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):

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

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 (2.8) with (2.7) 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 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 continous 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 continous 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 orginal 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 travelling 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 estimate 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.