, it is not rejected by the test. This is called an error of second kind or false positive. The corresponding probability is usually denoted by β. To specify β the alternative hypothesis have to be known,

i.e. we have to make assumptions about what is truly the case instead of the null hypothesis. $1-\beta$ is then also called the power of the test against this alternative hypothesis. If only the null hypothesis is specified this error is not determined.
So, in general to perform a test we need a test statistic T and we need its distribution under the assumption that the null hypothesis is valid. Among all the sets of possible values, we must choose one that we think represents the most extreme evidence against the hypothesis. That is called the critical region of the test statistic. The probability of the test statistic falling in the critical region when the null hypothesis is correct, is the α value (or size) of the test. The test has to be designed in such a way that its power against the possible alternatives is maximized.
Let us assume that we know that our test statistic is normally distributed. It is then called a z-statistic and the corresponding test z-test.

$$z = \frac{\hat{\mu} - \mu_0}{\sigma/\sqrt{n}} \ .$$

If the possible alternatives are only distributions with positive means, we can define the critical region as $x \geq x_\alpha$ with

$$F(x_\alpha) = 1 - \alpha \ ,$$

and asking whether z is larger than x_α would be a one-sided test with $x_\alpha \approx 1.6449$. If we want to perform a two-sided test, we have to require that $-x_{\alpha/2} < z < x_{\alpha/2}$ with $x_{\alpha/2} \approx 1.96$ (use *norminv* in MATLAB).

1.2.1 The χ^2 distribution

An important distribution for testing hypothesis of Gaussian distributed random variables is the χ^2 distribution. Let us assume we have n samples drawn from the same Gaussian distribution with mean μ and variance σ^2. The sum of the squares of the samples is then distributed according to the so called χ^2 distribution:

$$\chi^2 = x_1^2 + x_2^2 + \ldots + x_n^2$$

$$F(\chi^2) = \frac{1}{\Gamma(\lambda)2^\lambda} \int_0^{\chi^2} u^{\lambda-1} e^{-\frac{1}{2}u} du$$

with $\lambda = \frac{1}{2}n$ and n called the number of degrees of freedom.
The importance of this distribution comes from the fact that it describes the distribution of the normalized estimator of the variance of a sample $\frac{(n-1)s_n^2}{\sigma^2}$ is χ^2 distributed with $n-1$ degrees of freedom.

1.2.2 t-Test

All tests with a test statistic distributed according to students t-distribution are called t-tests. The test statistic in the simplest case for testing the sample mean against a given value μ_0 is the t-statistic

$$t = \frac{\hat{\mu} - \mu_0}{s_n/\sqrt{n}}$$

with $df = n - 1$ degrees of freedom and with the density function

$$F(t) = \frac{\Gamma(\frac{1}{2}(df+1))}{\Gamma(\frac{1}{2}df)\sqrt{df}} \int_{-\infty}^{t} \left(1 + \frac{t^2}{df}\right)^{-\frac{1}{2}(df+1)}$$

Most of the analytic results for parametric tests in statistics start with the assumption of normal distributed measurements. If this is not the case one can use non-parametric tests based on rank order statistics. Or one uses Monte Carlo procedures were one generates samples from distribution corresponding to the null hypothesis.

Chapter 2

Linear models

2.1 Overview

Linear process: A process $\{X_n\}$ is a linear process if it has the representation

$$X_n = \sum_{j=0}^{\infty} b_j \epsilon_{n-j}$$

for all n, where $\epsilon_n \propto N(0, \sigma^2)$ (Gaussian distributed with zero mean and variance σ^2 and $\sum_{j=0}^{\infty} b_j^2 < \infty$. Thus a time series of a linear process could be generated by applying a linear filter to Gaussian noise.
Linear processes are modeled using the following model classes:

Moving average(MA-)model of order q:

$$X_n = \epsilon_n + \sum_{l=1}^{q} b_l \epsilon_{n-l}$$

By setting $b_0 = 1$ this can be written as

$$X_n = \sum_{l=0}^{q} b_l \epsilon_{n-l} \ .$$

Using the shift operator $Bx_n = x_{n-1}$ we can write

$$X_n = (1 + \sum_{l=1}^{q} b_l B^l)\epsilon_n \ . \tag{2.1}$$

Autoregressive(AR-)model of order p:

$$X_n = \sum_{k=1}^{p} a_k X_{n-k} + \epsilon_n$$

or

$$(1 - \sum_{k=1}^{p} a_k B^k) X_n = \epsilon_n$$

ARMA-models of order (p, q):

$$x_n = \sum_{k=1}^{p} a_k X_{n-k} + \epsilon_n + \sum_{l=1}^{p} q_l \epsilon_{n-l}$$

with

$$(1 - \sum_{k=1}^{p} a_k B^k) X_n = (1 + \sum_{l=1}^{q} b_l B^l) \epsilon_n \tag{2.2}$$

State space models: They re equivalent to the the ARMA model class and are written as

$$\begin{aligned} \boldsymbol{x}_n &= \boldsymbol{A}\boldsymbol{x}_{n-1} + K\boldsymbol{\epsilon}_n \\ \boldsymbol{y}_n &= \boldsymbol{C}\boldsymbol{x}_n + \boldsymbol{\epsilon}_n \end{aligned}$$

in the so called innovation representation.

Basic properties of linear models:

- If the inputs ϵ are Gaussian iid noise then the x values are Gaussian distributed too.
- Any stationary process can be represented by a linear model with infinite model order and uncorrelated residuals ϵ_n, which, however, are only independent, if the process is real a linear one.

2.1.1 The autocorrelation function

We introduced already the autocorrelation

$$\rho(t_1, t_2) = \frac{\mathrm{Cov}[X_{t_1}, X_{t_2}]}{\sqrt{\mathrm{Cov}[X_{t_1}, X_{t_1}]\mathrm{Cov}[X_{t_2}, X_{t_2}]}} \ .$$

Under the assumption of stationarity this is equal to

$$\rho(\tau) = \frac{\mathrm{Cov}[X_t, X_{t+\tau}]}{\mathrm{Cov}[X_t, X_t]} = \frac{C(\tau)}{\sigma^2} \ .$$

with $C(\tau)$ denoting the autocovariance function.
Because the covariance is symmetric we have $C(\tau) = C(-\tau)$ and $C(0) = 1$.
The autocorrelation function is the normalized autocovariance function

$$\rho(\tau) = \frac{C(\tau)}{C(0)} \ .$$

The estimator of the autocorrelation function estimated from a time series is called the sample autocorrelation function $\hat{\rho}(\tau)$. In practice there are used more than one estimator for the autocovariance function:

Unbiased estimate:

$$\hat{C}(\tau) = \frac{1}{N-\tau} \sum_{k=1}^{N-\tau} (x_k - \hat{\mu})(x_{k+\tau} - \hat{\mu})$$

The problem is that for large τ only a very few samples enter.

Biased estimate:

$$\hat{C}(\tau) = \frac{1}{N} \sum_{k=1}^{N-\tau} (x_k - -\hat{\mu})(x_{k+\tau} - \hat{\mu})$$

How can we test that some data are uncorrelated, as e.g. the residuals $\{\epsilon_n\}$ of our linear models should be? There are a lot of proposals in the literature, however they assume that the data are not only uncorrelated but iid. i.e. independent.
The most simple one is to use the fact that for large N the sample autocorrelations of an iid sequence with finite variance are approximately iid, normal distributed with a variance $1/N$ ($\tau \ll N$), thus 95% of the sample autocorrelations should fall into the interval $\pm 1.96/\sqrt{N}$. If there more than 5% of the values fall outside this bound, we should think about rejecting the hypothesis.
Another possibility is the Portmanteau test with the test statistic

$$Q = N \sum_{j=1}^{m} \hat{\rho}^2(j)$$

which is distributed according to a χ^2-distribution with m degrees of freedom.

2.1.2 Autocorrelation function of MA-models

In the case of the MA-models the autocorrelation gives us the order of the model, because

$$\begin{aligned} C(\tau) &= \mathrm{Cov}[X_n, X_{n+\tau}] \\ &= \begin{cases} 0 & \text{if } \tau > q \\ \sigma^2 \sum_{k=0}^{q-\tau} b_k b_{k+\tau} & \text{if } \tau \leq q \end{cases} \end{aligned}$$

Thus for any process with a non-vanishing correlation function for larger τ the moving average model might be a bad choice for the model class.

2.2 Autoregressive models

2.2.1 AR(1)-model

Let us first look at an example. The simplest autoregressive model is the AR(1)-model

$$x_n = a x_{n-1} + \epsilon_n \tag{2.3}$$

containing only one parameter a. The deterministic part describes an exponentially damped motion with the fixed point $x = 0$. The invariant distribution results from this damping toward the origin and the simultaneous excitation by the noise. If the noise ϵ is Gau"ssian also the state variable x is Gaussian distributed and can be characterized by its mean and variance. Because ϵ_i has zero mean, also $\mu = E(x) = 0$. The variance can be estimated easily from (2.3) by squaring both sides and building the expectation taking into account that ϵ_n and x_{n-1} are uncorrelated:

$$E(x^2) = a^2 E(x^2) + E(\epsilon^2)$$

leads to

$$\sigma^2(x) = \frac{\sigma_\epsilon^2}{1 - a^2} \ .$$

In particular, we see that the variance will diverge if a is approaching 1, i.e. if the deterministic dynamics becomes unstable.
In the last chapter we considered iid samples, i.e. to subsequently measured samples should be statistically independent. Now we have temporal correlations. Multiplying both sides of (2.3) with x_{n-1} and taking the expectation value we get

$$E(x_n x_{n-1}) = a E(x_{n-1}^2)$$

or

$$a = \frac{E(x_n x_{n-1})}{E(x_n^2)}$$

i.e. the model parameter a is given by the value of the normalized autocorrelation function for one time step delay. Thus it seems obvious to estimate the model parameter using estimates of the autocorrelation function. This can be generalized for autoregressive models of arbitrary order and is known as the Yule-Walker algorithm.
How would the AR(1)-process looks like if we would represent it by a MA-model? By recursively inserting (2.3) we get

$$\begin{aligned} x_n &= a^2 x_{n-2} + \epsilon_n + a\epsilon_{n-1} \\ &= a^3 x_{n-3} + \epsilon_n + a\epsilon_{n-1} + a^2\epsilon_{n-3} \\ &= \dots \\ &= \sum_{k=0}^{\infty} a^k \epsilon_{n-k} \end{aligned}$$

i.e. $b_k = a^k$.

2.2.2 Stability of AR-models

Let us now consider the general AR-model

$$X_n = \sum_{k=1}^{p} a_k X_{n-k} + \epsilon \; .$$

In order to study its stability we rewrite it in matrix form

$$\boldsymbol{X}_n = \boldsymbol{A}\boldsymbol{X}_{n-1} + \boldsymbol{\epsilon}$$

with $\boldsymbol{X}_{n-1} = (X_{n-1}, \dots, X_{n-p})^T$, $\boldsymbol{\epsilon}_n = (\epsilon_n, 0, \dots, 0)^T$

$$A = \begin{pmatrix} a_1 & a_2 & \dots & a_{p-1} & a_p \\ 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 & 0 \end{pmatrix}$$

The model is stable, if the absolute vale of the eigenvalues of $\boldsymbol{A}$ is smaller than 1. The eigenvalues are given by the zeros of the characteristic polynomial

$$z^p - z^{p-1}a_1 - \dots - z a_{p-1} - a_p = 0 \; .$$

Complex zeros correspond to damped oscillatory behavior, real zeros to pure relaxatory behavior as in the AR(1)-model. If the zero are

$$z_k = r_k e^{-i\phi_k} \qquad f_k = \frac{\phi_k}{2\pi} \cdot f_s \qquad \gamma = -f_s \ln r$$

the model is stable if $r_k < 1$ for all k.

2.2.3 Estimating the AR-parameters

Least square estimation

(*ar-model* in TISEAN, *lpc* in MATLAB)
The most common way to test the quality of a model is to use it as a predictor and to calculate the prediction error by the mean square error, i.e.

$$MSE = \frac{1}{N-p} \sum_{n=p+1}^{N} (x_n - \hat{x}_n)^2 \tag{2.4}$$

with estimating $\hat{x}_n$ by the linear predictor

$$\hat{x}_n = \sum_{k=p+1}^{p} a_k x_{n-k} \; . \tag{2.5}$$

Thus an obvious way to estimate the parameter a_k from data would be to minimize the prediction error

$$\begin{aligned} 0 &\stackrel{!}{=} \frac{\partial 1/N \sum_{k=1}^{N} (x_n - \hat{x}_n)^2}{\partial a_k} \\ 0 &= \frac{1}{N-p} \sum_{n=p+1}^{N} (x_n - \sum_{k'=1}^{p} a'_k x_{n-k'}) x_{n-k} \end{aligned}$$

leading to a system of linear equations:

$$\frac{1}{N-p} \sum_{n=p+1}^{N} x_n x_{n-k} = \sum_{k'=1}^{p} a'_k \frac{1}{N-p} \sum_{n=p+1}^{N} x_{n-k'} x_{n-k} \tag{2.6}$$

which can be solved using standard techniques. The resulting estimator for the a_k is also known as least squares estimator. Recognizing that he equations contains some kind of sample autocorrelation function it can be written as

$$\hat{C}'(k) = \sum_{k'=1}^{p} a'_k \hat{C}'(k-k') \tag{2.7}$$

with asymptotically for large N $\hat{C}'(k-k') = \hat{C}(k-k')$.

Yule-Walker algorithm

(*aryule* in MATLAB)
Another possibility to derive an estimator starts directly from the model

$$x_n = \sum_{k=1}^{p} a_k x_{n-k} + \epsilon_n \ .$$

Multiplying both sides with $x_{n-k'}$ and calculating the expectation value leads to

$$E(x_n x_{n-k'}) = \sum_{k=1}^{p} a_k x_{n-k} x_{n-k'} \ .$$

by taking into account that $E(x_k \epsilon_m) = 0$ for $k < m$. Moreover, because $E(x_n) = 0$, we get

$$C(k') = \sum_{k=1}^{p} a_k C(k - k') \ . \tag{2.8}$$

These equations for the autocorrelation function are called Yule-Walker equations. If we replace the autocorrelation function by its sample estimate and solve the equations for the a_k we get the Yule-Walker estimate for the parameters. This estimates is as good as our sample estimate is for the autocorrelation function.
Comparing (Yule-Walker-Eq) with (LS-Eq-2) we recognize that they differ only in estimating the correlation function of the right hand side and that the coincide asymptotically for $N \to \infty$.
Not that (2.8) implies that the autocorrelation function contains all information about the model parameters. Or in other words: A linear process is fully specified by its autocovariance function. We will use this property later for constructing tests for non-linearity of time series.

Burg algorithm

(*arburg* in MATLAB)
A third algorithm for parameter estimation is the Burg algorithm. Here not only the the forward prediction error is minimized, but also the backward prediction error. This is based on the fact that linear processes are invariant with respect to time reversal. The probabilities $p(x_n|x_{n-1}, \ldots, x_{n-k}) = p(x_{n-k}|x_{n-k+1}, \ldots, x_1)$. The main advantage of this algorithm is that it always provides stable models.

Maximum Likelihood Estimation

While minimizing the the mean square error is a reasonable pragmatic strategy there is a more systematic approach to the problem of an optimal parameter estimate. For instance, we can as, how likely it is, that given certain values of the parameters, the data were produced by the given model, i.e. p(data—parameter). We can ask for the values of the parameter, for which the observed data were most likely. An estimator, which maximizes this likelihood is called maximum likelihood estimator. How does it looks like for the autoregressive model? We start with the assumption of independent Gaussian distributed residuals ϵ_n. The probability of the sequence of residuals is given by

$$\begin{aligned} L &= \prod_{i=p+1}^{N} p(\epsilon_i) \\ L &= \prod_{i=p+1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x_i - \sum_{k=1}^{p} a_k x_{i-k})^2\right) \\ -2\log L &= (N-p)\log(2\pi\sigma^2) + \frac{1}{\sigma^2}\sum_{i=p+1}^{N}(x_i - \sum_{k=1}^{p} a_k x_{i-k})^2 \ . \end{aligned}$$

Thus, maximizing the likelihood or the log-likelihood corresponds to minimizing the mean square errors, i.e. to least squares estimation in this case. But even the maximum likelihood approach could be criticized because it assumes that we observed a typical data set and so one has for instance problems with outliers. Because by calling a data point an outlier we say it is very unlikely that our system under study produces such a data point. How can we incorporate this kind of knowledge in our analysis? This is done in **Bayesian** statistics. Here we do not maximize the likelihood of the data, but we ask, how likely is a given parameter given the data. Because we only know the likelihood $p(data|parameter)$ we use Bayes' rule to estimate the probability of the parameter values given the data:

$$p(\text{parameter}|\text{data}) = \frac{p(\text{data}|\text{parameter})p(\text{parameter})}{\sum_{\text{parameter}} p(\text{data}|\text{parameter})p(\text{parameter})}$$

We can then use either the most probable parameter value or the conditioned expectation as an estimate. The main difference to the maximum likelihood estimate is the so called "prior" $p(\text{parameter})$, which contains our assumptions about reasonable models. In particular, the posterior probability $p(\text{parameter}|\text{data})$ cannot be non-zero for parameter values with zero

prior probability. The maximum posterior estimate is equal to the maximum likelihood estimate if we assume a constant prior.

2.2.4 Estimating the parameters of ARMA-models

While the parameter estimation in the case of the AR-models led to the problem of solving a system of linear equations, this is not the case anymore for ARMA and state space models. Therefore nonlinear, usually iterative procedures or approximations are necessary.

The Hannan-Rissanen algorithm

Here the parameter estimation is divided into two steps:

1. A high-order AR(m)-model is fitted to the data, with $m > \max(p, q)$. This model is used to estimate the noise terms

$$\epsilon_n = X_n - \sum_{k=1}^{m} \hat{a}_k X_{n-k} \; .$$

2. In a second step the parameters of the ARMA(p,q)-model are estimated by a least squares linear regression of X_n onto $(X_{n-1}, \ldots, X_{n-p}, \epsilon_{n-1}, \ldots, \epsilon_{n-q})$

2.2.5 Order selection

Before estimating the parameters of the model we have to specify the order p. Increasing the order p usually leads to smaller prediction errors. Does it mean that is also produces the better model? No, this is not the case.
From a statistical point of view and starting from the assumption of an underlying "true model" one has to note that at least the variance (and perhaps also the bias) of the estimator increases if I increase the model order for a fixed number of data and thus the probability that the true values are near the estimated ones decreases. We can, however, also adopt another point of view without referring to the "true model": Modeling a time series usually intends to build a model of the system, which generated the time series. Thus, we do not only want to describe the given time series, but the model should be a good model for any time series produced by this system, i.e. the model should generalize. In order to do so succesfully we have to distinguish between the regularities in the time series and the noise. Increasing the the model order increase the possibility that we do not fit the

regularities produced by the system, but only the noise. This is also called "overfitting". To avoid this we have different possibilities, depending on our prior knowledge about the system.

In sample and Out-of-sample error: If there are enough data available the data set can be splitted into a training data set and a test data set. The parameters are estimated on the training set leading to the in-sample prediction error. The the estimated model is used to predict the test data giving the out-of-sample prediction error.

Final prediction error: The FPE criterion was developed by Akaike Akaike (1969) by implementing the above idea for autoregressive processes, which led to an out-of-sample prediction error estimate

$$\mathrm{FPE}_p = \hat{\sigma}^2 \frac{n+p}{n-p}$$

with σ^2 being the mean square in-sample prediction error.

More general criteria based on estimations of the likelihood of the test data given the model estimated using the training data. There are the AIC (Akaike information criterion), its bias corrected version AICC or the BIC (Bayes information criterion). All these criteria have to be applied with caution, but they are often provided by software packages and can be used to give at least an orientation.

2.3 Spectral analysis

Performing spectral analysis represents the data as sum (or integral) of components at a single frequency. If we consider a time continuous signal $x(t)$ of infinite length we can define the Fourier transform

$$x(f) = \int_{-\infty}^{\infty} dt x(t) e^{-i2\pi ft} \qquad x(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} df x(f) e^{i2\pi ft}$$

The spectral power or power spectrum is the given by the absolute value of the fourier component at frequency f, i.e.

$$S(f) = |x(f)|^2$$

The Fourier transform of the convolution of two functions in time is the product of their Fourier transforms:

$$z(t) = \int_{\infty}^{+\infty} d\tau y(t-\tau) x(\tau) \quad \Rightarrow \quad z(f) = y(f)x(f) \; .$$

The inverse relationship is called modulation:

$$v(t) = x(t)y(t) \quad \Rightarrow \quad v(f) = \frac{1}{2\pi}\int_{\infty}^{+\infty} df' y(f-f')x(f)$$

The power spectrum is directly related to the autocorrelation function by the Wiener-Khinchin theorem:

$$C(t) = \int_{-\infty}^{\infty} d\tau x(t+\tau)x(\tau) \quad \Rightarrow \quad C(f) = S(f) = |x(f)|^2 \, .$$

The discrete Fourier transform of the time series sampled at discrete times can be written as

$$\hat{x}(f_k) \quad = \quad \sum_{n=0}^{N-1} x_n e^{-i2\pi f_k/f_s n} \tag{2.9}$$

which is the discrete Fourier transform for $f_k = f_s k/N$ and $k = 0, \ldots, N-1$. The inverse transform is then

$$x_n = \frac{1}{N}\sum_{n=0}^{N-1} x(f_k) e^{2\pi i k n/N} \, .$$

If one considers a given time series as a sample from a process which is continuous in time we can ask, under which conditions the time series represents the original process. This question is answered by the Nyquist-Shannon sampling theorem, saying that the sampling frequency f_s has to be twice as large as the highest frequency contribution. Half of the sampling frequency is also called Nyquist frequency $f_{Nyquist}$. This theorem is related to the problem of aliasing. Aliasing means that a high frequency component ($f > f_{Nyquist}$) of the original signal appears in the sampled signal as a low frequency component. A common example of temporal aliasing in film is the appearance of vehicle wheels traveling backwards, the so-called Wagon-wheel effect.
The second problem is that we have only a finite time series available (windowing). Both problems can be analyzed using the modulation property of the Fourier transform. Let us consider the following periodic function

$$s(t) = \sum_{n=-\infty}^{\infty} \delta(t - n\Delta) \qquad \Delta = 1/f_s$$

which can be represented in a FOUREIER series with the coefficients

$$c_n = \frac{1}{\Delta}\int_{\Delta} dt s(t) e^{-2\pi f_s i n t} = \frac{1}{\Delta}$$

$$s(t) = \sum_{-\infty}^{\infty} f_s e^{2\pi i n f_s t}$$

Applying the Fourier tranform we get

$$s(\omega) = \omega_s \sum_{n=-\infty}^{\infty} \delta(\omega - n\omega_s) \qquad \omega_s = 2\pi f_s$$

If we represent sampling the continuous function $x(t)$ at discrete times by multiplying with (2.3) we see that the resulting Fourier transform is a convolution of the original transform with the transform of (2.3). This results in a new transform

$$\tilde{x}(\omega) = f_s \sum_{n=-\infty}^{\infty} x(\omega + n\omega_s) \; .$$

The effect of finite time can be analyzed similarly by multiplying the signal with a window function. The rectangular window

$$w_R(t) = \left(\begin{array}{ll} 1 & \text{if} \qquad -\frac{\Delta}{2} \leq t < (N - \frac{1}{2})\Delta \\ 0 & \qquad\quad \text{otherwise} \end{array} \right)$$

has the Fourier transform

$$w_R(\omega) = N\Delta \frac{\sin \omega N\Delta/2}{\omega N\Delta/2} \exp -i\omega(1/2 - N)\Delta$$

with its main contribution at $\omega = 0$ but with a lot of side maxima which distort the original spectrum. Therefore one uses other windows, which taper smoothly to zero at both ends, such as the Bartlett, Welch, Hann or Hamming windows.

2.3.1 The periodogram

The periodogram of a time series $\{x_1, \ldots, x_N\}$ is the function

$$S_n(f_k) = \frac{1}{N} \left| \sum_{n=0}^{(} N - 1) x_n e^{-2\pi i f_k n} \right| \; .$$

Note that there is no consensus regarding the normalization. Thus one has to be check the normalization if one uses routines from program packages. In order to estimate the power spectrum of the underlying stochastic process

there is the problem that the periodogram is not a consistent estimator. In fact, the values of the estiamte are approximately distributed as exponential random numbers, i.e. their variance is equal to the mean. Increasing the number of data points increases the number of frequency values for which we estimate a value but the single estimates do not become better. There are two possibilities to overcome this problem:

1. Average over different frequency bins, which leads to spectral average estimators. This is for instance implemented in the TISEAN routine *spectrum*.

2. Welch' method: Split the data set into possibly overlapping segments and average the estimated periodograms. This is e.g. implemented in MATLAB's estimators of the power spectrum (*pwelch, spectrum.welch*).

2.3.2 Estimating the spectrum using ARMA models

A principal alternative to the periodogram is the estimation of the spectral density of a stochastic process fitting a linear model to the data and using the known spectral densitiy of this model as an estimate. Let us consider the time shift operator $BX_n = X_{n-1}$. It corresponds in Fourier space a Multiplikation with $z = e^{-2\pi i f_k}$. For an ARMA(p,q)-model written as

$$(1 - \sum_{k=1}^{p} a_k B^k) x_n = (1 + \sum_{l=1}^{q} b_l B^l) \epsilon_n$$

we get the spectral density

$$x(f_k) = \frac{\sigma^2 (1 + \sum_{l=1}^{q} b_l z^l)}{1 - \sum_{k=1}^{p} a_k z^k} \ .$$

The autoregressive part appears in the denominator, thus small values of it lead to high power at these frequencies. We discussed already the interpretation of the autoregressive part as a set of harmonic oscillators or linear relaxators, respectively. The frequencies of this oscillators correspond to the inverse zeros of the polynomial $(1 - \sum_{k=1}^{p} a_k z^k) = z^p((1/z)^p - \sum_{k=1}^{p} a_k (1/z)^k$. For the spectral density the polynomial is evaluated on the unit circle only, thus we see the nearer the poles are to the unit circle the higher and sharper is the maximum in the power spectrum. Let us consider the example of the

AR(2)-model.

$$\begin{aligned} X_n &= a_1 X_{n-1} + a_2 X_{n-2} + \epsilon_n \\ z^2 - a_1 z - a_2 &= (z - z_p)(z - z_p^*) \\ z_p &= r e^{i\phi} \qquad \phi = \omega\Delta \quad \Delta = 1/f_s \\ a_1 &= 2r\cos\phi \qquad a_2 = -r^2 \\ S(\omega) &= \frac{\sigma_\epsilon^2}{2\pi} \frac{1}{(1-r^2)^2 + 4r^2(\cos^2\phi + \cos^2\omega\Delta) - 4r(1+r^2)\cos\phi\cos\omega\Delta} \end{aligned}$$

with the maximum at

$$\cos(\omega_{max}\Delta) = \frac{1+r^2}{2r}\cos(\phi)$$

i.e. for $r < 1$ the maximum is not exactly at the position of the oscillator frequency.

Chapter 3

Nonlinear time series analysis

3.1 Deterministic dynamical systems

While statisticians, when trying to explain the real world, are starting from a "random world" by introducing correlations or dependencies, respectively, physicians often think about the world as a deterministic one[1] and stochasticity (noise) is introduced as an approximation of effects which are either too high—dimensional or fluctuate too fastly to take them explicitly into account. So our starting point is a deterministic dynamic system living in a state space X which should be at the moment finite dimensional. The dynamics is either defined for discrete times

$$\boldsymbol{x}_{n+1} = F(\boldsymbol{x}_n) \tag{3.1}$$

thus defining a map or for continuous times

$$\dot{\boldsymbol{x}}(t) = f(\boldsymbol{x}) \; . \tag{3.2}$$

by a system of coupled ordinary differential equations which defines a flow $\boldsymbol{x}(t+t_0) = \phi^t(\boldsymbol{x}(t_0))$. There are several possibilites to relate the two descriptions to each other. Very often one considers the stroboscopic map of (3.2) for a given time T with $\boldsymbol{x}_n = \phi^T(\boldsymbol{x}_{n-1})$, e.g. in the case of periodically driven systems, or the Poincare surface of section (Poincare map) - the section of the flow with a hyperplane transversal to the flow. Formally it is defined

[1] With the exception of quantum mechanics, but even there the evolution of the wave function is deterministic

in the neighbourhood of a periodic orbit, but often it can be extented to the whole phase space. A simple way to generate the hyperplane, is to set one coordinate of the dynamical system to a fixed value (TISEAN program: *poincare*). In the following we only consider maps F, be it generically maps, maps generated by sampling flow data with a fixed sampling interval or Poincare maps.

3.1.1 Characterization — Dynamical invariants

One of the objectives of time series analysis is the characterisation of the system which generated the time series in question. In the case of deterministic dynamic systems there are quantities available which are better suited for this task than simply taking the model parameters. Deterministic dynamical systems can be charaterised by quantities which are invariant with respect to coordinate transformations and therefore independent of the "channel" by which we observe the system. We will came back to that in 3.1.2.

Attractor dimension

The first invariant is the *attractor dimension*. There are several definitions of attractors of dynamical systems around. Intuitively an attractor is the set of points in the phase space which are visited by the system asymptotically if the transient is discarded. A little bit more mathematically one could say that an attractor is an invariant set, which is attracting - in contrast to a repellor or a saddle point. To be attracting the set A must be a subset of an open set U, its neighbourhood, with

$$\lim_{n\to\infty} \inf_{y\in A} ||F^n(x) - y|| \to 0 \qquad \forall x \in U \ .$$

Sometimes it is only required that A attracts a set of positive measure, which leads to the different concept of Milnor attractors.
With respect to the dimension we can distinguish between dimensions of a set or dimensions of a measure. The first simply considers all points of a set, the latter also takes into account how often this points are visited by the system.
Let us first consider the box-counting dimension, which is an example of the first, but can be considered also in the more general framework of the latter.

Box-counting dimension of a set A: There are several equivalent definitions of this dimension. One possibility is to partition the phase space of our system by hypercubes with a side length ϵ. Then we call $N_\epsilon(A)$ the number of cells, which are intersected by the attractor A. The box-counting dimension D_0 is then defined as

$$D_0 = \lim_{\epsilon \to 0} - \frac{\log N_\epsilon(A)}{\log \epsilon} \ .$$

This is a property of the set only. It is invariant with respect to smooth invertible transformations of the phace space.
Examples: Fixed point attractors have dimension zero, limit circles 1, quasiperiodic motion on a torus 2. Middle thirs cantor set: Repellor of

$$x_{n+1} = \begin{array}{ll} 3x & \text{if} \quad x \leq = 1/2 \\ 3 - 3x & \text{if} \quad x > 1/2; . \end{array}$$

is a Cantor set. With $\epsilon = 3^{-n}$ and $N_\epsilon = 2^n$ we get $D_0 = \log 2 / \log 3 = log_3 2$.
Before considering dimensions that take the measure into account, let us first discuss the entropy.

KS-entropy

While the dimension gives us information about the number of active degrees of freedom of the dynamical system, there is a second, complementary quantitiy, the metric or Kolmogrov-Sinai entropy which tells us about the randomness or irregularity of the dynamics. Basically it measures the uncertainty of the next observation given all the observations from the past.
To describe this we use the notion of an invariant measure. Remeber the probability space $(\Omega, \mathcal{B}, P)$ containing of a set of possible events Ω, a σ-algebra of subsets $\mathcal{B}$ (Set of subsets of Ω) and the probability measure P. Each set of events $A \subseteq \mathcal{B}$ has a probability $0 \leq P(A) \leq 1$, $P(\Omega) = 1$. Now our set of events is the phase space X of our dynamical system. We say a measure μ is invariant under a transformation $F : X \to X$, or F is a measure preserving transformation wrt to μ if

$$\mu(F^{-1}A) = \mu(A) \quad \forall \ A \in \mathcal{B} \ . \tag{3.3}$$

Let us consider some probability space $(X, \mathcal{B}, \mu)$ and a finite or countable index set I. A collection of measurable subsets, $\xi = \{C_\alpha \in \mathcal{B} | \alpha \in \mathcal{I}\}$ is called a **measurable partition** of X if

1. $\mu(X \setminus \cup_{\alpha \in I} C_\alpha) = 0$, i.e. the partition "contains" the whole measure.

2. $\mu(C_{\alpha_1} \cap C_{\alpha_2}) = 0$ if $\alpha_1 \neq \alpha_2$, i.e. the cells C_α of the partition are disjoint.

The entropy of μ with resepect to the partition ξ is then

$$H(\xi) := H_\mu(\xi) = -\sum_{\alpha \in I} \mu(C_\alpha) \log \mu(C_\alpha) \geq 0 \; . \tag{3.4}$$

Example: Logistic map

$$x_{n+1} = 1 - 2x^2 \tag{3.5}$$

with the partition $C_1 = [-1, 0)$,$C_2 = [0, 1]$. $\mu(C_1) = \mu(C_2) = 1/2$. Therefore $H(\xi) = \log 2$.
Now let us consider two partitions $\xi = \{C_\alpha | \alpha \in I\}$ and $\eta = \{D_\beta | \beta \in J\}$. Then the joint partition $\xi \vee \eta$ is defined as

$$\xi \vee \eta := \{C \cap D | C \in \xi, D \in \eta, \mu(C \cap D) > 0\}$$

It is also possible to define the conditional entropy of ξ given η using the notation $\mu(A|B) = \mu(A \cap B)/\mu(B)$ as

$$H(\xi|\eta) := -\sum_{\beta \in J} \mu(D_\beta) \sum_{\alpha \in I} \mu(C_\alpha|D_\beta) \log \mu(C_\alpha|D_\beta) \tag{3.6}$$

which can be written alternatively

$$H(\xi|\eta) = H(\xi \vee \eta) - H(\eta) \; .$$

Now we are able to define the entropy of the transformation F with respect to the partition ξ. First we introduce the joint partition of ξ and its preimages under F

$$\xi_{-n}^F := \xi \vee F^{-1}(\xi) \vee \ldots \vee F^{-n+1}(\xi) \; .$$

Example: ξ_{-2}^F for the logistic map (3.5) consists of the intervals between the points $-1, -\sqrt{(1/2)}, 0, \sqrt{(1/2)}, 1$, with $H(\xi_{-2}^F) = \log 4$ and $H(\xi_{-2}^F|\xi_{-1}^F) = \log 2$.

A this point we can also employ a complementary way to introduce these entropies, namely as entropies of a symbol sequence. Think of using the partition ξ to encode the phase space of the dynamical system. The trajectory $\{x_1, \ldots, x_n\}$ is encoded by a symbol sequence $\{\alpha_1, \ldots, \alpha_n\}$, if $x_1 \in C_{\alpha_1}$,$x_2 \in C_{\alpha_2}$ and so on. If we denote the probability to observe a certain symbol by $p(\alpha) = \mu(C_\alpha)$ we get for the entropy

$$H(\xi) = H(\alpha) = -\sum_{\alpha \in I} p(\alpha) \log p(\alpha) \; .$$

with α denoting the random variable which can have the value α with probability $p(\alpha)$. What corresponds then to ξ_{-n}^F? Being in a cell of this partition means that the trajectory was at time n in C_{α_n}, at $n-1$ in $C_{\alpha_{n-1}}$ and so on. Thus the measure of one cell of this partition corresponds to the joint probability $p(\alpha_n, \alpha_{n-1}, \ldots, \alpha_1)$, i.e. the probability of a certain subsequence of the string. Consequently the conditional entropy

$$H(\xi_{-2}^F|\xi_{-1}^F) = H(\alpha_2|\alpha_1) = -\sum_{\alpha_1,\alpha_2 \in I} p(\alpha_2, \alpha_1) \log p(\alpha_2|\alpha_1)$$

is denoting the uncertainty of observing the symbol α_2 after α_1 was seen. The **metric entropy** of the transformation F relative to the partition ξ (sometimes also called the entropy rate of the process generated by F) is defined as

$$h(F, \xi) := h_\mu(F, \xi) := \lim_{n\to\infty} \frac{1}{n} H(\xi_{-n}^F) \tag{3.7}$$

which is equivalent to

$$h(F, \xi) = \lim_{n\to\infty} H(\xi|F^{-1}(\xi_{-n}^F) \; . \tag{3.8}$$

$H(\xi|F^{-1}(\xi_{-n}^F)$ is monotonically decreasing. This can be shown using the representation via the symbol sequences:

$$H(\xi|F^{-1}(\xi_{-n}^F) = H(\alpha_0|\alpha_{-1}, \ldots, \alpha_{-n+1}) := h_n$$

Then

$$\begin{aligned} h_n - h_{n+1} &= H(\alpha_0|\alpha_{-1}, \ldots, \alpha_{-n+1}) - H(\alpha_0|\alpha_{-1}, \ldots, \alpha_{-n}) \\ &= MI(\alpha_0 : \alpha_n|\alpha_{-1}, \ldots, \alpha_{-n+1}) \geq 0 \end{aligned}$$

is a conditional mutual information.

The *KS-entropy* of F with respect to μ is then defined as the supremum over all partitions:

$$h_{KS}(F) := h_\mu(F) := \sup_{\xi, h(\xi)<\infty} h_\mu(F, \xi) \; . \tag{3.9}$$

A generating partition ξ_g is a partition for which the metric entropy is maximal, i.e.

$$h(F, \xi_g) = h_{KS}(F) \; .$$

There is however, in general no algorithm to find generating partitions for arbitrary dynamical systems. For 1-dimensional maps it is known how to

find them and for 2-d also an algorithm exists, which allowed to determine the generating partitions for well known systems, such as the henon map Grassberger and Kantz (1985) or the standard map Christiansen and Politi (1995).
But if we cannot find a generating partition, is it possible to estimate the KS-entropy? Yes, because in most cases (nonatomic Borel measure on a compact metric space) a finer and finer refinement of the partition allows to get better and better estimates. Or more formally: If I consider a sequence of partitions ξ_i with $\mathrm{diam}(\xi_i) \to 0$ ($\mathrm{diam}(\xi_i) := \sup_{C\in\xi} \mathrm{diam}(C)$), then $h(F, \xi_i) \to h_{KS}(F)$. An important property of the KS-entropy is that

$$h_{KS}(F^k) = k h_{KS}(F);. \tag{3.10}$$

This should be taken into account, when estimating entropies from flow data using a delay embedding.

Lyapunov exponents

The central property of chaotic dynamics is its sensitive dependence on the initial conditions, i.e. the exponential divergence of initially neighbouring trajectories. In order to keep the dynamics bounded, however, this "stretching" of the attractor has to be complemented by a folding mechanism, which brings points together which were far away from each other. If we look only locally at the dynamics wo only see the stretching. So, if we denote the distance between two trajectories at time n by $\Delta_n = ||\boldsymbol{x}_n - \boldsymbol{x}'_n|| = ||F^n(\boldsymbol{x}_0) - F^n(vx'_0)||$ then we expect

$$\Delta_n \propto e^{\lambda n}$$

with the Lyapunov exponent

$$\lambda = \lim_{n\to\infty} \lim_{\Delta_0 \to 0} \frac{1}{n} \log \Delta_n \,. \tag{3.11}$$

As we will see in a moment, one can define a whole spectrum of exponents and (3.11) ist the largest one. This Lyapunov exponents allow already a classification of deterministic dynamical systems:

- stable fixed point: $\lambda < 0$
- stable limit cycle: $\lambda = 0$
- chaotic behaviour: $\lambda > 0$

Note that, however, for a diffusion process (random walk) $\Delta_n \propto \sqrt{(}n)$, i.e. $\lambda \propto \frac{log(n)}{n} \to 0$ for $n \to \infty$.
Now let us analyze the dynamics of the difference between two trajectories $\boldsymbol{x}_n$ and $\boldsymbol{y}_n = \boldsymbol{x}_n + \boldsymbol{\Delta}_n$ in more detail. Because we consider in the end infinitesimal differences, their dynamics is governed by the linearization of the map $\boldsymbol{F}$ (3.1),i.e. its Jacobian

$$\boldsymbol{J}(\boldsymbol{x}_n) = \left(\frac{\partial \boldsymbol{F}}{\partial \boldsymbol{x}}\right)_{\boldsymbol{x}=\boldsymbol{x}_n} \qquad J_{ij}(\boldsymbol{x}_n) = \left(\frac{\partial F_i}{\partial x_j}\right)_{\boldsymbol{x}=\boldsymbol{x}_n} .$$

This leads to a linear dynamical system with time dependent coefficients for the perturbations $\boldsymbol{\Delta}$

$$\boldsymbol{\Delta}_{n+1} = \boldsymbol{J}(\boldsymbol{x}_n)\boldsymbol{\Delta}_n .$$

The long term dynamics is the governed by the eigenvalues Λ_i of the product of th Jacobians

$$\left(\prod_{n=1}^{N} \boldsymbol{J}(\boldsymbol{x}_n)\right) \boldsymbol{u}_i^{(N)} = \Lambda_i^{(N)} \boldsymbol{u}_i^{(N)} . \tag{3.12}$$

with $\boldsymbol{u}_i^{(N)}$ denoting the eigenvectors of the product of the N Jacobians.
The Lyapunov exponent λ_i is then defined as the normalized logarithm of the modulus of the ith eigenvalue Λ_i of the product of all Jacobians along the trajectory (in time order) in the limit of an infinitely long trajectory:

$$\lambda_i = \lim_{N\to\infty} \frac{1}{N} \log |\Lambda_i^{(N)}| \tag{3.13}$$

Usually the eigenvalues are ordered according their magnitude, starting with the largest. The fact that the limit (3.13) exists and is unique was schown by Osedelec (1968) and is known as multiplicative ergodic theorem. This is a highly non-trivial result because the multiplication of matrices is non-commutative and the logarithm cannot be exchanged with the formation of the eigenvalues. In the case of one-dimensional maps, however, the definition reduces to

$$\lambda = \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \log |F'(x_n)|$$

and the existence and uniqueness is established by the usual (Birkhoff) ergodic theorem.
Some properties:

- The Lyapunov exponents are invariant under smooth transformations of the phase space.
$$\tilde{\boldsymbol{F}}(\tilde{\boldsymbol{x}}) = \boldsymbol{g} \circ \boldsymbol{F} \circ \boldsymbol{g}^{-1}(\tilde{\boldsymbol{x}})$$
Then
$$\prod_{n=1}^{N} \tilde{\boldsymbol{J}}_n(\boldsymbol{x}_n) = \tilde{\boldsymbol{J}}_N^{(\boldsymbol{g})} \prod \boldsymbol{J}_n \tilde{\boldsymbol{J}}_1^{(\boldsymbol{g}^{-1})}$$
yields in the limit $N \to \infty$ the same eigenvalues and thus the same Lyapunov spectrum as the original dynamics. This ensures that the Lyapunov exponents are indeed invariants of a dynamical system.

- If μ is invariant under $\boldsymbol{F}$ then it is also under $\boldsymbol{F}^{-1}$. The absolute values of the Lyapunov exponents of $\boldsymbol{F}^{-1}$ remain the same but the sign of the exponents becomes reversed.

- Flow data have always at least one $\lambda_j = 0$.

- The Lyapunov spectra of Hamiltonian systems are symmetric wrt to zero, because the dynamics remains invariant wrt to time reversal.

Relation between the invariants

In many cases the Lyapunov spetrum contains all informations about the invariants of a dynamical system: The entropy is equal to the sum of the positiv Lyapunov exponents, the so called PESIN identity

$$h_{KS} = \sum_{\lambda_k > 0} \lambda_k \; . \tag{3.14}$$

The KAPLAN-YORKE formula makes the connection between the Lyapunov exponents and the fractal dimension of the attractor. If there is only one fractal diraction, one has

$$D_{KY} = n + \frac{\sum_{i=1}^{n} \lambda_i}{|\lambda_{n+1}|} \; , \sum_{i=1}^{n} \lambda_i \geq 0 > \sum_{i=1}^{n+1} \lambda_i$$

D_{KY} is called the KAPLAN-YORKE or also the LYAPUNOV dimension. A more general theorem was proven by Ledrappier and Young (1985):

$$\sum_i D_i \lambda_i = 0.$$

with D_i being partial dimensions, i.e. dimensions in a certain direction, with values between 0 and 1.

Some Examples

- Linear Systems: What about the deterministic parts of the linear systems considered by the statisticians? If they are stable, they have fixed point attractors, i.e. $D = 0$, only negative Lyapunov exponents and thus zero entropy.
- One dimensional maps: They have only one Lyapunov exponent. If $\lambda > 0$, $h_{KS} = \lambda$, $D_{KS} = 1$.
- Two dimensional maps: If $\lambda_1 > 0 > \lambda_2$:

$$D_{KS} = \begin{cases} 2 & \text{if} \quad |\lambda_1| > |\lambda_2| \\ 1 + \frac{\lambda_1}{|\lambda_2|} & else \end{cases}$$

3.1.2 Phase space recontruction — embedding theorems

Very often one can only observe one or a few variables of a higher dimensional dynamics. The question then is: Can we reconstruct the phase space of the underlying dynamical systems in order to estimate dimension, entropy and the Lyapunov exponents? The answer is yes and is founded on the embedding theorems by Whitney (1936), Takens (1980) and its extension by Sauer et al. (1991). The basic idea is the following: If we oberve a dynamical system

$$\boldsymbol{x}_{n+1} = F(\boldsymbol{x}_n)$$

via an observation function $y = h(\boldsymbol{x})$, the dynamical system (3.1) gives rise to a dynamic of y. Takens proposed to reconstruct the original phase space using the so called *delay coordinates* $\boldsymbol{y}_n = (y_n, y_{n-1}, \ldots, y_{n-m+1}$. The question is now under which conditions there exists a deterministic dynamical system G for the dynamics of $\boldsymbol{y}$ and how it is related to F? Obviously, $\boldsymbol{F}$ induces a dynamics for $\boldsymbol{y}$ because

$$y_{n-k} = h(\boldsymbol{x}_{n+k}) = h(F^k(\boldsymbol{x}_n)) \ .$$

thus we have But is this map also invertible, i.e. will we have a one to one relationship between $\boldsymbol{x}$ and $\boldsymbol{y}$? The answer is that under generic conditions m has to be large enough to ensure this one to one relationship. Whitney proved that every D-dimensional smooth manifold can be embedded in the $\mathbb{R}^{2D+1}$, and that the set of maps forming an embedding is a dense and open set in the space of C^1 (continously differentiable) maps. Thus for an arbitrary map $\in C^1$ there exists an embedding in its neighborhood. Takens

applied this to attractor reconstruction using delay coordinates. Sauer et al. improved the result of Takens and extented it to more general situations. Their central result is, that the recontructed state space has to be at least of dimension $m > 2D_0$, with D_0 the box counting dimension of the attractor, in order to have almost every embedding of the original phase space being one to one for the states and the Jacobian (Immersion).
Sauer et al. also considered the question, whether filtering the data could affect to possibility of a proper embedding. The result was, that the application of finite impuls response (FIR) filters to the delay coordinates would still allow an embedding, as long as enough independent observables will be considered. On the other hand, IIR filters might change the properties of the dynamical system (they are a dynamical system by themselves) and therefore affect the dimension and entropies of the whole system. Consider for instance the following extented Henon map:

$$x_{n+1} = 1 - Ax_n^2 + By_n \tag{3.15}$$

$$y_{n+1} = x_n \tag{3.16}$$

$$z_{n+1} = \alpha z_n + x_n \tag{3.17}$$

Even for $|\alpha| < 1$ this additional degree of freedon can increase the attractor dimension.
Up to this point we only discussed to which extend the properties of a given dynamical system can be recovered in the reconstructed phase space, e.g. by using a delay embedding. Here two remarks are in order:

1. For practical applications there might be better or worse phase space reconstruction. For instance, in the case of the delay embedding the delay time τ has to be selected, which we set to 1 so far, but which can be set arbitrarily in principle — with some exceptions for periodic processes, remember the discussion of the aliasing problem. Also, if more then one observable is available, one can ask, which coordinates should be used, delay coordinates from only one, or some mixed delay vector of the two, but which one? There is up to now no general method to find optimal state space reconstructions, but there are some pragmatical approaches available, which we will discuss later.

2. Up to now we startet with a dynamical system and a given "true" state space. This is, however, not the situation, which we will find in practice. There we want to characterize the system, which has produced the data, but there is nothing like a "true" state space - there are only equivalent representations of one physical system and one of

them is our state space reconstruction. There might be, however, some of them easier to interpret than others.

False nearest neighbors

TISEAN program: *false_nearest*
How can we detect a sufficiently large embedding dimension m? One Possibility is to look for so called *false nearest neighbors* (Kennel et al. (1992)). The idea is to use the geometrical structure induced by the deterministic character of the dynamics, i.e. the fact that the attractor lies in a low-dimensional manifold. As long as the embedding dimension is too low, there is no one to one embedding of the attractor and neighbouring points in the embedding space might not be neighbours in the phase space. Thus if a point $\boldsymbol{x}_i$ is a nearest neighbour to $\boldsymbol{x}_j$ in m dimensions, but not in $m+1$ dimensions, it is called a false neighbor [2] .
Then with increasing m the fraction of false nearest neighbours is estimated. If this fraction drops for some m^* this is a good candidate for a minimal embedding dimension. Usually, it drops already for $m > D_0$, which might not be sufficient as an embedding dimension, depending what one wants to analyze. There are, however, some pitfalls of this algorithm, which one has to aware of:

- In chaotic systems also true neighbours become more seperated when increasing the embedding dimension due to the effect of the chaotic dynamics.
- If the data are noisy the signature of the determinism becomes weakened.
- If the attractor is strongly folded in the reconstructed phase space the neighborhood size has to be small enough to separate several sheets of the folded attractor.

3.1.3 Dimension and entropy estimation

Box-counting dimensions and — entropies

TISEAN implementation: *boxcount*.
Univariate data: Let us start with a time series of N data points $\{x_1, \ldots, x_N\}$.

[2] In *false_nearest* a slighty different criterion is used: if the distance in $m+1$ is larger than *factor* times the distance in m dimensions it is considered a false nearest neighbor.

If we have data from an interval $[x_{min}, x_{max}]$ encoding the data with k-symbols corresponds to a partition of the reconstructed phase space with hypercubes of side length $\epsilon = \frac{x_{max}-x_{min}}{k}$. In the m-dimensional reconstructed phase space spanned by the points $\boldsymbol{x}_n = (x_n, x_{n-1}, \ldots, x_{n-m+1})$ we can count how often each of the hypercubes is visited. The relative frequencies defines a probability distribution on the cells of this partition and we can estimate its Shannon entropy

$$H(m, \epsilon) = -\sum p_j \log p_j \quad \text{with} \quad p_j = \frac{n_j}{N} \tag{3.18}$$

The information dimension can then be estimated by looking at the slope of $H(m, \epsilon)$ with respect to $-\log \epsilon$, because

$$H(m, \epsilon) = \text{const} - D_1 \log \epsilon + \mathcal{O}(\epsilon) \; .$$

Clearly, for $k = 1$ and therefore $\epsilon = x_{max} - x_{min}$ only one box is filled, with $p = 1$ and $H(\epsilon) = 0$. On the other hand, for sufficciently small ϵ, each cell of the partition contains only one point therefore $p_j = 1/N$ and $H(\epsilon) = \log N$. This is clearly a finite sample effect. The entropy (3.18) is only a good estimate of the entropy of the invariant measure if ϵ is not too small, or N is large enough, respectively. For a more detailed discussion of finite sample effects and its correction see Grassberger (2003).
The dimension

$$D_1 = \lim_{\epsilon \to 0} -\frac{H_\epsilon}{\log \epsilon} \tag{3.19}$$

is called the *information dimension*. It is possible to introduce a whole family of D_q, the so called Renyi dimensions, using the Renyi entropies

$$H^{(q)}(\epsilon) = \frac{1}{1-q} \log \sum_j p_j^q \tag{3.20}$$

and corresponding dimensions

$$D^{(q)}(m, \epsilon) = \lim_{\epsilon \to 0} -\frac{H_\epsilon^{(q)}}{\log \epsilon} \; . \tag{3.21}$$

Exercise: Show using the rule of l'Hospital that

$$\lim_{q \to 1} H^{(q)}(\epsilon) = -\sum p_j \log p_j \; .$$

If the D_q are different the system is called *multifractal*. Several methods, like multifractal analysis and the thermodynamic formalism builds upon the Renyi entropies and dimensions, respectively.

Estimating the KS-entropie

Estimating the KS-entropy using the box-counting entropy estimates $H(m, \epsilon)$ is straightforward:

1. Find an embedding dimension m_0, which is large enough, at least $m_0 > D_0$.

2. Estimate $H(m, \epsilon)$ for some values of $m \geq m_0$. Estimate the conditional entropies
$$h(m, \epsilon) = H(m + 1, \epsilon) - H(m, \epsilon) \tag{3.22}$$
plot $h(m, \epsilon)$ as a function of $\log \epsilon$ and look for a plateau $h(m, \epsilon) \approx \text{const}$ at some ϵ range. The ϵ has to be large enough to minimize finite sample effects, but also small enough to resolve the deterministic structure.

3. If the $h(m, \epsilon)$ remains also constant for increasing m the value might be used as an estimate for h_{KS}.

There are, however, severe problems with this procedure. Although the $h(m, \epsilon)$ are monotonically decreasing, i.e. $h(m, \epsilon) \geq h(m + 1, \epsilon)$ we cannot expect that $h(m, \epsilon)$ estimated from the data gives an upper bound for the $h(\infty, \epsilon)$, because the finite sample effects lead to an underestimation of the conditional entropies. Thus also alternative methods should be used to estimate the KS-entropy from a time series, such as the correlation entropy and the Lyapunov exponents.

The correlation dimension

TISEAN implementation *d2*
The most popular quantity from nonlinear time series analysis is the correlation dimension. For low dimensional data it can be reliably estimated already from relatively short data sets with a relatively simple algorithm. Mathematically the correlation dimension of a measure μ is defined as follows:
$$D_2 = -\lim_{\epsilon \to 0} \frac{\log \int_X \mu(B(\boldsymbol{x}, \epsilon)) d\mu(\boldsymbol{x})}{\log \epsilon} \tag{3.23}$$
with $B(\boldsymbol{x}, \epsilon)$ denoting the Ball of radius ϵ centered at point $\boldsymbol{x}$, i.e. the set of points $\boldsymbol{y}$ with $||\boldsymbol{x} - \boldsymbol{y}|| < \epsilon$. It is estimated from N data points via the correlation sum
$$C(\epsilon) = \frac{2}{N(N-1)} \sum_{i=1}^{N} \sum_{j=i+1}^{N} \Theta(\epsilon - ||\boldsymbol{x}_i - \boldsymbol{x}_j||)$$

with Θ being the Heaviside step function $\Theta(x) = 0$ if $x \leq 0$ and $\Theta(x) = 1$ if $x > 1$. That means we count the fraction of distances between data points in the phase space, which is smaller than ϵ. In the limit $N \to \infty$ we expect C to scale like a power law, $C(\epsilon) \propto \epsilon^D$, and we can define the correlation dimension by

$$D_2(N, \epsilon) = \frac{\partial C(\epsilon, N)}{\partial \log \epsilon} \tag{3.24}$$

$$D_2 = \lim_{\epsilon \to 0} \lim_{N \to \infty} D_2(N, \epsilon) \tag{3.25}$$

In practice, however, we have only a finite amount of data, so we cannot perform the limits and so he have to estimate the dimension at finite resolution ϵ. Therefore one usually plots $D_2(m, \epsilon)$ via $\log \epsilon$ (see Fig. **??**) for the example of the henon map. Then one has to identify a region, where it is approximately constant and can then estimate it by fitting a stright line in the log-log plot of $C(\epsilon)$. This plateau or scaling range is limited on the large scales, because if ϵ is too large, the structure of the attractor cannot be resolved and usually the dimension is overestimated[3], while the lower end might be determined by the accuracy of the measurement (how many digits), the number of data points and/or the amount of noise.
If the embedding dimension m is too small and the amount of data is sufficiently large the plateau should appear at the value of the embedding dimension $D_2 = m$. Only if the embedding dimension is larger than D_2 we can expect to find a plaeau et the value of the attractor dimension. This happens usually alread for $m > D_0$ and not only for $m > 2D_0$, the correct embedding dimension. The explanation is that the self-intersections of the attractor have zero measure and therefore do not affect our dimension estimates. However, for prediction or modelling this self-intersections are important and that m might be too small.

Temporal correlations and the Theiler correction

There is an important practical problem, which lead to many spurious dimension estimates in the past, the problem of temporal correlations. We use the number of neighbours of $\boldsymbol{x}$ with a distance smaller than ϵ to estimate the measure of $\mu(B(\boldsymbol{x}, \epsilon))$, i.e. the probability to find a point in the ϵ neighbourhood of $\boldsymbol{x}$. If now the actual neighbours of these points are not only

[3]Note that this might be totally different for strongly correlated data, such as highly sampled flow data.

neighbours in the phase space but also neighbours in time, we get obviously a biased estimate. There might be even contributions to this bias from the other points in the neighbourhood and their temporal correlated neighbours if they are also neighbours of the first point. Theiler (1986) proposed therefore to exclude all points in a teporal window around the refence point from the calculation. This is sometimes called the "Theiler window" n_{TW}. The formula for the correlation sum then reads

$$C(\epsilon) = \frac{2}{(N - n_{TW}(N - 1 - n_{TW})} \sum_{i=1}^{N} \sum_{j=i+n_{TW}+1}^{N} \Theta(\epsilon - ||\boldsymbol{x}_i - \boldsymbol{x}_j||) \; .$$

To determine a good value of this window Provenzale et al. (1992) introduced the so called space time separation plot (in TISEAN *stp*).

3.1.4 Estimating the Lyapunov exponents

The largest Lyapunov exponent

TISEAN: *lyap_k, lyap_r* For the restimation of the largest Lyapunov exponent the expansion rate has to be estimated. This is done by calculating the logarithm of the mean difference between points which were initially in the neighbourhood of a reference point and finally also averaging over these reference points:

$$S(\Delta n) = \frac{1}{N - \Delta n} \sum_{n=1}^{N-\Delta n} \log \left(\frac{1}{|\mathcal{U}(\boldsymbol{x}_\prime)|} \sum_{\boldsymbol{y}_n \in \mathcal{U}(\boldsymbol{x}_\prime)} |vy_{n+\Delta n} - \boldsymbol{x}_{n+\Delta n}| \right) \quad (3.26)$$

Then the linear slope of $S(\Delta n)$ should be an estimate of the largest Lyapunov exponent, becaue the difference will be dominated by the largest exponent. Beside the usual embedding parameters one has also to specify the neighbourhood, either by its diameter ϵ or by the number of neighbours. The program *lyap_k* estimates the stretching factor for a set of neighbourhood sizes and provides some statistics about the numbers of neighbours found.

Lyapunov spectrum

A reliable estimation of the Lyapunov spectrum is in most cases only possible if the system equations are known or a global model is available for a given data set. In this case we can estimate the Jacobian from the equations. Nevertheless, the product of the Jacobians will become singular, so it cannot

be evaluated. Therefore usually a procedure introduced by Bennetin et al. (1978) is used: A set of orthogonal vecors, spanning the phase space is iterated using the linearized dynamics. After a few steps the vectors become more and more aligned in the direction of the largest Lyapunov exponent. Therefore the vectors are iteratively othonormalized, beginning with the largest one and the scaling factors are stored. The Lyapunov exponents are then estimated by the averages of the logarithms of the scaling factors.
One possibility to estimate the Lyapunov spectra from data would be to estimate the Jacobians directly from the data (TISEAN: *lyap_spec*). This corresponds to fitting loacally linear models, which will be discussed later. At this point we will only mentions some problems of this approach:

- The local neighbourhoods used for the linear fit has to be large enough to avoid fitting the pecularities of the noise.
- On the other hand side the local neighbourhood has to be small enough not to smear out the nonlinear structures of the attractor.
- Very often this method does not provide robust estimates of the exponents. Thus, although conceptionally appealing because it contains all information about the invariants, only estimating the Lyapunov spectrum from data might be actually a bad idea.

Bibliography

Akaike, H., 1969. Fitting autoregressions for prediction. Ann. Inst. Statist. Math 21, 243–347.

Bennetin, G., Galgani, L., Giorgilli, A., Strelcyn, J. M., 1978. Lyapunov characteristic exponents for smooth dynamical systems and for Hamiltoniansystems: a method for computing all of them. C. R. Academie Sci. Paris A 206, 431.

Christiansen, Politi, May 1995. Generating partition for the standard map. Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics 51 (5), R3811–R3814.

Grassberger, P., 2003. Entropy estimates from insufficient samplings. Tech. rep., arXiv:physics/0307138v1.

Grassberger, P., Kantz, H., 1985. Generating partitions for the dissipative Hénon map. Phys. Lett. A 113 (5), 235–238.

Kennel, M. B., Brown, R., Abarbanel, H. D. I., 1992. Determining embedding dimension for phase-space reconstruction using a geometrical construction. Phys. Rev. A 45, 3403–3411.

Ledrappier, F., Young, L.-S., 1985. The metric entropy of diffeomorphisms part I and II. Ann. Math. 122, 509.

Osedelec, V. I., 1968. A multiplicatice ergodic theorem. lyapunov characteristic numbers for dynamical systems. Tran. Moscow Math. Soc. 19, 197.

Provenzale, A., Smith, L., Vio, R., Murante, G., 1992. Distinguishing between low-dimensional dynamics and randomness in measured time series. Physica D 58, 31–49.

Sauer, T., Yorke, J. A., Casdagli, M., 1991. Embedology. J. Stat. Phys. 65, 579–616.

Takens, F., 1980. Detecting strange attractors in turbulence. In: Rand, D. A., Young, L.-S. (Eds.), Dynamical Systems and Turbulence (Warwick 1980). Vol. 898 of Lecture Notes in Mathematics. Springer-Verlag, Berlin, pp. 366–381.

Theiler, J., 1986. Spurious dimensions from correlation algorithms applied to limited time-series data. Phys. Rev. A 34, 2427–2432.

Whitney, H., 1936. Differentiable manifolds. Ann. Math. 37, 645.