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:

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

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:

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

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.