STATS 207: Time Series Analysis
Lecture Notes

Dominik Rothenhäusler

April 4, 2020

Thanks a lot to Merle Behr, Aditya Guntuboyina, Nicolai Meinshausen and Hans-Ruedi Künsch. These notes build to a very large extend on their material. These notes are intended to just give a quick summary of what we discussed in the course. For examples and illustrations of the concepts and methods, you should look at the slides and the *R*-demonstrations which are on the course web page and the examples in the book Shumway & Stoffer. There are bound to be errors – I would appreciate if you point them out to me so I can get a corrected version to everybody.

Contents

Objectives of time series analysis

Goals of time series analysis can be classified in one of the following

i) Compact description of data as $X_t = m_t + s_t + W_t$, where X_t is the observed time-series, m_t a trend, s_t a seasonal component and W_t white noise. This can aid with interpretation for example by seasonal adjustment of unemployment figures.

ii) Hypothesis testing. We might for example want to test whether the trend component m_t vanishes for summer rainfall figures in Washington over the last 10 years.

iii) Prediction. Examples are: predict COVID-19 cases/ unemployment data/ strength of El Nino / airlines passenger numbers or next word in a text. Might sometimes only be possible via simulation, as when trying to forecast hurricane intensity for the next decade at a specific location.

iv) Control/Causality/Reinforcement learning. One example is impact of monetary policy (interest rates) on inflation, where causal impact is quite different (possibly even different sign) to pure observational correlation. Or optimal filling and draining of lakes for energy storage.

1 Introduction

A stochastic process is a mathematical model for a time series.
Stochastic process = Collection of random variables $(X_t(\omega); t \in T)$. Alternative view: Stochastic process as a random function from T to $\mathbb{R}$.
A basic distinction is between continuous and discrete equispaced time T. Models in continuous time are prefered for irregular observation points. In this course we will restrict ourselves mostly to discrete equispaced time and, if not stated otherwise, use $T = \mathbb{Z}$.
In all interesting cases, there is dependence between the random variables at different times. Hence need to consider joint distributions, not only marginals. Gaussian stochastic processes have joint Gaussian distribution for any number of time points.
A stochastic process describes how different time series (when different ω's are drawn) could look like. In most cases, we observe only one realization $x_t(\omega)$ of the stochastic process (a single ω). Hence it is clear that we need additional assumptions, if we want to draw conclusions about the joint distributions (which involves many ω's) from a single realization. The most common such assumption is stationarity.
Stationarity means the same behavior of the observed time series in different time windows. Mathematically, it is formulated as invariance of (joint) distributions when time is shifted. Stationarity justifies taking of averages (mathematically, one needs ergodicity in addition).
Some examples of time series:

a) IID noise $X_t = W_t$, where $W_t \sim F$ i.i.d. for some distribution F with mean 0 and variance σ^2. Special case is Gaussian noise, where $F = \Phi$.

b) Harmonic oscillations plus (white) noise,

$$X_t = \sum_{k=1}^{K} \alpha_k \cos(\lambda_k t + \phi_k) + W_t,$$

where W_t as above is a IID noise process and K, α, λ, ϕ unknown parameters.

c) Moving averages. For example

$$X_t = \frac{1}{3}(W_t + W_{t-1} + W_{t-2}),$$

where W_t is Gaussian noise.

d) Auto-regressive processes. For example

$$X_t = 0.9X_{t-1} + W_t,$$

plus initial conditions.

e) Random Walk (special case of an auto-regressive process)

$$X_t = X_{t-1} + W_t$$

or, with drift,

$$X_t = X_{t-1} + 0.2 + W_t.$$

f) Auto-regressive conditional heteroscedastic models

$$X_t = \sqrt{1 + 0.9X_{t-1}^2} \cdot W_t,$$

where again W_t is a Gaussian noise process.

We will start with very simple time series models and add more complex structure step by step.

2 White Noise

Definition 2.1.
Random variables $X_1, \ldots, X_n$ will be denoted as

1. ***white noise*** *if they have mean zero, variance σ^2 and are uncorrelated,*
2. ***IID noise*** *if they are white noise and are independent and identically distributed (IID),*
3. ***Gaussian noise*** *if they are IID noise and are normally distributed $X_i \sim \mathcal{N}(0, \sigma^2)$.*

How to check if white noise is a good model for a given dataset? In white noise, for any t, the random variable X_t is uncorrelated from X_{t+1}, X_{t+2}, and so on. That is, for white noise, X_t cannot help for (linearly) predicting any of X_{t+1}, $X_{t+2}, \ldots$.

Definition 2.2 (Autocorrelation)**.**
For random variables $X_1, \ldots, X_n$ the autocorrelation (ACF) is defined as

$$\rho(s,t) := \frac{\mathrm{Cov}(X_s, X_t)}{\sqrt{\mathrm{Var}(X_s)\,\mathrm{Var}(X_t)}} = \frac{\mathrm{E}((X_s - \mathrm{E}(X_s))(X_t - \mathrm{E}(X_t))}{\sqrt{\mathrm{E}((X_s - \mathrm{E}(X_s))^2)\,\mathrm{E}((X_t - \mathrm{E}(X_t))^2)}}.$$

Analog, we also define the **autocovariance** as $\gamma(s,t) := \mathrm{Cov}(X_s, X_t)$. Note that for white noise $\rho(s,t) = 0$ for all $s \neq t$ and $\rho(t,t) = 1$. In particular, we have that $\rho(s,t) = \rho(|s-t|)$. Thus, for a set of observations $X_1, \ldots, X_n$ which is white noise, a natural way of estimating $\rho(k)$ for some lag $k = 0, \ldots, n-1$ is via the following.

Definition 2.3 (Sample autocorrelation function)**.**
For observations $X_1, \ldots, X_n$ the sample autocorrelation at lag k is defined as

$$r_k = \frac{\sum_{t=1}^{n-k}(X_t - \bar{X})(X_{t+k} - \bar{X})}{\sum_{t=1}^{n}(X_t - \bar{X})^2} \qquad \text{for } k = 1, 2, \ldots$$

For white noise with n large enough, we expect r_k to be a good estimate of $\rho(k)$ and hence, $r_k \approx 0$. This can be made mathematically more precise, as the following theorem shows (the precise formulation of the theorem and its proof are given in (Shumway and Stoffer, 2006, Theorem A.7, see also Property P1.1)).

Theorem 2.4.
Under some general conditions, if $X_1, \ldots, X_n$ is white noise, then for any fixed lag k and n large enough the sample autocorrelations $r_1, r_2, \ldots, r_k$ are approximately independent normally distribution with mean zero and variance $1/n$, that is,

$$\sqrt{n} \begin{pmatrix} r_1 \\ \vdots \\ r_k \end{pmatrix} \to \mathcal{N}(0, I) \qquad \text{as } n \to \infty, \tag{1}$$

where I denotes the $k \times k$ identity matrix.

Note that the variance decreases to zero as n increases and the mean is zero. Thus, for large n and reasonably small k, the sample autocorrelations $r_1, \ldots, r_k$ should be close to zero. Also note that for large n the sample autocorrelations for different lags are approximately independent.
Therefore, one way of checking if the white noise model is a good fit to the data is to plot the sample autocorrelations. This plot is known as the **correlogram**.
Use the function **acf** in **R** to get the correlogram. The blue bands in the correlogram correspond to levels of $\pm 1.96 n^{-1/2}$, where 1.96 is the 97.5% quantile of the standard normal distribution.
When $X_1, \ldots, X_n$ is white noise, then for any k

$$\mathbf{P}(|r_k| > 1.96 n^{-1/2}) = \mathbf{P}(n^{1/2}\,|r_k| > 1.96) \approx \mathbf{P}(|\mathcal{N}(0,1)| > 1.96) = 5\%.$$

Thus, a value of r_k outside the blue bands is **significant** i.e., it gives evidence against pure white noise.
However, the overall probability of getting at least one r_k outside the bands increases with the number of coefficients plotted! For example, if 20 r_k's are plotted, one expects to get one ($= 20 * 5\%$) significant value under pure noise.

3 Trend and Seasonality Models

3.1 Trend Models

Many time series datasets show an increasing or decreasing trend. A simple model for such datasets is obtained by adding a deterministic trend function of time to white noise:

$$X_t = m_t + Z_t. \tag{2}$$

Here m_t is a deterministic trend function and Z_t is white noise. Common techniques for fitting this model to the data, that is, estimating m_t and removing the noise Z_t from the observations X_t are the following.

- **Parametric form for m_t and linear regression:**
  Assume a simple parametric form for m_t, say linear or quadratic, and fit it via linear regression.

- **Smoothing:**
  It is well-known that noise is eliminated by averaging. Consider

  $$\hat{m}_t = \frac{1}{2q+1} \sum_{j=-q}^{q} X_{t+j}. \tag{3}$$

  If m_t is linear on the interval $[t-q, t+q]$, then check that

  $$\frac{1}{2q+1} \sum_{j=-q}^{q} m_{t+j} = m_t.$$

  Thus if m_t is approximately linear over $[t-q, t+q]$ and q is sufficiently large, then

  $$\hat{m}_t \approx m_t + \frac{1}{2q+1} \sum_{j=-q}^{q} Z_{t+j} \approx m_t.$$

  $\hat{m}_t$ is also called the *Simple Moving Average* of X_t.

How to chose the parameter q for smoothing? Observe that

$$\hat{m}_t = \frac{1}{2q+1} \sum_{j=-q}^{q} m_{t+j} + \frac{1}{2q+1} \sum_{j=-q}^{q} Z_{t+j}.$$

If q is very small, then the second term above is not quite small and so the trend estimate will also involve some noise component and therefore $\hat{m}_t$ will be very noisy. On the other hand, if q is large, then the assumption that m_t is linear on $[t-q, t+q]$ may not be quite true and thus, $\hat{m}_t$ may not be close to m_t. This is often referred to as the *Bias-Variance tradeoff*. Therefore q should be neither too small nor too large.

Parametric Curve Fitting versus Smoothing: In the following we summarize the main advantages and disadvantages of parametric curve fitting versus smoothing.

	Smoothing	Parametric Fitting
Pro	Approximate linearity on intervals $[t-q, t+q]$ for small q is a quite **weak assumption**.	Estimation of m_t is based on **all observations** and hence, more precise.
Contra	Estimation of m_t is based **only on** $2q+1 < n$ **observations**.	Linear/quadratic form for m_t might be a **strong assumption** violated in practice.

3.1.1 Filtering for Trend Estimation

The smoothing estimate (3) of the trend function m_t is a special case of *linear filtering*. A linear filter converts the observed time series X_t into an estimate of the trend $\hat{m}_t$ via the linear operation:

$$\hat{m}_t = \sum_{j=-q}^{s} a_j X_{t+j}. \tag{4}$$

The numbers $a_{-q}, a_{-q+1}, \ldots, a_{-1}, a_0, a_1, \ldots, a_s$ are called the *weights of the filter*. The Smoothing method is clearly a special instance of filtering with $s = q$ and $a_j = 1/(2q+1)$ for $|j| \leq q$ and 0 otherwise. In addition, there are other choice of filters that people commonly use:

Binomial Weights When we are estimating the value of the trend m_t at t, it makes sense to give a higher weight to X_t compared to $X_{t\pm1}$ and a higher weight to $X_{t\pm1}$ compared to $X_{t\pm2}$ and so on. An example of such weights are:

$$a_j = 2^{-q}\binom{q}{q/2+j} \qquad \text{for } j = -q/2, -q/2+1, \ldots, -1, 0, 1, \ldots, q/2.$$

As in usual smoothing, choice of q is an issue here.

Spencer's 15 point moving average We have seen that simple moving average filter leaves linear functions untouched. Is it possible to design a filter which leaves higher order polynomials untouched? For example, can we come up with a filter which leaves all quadratic polynomials untouched? Yes! For a filter with weights a_j to leave all quadratic polynomials untouched, we need the following to be satisfied for every quadratic polynomial m_t:

$$\sum_j a_j m_{t+j} = m_t \qquad \text{for all } t$$

In other words, if $m_t = \alpha t^2 + \beta t + \gamma$, we need

$$\sum_j a_j \left(\alpha(t+j)^2 + \beta(t+j) + \gamma\right) = \alpha t^2 + \beta t + \gamma \qquad \text{for all } t.$$

Simplify to get

$$\alpha t^2 + \beta t + \gamma = (\alpha t^2 + \beta t + \gamma)\sum_j a_j + (2\alpha t + \beta)\sum_j j a_j + \alpha \sum_j j^2 a_j \qquad \text{for all } t.$$

This will clearly be satisfied if

$$\sum_j a_j = 1 \qquad \sum_j j a_j = 0 \qquad \sum_j j^2 a_j = 0. \tag{5}$$

An example of such a filter is Spencer's 15 point moving average defined by

$$a_0 = \frac{74}{320}, a_1 = \frac{67}{320}, a_2 = \frac{46}{320}, a_3 = \frac{21}{320},$$
$$a_4 = \frac{3}{320}, a_5 = \frac{-5}{320}, a_6 = \frac{-6}{320}, a_7 = \frac{-3}{320}$$

and $a_j = 0$ for $j > 7$. Also the filter is symmetric in the sense that $a_{-1} = a_1, a_{-2} = a_2$ and so on. Check that this filter satisfies the condition (5). Because this is a symmetric filter, it can be checked that it allows all cubic polynomials to pass unscathed as well.

Exponential Smoothing To obtain $\hat{m}_t$ in this method, only the *previous* observations $X_t, X_{t-1}, X_{t-2}, \ldots$ are used. The weights assigned to these observations *exponentially decrease* the further one goes back in time. A natural way of estimating $\hat{m}_t$ using only the previous observations is to use

$$wX_t + w^2 X_{t-1} + w^3 X_{t-2} + \ldots.$$

Because we would not like to change the scale of X_t, it is necessary that the sum of the weights is equal to one. Because $\sum_{j=1}^{\infty} w^j = w/(1-w)$, a proper way to use the weighted average is to consider

$$\hat{m}_t := \frac{1-w}{w}\left[wX_t + w^2 X_{t-1} + w^3 X_{t-2} + \ldots\right].$$

Check that now the weights add up to 1. w is a parameter that determines the amount of smoothing (w here is analogous to q in smoothing). If w is close to 0, there is very little smoothing and vice versa.

3.1.2 Isotonic Trend Estimation

Isotonic estimation is a very elegant way of estimating **monotone** trends, where in (2) $m_1 \leq m_2 \leq \ldots \leq m_n$. The Isotonic estimator $\hat{m}_t$ for m_t is the solution to the following minimization problem:

$$\text{Minimize } \sum_{t=1}^{n}(X_t - a_t)^2 \text{ under the constraint } a_1 \leq \cdots \leq a_n. \tag{6}$$

This is a convex optimization problem, which can be solved highly efficiently. Isotonic estimator is one of the early examples of estimation procedures in statistics that are based on convex optimization.

Advantages over smoothing methods:

1. No need to select a smoothing parameter q. Smoothing parameter selection is a very tricky issue in general and isotonic methods completely avoid this issue.

2. One gets $\hat{m}_t$ for all t unlike moving average smoothing which does not yield $\hat{m}_t$ at end-points.

The problem however is that this method only works in situations when we know the trend to be either non-decreasing or non-increasing.

3.1.3 Differencing for Trend Elimination

So far, we looked at trend models: $X_t = m_t + Z_t$ where m_t is a deterministic trend function and $\{Z_t\}$ is white noise. The residuals obtained after fitting the trend function m_t in the model $X_t = m_t + Z_t$ are studied to see if they are white noise or have some dependence structure that can be exploited for prediction. Suppose that the goal is just to produce such detrended residuals. Differencing is a simple technique which produces such de-trended residuals. One just looks at $Y_t = X_t - X_{t-1}, t = 2, \ldots, n$. If the trend m_t in $X_t = m_t + Z_t$ is linear, then this operation simply removes it because if $m_t = \alpha t + b$, then $m_t - m_{t-1} = \alpha$ so that $Y_t = \alpha + Z_t - Z_{t-1}$.
Suppose that the first differenced series Y_t appears like white noise. What then would be a reasonable forecast for the original series: X_{n+1}? Because Y_t is like white noise, we forecast Y_{n+1} by the sample mean $\bar{Y} := (Y_2 + \cdots + Y_n)/(n-1)$. But since $Y_{n+1} = X_{n+1} - X_n$, this results in the forecast $X_n + \bar{Y}$ for X_{n+1}.
Sometimes, even after differencing, one can notice a trend in the data. In that case, just difference again. It is useful to follow the notation ∇ for differencing:

$$\nabla X_t = X_t - X_{t-1} \qquad \text{for } t = 2, \ldots, n$$

and second differencing corresponds to

$$\nabla^2 X_t = \nabla(\nabla X_t) = \nabla X_t - \nabla X_{t-1} = X_t - 2X_{t-1} + X_{t-2} \qquad \text{for } t = 3, \ldots, n.$$

It can be shown that quadratic trends simply disappear with the operation ∇^2. Suppose the data $\nabla^2 X_t$ appear like white noise, how would you obtain a forecast for X_{n+1}?
Differencing is a quick and easy way to produce detrended residuals and is a key component in the ARIMA forecasting models (later). A problem however is that it does not result in any estimate for the trend function m_t.

Stochastic Trend One can also consider models $X_t = m_t + Z_t$ where m_t is a stochastic trend function as opposed to a deterministic trend function. A popular choice is

$$m_t = m_{t-1} + \delta + W_t$$

where $\delta \in \mathbb{R}$ is a fixed constant and W_t is white noise. This is an example of a *state space model*. When $\delta = 0$, this is called the *local level model*.
The noise $\{W_t\}$ is called *evolution error* and the noise $\{Z_t\}$ is known as *observational error*. It is assumed that these two error processes are independent. In some applications, this model might be more suitable than any deterministic trend model.
Observe that the differenced series for X_t is:

$$\nabla X_t = X_t - X_{t-1} = m_t - m_{t-1} + Z_t - Z_{t-1} = \delta + W_t + Z_t - Z_{t-1}.$$

∇X_t is thus a detrended series. Therefore, differencing also works with this particular stochastic model for trend.

3.2 Seasonality Models

Many time series datasets exhibit seasonality. Simplest way to model this is

$$X_t = s_t + Z_t \tag{7}$$

where s_t is a periodic function of a known period d, that is, $s_{t+d} = s_t$ for all t and Z_t is white noise. These models are appropriate, for example, to monthly, quarterly or weekly data sets that have a seasonal pattern to them. Just like the trend case, there are three different approaches to dealing with seasonality: fitting parametric functions, smoothing and differencing.

- **Fitting a parametric seasonality function:**
  The simplest periodic functions of period d are $a\cos(2\pi ft/d)$ and $a\sin(2\pi ft/d)$. Here f is a positive integer. The quantity a is called *Amplitude* and f/d is called *frequency* and its inverse, d/f is called *period*. The higher f is, the more rapid the oscillations in the function are. More generally,

  $$s_t = a_0 + \sum_{f=1}^{k} (a_f \cos(2\pi ft/d) + b_f \sin(2\pi ft/d)) \tag{8}$$

  is a periodic function of period d. Choose a value of k (not too large) and fit this to the data. Note that there is no need to consider values of k that are more than $d/2$, as every periodic seasonal component $s_1, \ldots, s_n$ with period d can be written in the form (8) with $k = d/2$.

- **Nonparametric seasonality function estimation**
  Because of periodicity, the function s_t only depends on the d values $s_1, s_2, \ldots, s_d$. Clearly, s_t can be estimated by

  $$\hat{s}_i := \text{average of } X_i, X_{i+d}, X_{i+2d}, \ldots \tag{9}$$

  Note that, here we are fitting d parameters from n observations. Thus, if n is not sufficiently large compared to d, this might lead to overfitting.

- **Differencing**
  One can obtain residuals adjusted for seasonality from the data (7) without explicitly fitting a seasonality function, via

  $$X_t - X_{t-d} = s_t - s_{t-d} + Z_t - Z_{t-d} = Z_t - Z_{t-d}. \tag{10}$$

  The lag-d differenced data $X_t - X_{t-d}$ do not display any seasonality. This method of producing deseasonalized residuals is called *Seasonal Differencing*.

3.3 Both Trend and Seasonality

Finally, consider the situation where the datasets has both trend and seasonal component, that is

$$X_t = m_t + s_t + W_t, \tag{11}$$

with trend component m_t, seasonal component s_t with period d, and white noise W_t. Again, one can apply three basic methods to remove both the trend and the seasonal component:

- **Fit Parametric Functions:** For example a linear or quadratic function for m_t and a sinusoid for s_t.

- **Smoothing:** The local smoothing method from Section 3.1 suggests to take a local average of $X_{t-q}, \ldots, X_{t+q}$ to estimate the trend m_t. However, with an additional seasonal component, this will be biased when $[t-q, t+q]$ covers only a fraction of a period.

  For example, Suppose we want to estimate m_t for December, 2011. The smoothing method suggests taking a local average of X values near December, 2011. How near? If we, for instance, take an average from Oct, 2011 to Mar, 2012, then we might get a value which is affected by the fact that the average is over the winter period. To overcome this fact, the sensible thing to do is average values either from June, 2011 to May, 2012 or from July, 2011 to June 2012. Both of these are viable options and the seasonality effect is now taken care of because all months (and hence all seasons) are represented equally in this average. There is still the minor question of which of these two averages to use. One can just take a further average of these two averages.

  Thus, when d is even, we estimate

$$\hat{m}_t := \frac{0.5X_{t-d/2} + X_{t-(d/2)+1} + \cdots + X_{t+(d/2)-1} + 0.5X_{t+d/2}}{d}$$

  and similar when d is odd. The observations $Y_t = X_t - \hat{m}_t$ then correspond to detrended data. One can then estimate

$$\hat{s}_i := \text{average of } Y_i, Y_{i+d}, Y_{i+2d}, \ldots. \tag{12}$$

- **Differencing:** Because the trend function satisfies $s_t = s_{t-d}$, the difference $X_t - X_{t-d}$ equals $m_t - m_{t-d} + W_t - W_{t-d}$ and hence this differenced data does not have any seasonal component. The trend $m_t - m_{t-d}$ can then be eliminated using the methods previously studied, in particular, by differencing.

3.4 Variance Stabilizing Transform

In the models (2), (7), and (11) with deterministic trend, seasonal, and trend+seasonal component and additive white noise Z_t, one implicitly assumes that the observations X_t have a constant variance. This is denoted as *homoscedasticity*. Now suppose that the variability of the time series data set appears to be non-constant. In particular, consider the situation where the variability of the data X_t changes over time with its mean $\mathrm{E}(X_t) = \mu_t$, that is,

$$\mathrm{Var}(X_t) = g(\mu_t) \text{ for some function } g.$$

Recall that for the trend model (2) $\mu_t = m_t$ and for the seasonal model (7) $\mu_t = s_t$. Then, one can often transform the data with some function f and consider observations $f(X_t)$ to obtain (approximate) homoscedasticity. This is denoted as a *Variance Stabilizing Transform*.
To this end, consider a first order Taylor approximation of $f(X_t)$ around the mean μ_t

$$f(X_t) \approx f(\mu_t) + f'(\mu_t)(X_t - \mu_t), \tag{13}$$

such that

$$\mathrm{Var}(f(X_t)) \approx (f'(\mu_t))^2 \,\mathrm{Var}(X_t) = (f'(\mu_t))^2 g(\mu_t). \tag{14}$$

If we chose f such that the function $(f'(\cdot))^2 g(\cdot)$ is constant, than the variance of $f(X_t)$ will be approximately constant over time and $f(X_t)$ approximately homoscedastic.
Examples:

- When the variance increases **linear** with $\mathrm{Var}(X_t) = C\mu_t$, then for $f(x) = \sqrt{x}$ we find that $\mathrm{Var}(\sqrt{X_t}) \approx C/4$.
  (For example, count data are often modeled via Poisson Ransom variables, where the variance equals the mean.)

- When the variance increases **quadratic** with $\mathrm{Var}(X_t) = C\mu_t^2$, then for $f(x) = \log x$ we find that $\mathrm{Var}(\log X_t) \approx C$.

- The above examples are both special cases of the **Box-Cox transformation** with parameter λ, which considers the function

$$f(x) = f_\lambda(x) = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \log(x) & \text{if } \lambda = 0, \end{cases} \tag{15}$$

  where square root essentially corresponds to $\lambda = 1/2$.

4 Stationary Time Series

The concept of a stationary time series is the most important thing in this course. Stationary time series are typically used to model the residuals after trend and seasonality have been removed. Stationarity allows a systematic study of time series forecasting, as it provides the basis to learn the dependence structure of a time series as the number of observations n increases. For the following definitions, it is convenient of think of the random variables $\{X_t\}$ as forming a doubly infinite sequence:

$$\ldots, X_{-2}, X_{-1}, X_0, X_1, X_2, \ldots.$$

The notion of stationarity will apply to the doubly infinite sequence of random variables $\{X_t\}$. Stationary essentially means that the dependence is invariant over time and hence, we can learn while observing more and more data.

Definition 4.1 (Strict or Strong Stationarity)**.**
*A doubly infinite sequence of random variables $\{X_t\}$ is **strictly stationary** if for every choice of times $t_1, \ldots, t_k$ and lag h the joint distribution of $(X_{t_1}, X_{t_2}, \ldots, X_{t_k})$ is the same as the joint distribution of $(X_{t_1+h}, X_{t_2+h}, \ldots, X_{t_k+h})$.*

Stationarity means that the joint distribution of the random variables remains constant over time. For example, under stationarity, the joint distribution of today's and tomorrow's random variables is the same as the joint distribution of the variables from *any* two successive days (past or future).
Note how stationarity makes the problem of forecasting or prediction feasible. From the data, we can study how a particular day's observation depends on the those of the previous days and because under stationarity, such a dependence is assumed to be constant over time, one can hope to use it to predict future observations from the current data.
Many of the things that we shall be doing with stationarity actually go through even with the following notion that is weaker than strong stationarity.

Definition 4.2 (Second-Order or Weakly or Wide-Sense Stationarity)**.**
*A doubly infinite sequence of random variables $\{X_t\}$ is **weak stationary** if*

1. *The mean of the random variable X_t, denoted by $\mathrm{E}\, X_t$, is the same for all times t.*
2. *The covariance between X_t and X_s is the same as the covariance between X_{t+h} and X_{s+h} for every choice of times t, s and lag h.*

Weak stationarity means that the *second order properties* (means and covariances) of the random variables remain constant over time. Unlike strong stationarity, the joint distribution of the random variables may well change over time.
Another way of phrasing the condition in 2. namely

$$\mathrm{Cov}(X_t, X_s) = \mathrm{Cov}(X_{t+h}, X_{s+h}) \qquad \text{for all } t, s \text{ and } h,$$

is that the covariance between the two random variables only depends on the *time lag* between them and (for time lag $h = 0$) that the variance remains constant, that is, $\mathrm{Var}(X_t) = \mathrm{Var}(X_0)$. In other words, autocovariance $\gamma(s,t) = \mathrm{Cov}(X_s, X_t)$ and autocorrelation $\rho(s,t) = \gamma(s,t)/\sqrt{\mathrm{Var}(X_s)\,\mathrm{Var}(X_t)}$ as in Definition 2.2 only depend on the time lag $|t-s|$ between them.

Definition 4.3.
For weakly stationary sequences $\{X_t\}$ *we define the* ***autocovariance function*** *(acvf)*

$$\gamma(h) = \gamma(t, t+h) = \mathrm{Cov}(X_t, X_{t+h})$$

and the ***autocorrelation function*** *(acf)*

$$\rho(h) = \gamma(h)/\gamma(0).$$

Note that the concept of stationarity (both weak and strong), as well as the notion of autocovariance and autocorrelation functions $\gamma(h)$ and $\rho(h)$ applies to the random variables $\{X_t\}$, not to a specific data set (which is a single realization of the random variables $\{X_t\}$). Thus, strictly speaking, we cannot say that a particular data set is stationary. We can only say that a particular data set is a realization of stationary time series random variables. Also note that, because $\rho(h)$ is a correlation, it follows that $|\rho(h)| \leq 1$. Also $\rho(0)$ equals 1. Any strong stationary time series with existing means and autocovariances is also weakly stationary. The other direction does not hold, in general. An exception, where weak and strong stationarity are equivalent are Gaussian processes.

Definition 4.4 (Gaussian Process)**.**
The sequence $\{X_t\}$ *is said to be a* ***Gaussian process*** *if for every choice of times* $t_1, \ldots, t_k$ *the joint distribution of* $(X_{t_1}, \ldots, X_{t_k})$ *is multivariate normal.*

Recall that $(X_{t_1}, \ldots, X_{t_k})$ is multivariate normal if and only if every **linear combination** of $(X_{t_1}, \ldots, X_{t_k})$ is univariate normal. In particular, it is much stronger than saying that each of $X_{t_1}, \ldots, X_{t_k}$ has a univariate normal distribution. The multivariate normal distributions are uniquely determined by their means and covariances. Hence, we obtain

$$\text{weak stationarity} + \text{Gaussian Process} \implies \text{Strong Stationarity.}$$

In the rest of this course, when talking about *stationarity* we will always mean *weak stationarity.*

Example 4.5 (White Noise)**.**
Let $\{X_t\}$ *be a white noise sequence as in Definition 2.1, that is, all* X_t *have mean zero, variance* σ^2 *and are pairwise uncorrelated. In particular for any two time points* t, s

$$\mathrm{Cov}(X_t, X_s) = \begin{cases} 0 & \text{if } t \neq s \\ \sigma^2 & \text{otherwise.} \end{cases}$$

Thus, any white noise is stationary with acvf

$$\gamma(h) = \begin{cases} 0 & \text{if } h \neq 0 \\ \sigma^2 & \text{if } h = 0 \end{cases}$$

and acf

$$\rho(h) = \begin{cases} 0 & \text{if } h \neq 0 \\ 1 & \text{if } h = 0 \end{cases}$$

White noise is only a very special example of a stationary time series. Stationarity allows for considerable dependence between successive random variables in the series. The only requirement is that the dependence should be constant over time.

Example 4.6 (Moving Average Process of Order 1)**.**
Given a white noise series Z_t with variance σ^2 and a number θ, set

$$X_t = Z_t + \theta Z_{t-1}.$$

This is called a moving average of order 1. The series is stationary with mean zero and acvf

$$\begin{aligned} \gamma_X(h) &= \sigma^2(1+\theta^2) \qquad \text{if } h = 0 \\ &= \theta\sigma^2 \qquad \text{if } h = 1 \\ &= 0 \qquad \text{otherwise.} \end{aligned} \tag{16}$$

As a consequence, X_{t_1} and X_{t_2} are uncorrelated whenever t_1 and t_2 are two or more time points apart. This time series has short memory.
The autocorrelation function, acf, for $\{X_t\}$ is given by

$$\rho_X(h) = \frac{\theta}{1+\theta^2}$$

for $h = 1$ and 0 for $h > 1$. What is the maximum value that $\rho_X(1)$ can take?

5 Moving average and autoregressive models

5.1 Moving average models

Definition 5.1.
*Let $\ldots, Z_{-2}, Z_{-1}, Z_0, Z_1, Z_2, \ldots$ be a double infinite white noise sequence. The **moving average model** of order q or **MA(q)** model is defined as*

$$X_t = Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2} + \ldots + \theta_q Z_{t-q},$$

where $\theta_1, \ldots, \theta_q$ with $\theta_q \neq 0$ are parameters.

Autocovariance and autocorrelation functions of an MA(q) time series: The MA(q) model can be concisely written as $X_t = \sum_{j=0}^{q} \theta_j Z_{t-j}$ where we take $\theta_0 = 1$.
The mean of X_t is clearly 0. For $h \geq 0$, the covariance between X_t and X_{t+h} is given by

$$\text{cov}(X_t, X_{t+h}) = \text{cov}(\sum_{j=0}^{q} \theta_j Z_{t-j}, \sum_{k=0}^{q} \theta_k Z_{t+h-k}) = \sum_{j=0}^{q} \sum_{k=0}^{q} \theta_j \theta_k \text{cov}(Z_{t-j}, Z_{t+h-k}).$$

Because $\{Z_t\}$ is white noise, the covariance between Z_{t-j} and Z_{t+h-k} is non-zero and equal to σ^2 if and only if $t - j = t + h - k$ i.e., if and only if $k = j + h$. But because k has to lie between 0 and q, we must have that j has to lie between 0 and $q - h$. We thus get:

$$\begin{aligned} \gamma_X(h) &= \sigma^2 \sum_{j=0}^{q-h} \theta_j \theta_{j+h} \qquad \text{if } h = 0, 1, \ldots, q \\ &= 0 \qquad \text{if } h > q. \end{aligned} \tag{17}$$

For the autocorrelation function we thus get

$$\begin{aligned} \rho_X(h) &= \frac{\sum_{j=0}^{q-h} \theta_j \theta_{j+h}}{\sum_{j=0}^{q} \theta_j^2} \qquad \text{if } h = 0, 1, \ldots, q \\ &= 0 \qquad \text{if } h > q. \end{aligned} \tag{18}$$

Note that the autocovariance and the autocorrelation functions *cut off* after lag q.
Moreover, we find that $\text{cov}(X_t, X_{t+h})$ does not depend on t and hence, we deduce the following.

Theorem 5.2.
Let $\ldots, X_{-2}, X_{-1}, X_0, X_1, X_2, \ldots$ be a time series which follows an MA(q) model. Then $\{X_t\}$ is weakly stationary.

Backshift Notation: A convenient piece of notation avoids the trouble of writing huge expressions in the sequel. Let B denote the *backshift operator* defined by

$$BX_t = X_{t-1}, B^2 X_t = X_{t-2}, B^3 X_t = X_{t-3}, \ldots$$

and similarly

$$BZ_t = Z_{t-1}, B^2 Z_t = Z_{t-2}, B^3 Z_t = Z_{t-3}, \ldots.$$

Also let I denote the identity operator: $IX_t = X_t$. More generally, we can define polynomial functions of the Backshift operator by, for example,

$$(I + B + 3B^2)X_t = IX_t + BX_t + 3B^2X_t = X_t + X_{t-1} + 3X_{t-2}.$$

In general, for every polynomial $f(z)$, we can define $f(B)$. One can even extend this notation to negative powers of B which correspond to forward shifts. For example, $B^{-1}X_t = X_{t+1}, B^{-5}X_t = X_{t+5}$ and $(B^3 + 9B^{-2})X_t = X_{t-3} + 9X_{t+2}$ etc.
For example, in this notation, the defining equation $X_t = Z_t + \theta Z_{t-1}$ for the MA(1) process can be written as $X_t = \theta(B)Z_t$ for the polynomial $\theta(z) = 1 + \theta_1 z$.

Definition 5.3.
*For parameters $\theta_1, \dots, \theta_q$ with $\theta_q \neq 0$ define the **moving average operator** of order q as*

$$\theta(B) = 1 + \theta_1 B + \dots \theta_q B^q.$$

With the moving average operator we can write the MA(q) model from Definition 5.1 as

$$X_t = \theta(B)Z_t, \tag{19}$$

for a white noise process $\{Z_t\}$.

Invertibility:

Example 5.4.
Consider the case of the MA(1) model whose acvf is given by

$$\gamma_X(0) = \sigma_Z^2(1 + \theta^2)$$
$$\gamma_X(1) = \theta\sigma_Z^2$$
$$\gamma_X(h) = 0 \text{ for all } h \geq 2.$$

*It is easy to see that for $\theta = 5, \sigma_Z^2 = 1$, we get the same acvf as for $\theta = 1/5, \sigma_Z^2 = 25$. In other words, there exist different parameter values that give the same acvf. Now assume the white noise $\{Z_t\}$ to be Gaussian. The Gaussian distribution is uniquely defined by its mean and variance, this implies that the parameter pairs (θ, σ_Z^2) and $(1/\theta, \theta^2\sigma_Z^2)$ correspond to exactly the same time series $\{X_t\}$. This implies that one **can not uniquely** estimate the parameters of an MA(1) model from data.*
Note that the two pairs (θ, σ_Z^2) and $(1/\theta, \theta^2\sigma_Z^2)$ correspond to a time-inversion of the process $\{Z_t\}$. The white noise process $Z_t, Z_{t+1}, \dots$ does not have any structure over time. Hence, if, instead we considered the time-reversed process $Z_t, Z_{t-1}, \dots$ this still corresponds to the same white noise. Thus, the process $X_t = Z_t + \theta Z_{t-1}$ has the same distribution as the process $X_t = Z_{t-1} + \theta Z_t = \tilde{Z}_t + 1/\theta \tilde{Z}_{t-1}$, where $\tilde{Z}_t = \theta Z_t$ is white noise with variance $\theta^2\sigma_Z^2$. A natural fix is to consider only those MA(1) for which $|\theta| < 1$. This condition is called invertibility. The condition $|\theta| < 1$ for the MA(1) model is equivalent to stating that the moving average polynomial $\theta(z) = 1 + \theta z$ has all roots of magnitude strictly larger than one.

For general MA(q) processes, the distribution of $\{X_t\}$ is invariant under time shifts and time inversions of the underlying white noise process $\{Z_t\}$. Thus, the parameters of an MA(q) process are not uniquely determined by the distribution of $\{X_t\}$, in general. To overcome this problem, one imposes the invertibility condition on the MA(q) process.

Definition 5.5.
*An MA(q) model $X_t = \theta(B)Z_t$ is said to be **invertible**, if $\theta(z) \neq 0$ for $|z| \leq 1$.*

Equivalently one can define the invertibility condition as follows. The proof is given in (Shumway and Stoffer, 2006, Appendix B).

Theorem 5.6.
An MA(q) model $X_t = \theta(B)Z_t$ is invertible if and only if the time series $\{X_t\}$ and the white noise $\{Z_t\}$ can be written as

$$Z_t = \pi(B)X_t = \sum_{j=0}^{\infty} \pi_j X_{t-j},$$

where $\pi(B) = \sum_{j=0}^{\infty} \pi_j B^j$ and $\sum_{j=0}^{\infty} |\pi_j| < \infty$ and $\pi_0 = 1$.

Infinite Order Moving Average: We can extend the definition of moving average processes to even infinite order by taking:

$$X_t = Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2} + \cdots + \theta_q Z_{t-q} + \theta_{q+1} Z_{t-q-1} + \ldots. \tag{20}$$

Here, once again, $\{Z_t\}$ represents white noise with mean zero and variance σ^2. We will write this expression succinctly via $X_t = \sum_{j=0}^{\infty} \theta_j Z_{t-j}$ with θ_0 taken to be 1. This is an $MA(\infty)$ model.
The right hand side in (20) is an infinite sum and hence we need to address convergence issues. A sufficient condition which ensures that the infinite sum is finite (almost surely) is $\sum_j |\theta_j| < \infty$. In this class, we will always assume this condition when talking about the infinite series $\sum_{j>0} \theta_j Z_{t-j}$. See (Shumway and Stoffer, 2006, Appendix A) for more details on how exactly this infinite sum is defined.
It turns out that $X_t = \sum_{j=0}^{\infty} \theta_j Z_{t-j}$ is a stationary process because

$$\mathrm{E}\, X_t = \mathrm{E}\left(\sum_{j=0}^{\infty} \theta_j Z_{t-j}\right) = \sum_{j=0}^{\infty} \theta_j \,\mathrm{E}\, Z_{t-j} = 0$$

and

$$\begin{aligned} Cov(X_t, X_{t+h}) &= Cov\left(\sum_{j=0}^{\infty} \theta_j Z_{t-j}, \sum_{k=0}^{\infty} \theta_k Z_{t+h-k}\right) \\ &= \sum_{j=0}^{\infty}\sum_{k=0}^{\infty} \theta_j \theta_k Cov(Z_{t-j}, Z_{t+h-k}) = \sigma^2 \sum_{j=0}^{\infty} \theta_j \theta_{j+h}. \end{aligned}$$

We could freely interchange the expectation and covariance operators above with the infinite sum because of the condition $\sum_j |\theta_j| < \infty$.

Note that the expectation $\mathrm{E}\, X_t$ and the covariance $Cov(X_t, X_{t+h})$ do not depend on t and the autocovariance is given by

$$\gamma_X(h) = \sigma^2 \sum_{j=0}^{\infty} \theta_j \theta_{j+h}. \tag{21}$$

In particular, we get the following.

Theorem 5.7.
Let $\ldots, X_{-2}, X_{-1}, X_0, X_1, X_2, \ldots$ be a time series which follows an MA(∞) model. Then $\{X_t\}$ is weakly stationary.

Fix ϕ with $|\phi| < 1$. The choice of weights $\theta_j = \phi^j$ in the $MA(\infty)$ model (20) leads to a very interesting model. With this choice of weights, the resulting process can be written as $X_t = \sum_{j=0}^{\infty} \phi^j Z_{t-j}$. Its autocovariance function is given (using the formula (21)) by

$$\gamma(h) = \sigma^2 \sum_{j=0}^{\infty} \phi^j \phi^{j+h} = \sigma^2 \phi^h \sum_{j=0}^{\infty} \phi^{2j} = \frac{\phi^h \sigma^2}{1-\phi^2} \qquad \text{for } h \geq 0$$

The autocorrelation function is therefore given by $\rho(h) = \phi^h$ for $h \geq 0$. Note that unlike the case of the Moving Average Model of Order 1, this acf is strictly non-zero for all lags. But, since $\rho(h)$ drops exponentially as lag increases, this is also taken to be an example of a stationary time series with short range dependence. Note that if ϕ is negative, the acf $\rho(h)$ osillates as h increases.
Here is an important property of this process X_t:

$$\begin{aligned} X_t = Z_t + \phi Z_{t-1} + \phi^2 Z_{t-2} + \ldots &= Z_t + \phi \left(Z_{t-1} + \phi Z_{t-2} + \phi^2 Z_{t-3} + \ldots \right) \\ &= Z_t + \phi X_{t-1} \qquad \text{for every } t = \ldots, -1, 0, 1, \ldots. \end{aligned}$$

Thus X_t satisfies the following first order *difference equation*:

$$X_t = \phi X_{t-1} + Z_t. \tag{22}$$

For this reason, $X_t = \sum_{j=0}^{\infty} \phi^j Z_{t-j}$ is called the *Stationary Autoregressive Process of order one.*

5.2 Autoregressive models

Definition 5.8.
*Let $\ldots, Z_{-2}, Z_{-1}, Z_0, Z_1, Z_2, \ldots$ be a double infinite white noise sequence. The **autoregressive model** of order p or **AR(p)** model is of the form*

$$X_t = Z_t + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p},$$

where $\phi_1, \ldots, \phi_p$ with $\phi_p \neq 0$ are parameters.

Analog to the moving average operator, we define the autoregressive operator

Definition 5.9.
*For parameters $\phi_1, \ldots, \phi_p$ with $\phi_p \neq 0$ define the **autoregressive operator** of order p as*

$$\phi(B) = 1 - \phi_1 B - \ldots \phi_p B^p.$$

With the moving average operator we can write the AR(p) model from Definition 5.8 as

$$\phi(B) X_t = Z_t, \tag{23}$$

for a white noise process $\{Z_t\}$.

AR(1) process: We will first look at AR(1) processes which satisfy the difference equation (22). We have seen in the previous section that when $|\phi| < 1$ the MA(∞) process $X_t = \sum_{j=0}^{\infty} \phi^j Z_{t-j}$ solves this difference equation (22).
Is $X_t = \sum_{j=0}^{\infty} \phi^j Z_{t-j}$ the only solution to the difference equation (22) for $|\phi| < 1$? No! Define X_0 to be an arbitrary random variable that is uncorrelated with the white noise series $\{Z_t\}$ and define $X_1, X_2, \ldots$ as well as $X_{-1}, X_{-2}, \ldots$ using the difference equation (22).
The resulting sequence surely satisfies (22). Is it stationary? NO! Because $X_{-1} = X_0/\phi - Z_0/\phi$ and since $|\phi| < 1$ and X_0 and Z_0 are uncorrelated, this would give $\text{var}(X_{-1}) > \text{var}(X_0)$ contradicting stationarity.
Even though the difference equation (22) for $|\phi| < 1$ has many solutions it has only one stationary solution.

Theorem 5.10.
For some white noise process $\{Z_t\}$ and fixed parameter $|\phi| \neq 1$ there exists exactly one time series process $\{X_t\}$ with mean zero which is stationary and solves the difference equation

$$X_t - \phi X_{t-1} = Z_t.$$

Before we prove this theorem, let us analyze what the unique stationary solution of the difference equation (22) is in a rather more heuristic way. The difference equation (22) can be rewritten as $\phi(B)X_t = Z_t$ where $\phi(B)$ is given by the polynomial $\phi(z) = 1 - \phi z$. Therefore, it is natural that the solution of this equation is

$$X_t = \frac{1}{\phi(B)} Z_t.$$

First consider $|\phi| < 1$. From the formula for the sum of a geometric series, we have

$$\frac{1}{\phi(z)} = (1 - \phi z)^{-1} = 1 + \phi z + \phi^2 z^2 + \phi^3 z^3 + \dots$$

As a result, we expect as a stationary solution

$$X_t = \frac{1}{\phi(B)} Z_t = \left(I + \phi B + \phi^2 B^2 + \dots\right) Z_t = Z_t + \phi Z_{t-1} + \phi^2 Z_{t-2} + \dots = \sum_{j=0}^{\infty} \phi^j Z_{t-j}.$$

Second consider $|\phi| > 1$. Here, we can write

$$\frac{1}{\phi(z)} = \frac{1}{1 - \phi z} = \frac{-1}{\phi z}\left(1 - \frac{1}{\phi z}\right)^{-1} = -\frac{1}{\phi z} - \frac{1}{\phi^2 z^2} - \frac{1}{\phi^3 z^3} - \dots = -\frac{z^{-1}}{\phi} - \frac{z^{-2}}{\phi^2} - \frac{z^{-3}}{\phi^3} - \dots$$

As a result, we expect as a stationary solution

$$X_t = \left(-\frac{B^{-1}}{\phi} - \frac{B^{-2}}{\phi^2} - \frac{B^{-3}}{\phi^3} - \dots\right) Z_t = -\frac{Z_{t+1}}{\phi} - \frac{Z_{t+2}}{\phi^2} - \frac{Z_{t+3}}{\phi^3} - \dots \tag{24}$$

This is indeed true and we will prove this in the following. The strange part about (24) is that X_t depends on only future white noise values: $Z_{t+1}, Z_{t+2}, \dots$. As a result, autoregressive processes of order 1 for $|\phi| > 1$ are rarely used in time series modelling.

Proof. We only present to proof for $|\phi| < 1$. The case for $|\phi| > 1$ is analog.
We have seen above that $\sum_{j=0}^{\infty} \phi^j Z_{t-j}$ is one stationary solution of the difference equation. Suppose $\{Y_t\}$ is any other stationary sequence which also satisfies (22) i.e., $Y_t = \phi Y_{t-1} + Z_t$. In that case, by successively using this equation, we obtain

$$\begin{aligned} Y_t &= Z_t + \phi Y_{t-1} \\ &= Z_t + \phi Z_{t-1} + \phi^2 Y_{t-2} \\ &= Z_t + \phi Z_{t-1} + \phi^2 Z_{t-2} + \phi^3 Y_{t-3} \\ &= Z_t + \phi Z_{t-1} + \phi^2 Z_{t-2} + \phi^3 Z_{t-3} + \phi^4 Y_{t-4}. \end{aligned}$$

In general, one would have $Y_t = \sum_{i=0}^{k} \phi^i Z_{t-i} + \phi^{k+1} Y_{t-k-1}$ for *every* k. The idea is now to let k approach ∞. The first term on the right hand side is $\sum_{i=0}^{k} \phi^i Z_{t-i}$ which we have argued (while defining X_t) converges to $X_t = \sum_{i=0}^{\infty} \phi^i Z_{t-i}$ as k goes to infinity. We need the hypothesis that $\{Y_t\}$ is stationary to deal with the second term. Indeed, because of stationarity, one has

$$\mathrm{E}\left(\phi^{k+1} Y_{t-k-1}\right)^2 = \phi^{2k+2}\,\mathrm{E}\, Y_{t-k-1}^2 = \phi^{2k+2}\,\mathrm{E}\, Y_0^2 \to 0 \qquad \text{as } k \to \infty.$$

It follows therefore that Y_t and X_t are the same. □

Finally, consider the case $|\phi| = 1$? Here the difference equation becomes $X_t - X_{t-1} = Z_t$ for $\phi = 1$ and $X_t + X_{t-1} = Z_t$ for $\phi = -1$. These difference equations have **no** stationary solutions. Let us see this for $\phi = 1$ (the $\phi = -1$ case is similar). Note that $X_t - X_0 = Z_1 + \cdots + Z_t$ which implies that the variance of $X_t - X_0$ equals $t\sigma^2$ and hence grows with t. This cannot happen if $\{X_t\}$ were stationary.
Here is the AR(1) summary:

1. If $|\phi| < 1$, the difference equation (22) has a unique stationary solution given by $X_t = \sum_{j=0}^{\infty} \phi^j Z_{t-j}$. The solution clearly only depends on the present and past values of $\{Z_t\}$. It is hence called **causal**.

2. If $|\phi| > 1$, the difference equation (22) has a unique stationary solution given by $X_t = -\sum_{j=1}^{\infty} \phi^{-j} Z_{t+j}$. This is **non-causal**.

3. If $|\phi| = 1$, no stationary solution exists.

This summary can be reinterpreted in terms of the polynomial $\phi(z) = 1 - \phi z$. The root of this polynomial is $1/\phi$.

1. If the magnitude of the root of $\phi(z)$ is strictly larger than 1, then $\phi(B)X_t = Z_t$ has a unique **causal** stationary solution.

2. If the magnitude of the root of $\phi(z)$ is strictly smaller than 1, then $\phi(B)X_t = Z_t$ has a unique stationary solution which is **non-causal**.

3. If the magnitude of the root of $\phi(z)$ is exactly equal to one, then $\phi(B)X_t = Z_t$ has no stationary solution.

Causality: Analog as for the invertiblity condition for MA(q) processes in Definition 5.5 and Theorem 5.6 one can define the causality condition for general AR(p) processes.

Definition 5.11.
*An AR(p) model $\phi(B)X_t = Z_t$ is said to be **causal**, if $\phi(z) \neq 0$ for $|z| \leq 1$.*

Analog as for invertibility in Theorem 5.6 one gets the following equivalent definition. The proof is given in (Shumway and Stoffer, 2006, Appendix B).

Theorem 5.12.
An AR(p) model $\phi(B)X_t = Z_t$ is causal if and only if the time series $\{X_t\}$ and the white noise $\{Z_t\}$ can be written as

$$X_t = \psi(B)Z_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j},$$

where $\psi(B) = \sum_{j=0}^{\infty} \psi_j B^j$ and $\sum_{j=0}^{\infty} |\psi_j| < \infty$ and $\psi_0 = 1$.

5.3 ARMA models

The combination of ideas behind the Autoregressive and Moving Average processes give us ARMA processes, which provide a large class of stationary models for modelling residuals after trend and seasonality are removed.

Definition 5.13.
Let $\ldots, Z_{-2}, Z_{-1}, Z_0, Z_1, Z_2, \ldots$ be a double infinite white noise sequence. A (zero mean) ***ARMA(p,q)*** *model is of the form*

$$\phi(B)X_t = \theta(B)Z_t \tag{25}$$

where $\phi(B)$ and $\theta(B)$ are the AR and MA operators from Definition 5.9 and 5.3 with parameters $\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q$ with $\phi_p, \theta_q \neq 0$.

Remark 5.14 (ARMA with non-zero mean)**.**
In the following, we will always assume that an ARMA process $\{X_t\}$ in Definition 5.13 has mean zero. If we want to study a stationary process with mean $\mu \neq 0$, we simply have to replace (25) by

$$(X_t - \mu) - \phi_1(X_{t-1} - \mu) - \cdots - \phi_p(X_{t-p} - \mu) = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q},$$

that is, $\{X_t - \mu\}$ is ARMA as in Definition 5.13.

Inserting the definitions of MA and AR operators we find that for an ARMA(p,q) process

$$X_t = \phi_1 X_{t-1} + \ldots + \phi_p X_{t-p} + Z_t + \theta_1 Z_{t-1} + \ldots + \theta_q Z_{t-q}.$$

This class of models includes the examples of stationary models that we have so far studied:

1. **White noise**: Corresponds to $\phi(z) = 1$ and $\theta(z) = 1$, that is ARMA(0,0).
2. **Moving Average**: Corresponds to $\phi(z) = 1$ and $\theta(z) = 1 + \theta_1 z + \theta_2 z^2 + \cdots + \theta_q z^q$, that is ARMA(0,q).
3. **Autoregressive Process**: Corresponds to $\theta(z) = 1$ and $\phi(z) = 1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p$, that is ARMA(p,0).

See (Shumway and Stoffer, 2006, Example 3.6) for the following example.

Example 5.15 (Parameter Redundancy)**.**
Consider a white noise process $X_t = Z_t$. Clearly, this satisfies the equation

$$X_t = 0.5X_{t-1} + Z_t - 0.5Z_{t-1}, \tag{26}$$

which looks like an ARMA(1,1) model, although X_t is just white noise. This is hidden because of the parameter redundancy. In operator form $\phi(B)X_t = \theta(B)Z_t$ (26) becomes

$$(1 - 0.5B)X_t = (1 - 0.5B)Z_t,$$

in particular, the polynomials $\phi(z)$ and $\theta(z)$ have a common factor, namely $(1 - 0.5z)$. Discarding the common factor in each leaves $\phi(z) = \theta(z) = 1$, which shows that the model is actually white noise. Taking parameter redundancy into account becomes crucial when we considering parameter estimation in ARMA models. As this example shows, one might fit an ARMA(1,1) model to white noise data.

The previous example shows that, for ARMA(p,q) models one should always remove any common factors of $\phi(z)$ and $\theta(z)$!
As MA and AR models are special cases of of ARMA models, the concepts of invertible and causal conditions from MA and AR models, respectively, carry over the ARMA models.

Definition 5.16.
An ARMA(p,q) model $\phi(B)X_t = \theta(B)Z_t$ is said to be

1. ***invertible*** *if $\theta(z) \neq 0$ for any $|z| \leq 1$,*
2. ***causal*** *if $\phi(z) \neq 0$ for any $|z| \leq 1$.*

Analog to Theorem 5.6, 5.10, and 5.12 we have the following theorem.

Theorem 5.17.
Let $\{Z_t\}$ be a white noise model. An ARMA(p,q) model $\phi(B)X_t = \theta(B)Z_t$

1. *has a unique **stationary solution** if and only if $\phi(z) \neq 0$ for any $|z| = 1$*
2. *is **causal** if and only if the time series $\{X_t\}$ and the white noise $\{Z_t\}$ can be written as*

$$X_t = \psi(B)Z_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j},$$

as in Theorem 5.12, where $\psi(z)$ can be determined by solving

$$\psi(z) = \sum_{j=0}^{\infty} \psi_j z^j = \frac{\theta(z)}{\phi(z)}, \quad |z| \leq 1,$$

3. *is **invertible** if and only if the time series $\{X_t\}$ and the white noise $\{Z_t\}$ can be written as*

$$Z_t = \pi(B)X_t = \sum_{j=0}^{\infty} \pi_j X_{t-j},$$

as in Theorem 5.6, where $\pi(z)$ can be determined by solving

$$\pi(z) = \sum_{j=0}^{\infty} \pi_j z^j = \frac{\phi(z)}{\theta(z)}, \quad |z| \leq 1,$$

In order to find the unique causal solution $X_t = \psi(B)W_t$ (if it exists), we need to divide the two polynomials $\theta(z)/\phi(z) = \psi(z)$. There are two different strategies to do this:

1. Because $\psi(z) = \theta(z)/\phi(z)$, we have

$$(1 - \phi_1 z - \cdots - \phi_p z^p)(\psi_0 + \psi_1 z + \dots) = 1 + \theta_1 z + \cdots + \theta_q z^q.$$

Equate the coefficients of z^j on both sides for $j = 0, 1, 2, \dots$ to get

$1 = \psi_0$, $\theta_1 = \psi_1 - \psi_0\phi_1$, $\theta_2 = \psi_2 - \psi_1\phi_1 - \psi_0\phi_2$, $\theta_3 = \psi_3 - \phi_1\psi_2 - \phi_2\psi_1 - \phi_3\psi_0$, $\dots$

2. Another way is to write $\phi(z) = (1 - a_1 z)(1 - a_2 z) \dots (1 - a_p z)$ where $1/a_1, \dots, 1/a_p$ are the (possibly complex) roots of $\phi(z)$ each satisfying $|a_i| < 1$ so that

$$\begin{aligned}\psi(z) &= \frac{\theta(z)}{\phi(z)} \\ &= \frac{\theta(z)}{(1 - a_1 z) \dots (1 - a_p z)} \\ &= \theta(z)(1 - a_1 z)^{-1} \dots (1 - a_p z)^{-1} \\ &= \theta(z)(1 + a_1 z + a_1^2 z^2 + \dots)(1 + a_2 z + a_2^2 z^2 + \dots) \dots (1 + a_p z + a_p^2 z^2 + \dots).\end{aligned}$$

   The product above can be multiplied out.

Both these techniques for determining $\psi_1, \psi_2, \dots$ can be very tedious in some cases.

Example 5.18.
Consider the following ARMA(1, 1) difference equation:

$$X_t - 0.5 X_{t-1} = Z_t + 0.4 Z_{t-1}$$

where $\{Z_t\}$ is white noise. Does this have a unique stationary solution? Is it causal? Find the solution.
The autoregressive polynomial is $\phi(z) = 1 - 0.5z$. The moving average polynomial is $\theta(z) = 1 + 0.4z$. ϕ has only one root: $z = 2$. This root has magnitude $\neq 1$; hence there exists a unique stationary solution. Moreover, the root also has magnitude > 1; hence the unique stationary solution is causal. To find the solution, we need to find $\psi(z) = \theta(z)/\phi(z)$. In other words:

$$(1 - 0.5z)(\psi_0 + \psi_1 z + \psi_2 z^2 + \dots) = (1 + 0.4z).$$

Equate coefficients of z^j on both sides.

See (Shumway and Stoffer, 2006, Example 3.7 and 3.8) for further examples.

5.4 Autocovariance of ARMA processes

From now on, we will consider causal, stationary and invertible ARMA processes as in Theorem 5.17, that is $\phi(B)X_t = \theta(B)Z_t$ with $\phi(z) \neq 0$ for any $|z| \leq 1$ and $\theta(z) \neq 0$ for any $|z| \leq 1$.
In the following, we want to present to different approaches, how to determine the acvf $\gamma(h)$ and the acf $\rho(h)$ for such a process:

5.4.1 Dividing polynomials

A causal, stationary ARMA process can be explicitly written as (recall Theorem 5.17)

$$X_t = \psi(B)Z_t = \psi_0 Z_t + \psi_1 Z_{t-1} + \psi_2 Z_{t-2} + \dots \tag{27}$$

where $\psi(z) = \theta(z)/\phi(z)$. Note that ψ_0 will always equal one. This explicit representation can be used to calculate the acvf:

$$\gamma_X(h) = \text{cov}(X_t, X_{t+h}) = \sigma_Z^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+h} \qquad \text{for } h \geq 0, \tag{28}$$

where σ_Z^2 is the variance of the white noise process Z_t. The acf (autocorrelation function) can be calculated from this via $\rho_X(h) = \gamma_X(h)/\gamma_X(0)$.
This method requires to explicit calculate the function $\psi(z)$ by dividing the two polynomials $\theta(z)$ and $\phi(z)$. Recall from the previous section, how this can be done.

Example 5.19 (MA(2) process)**.**
For any MA process we have that $\psi(z) = \theta(z)$. In particular, for the second order MA process: $X_t = Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2}$, we have $\psi_0 = 1, \psi_1 = \theta_1, \psi_2 = \theta_2$ and $\psi_j = 0$ for $j \geq 3$. The identity (28) with these ψ_js gives

$$\gamma_X(0) = \sigma_Z^2(1 + \theta_1^2 + \theta_2^2) \quad \gamma_X(1) = \sigma_Z^2\theta_1(1 + \theta_2) \quad \gamma_X(2) = \sigma_Z^2\theta_2 \quad \gamma_X(h) = 0 \text{ for } h \geq 3.$$

The corresponding autocorrelations are given by

$$\rho_X(1) = \frac{\theta_1(1 + \theta_2)}{1 + \theta_1^2 + \theta_2^2} \qquad \rho_X(2) = \frac{\theta_2}{1 + \theta_1^2 + \theta_2^2} \qquad \rho_X(h) = 0 \quad \text{for } h \geq 3.$$

Example 5.20 (AR(1) process)**.**
Let us now look at AR(1) given by $X_t - \phi X_{t-1} = Z_t$. For this to be stationary and causal, we have seen that the condition $|\phi| < 1$ is necessary and sufficient. In this case, $X_t = Z_t + \phi Z_{t-1} + \phi^2 Z_{t-2} + \dots$. This is same as (27) if we take $\psi_j = \phi^j$ for $j = 0, 1, 2, \dots$. Thus formula (28) for the autocovariance gives:

$$\gamma_X(h) = \sigma_Z^2 \sum_{j=0}^{\infty} \phi^j \phi^{j+h} = \sigma_Z^2 \phi^h \sum_{j=0}^{\infty} \phi^{2j} = \sigma_Z^2 \frac{\phi^h}{1 - \phi^2} \qquad \text{for } h > 0,$$

which results in the simple expression for the autocorrelation: $\rho_X(h) = \phi^h$ for $h > 0$.

5.4.2 Solving difference equations

The idea of the second approach is to derive difference equations for $\gamma(h)$ which then can be solved. For any ARMA(p,q) process $\{X_t\}$ with $\phi(B)X_t = \theta(B)Z_t$ and any $k \geq 0$ we have that

$$\operatorname{Cov}(\phi(B)X_t, X_{t-k}) = \operatorname{Cov}(\theta(B)Z_t, X_{t-k}). \tag{29}$$

Moreover

$$\begin{aligned} &\operatorname{Cov}(\phi(B)X_t, X_{t-k}) \\ =&\operatorname{Cov}(X_t - \phi_1 X_{t-1} - \ldots - \phi_p X_{t-p}, X_{t-k}) \\ =&\operatorname{Cov}(X_t, X_{t-k}) - \phi_1 \operatorname{Cov}(X_{t-1}, X_{t-k}) - \ldots - \phi_p \operatorname{Cov}(X_{t-p}, X_{t-k}) \\ =&\gamma_X(k) - \phi_1\gamma_X(k-1) - \cdots - \phi_p\gamma_X(k-p). \end{aligned} \tag{30}$$

To evaluate the right hand side of (29), we use the expression $X_t = \psi_0 Z_t + \psi_1 Z_{t-1} + \ldots$ (recall Theorem 5.17) to get

$$\begin{aligned} \operatorname{cov}(\theta(B)Z_t, X_{t-k}) &= \operatorname{cov}(Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q}, \psi_0 Z_{t-k} + \psi_1 Z_{t-k-1} + \ldots) \\ &= \begin{cases} (\psi_0\theta_k + \psi_1\theta_{k+1} + \cdots + \psi_{q-k}\theta_q)\,\sigma_Z^2 & \text{if } k \leq q \\ 0 & \text{if } k > q \end{cases} \end{aligned} \tag{31}$$

Equating (30) and (31) as in (29), we get for all $k \geq 0$

$$\gamma_X(k) - \phi_1\gamma_X(k-1) - \cdots - \phi_p\gamma_X(k-p) = c_k \tag{32}$$

with

$$c_k = \begin{cases} (\psi_0\theta_k + \psi_1\theta_{k+1} + \cdots + \psi_{q-k}\theta_q)\,\sigma_Z^2 & \text{for } 0 \leq k \leq q \\ 0 & \text{for } k > q. \end{cases} \tag{33}$$

The autocovariance function γ_X can be obtained by solving these equations. Note that in contrast to the previous method, here, we only need to find out $\psi_1, \ldots, \psi_q$ (recall that we always have $\psi_0 = 1$) instead of the whole sequence $\{\psi_j\}$.
How exactly do we solve this equations? Note that, $\gamma(k) = \gamma(-k)$. Hence, using equation (32) for $k = 0, \ldots, p$ we can build a linear system of $p+1$ equations which we can solve for the $p+1$ unknowns $\gamma(0), \ldots, \gamma(p)$. Then, for $k > p$ we can recursively compute as in (32)

$$\gamma_X(k) = c_k + \phi_1\gamma_X(k-1) + \cdots + \phi_p\gamma_X(k-p). \tag{34}$$

Remark 5.21 (Yules-Walker equations)**.**
For AR(p) processes we have that $q = 0$ and hence, c_k in (32) has a very simple form (recall that $\psi_0 = \theta_0 = 1$), namely

$$c_k = \begin{cases} \sigma_Z^2 & \text{if } k = 0 \\ 0 & \text{if } k > 0. \end{cases}$$

In this case, the equations in (32) are also denotes as Yules-Walker equations. We will also be useful later in the course, when we discuss estimation of the AR coefficients $\phi_1, \ldots, \phi_p$ from data.

Example 5.22 (ARMA(1,1)).
$\{X_t\}$ satisfies $X_t - \phi X_{t-1} = Z_t + \theta Z_{t-1}$. As we assume the process to be causal, we have $|\phi| < 1$. Here, $p = 1, \phi_0 = 1, \phi_1 = \phi$ and $q = 1, \theta_0 = 1, \theta_1 = \theta$.
The equations (32) *gives*

$$\gamma_X(0) - \phi\gamma_X(-1) = \gamma_X(0) - \phi\gamma_X(1) = (1 + \psi_1\theta_1)\sigma_Z^2 \qquad \text{for } k = 0,$$
$$\gamma_X(1) - \phi\gamma_X(0) = \theta\sigma_Z^2 \qquad \text{for } k = 1,$$

and

$$\gamma_X(k) = \phi\gamma_X(k-1) \qquad \text{for } k \geq 2.$$

The number ψ_1 is the coefficient of z in $(1 + \theta z)(1 + \phi z + \phi^2 z^2 + \dots)$ and equals $\theta + \phi$. Solving the first two equations, we get

$$\gamma_X(0) = \sigma_Z^2 \frac{1 + \theta^2 + 2\phi\theta}{1 - \phi^2} \quad \gamma_X(1) = \sigma_Z^2 \frac{(\theta + \phi)(1 + \theta\phi)}{1 - \phi^2}.$$

Thus, clearly the last equations gives

$$\gamma_X(k) = \phi^{k-1}\sigma_Z^2 \frac{(\theta + \phi)(1 + \theta\phi)}{1 - \phi^2}.$$

This results in the autocorrelations:

$$\rho_X(k) = \frac{(\theta + \phi)(1 + \theta\phi)}{1 + \theta^2 + 2\phi\theta}\phi^{k-1} \qquad \text{for } k \geq 1.$$

The autocorrelation at lag one is not equal to ϕ. After lag one, subsequent autocorrelations decay exponentially with factor ϕ. When $\theta = 0$, we get back the autocorrelations for AR(1).

In the previous example we have seen that although (34) just yields a recursive formula to compute the values of $\gamma_X(k)$ successively, for the ARMA(1,1) case, it also yields an explicit expression for $\gamma_X(k)$ for all $k \geq 0$. One could ask whether this is generally the case. The answer is yes! More generally, using theory about difference equations, one can show that, given the initial values $\gamma(0), \dots, \gamma(p)$ (which can be obtain as explained above), one can obtain a closed form expression for $\gamma_X(k)$ which depends on the (possibly complex) zeros of the polynomial $\phi(z)$. As an example, we will just give the result for second order difference equations. The explicit form for general difference equations is similar (see (Shumway and Stoffer, 2006, Chapter 3.3)).

Theorem 5.23.
For some constants $\alpha_1, \alpha_2 \in \mathbb{R}$ with $\alpha_2 \neq 0$ consider the second order (homogeneous) difference equation

$$u_k - \alpha_1 u_{k-1} - \alpha_2 u_{k-2} = 0 \qquad \text{for } k = 2, 3, \dots$$

with initial conditions $u_0 = b_0$ and $u_1 = b_1$. Further, let z_1, z_2 be the two (possibly complex) roots of the polynomial $f(z) = 1 - \alpha_1 z - \alpha_2 z^2$.

1. *When $z_1 \neq z_2$ and z_1, z_2 both real then*

$$u_k = c_1 z_1^{-k} + c_2 z_2^{-k},$$

where $c_1, c_2 \in \mathbb{R}$ are determined by the initial conditions

$$c_1 + c_2 = b_0 \qquad \text{and} \qquad c_1 z_1^{-1} + c_2 z_2^{-1} = b_1.$$

2. *If $z_1 = z_2$ (and hence, z_1 must be real) then*

$$u_k = z_1^{-k}\,(c_1 + c_2 k)$$

where $c_1, c_2 \in \mathbb{R}$ are determined by the initial conditions:

$$u_0 = c_1 = b_0 \quad \text{and} \quad u_1 = (c_1 + c_2) z_0^{-1} = b_1.$$

3. *If z_1 and z_2 are non-real (thus $z_2 = \bar{z}_1$), then*

$$u_k = c_1 z_1^{-k} + \bar{c}_1 \bar{z}_1^{-k}$$

where the complex number c_1 (which has a real part and an imaginary part) is determined by

$$u_0 = c_1 + \bar{c}_1 = b_0 \quad \text{and} \quad u_1 = c_1 z_1^{-1} + \bar{c}_1 \bar{z}_1^{-1} = b_1.$$

You can simply proof this theorem checking that the difference equation are fulfilled by the given solutions.

Example 5.24.
Consider the following AR(2) process

$$X_t - X_{t-1} + 0.5 X_{t-2} = Z_t$$

where $\{Z_t\}$ is white noise with variance σ_Z^2. By (32) we know that $u_k = \gamma_X(k)$ fulfills a second order difference equation as in Theorem 5.23 for $k \geq 1$ with $\alpha_1 = 1$ and $\alpha_2 = -0.5$. The polynomial $f(z) = 1 - x + 0.5z^2$ has the two roots $1 \pm i$. We can write these roots as $z_1 = \sqrt{2}\exp(i\pi/4)$ and $z_2 = \sqrt{2}\exp(-i\pi/4)$. Thus, Theorem 5.23 yields that for some complex number $c_1 = ae^{ib}$ we have

$$\begin{aligned} \gamma_X(k) &= ae^{ib}(\sqrt{2})^{-k}\exp(-ik\pi/4) + ae^{-ib}(\sqrt{2})^{-k}\exp(ik\pi/4) \\ &= (\sqrt{2})^{-k} a \left(e^{i(b-k\pi/4)} + e^{-i(b-k\pi/4)}\right) = 2(\sqrt{2})^{-k} a \cos(b - k\pi/4) \end{aligned} \tag{35}$$

for $k \geq 1$ (recall Euler's formula $e^{ib} = \cos(b) + i\sin(b)$). Finally, we can use the solution (35) for $k = 1, 2$ and solve the three equations in (32) for $k = 0, 1, 2$ to determine the three unknowns a, b and $\gamma(0)$.

5.5 Approximate distribution of sample autocorrelations

How can one decide whether ARMA(p,q) for some orders p, q is a good model for an observed time series $\{X_t\}$? Remember that for the white noise model in Section 2 we could simply plot the sample autocorrelations $r_1, \dots, r_k$ (denoted as correlogram) and check whether this looks like i.i.d. Gaussian with variance $1/n$ (recall Theorem 2.4). We can do something similar for ARMA(p,q) as the following theorem shows. The proof is not easy and can be found for example in (Brockwell and Davis, 1991, Theorem 7.2.2, see also Remark 1, 2).

Theorem 5.25 (Bartlett's formula)**.**
Under some general conditions[1] *on the white noise process $\{Z_t\}$, if $\{X_t\}$ is an causal and invertible ARMA process $\phi(B)X_t = \theta(B)Z_t$, then for any fixed lag k and n large enough the sample autocorrelations $(r_1, r_2, \dots, r_k)$ are approximately multivariate normal distributed with mean $(\rho_X(1), \dots, \rho_X(k))$ and covariance matrix W/n with (i,j)th entry equal to $W_{ij} =$*

$$\sum_{m=1}^{\infty} \left(\rho_X(m+i) + \rho_X(m-i) - 2\rho_X(i)\rho_X(m)\right)\left(\rho_X(m+j) + \rho_X(m-j) - 2\rho_X(j)\rho_X(m)\right),$$

that is,

$$\sqrt{n}\left(\begin{pmatrix} r_1 \\ \vdots \\ r_k \end{pmatrix} - \begin{pmatrix} \rho_X(1) \\ \vdots \\ \rho_X(k) \end{pmatrix}\right) \to \mathcal{N}(0, W) \qquad \text{as } n \to \infty. \tag{36}$$

Check that for white noise $X_t = Z_t$ Theorem 5.25 yields the same as Theorem 2.4.
In particular, for each individual r_i Theorem 5.25 yields that under an ARMA(p,q) model

$$\mathbf{P}(|r_i - \rho_X(i)| \geq 1.96\sqrt{W_{ii}/n}) \approx 5\%.$$

Analog, for the expected number of $r_i's$ which lie outside of the $\rho_X(i) \pm 1.96\sqrt{W_{ii}/n}$ band we find that

$$\mathrm{E}\left(\#\{i = 1, \dots, k \ : \ |r_i - \rho_X(i)| \geq 1.96\sqrt{W_{ii}/n}\}\right) \approx k \cdot 5\%.$$

Example 5.26 (MA(1) Process)**.**
Suppose $X_t = Z_t + \theta Z_{t-1}$. We have seen that $\rho_X(1) = \theta/(1+\theta^2)$ and $\rho_X(h) = 0$ for higher lags h. Bartlett's formula says that the variance of r_i is approximately W_{ii}/n where

$$W_{ii} = \sum_{m=1}^{\infty} \left(\rho(m+i) + \rho(m-i) - 2\rho(i)\rho(m)\right)^2.$$

For $i = 1$ i.e., when we consider the first order sample autocorrelation, this formula gives

$$\mathrm{Var}(r_1) \approx W_{11}/n = (1 - 3\rho^2(1) + 4\rho^4(1))/n < 1/n$$

(to see the last inequality note that for any θ we have $\rho^2(1) = \frac{\theta^2}{(1+\theta^2)^2} \leq 1/4$). In other words, r_1 for MA(1) is less variable than r_1 for white noise.

[1] IID is always sufficient, but it also works with more general moment conditions.

For higher values of i, the formula gives

$$W_{ii} = \sum_{m=1}^{\infty} \rho^2(m-i) = 1 + 2\rho^2(1) > 1.$$

In other words, r_k for $k \geq 2$ are more variable for MA(1) than for white noise. Thus we can expect to see more r_k's sticking out the horizontal blue lines for MA(1).

A general strategy to find out whether ARMA(p,q) is a good model for data is as follows:

1. Plot the correlogram $(r_1, \ldots, r_k)$.
2. Compare this with the theoretical ACF $\rho_X(h)$ of the ARMA(p,q) model.
3. Keep in mind the variability of the r_k's given by Bartlett's formula (Theorem 5.25).

For example, when the sample autocorrelation after lag q drop off and lie between the band for MA(q) given by Bartlett's formula, the MA(q) model might be appropriate.
For AR(p) models the ACF does not drop to zero for large lags. Thus, it is more difficult to choose the order of an appropriate AR process for data by looking at the sample ACF. We will later introduce the *partial autocorrelation function (PACF)*, which will be more helpful for AR(p) models: the PACF of an AR(p) model is zero for lags strictly larger than p.

5.6 Prediction

In the following, we want to study how to predict a future observation X_{n+1} from a given time series data set $X_1, \dots, X_n$. We will assume that the stationary time series $\{X_t\}$ has mean zero. Otherwise, when $\mathrm{E}(X_t) = \mu \neq 0$ we can just consider $\{X_t - \mu\}$ instead. When μ is unknown, it can be estimated by the sample mean $\hat{\mu} = \overline{X} = \sum_{i=1}^n X_i/n$.
First, we study the general problem of predicting the outcome of a random variable Y based on some other zero mean random variables $W_1, \dots, W_n$.

Theorem 5.27 (Best prediction)**.**
Let $Y, W_1, \dots, W_n$ be random variables. Then for the best mean squared error prediction $f^\star(W_1, \dots, W_n)$ of Y, that is

$$\mathrm{E}(Y - f^\star(W_1, \dots, W_m))^2 = \min_f \mathrm{E}(Y - f(W_1, \dots, W_m))^2,$$

it holds that

$$f^*(W_1, \dots, W_n) := \mathrm{E}\left(Y | W_1, \dots, W_m\right).$$

The proof can be done as an exercise (see (Shumway and Stoffer, 2006, Problem 3.14)). A problem with this best predictor, however, is that one needs to, in general, know the entire joint distribution of $Y, W_1, \dots, W_m$ in order to compute it. On the other hand, it is much easier to compute the best **linear** prediction of Y in terms of $W_1, \dots, W_m$. To this end, assume that W_i and Y all have finite second moments and let Δ denote the covariance matrix of $W = (W_1, \dots, W_n)$, which we assume to be invertible (this just exclude the situation that a linear combination of the W_i's has variance zero), that is

$$\Delta_{ij} = \mathrm{Cov}(W_i, W_j) \quad \text{and} \quad \zeta_i = \mathrm{Cov}(Y, W_i).$$

Theorem 5.28 (Best linear prediction)**.**
Let $Y, W_1, \dots, W_n$ be zero mean random variables with finite second moments. Then for the best mean squared error linear prediction $a_1 W_1 + \ldots + a_n W_n$ of Y, that is

$$\mathrm{E}(Y - (a_1^\star W_1 + \ldots + a_n^\star W_n))^2 = \min_a \mathrm{E}(Y - (a_1 W_1 + \ldots + a_n W_n))^2,$$

it holds that

$$(a_1^\star, \dots, a_n^\star)^\top = \Delta^{-1}\zeta.$$

Proof. We have that

$$\begin{aligned} F(\mathbf{a}) &:= \mathrm{E}\left(Y - a_1 W_1 - \cdots - a_n W_n\right)^2 \\ &= \mathrm{E}\left(Y - a^T W\right)^2 \\ &= \mathrm{E}\, Y^2 - 2\,\mathrm{E}((a^T W) Y) + \mathrm{E}(a^T W W^T a) \\ &= \mathrm{E}\, Y^2 - 2a^T\zeta + a^T \Delta a. \end{aligned}$$

Differentiate with respect to a and set equal to zero to get

$$-2\zeta + 2\Delta a = 0$$

or $a = \Delta^{-1}\zeta$. Therefore the best linear predictor of Y in terms of $W_1, \dots, W_n$ equals $\zeta^T \Delta^{-1} W$. □

There exists a very useful equivalent characterization of the best linear predictor in Theorem 5.28.

Theorem 5.29 (Characterization of best linear prediction)**.**
The best linear predictor $(a_1^\star, \dots, a_n^\star)^\top$ *in Theorem 5.28 is uniquely characterized by the property that*

$$\mathrm{Cov}(Y - a_1 W_1 - \cdots - a_m W_n, W_i) = 0 \qquad \text{for all } i = 1, \dots, n.$$

The proof of this theorem is an easy exercise.

Example 5.30 (Best linear prediction for $n = 1$)**.**
The special case of this for $n = 1$ *(when there is only one predictor* W_1*) may be more familiar. When* $n = 1$*, we have* $\zeta_1 = cov(Y, W_1)$ *and* $\Delta_{11} = \mathrm{Var}(W_1)$*. Thus, the best predictor or* Y *in terms of* W_1 *is*

$$\frac{\mathrm{Cov}(Y, W_1)}{\mathrm{Var}(W_1)} W_1.$$

Prediction of a stationary process Now for prediction of a stationary zero mean time series $\{X_t\}$ with ACVF $\gamma_X(h)$, we can easily apply Theorem 5.28 to obtain that the best linear predictor of X_{n+1} in terms of k previous observations $X_n, X_{n-1}, \ldots, X_{n-k+1}$ is

$$\text{predict} \quad X_{n+1} \quad \text{by} \quad (X_n, X_{n-1}, \ldots, X_{n-k+1})\Delta^{-1}\zeta$$

with

$$\begin{aligned} \Delta_{ij} &= \mathrm{Cov}(X_{n-i+1}, X_{n-j+1}) = \gamma_X(i-j) \quad \text{for } i,j = 1,\ldots,k, \\ \zeta_i &= \mathrm{Cov}(X_{n+1}, X_{n-i+1}) = \gamma_X(i) \quad \text{for } i = 1,\ldots,k. \end{aligned}$$

Note that for a stationary process, the matrix Δ has a very specific structure, namely it is of so called *Toeplitz* form, namely

$$\Delta = \begin{pmatrix} \gamma_0 & \gamma_1 & \gamma_2 & \cdots & \cdots & \gamma_{k-1} \\ \gamma_1 & \gamma_0 & \gamma_1 & \ddots & & \vdots \\ \gamma_2 & \gamma_1 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \gamma_1 & \gamma_2 \\ \vdots & & \ddots & \gamma_1 & \gamma_0 & \gamma_1 \\ \gamma_{k-1} & \cdots & \cdots & \gamma_2 & \gamma_1 & \gamma_0 \end{pmatrix}$$

Linear systems of equation $\Delta a = \zeta$, where Δ is of Toeplitz form, can be solved very efficiently with iterative algorithms (which do not invert the matrix Δ explicitly). One example is the *Durbin-Levinson algorithm*, see for example, (Shumway and Stoffer, 2006, Property 3.4)

Example 5.31 (AR(p))**.**
Consider the special case of a zero mean AR(p) model: $X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = Z_t$. It follows from causality that $X_n - \phi_1 X_{n-1} - \cdots - \phi_p X_{n-p} = Z_n$ is uncorrelated with $X_{n-1}, X_{n-2}, \ldots, X_1$. Thus, from the defining equation in Theorem 5.29 we deduce that when $n > p$ the best linear predictor of X_n in terms of $X_{n-1}, X_{n-2}, \ldots, X_1$ equals $\phi_1 X_{n-1} + \phi_2 X_{n-2} + \cdots + \phi_p X_{n-p}$.
Note that for AR(p) models, where the white noise process $\{Z_t\}$ is i.i.d. and $n > p$, the Best Linear Prediction coincides with the Best Prediction. To see this, just take the conditional expectation on both sides of the AR(p) equation.

Remark 5.32 (Prediction in ARMA models)**.**
In general, for prediction in causal ARMA(p,q) models one can consider their AR(∞) representation $X_t = -\sum_{j=1}^{\infty} \pi_j X_{t-j} + Z_t$. Then, for n large enough, this is well approximated by the AR(n) model $X_t \approx -\sum_{j=1}^{n} \pi_j X_{t-j} + Z_t$ (recall that $\sum_j |\pi_j| < \infty$ and thus $|\pi_j| \to 0$ as $n \to \infty$). Hence, the Best Linear Prediction (and for i.i.d. noise $\{Z_t\}$ also the Best Prediction) of X_{n+1} in terms of $X_n, X_{n-1}, \ldots, X_1$ is well approximated by $-\sum_{j=1}^{n} \pi_j X_{n-j+1}$. Further details on prediction, and also prediction intervals, in causal and invertible ARMA models can be found in (Shumway and Stoffer, 2006, page 107 ff.).

The Best Linear Prediction is in general worse than the Best Prediction. However, it is much easier to compute because it only requires knowledge of the covariances between the variables while the best predictor requires knowledge of the entire joint distribution. In the special case when $Y, W_1, \ldots, W_m$ are jointly Gaussian, one can show that the best prediction and best linear prediction coincide, see (Shumway and Stoffer, 2006, Theorem B.3).

5.7 Partial autocorrelation function

Definition 5.33.
Let $\{X_t\}$ be a mean zero stationary process. The Partial Autocorrelation at lag h, denoted by pacf(h) is defined as the coefficient of X_{t-h} in the best linear predictor for X_t in terms of $X_{t-1}, \dots, X_{t-h}$.

Check that $pacf(1)$ is the same as the autocorrelation at lag one, $\rho(1)$. But $pacf(h)$ for $h > 1$ can be quite different from $\rho(h)$. Recall that for an AR(p) model we have $X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = Z_t$ and hence by Theorem 5.29 we immediately get the following.

Theorem 5.34.
For the partial autocorrelation function of a causal AR(p) model $\phi(B)X_t = Z_t$ it holds that $pacf(p) = \phi_p$ and $pacf(h) = 0$ for $h > p$.

From the definition, it is not quite clear why this is called a correlation. We make this more clear in the following. Let $a_1 X_{t-1} + \cdots + a_{h-1} X_{t-h+1}$ denote the best linear predictor of X_t in terms of $X_{t-1}, \dots, X_{t-h+1}$. By stationarity, the two sequences

$$X_t, X_{t-1}, \dots, X_{t-h+1}$$

and

$$X_{t-h}, X_{t-h+1}, \dots, X_{t-1}$$

have the same covariance matrix. Therefore, the best linear prediction of X_{t-h} in terms of $X_{t-h+1}, \dots, X_{t-1}$ equals $a_1 X_{t-h+1} + \cdots + a_{h-1} X_{t-1}$. One can show that (see e.g. (Brockwell and Davis, 1991, Corollary 5.2.1))

$$\begin{aligned} &pacf(h) = \\ &\operatorname{corr}\left(X_t - a_1 X_{t-1} - \cdots - a_{h-1} X_{t-h+1}, X_{t-h} - a_1 X_{t-h+1} - \cdots - a_{h-1} X_{t-1}\right). \end{aligned} \tag{37}$$

In other words, $pacf(h)$ is the correlation between the errors in the best linear predictions of X_t and X_{t-h} in terms of the intervening variables $X_{t-1}, \dots, X_{t-h+1}$. That is, the correlation between X_t and X_{t-h} with the effect of the intervening variables $X_{t-1}, X_{t-2}, \dots, X_{t-h+1}$ removed.
The fact that $pacf(h)$ equals zero for lags $h > p$ for an AR(p) model can also be deduced from the equivalent definition in (37) of $pacf$. To see this, note that for $h > p$, the best linear predictor for X_t in terms of $X_{t-1}, \dots, X_{t-h+1}$ equals $\phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p}$. In other words, $a_1 = \phi_1, \dots, a_p = \phi_p$ and $a_i = 0$ for $i > p$. Therefore for $h > p$, we have by causality

$$\begin{aligned} pacf(h) &= \operatorname{corr}\left(X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p}, X_{t-h} - \phi_1 X_{t-h+1} - \cdots - \phi_p X_{t-h+p}\right) \\ &= \operatorname{corr}\left(Z_t, X_{t-h} - \phi_1 X_{t-h+1} - \cdots - \phi_p X_{t-h+p}\right) = 0. \end{aligned}$$

Estimating $pacf$ from data: How does one estimate $pacf(h)$ from data for different lags h? The coefficients $a_1, \dots, a_h$ of $X_{t-1}, \dots, X_{t-h}$ in the best linear predictor of X_t are obtained by solving an equation of the form $\Delta a = \zeta$. Now all the elements of Δ and ζ are of the form $\gamma_X(i-j)$ for some i and j. Therefore, a natural method of estimating $pacf(h)$ is to estimate the entries in Δ and ζ by the respective sample autocorrelations to obtain $\hat{\Delta}$

and $\hat{\zeta}$ and then to solve the equation $\hat{\Delta}\hat{a} = \hat{\zeta}$ for $\hat{a}$. Note that $pacf(h)$ is precisely a_h. When we want to check whether an AR(p) is a good fit for time series data, a natural approach is to plot the sample pacf. As the true pacf for an AR(p) model is zero for lags larger than p, the sample pacf should be close to zero for lags larger than p. Just as Theorem 5.25 for the autocorrelation function, one can quantify the variability of the pacf function precisely, as the following theorem shows (see (Brockwell and Davis, 1991, Theorem 8.1.2)

Theorem 5.35.
Let $\{X_t\}$ a causal AR(p) process with i.i.d. noise $\{Z_t\}$. Let p_k denote the sample pacf at lag k defined above. Then for $k > p$ we have that the p_k's are approximately independent normally distributed with mean zero and variance $1/n$.

Thus for $h > p$, bands at $\pm 1.96 n^{-1/2}$ can be used for checking if an AR(p) model is appropriate.
Summing up, for an $MA(q)$ model, the autocorrelation function $\rho_X(h)$ equals zero for $h > q$. Also for $h > q$, the sample autocorrelation functions r_h are approximately normal with mean 0 and variance w_{hh}/n where $w_{hh} := 1 + 2\rho^2(1) + \cdots + 2\rho^2(q)$ by Theorem 5.25. For an $AR(p)$ model, the partial autocorrelation function $pacf(h)$ equals zero for $h > p$. Also for $h > p$, the sample partial autocorrelations are approximately normal with mean 0 and variance $1/n$ by Theorem 5.35. If the sample acf for a data set cuts off at some lag, we use an MA model. If the sample pacf cuts off at some lag, we use an AR model. What if neither of the above happens? In principle this is a model selection problem. We will study general strategies on this later in the course.

5.8 Parameter estimation

In the following we want to discuss the problem of fitting an ARMA(p,q) model to data. That is, for a given time series data, assuming that it was generated by an ARMA(p,q) model, how can we estimate the parameters $\theta_1, \dots, \theta_q$ and $\phi_1, \dots, \phi_p$. We will look at three different methods: Method of moments (Yule-Walker), least squares (LS), and Maximum Likelihood (MLE). These methods are best understood, if we first look at the more simple AR(p) model.

5.8.1 Parameter estimation in AR(p) models

Assume our given data $x_1, \dots, x_n$ was generated by a causal AR(p) model with mean μ, that is,

$$(X_t - \mu) - \phi_1(X_{t-1} - \mu) - \cdots - \phi_p(X_{t-p} - \mu) = Z_t.$$

with a white noise process $\{Z_t\}$ with variance σ_Z^2. We are interested in finding estimates $\hat{\mu}, \hat{\phi}_1, \dots, \hat{\phi}_p, \hat{\sigma}_Z^2$ the parameters $\mu, \phi_1, \dots, \phi_p, \sigma_Z^2$.

Method of Moments or Yule-Walker Method For all t we have that $\mathrm{E}(X_t) = \mu$. Therefore, the method of moments simply estimates μ by the sample mean:

$$\hat{\mu} = \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

For estimating the other parameters $\phi_1, \dots, \phi_p$ and and σ_Z^2, recall the Yule-Walker equations from Remark 5.21

$$\gamma_X(0) - \phi_1\gamma_X(1) - \cdots - \phi_p\gamma_X(p) = \sigma_Z^2, \tag{38}$$

and

$$\gamma_X(k) - \phi_1\gamma_X(k-1) - \cdots - \phi_p\gamma_X(k-p) = 0 \qquad \text{for } k \geq 1. \tag{39}$$

Previously, we considered solving these equations to write $\gamma_X(k)$ in terms of σ_Z^2 and $\phi_1, \dots, \phi_p$. But these same equations can be used to estimate σ_Z^2 and $\phi_1, \dots, \phi_p$ from the data $x_1, \dots, x_n$:

Definition 5.36 (Yule-Walker estimates)**.**
The Yule-Walker estimates $\hat{\phi}_1, \dots, \hat{\phi}_p, \hat{\sigma}_Z^2$ for the parameters $\phi_1, \dots, \phi_p, \sigma_Z^2$ in an AR(p) model are obtained by

1. *estimate the autocovariances $\gamma_X(h)$ by the sample autocovariances $\hat{\gamma}_X(h)$.*
2. *and solve* (38) *and* (39) *for the unknown parameters σ_Z^2 and $\phi_1, \dots, \phi_p$.*

Note that in Definition 5.36 we have an infinite set of equations in (39), but we only need to estimate $p+1$ parameters. So we will only use (38) and the first p of the equations in (39). This gives us $p+1$ equations to solve for the $p+1$ unknowns $\phi_1, \dots, \phi_p$ and σ_Z^2. Essentially, one is trying to find an AR(p) model whose autocovariance function equals the observed sample autocovariance function at lags $0, 1, \dots, p$. This is why this method is called the method of moments.

Example 5.37 (AR(1))**.**
For $p = 1$ i.e., the AR(1) case, we just have the two equations:

$$\hat{\gamma}_X(0) - \phi\hat{\gamma}_X(1) = \sigma_Z^2 \quad \textit{and} \quad \hat{\gamma}_X(1) = \phi\hat{\gamma}_X(0).$$

This of course gives

$$\hat{\phi} = \frac{\hat{\gamma}_X(1)}{\hat{\gamma}_X(0)} = r_1 \qquad \textit{and} \qquad \hat{\sigma_Z^2} := \hat{\gamma}_X(0)\left(1 - r_1^2\right).$$

Example 5.38 (AR(2))**.**
When $p = 2$ i.e., AR(2), we get the three equations:

$$\hat{\gamma}_X(0) - \phi_1\hat{\gamma}_X(1) - \phi_2\hat{\gamma}_X(2) = \sigma_Z^2$$

and

$$\hat{\gamma}_X(1) - \phi_1\hat{\gamma}_X(0) - \phi_2\hat{\gamma}_X(1) = 0 \quad \textit{and} \quad \hat{\gamma}_X(2) - \phi_1\hat{\gamma}_X(1) - \phi_2\hat{\gamma}_X(0) = 0$$

The last two equations can used to solve for ϕ_1 and ϕ_2 to yield:

$$\hat{\phi}_1 = \frac{r_1(1 - r_2)}{1 - r_1^2} \quad \textit{and} \quad \hat{\phi}_2 = \frac{r_2 - r_1^2}{1 - r_1^2}.$$

Plugging these values for ϕ_1 and ϕ_2 into $\hat{\gamma}_X(0) - \phi_1\hat{\gamma}_X(1) - \phi_2\hat{\gamma}_X(2) = \sigma_Z^2$, we get an estimate for σ_Z^2.

Conditional least squares

Definition 5.39 (Conditional least squares estimates)**.**
The conditional least squares estimates for the parameters $\mu, \phi_1, \ldots, \phi_p$ in an AR(p) model are obtained by minimizing

$$S_c(\phi, \mu) = \sum_{i=p+1}^{n} \left(x_i - \mu - \phi_1(x_{i-1} - \mu) - \ldots - \phi_p(x_{i-p} - \mu)\right)^2 . \tag{40}$$

The variance σ_Z^2 is then estimated as

$$\hat{\sigma}_Z^2 = \frac{1}{n-p} S_c(\hat{\phi}, \hat{\mu}).$$

As we will see later, this is called conditional least squares estimation because this minimization arises when one maximizes the conditional likelihood of $x_{p+1}, \ldots, x_n$ given $x_1, \ldots, x_p$ under the iid gaussian assumption on $\{Z_t\}$.

Example 5.40 (AR(1))**.**
Consider the case $p = 1$. To minimize (40)*, let $\beta_0 = \mu(1-\phi)$ and $\beta_1 = \phi$ and rewrite it as*

$$\sum_{i=2}^{n} \left(x_i - \beta_0 - \beta_1 x_{i-1}\right)^2 .$$

Minimizing this now is exactly linear regression and the answers are given by

$$\hat{\beta}_1 = \frac{\sum_{i=2}^{n}(x_i - \bar{x}_{(2)})(x_{i-1} - \bar{x}_{(1)})}{\sum_{i=2}^{n}(x_{i-1} - \bar{x}_{(1)})^2}$$

where

$$\bar{x}_{(1)} := \frac{x_1 + \cdots + x_{n-1}}{n-1} \quad \text{and} \quad \bar{x}_{(2)} := \frac{x_2 + \cdots + x_n}{n-1}$$

and $\hat{\beta}_0 := \bar{x}_{(2)} - \hat{\beta}_1 \bar{x}_{(1)}$. This will give

$$\hat{\phi} = \frac{\sum_{i=2}^{n}(x_i - \bar{x}_{(2)})(x_{i-1} - \bar{x}_{(1)})}{\sum_{i=2}^{n}(x_{i-1} - \bar{x}_{(1)})^2} \quad \text{and} \quad \hat{\mu} := \frac{\bar{x}_{(2)} - \hat{\phi}\bar{x}_{(1)}}{1 - \hat{\phi}}.$$

The parameter σ_Z^2 is estimated by

$$\hat{\sigma_Z^2} := \frac{\sum_{i=2}^{n}\left(x_i - \hat{\mu} - \hat{\phi}(x_{i-1} - \hat{\mu})\right)^2}{n-1}.$$

It is easily seen that these estimates are very close to those obtained by the Yule-Walker method.

Maximum Likelihood MLE is a very general estimation technique in statistics. One just writes down the likelihood function of the observed data in terms of the unknown parameters and estimates the parameters by the maximizers of the likelihood (or its logarithm) over the unknown parameters.
To write a likelihood, we need a distribution assumption on $\{Z_t\}$. Most common assumption is that $\{Z_t\}$ are i.i.d normal with mean 0 and variance σ_Z^2. Then $(x_1, \dots, x_n)$ are distributed according to the multivariate normal distribution with mean $(\mu, \dots, \mu)$ and covariance matrix $\Gamma_n := \gamma_X(i-j)$, which has the likelihood function

$$f_{\mu,\Gamma_n}(x_1, \dots, x_n) = (2\pi)^{-n/2}|\Gamma|^{-1/2} \exp\left(-\frac{1}{2}(x-\mu)^T\Gamma^{-1}(x-\mu)\right). \tag{41}$$

Definition 5.41 (Maximum likelihood estimator)**.**
Under Gaussian noise assumption, the maximum likelihood estimator for the parameters $\mu, \phi_1, \dots, \phi_p$ in an AR(p) model are obtained by

- *writing down covariance matrix $\Gamma_n := \gamma_X(i-j)$ as a function of $\phi_1, \dots, \phi_p, \sigma_Z^2$,*

$$\Gamma_n = \Gamma_n(\phi_1, \dots, \phi_p, \sigma_Z^2)$$

- *and estimate $\mu, \phi_1, \dots, \phi_p$ by maximizing $f_{\mu,\Gamma_n(\phi_1,\dots,\phi_p,\sigma_Z^2)}(x_1, \dots, x_n)$ in (41).*

Example 5.42 (AR(1))**.**
In the AR(1) case, it is easy to simplify this likelihood. Decompose the joint density as:

$$f_{\mu,\phi,\sigma^2}(x_1, \dots, x_n) := f(x_1)f(x_2|x_1)f(x_3|x_1, x_2)\dots f(x_n|x_1, \dots, x_{n-1}).$$

Because of the Gaussian assumption on $\{Z_t\}$, it is easy to see that for $i \geq 2$, the conditional distribution of x_i given $x_1, x_2, \dots, x_{i-1}$ is normal with mean $\mu + \phi(x_{i-1} - \mu)$ and variance σ_Z^2. Moreover x_1 is distributed as a normal with mean μ and variance $\gamma(0) = \sigma_Z^2/(1-\phi^2)$. We thus get the following likelihood:

$$f_{\mu,\phi,\sigma_Z^2}(x_1, \dots, x_n) := (2\pi\sigma_Z^2)^{-n/2}(1-\phi^2)^{1/2} \exp\left(-\frac{S(\mu,\phi)}{2\sigma_Z^2}\right), \tag{42}$$

where

$$S(\mu,\phi) := (1-\phi^2)(x_1-\mu)^2 + \sum_{i=2}^{n}(x_i - \mu - \phi(x_{i-1}-\mu))^2. \tag{43}$$

This above sum of squares is called unconditional least squares. Maximizing the likelihood (42) or its logarithm results in a non-linear optimization problem. R solves it when you choose the method mle in the ar function. A compromise between maximum likelihood and the least squares technique (previous section) is to minimize the unconditional least squares $S(\mu, \phi)$. This also results in a non-linear optimization problem.

Summary: We have studied three different methods to estimate the parameters in an AR(p) model. Assuming that the order p is known, all three methods can be carried out in R by invoking the function $ar()$.

1. **Yule Walker or Method of Moments**:
   Finds the AR(p) model whose acvf equals the sample autocorrelation function at lags $0, 1, \dots, p$. Use *yw* for method in R.

2. **Conditional Least Squares**:
   Minimizes the conditional sum of squares: $\sum_{i=p+1}^{n}(x_i - \mu - \phi_1(x_{i-1} - \mu) - \cdots - \phi_p(x_{i-p} - \mu))^2$ over μ and $\phi_1, \ldots, \phi_p$. And σ^2 is achieved by the average of the squared residuals. Use *ols* for method in R. In this method, given data $x_1, \ldots, x_n$, R fits a model of the form $x_t - \bar{x} = intercept + \phi(x_{t-1} - \bar{x}) + residual$ to the data. The fitted value of intercept can be obtained by calling $\$x.intercept$. One can convert this to a model of the form $x_t = intercept + \phi x_{t-1} + residual$. Check the help page for the R function ar.ols.

3. **Maximum Likelihood**:
   Here one maximizes the likelihood function. The method is described below. The likelihood is relatively straightforward to write down but which requires an optimization routine to maximize. Use *mle* for method in R.

It is usually the case that all these three methods yield similar answers. The default method in R is Yule-Walker.

Asymptotic distribution of estimates We can make the similar behavior of the three methods discussed above more precise by studying their asymptotic distributions. Recall that an estimator $\hat{\phi}$ of a parameter ϕ is a function of the data $X_1, \ldots, X_n$, that is $\hat{\phi} = \hat{\phi}(X_1, \ldots, X_n)$. Thus, the estimator $\hat{\phi}$ is a random variable which depends on the sample size n. The following theorem gives the approximate distribution of the estimators discussed above when n is large. A proof and references are given in (Shumway and Stoffer, 2006, Appendix B.3, see also Property 3.8)

Theorem 5.43.
Assume a causal AR(p) process $\{X_t\}$ with acvf $\gamma_X(h)$ and define the $p \times p$ matrix Γ with entries $\Gamma_{ij} = \gamma_X(i-j)$. Let $\hat{\phi}$ be either the Yule-Walker from Definition 5.36, the conditional least squares estimator from Definition 5.39, or the maximum likelihood estimator as in Definition 5.41 .
Then, under some general conditions on the white noise process $\{Z_t\}$ with $\mathrm{Var}(Z_t) = \sigma_Z^2$, for n large enough, $\hat{\phi}$ is approximately multivariate normal distributed with mean $\phi = (\phi_1, \ldots, \phi_p)^\top$ and covariance matrix $n^{-1}\sigma_Z^2\Gamma^{-1}$, that is

$$\sqrt{n}(\hat{\phi} - \phi) \to \mathcal{N}(0, \sigma_Z^2\Gamma^{-1}) \quad \text{as } n \to \infty.$$

Proof sketch. Assume $\mu = 0$ for simplicity. It is easiest to work with the conditional least squares estimates. The AR(p) model is:

$$X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + Z_t.$$

We may write this model in matrix notation as:

$$X_t = \mathbb{X}_{t-1}^T \phi + Z_t$$

where $\mathbb{X}_{t-1}$ is the $p \times 1$ vector $\mathbb{X}_{t-1} = (X_{t-1}, X_{t-2}, \ldots, X_{t-p})^T$ and ϕ is the $p \times 1$ vector $(\phi_1, \ldots, \phi_p)^T$. The conditional least squares method minimizes the sum of squares:

$$\sum_{t=p+1}^{n} \left(X_t - \phi^T \mathbb{X}_{t-1}\right)^2$$

with respect to ϕ. The solution is:

$$\hat{\phi} = \left(\sum_{t=p+1}^{n} \mathbb{X}_{t-1}\mathbb{X}_{t-1}^T\right)^{-1} \left(\sum_{t=p+1}^{n} \mathbb{X}_{t-1} X_t\right).$$

Writing $X_t = \mathbb{X}_{t-1}^T \phi + Z_t$, we get

$$\hat{\phi} = \phi + \left(\sum_{t=p+1}^{n} \mathbb{X}_{t-1}\mathbb{X}_{t-1}^T\right)^{-1} \left(\sum_{t=p+1}^{n} \mathbb{X}_{t-1} Z_t\right).$$

As a result,

$$\sqrt{n}\left(\hat{\phi} - \phi\right) = \left(\frac{1}{n}\sum_{t=p+1}^{n} \mathbb{X}_{t-1}\mathbb{X}_{t-1}^T\right)^{-1} \left(\frac{1}{\sqrt{n}}\sum_{t=p+1}^{n} \mathbb{X}_{t-1} Z_t\right). \tag{44}$$

The following assertions are intuitive (note that $\mathbb{X}_{t-1}$ and Z_t are uncorrelated and hence independent under the gaussian assumption) and can be proved rigorously:

$$\frac{1}{n}\sum_{t=p+1}^{n} \mathbb{X}_{t-1}\mathbb{X}_{t-1}^T \to \Gamma_p \qquad \text{as } n \to \infty \text{ in probability}$$

and

$$\frac{1}{\sqrt{n}}\sum_{t=p+1}^{n} \mathbb{X}_{t-1}Z_t \to N\left(0, \sigma_Z^2\Gamma_p\right) \qquad \text{as } n \to \infty \text{ in distribution.}$$

These results can be combined with the expression (44) to prove that $\sqrt{n}\left(\hat{\phi} - \phi\right)$ converges in distribution to a normal distribution with mean 0 and variance covariance matrix $\sigma_Z^2\Gamma_p^{-1}$. □

Example 5.44 (AR(1))**.**
In the AR(1) case:

$$\Gamma_p = \Gamma_1 = \gamma_X(0) = \sigma_Z^2/(1-\phi^2).$$

Thus $\hat{\phi}$ is approximately normal with mean ϕ and variance $(1-\phi^2)/n$.

Example 5.45 (AR(2))**.**
For AR(2), using

$$\gamma_X(0) = \frac{1-\phi_2}{1+\phi_2}\frac{\sigma_Z^2}{(1-\phi_2)^2-\phi_1^2} \qquad \textit{and} \qquad \rho_X(1) = \frac{\phi_1}{1-\phi_2},$$

we can show that $(\hat{\phi}_1, \hat{\phi}_2)$ is approximately normal with mean (ϕ_1, ϕ_2) and covariance matrix is $1/n$ times

$$\begin{pmatrix} 1-\phi_2^2 & -\phi_1(1+\phi_2) \\ -\phi_1(1+\phi_2) & 1-\phi_2^2 \end{pmatrix}$$

Note that the approximate variances of both $\hat{\phi}_1$ and $\hat{\phi}_2$ are the same. Observe that if we fit AR(2) model to a dataset that comes from AR(1), then the estimate of $\hat{\phi}_1$ might not change much but the standard error will be higher. We lose precision.

5.8.2 Parameter estimation in ARMA models

Method of Moments or Yule-Walker Method The process, in principle, of solving some subset of equations for the unknown parameters $\theta_1, \ldots, \theta_q, \phi_1, \ldots, \phi_p$ and σ_Z^2 (and μ is estimated by the sample mean), by plugging in the sample acvf $\hat{\gamma}(k)$ as an estimate for the true acvf $\gamma(k)$, such as

$$\hat{\gamma}(k) - \phi_1\hat{\gamma}(k-1) - \cdots - \phi_p\hat{\gamma}(k-p) = (\psi_0\theta_k + \psi_1\theta_{k+1} + \cdots + \psi_{q-k}\theta_q)\sigma_Z^2$$

for $0 \leq k \leq q$ and

$$\hat{\gamma}(k) - \phi_1\hat{\gamma}(k-1) - \cdots - \phi_p\hat{\gamma}(k-p) = 0 \qquad \text{for } k > q$$

can in principle be applied for ARMA(p,q) models, as well. Note that ψ_j above are functions of $\theta_1, \ldots, \theta_q$ and $\phi_1, \ldots, \phi_p$.

Example 5.46 (MA(1))**.**
Consider an invertible MA(1) model $X_t = Z_t + \theta Z_t$. Here we have that

$$\gamma_X(0) = \sigma_Z^2(1+\theta^2) \quad \text{and } \gamma_X(1) = \sigma_Z^2\theta.$$

Thus, with the method of moments one would estimate θ by solving

$$r_1 = \hat{\rho}(1) = \frac{\hat{\gamma}(1)}{\hat{\gamma}(0)} = \frac{\hat{\theta}}{(1+\hat{\theta}^2)}$$

(two solutions exists, so we would pick the invertible one). The problem with this estimator is, that the above equation only has a solution when $|r_1| \leq 1/2$. Although $|\rho(1)| \leq 1/2$, because $\hat{\rho}(1)$ is just an estimate, it does not always hold true that $|r_1| \leq 1/2$. Try out some simulations in R with $|\rho(1)| \approx 1/2$ and check that $|r_1| > 1/2$ quite often.

In general, the method of moments for ARMA(p,q) models, has two major problems:

1. It is cumbersome (unless we are in the pure AR case): Solutions might not always exist to these equations (see Example 5.46).The parameters are estimated in an arbitrary fashion when these equations do not have a solution.

2. The estimators obtained are *inefficient*. The other techniques below give much better estimates (smaller standard errors).

Because of these problems, no one uses method of moments for estimating the parameters of a general ARMA model. R does not even have a function for doing this. Note, however, that both of these problems disappear for the case of the pure AR model.

Conditional least squares We start our discussion by looking at two examples.

Example 5.47 (MA(1))**.**
For an MA(1) model we have $X_t - \mu = Z_t + \theta Z_{t-1}$. We want to fit this model to data $x_1, \ldots, x_n$. If the data were indeed generated from this model, then

$$Z_1 = x_1 - \mu - \theta Z_0; Z_2 = x_2 - \mu - \theta Z_1; \ldots; Z_n = x_n - \mu - \theta Z_{n-1}.$$

If we set Z_0 to its mean 0, then for every fixed values of θ and μ, we can recursively calculate $Z_1, \ldots, Z_n$. We can then compute the sum of squares $\sum_{i=1}^n Z_i^2$. This value would change for different values of θ. We would then choose the value of θ for which it is small (this is accomplished by an optimization procedure). This is also called conditional least squares because this minimzation is obtained when one tries to maximize the conditional likelihood of the data conditioning on $z_0 = 0$.

Example 5.48 (ARMA(1,1))**.**
Here the model is $X_t - \mu - \phi(X_{t-1} - \mu) = Z_t + \theta Z_{t-1}$. Here it is convenient to set Z_1 to be zero. Then we can write

$$Z_2 = x_2 - \mu - \phi(x_1 - \mu); Z_3 = x_3 - \mu - \phi(x_2 - \mu) - \theta Z_2; \ldots; Z_n = x_n - \mu - \phi(x_{n-1} - \mu) - \theta Z_{n-1}.$$

After this, one forms the sum of squares $\sum_{i=2}^n Z_i^2$ which can be computed for every fixed values of θ, ϕ and μ. One then minimizes these resulting sum of squares over different values of the unknown parameters.

This leads to the following general definition.

Definition 5.49 (Conditional least squares for ARMA(p,q))**.**
Given some data $x_1, \ldots, x_n$ and $p, q \in \mathbb{N}$, define a function $S_c(\mu, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q)$ as follows:

1. *Set $Z_t = 0$ for all $t \leq p$.*
2. *Calculate recursively for $t = p+1, \ldots, n$*
$$Z_t = X_t - \mu - \phi_1(X_{t-1} - \mu) - \cdots - \phi_p(X_{t-p} - \mu) - \theta_1 Z_{t-1} - \cdots - \theta_q Z_{t-q}.$$
3. *Let $S_c(\mu, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q) = \sum_{t=p+1}^n Z_t^2$.*

Then the conditional last squares estimator $\hat{\mu}, \hat{\phi}_1, \ldots, \hat{\phi}_p, \hat{\theta}_1, \ldots, \hat{\theta}_q$ is defined by minimizing the conditional sum of squares

$$S_c(\hat{\mu}, \hat{\phi}_1, \ldots, \hat{\phi}_p, \hat{\theta}_1, \ldots, \hat{\theta}_q) = \min_{\mu, \phi, \theta} S_c(\mu, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q)$$

This is equivalent to writing the likelihood conditioning on $X_1, \ldots, X_p$ and $Z_t = 0$ for $t \leq p$. If $q = 0$ (AR models), minimizing the sum of squares is equivalent to linear regression and no iterative technique is needed. If $q > 0$, the problem becomes nonlinear regression and numerical optimization routines need to be used. In R, this method is performed by calling the function *arima*() with the method argument set to *CSS* (CSS stands for conditional sum of squares). As before, we can estimate the noise variance via

$$\hat{\sigma}_Z^2 = \frac{S_c(\hat{\mu}, \hat{\phi}, \hat{\theta})}{n - p}.$$

Maximum Likelihood This method is simple in principle. Assume that errors $\{Z_t\}$ are Gaussian. Write down the likelihood of the observed data $x_1, x_2, \ldots, x_n$ in terms of the unknown parameter values $\mu, \theta_1, \ldots, \theta_q$, $\phi_1, \ldots, \phi_p$ and σ_Z^2. Maximize over these unknown parameter values.
It is achieved in R by calling the function *arima*() with the method argument set to *ML* or *CSS-ML*. ML stands of course for Maximum Likelihood. R uses an optimization routine to maximize the likelihood. This routine is iterative and needs suitable initial values of the parameters to start. In CSS-ML, R selects these starting values by CSS.

Asymptotic distribution One can derive the asymptotic distribution of the maximum likelihood estimator, as well as for the conditional least squares estimator in Definition 5.49, see (Shumway and Stoffer, 2006, Property 3.10). Both estimator have the same asymptotic normal distribution, which is optimal, see (Shumway and Stoffer, 2006, Appendix B.3) for details.

5.9 Extensions of ARMA models

5.9.1 ARIMA models

ARIMA is essentially differencing plus ARMA. We have seen previously in Chapter 3, in particular, Chapter 3.1.3, that differencing can be used on time series data to remove trends and seasonality. For example, differencing can be used for

1. Removing polynomial trends: Suppose the data come from the model $Y_t = \mu_t + X_t$ where μ_t is a polynomial of order k and X_t is stationary, then differencing of order k: $\nabla^k Y_t = (I - B)^k Y_t$ results in stationary data to which an ARMA model can be fit.

2. Random walk models: Suppose that the data come from the random walk model: $Y_t = Y_{t-1} + X_t$ where X_t is stationary. Then $\nabla Y_t = X_t$ is stationary and an ARMA model can be fit to this difference data.

Such models, which after appropriate differencing, reduce to ARMA models are called ARIMA models.

Definition 5.50 (ARIMA)**.**
A process Y_t is said to be ARIMA(p, d, q) if $X_t = (I - B)^d Y_t$ is ARMA(p, q) with mean μ. In other words:

$$\phi(B)(X_t - \mu) = \theta(B)Z_t,$$

where $\{Z_t\}$ is white noise.

For fitting an ARIMA model in R one can employ the function *arima(dataset, order = c(p, d, q))*. The arima function gives you estimates of μ (under the name intercept), $\phi_1, \ldots, \phi_p$ and $\theta_1, \ldots, \theta_q$. It will also give you the estimated standard errors. An estimate of σ^2 is also provided. Just as for ARMA model (recall Chapter 5.6) one can compute predictions with the R function *predict*, see *help(predict.Arima)*.

5.9.2 Seasonal ARMA models

Definition 5.51 (Seasonal ARMA)**.**
The doubly infinite sequence $\{X_t\}$ is said to be a seasonal ARMA(P, Q) process with period s if it is stationary and if it satisfies the difference equation $\Phi(B^s)X_t = \Theta(B^s)Z_t$ where $\{Z_t\}$ is white noise and

$$\Phi(B^s) = 1 - \Phi_1 B^s - \Phi_2 B^{2s} - \cdots - \Phi_P B^{Ps}$$

and

$$\Theta(B^s) = 1 + \Theta_1 B^s + \Theta_2 B^{2s} + \cdots + \Theta_Q B^{Qs}.$$

Note that these can also be viewed as ARMA(Ps, Qs) models. However note that these models have $P + Q + 1$ (the 1 is for σ^2) parameters while a general ARMA(Ps, Qs) model will have $Ps + Qs + 1$ parameters. So these are much sparser models.
Unique stationary solution exists to $\Phi(B^s)X_t = \Theta(B^s)Z_t$ if and only if every root of $\Phi(z^s)$ has magnitude different from one. Causal stationary solution exists if and only if every root of $\Phi(z^s)$ has magnitude strictly larger than one. Invertible stationary solution exists if and only if every root of $\Theta(z^s)$ has magnitude strictly larger than one.
The ACF and PACF of these models are non-zero only at the seasonal lags $h = 0, s, 2s, \ldots$. At these seasonal lags, the ACF and PACF of these models behave just as the case of the unseasonal ARMA model: $\Phi(B)X_t = \Theta(B)Z_t$.

5.9.3 Multiplicative seasonal ARMA models

Sometimes it is useful to combine ARMA and seasonal ARMA (my multiplication) to obtain models with desirable properties of their acf functions.

Definition 5.52 (MSARMA)**.**
The Multiplicative Seasonal Autoregressive Moving Average Model ARMA(p, q) × (P, Q)$_s$ is defined as the stationary solution to the difference equation:

$$\Phi(B^s)\phi(B)X_t = \Theta(B^s)\theta(B)Z_t,$$

for some white noise process $\{Z_t\}$.

Example 5.53.
In the co2 dataset in R, for the first and seasonal differenced data (with period 12), we observe sample autocorrelations which are significantly non-zero at lags 1, 11, 12 and 13, only. We can use a MA(13) model to this data but that will have 14 parameters and therefore will likely overfit the data. We can get a much more parsimonious model for this dataset by combining the MA(1) model with a seasonal MA(1) model of period 12. Specifically, consider the model ARMA(0,1) × $(0,1)_{12}$ model

$$X_t = (1 + \Theta B^{12})(1 + \theta B)Z_t.$$

This model has the autocorrelation function:

$$\rho_x(1) = \frac{\theta}{1+\theta^2} \quad \textit{and} \quad \rho_X(12) = \frac{\Theta}{1+\Theta^2}$$

and

$$\rho_X(11) = \rho_X(13) = \frac{\theta\Theta}{(1+\theta^2)(1+\Theta^2)}.$$

At every other lag, the autocorrelation $\rho_X(h)$ equals zero. This is therefore a suitable model for the first and seasonal differenced data in the co2 dataset.

In general, when you get a stationary dataset whose correlogram shows interesting patterns at seasonal lags, consider using a multiplicative seasonal ARMA model. You may use the R function *ARMAacf* to understand the autocorrelation and partial autocorrelation functions of these models.

5.9.4 SARIMA models

Finally, we can combine differencing and differencing with multiplicative seasonal ARMA models.

Definition 5.54 (SARIMA)**.**
A process Y_t is said to be ARIMA(p, d, q) × (P, D, Q)$_s$, if after differencing d times and seasonal differencing D times, it follows a multiplicative seasonal ARMA model, that is, if it satisfies the difference equation:

$$\Phi(B^s)\phi(B)\nabla_s^D\nabla^d Y_t = \delta + \Theta(B^s)\theta(B)Z_t.$$

Recall that $\nabla_s^d = (1-B^s)^d$ and $\nabla^d = (1-B)^d$ denote the differencing operators. In R this model can be fit to the data by using the function *arima*() with the *seasonal* argument.

5.10 Model diagnostics and selection

In the previous sections we have learned about various models for time series data, how to estimate their parameters, and how to make forecasts based on them. However, a crucial question is, how one should decide which model to use. In this section we will discuss a few common techniques which help to decide whether a specific model is appropriate for a given data set.

Ljung-Box-Pierce test Assume that the data $X_1, \dots, X_n$ is generated from an invertible ARMA(p,q) model with parameters ϕ_i, θ_i. By invertibility, we can write

$$X_t = -\sum_{j \geq 0} \pi_j X_{t-j} + Z_t$$

and hence, it is easy to check, that the best linear prediction of X_t based on $X_{t-1}, X_{t-2}, \dots$ is given by

$$\hat{X}_t(\phi, \theta) = -\sum_{j \geq 0} \pi_j X_{t-j}.$$

Consequently, the residuals

$$R_t = \hat{X}_t(\phi, \theta) - X_t = Z_t$$

coincide with the white noise process $\{Z_t\}$. From Theorem 2.4 we thus know that the sample acf of the residuals R_t, denoted by $r_1, \dots, r_k$ for some maximal lag k, are approximately i.i.d. normal distributed with mean zero and variance $1/n$. Thus, the statistic

$$Q = n \sum_{i=1}^{k} r_i^2 \sim \chi_k^2$$

follows a chi-square distribution with k degrees of freedom. When the true parameters ϕ_i, θ_i are replaced by appropriate estimates $\hat{\phi}_i, \hat{\theta}_i$, the respective estimated residuals

$$\hat{R}_t = \hat{X}_t(\hat{\phi}, \hat{\theta}) - X_t$$

should still be approximately white noise[2]. One can check this, by plotting the correlogram, that is, the sample acf $\hat{r}_1, \dots, \hat{r}_k$ of the the estimated residuals $\hat{R}_1, \dots, \hat{R}_n$.
A more formal analysis can be performed by looking at the test statistic

$$\hat{Q} = n \sum_{i=1}^{k} \hat{r}_i^2,$$

which is denoted as *Box-Pierce* test statistic. Under an ARMA(p,q) model, one can show that for n large enough $\hat{Q}$ is approximately chi-square distributed with $k - p - q$ degrees of freedom

$$\hat{Q} \to \chi_{k-p-q}^2 \quad \text{for } n \to \infty.$$

In practice, one often considers a slightly modified version of the statistic $\hat{Q}$, namely

$$\tilde{Q} = n(n+2) \sum_{i=1}^{k} \frac{\hat{r}_i^2}{n-i},$$

[2]For their precise asymptotic distribution see (Brockwell and Davis, 1991, Chapter 9.4, page 308)

which is denoted as the *Ljung-Box-Pierce* test statistic. See (Brockwell and Davis, 1991, Chapter 9.4) for details. This yields the following test.

Definition 5.55 (Ljung-Box-Pierce test)**.**
Fix a maximum lag k (typically $k = 20$). Reject the hypothesis that data $x_1, \dots, x_n$ was generated from a causal and invertible ARMA(p,q) model if

$$\tilde{Q}(x_1, \dots, x_n) > q_{1-\alpha},$$

where $q_{1-\alpha}$ denotes the $(1-\alpha)$-quantile of the χ^2 distribution with $k-p-q$ degrees of freedom.

Akaike's information criterion The Ljung-Box-Pierce test and analysis of the estimated residuals, discussed in the previous paragraph, provide a strategy how to evaluate for given p, q whether or not an ARMA(p,q) model is appropriate for data $x_1, \dots, x_n$. But how should we choose the parameters p and q in the first place? Clearly every ARMA(p,q) model can be arbitrary well approximated by an ARMA(p', q') model with $p' > p$ and $q' > q$, as one can just just choose the values $\phi_{p+1}, \dots, \phi_{p'}$ and $\theta_{q+1}, \dots, \theta_{q'}$ arbitrary small. This is analog to fitting a polynomial in linear regression. The large we chose the degree of the polynomial, the better the fit. In the extreme case, when we have a 100 observations and fit a polynomial of degree 99 the fit will be perfect. However, the the model might be useless to predict future values. Generally, this is a problem of model selection: one wants the number of model parameters to be large enough, so that it can fit the data well. At the same time the number of model parameters should not be too large, which would result in overfitting. The Akaike information criterion (AIC) is a general strategy to chose a reasonable model size. It is used in various areas, not just in time series analysis. It is a model selection criterion that recommends choosing a model for which:

$$AIC = -2\log(\text{maximum likelihood}) + 2k$$

is the smallest. Here k denotes the number of parameters in the model. For example, in the case of an ARMA(p, q) model with a non-zero mean μ and noise variance σ^2, we have $k = p + q + 2$. The first term in the definition of AIC measures the fit of the model i.e., the model performance on the given data set. The term $2k$ serves as a penalty function which penalizes models with too many parameters. When you fit a model in R with the *arima()* function it will also output a value *aic*. So if you want to choose between a collection of different models, that is values for p and q you can chose the one where the AIC value is smallest.
There are various related criteria. For example the BIC (Bayesian Information Criterion) suggests choosing the model that minimizes

$$BIC = -2\log(\text{maximum likelihood}) + k\log n.$$

Note that the penalty above is larger than that of AIC. Consequently, BIC selects more parsimonious models compared to AIC. Another criteria is AICc (bias corrected AIC), which tends to perform better for small sample sizes. It suggests to minimize

$$AICc = AIC + \frac{2k(k+1)}{n-k-1}.$$

Similar as BIC, AICc selects more parsimonious models compared to AIC, but the difference between AIC and AICc vanishes as n increases.

Cross Validation Cross validation is another very popular technique for model selection and choice of tuning parameters, in general. The two articles on Rob Hyndman's blog[3] give an introduction to cross validation in general and for cross validation specific to time series.
There are many ways to do cross validation for time series. Suppose we have monthly data for m years $x_1, \ldots, x_n$ where $n = 12m$ and the objective is to predict the data for the next year. Suppose we have ℓ competing models $M_1, \ldots, M_\ell$ for the dataset. We can use cross-validation in order to pick one of these models in the following way:

1. Fix a model M_i. Fix $k < m$.
2. Fit the model M_i to the data from the first k years.
3. Using the fitted model, predict the data for the $(k+1)$st year.
4. Calculate the sum of squares of errors of prediction for the $(k+1)$st year.
5. Repeat these steps for $k = k_0, \ldots, m-1$ where k_0 is an arbitrary value of your choice.
6. Average the sum of squares of errors of prediction over $k = k_0, \ldots, m-1$. Denote this value by CV_i and call it the Cross Validation score of model M_i.
7. Calculate CV_i for each $i = 1, \ldots, \ell$ and choose the model with the smallest Cross-Validation score.

[3] See `http://robjhyndman.com/hyndsight/crossvalidation/` and `http://robjhyndman.com/hyndsight/tscvexample/`.

6 Frequency domain analysis of time series

In Section 5 we studied models for stationary processes which were directly constructed via the relationship of observations X_t at different time points. This is denoted as a *time domain approach*. In this section we will study a different approach for analyzing and modeling stationary processes. Namely, we study a stationary process as a composition of periodic components with different frequencies. This is quite natural for many time series data, which are often directly driven by periodic random events. One example is speech data, where different frequencies correspond to the respective opening and closing of the glottis.

6.1 Periodogram

In order to define the rate at which a process oscillates we first need to define basic periodic function.

Definition 6.1 (Sinusoids)**.**
We define the set of sinusoid functions as

$$\{g(t) = R\cos(2\pi ft + \Phi) \ : \ R \in \mathbb{R}_+, \ f \in \mathbb{R}_+, \ \Phi \in [0, 2\pi/f)\}, \tag{45}$$

*where R is called the **amplitude**, f is called the **frequency** and Φ is called the **phase**. The quantity $1/f$ is called the **period** and $2\pi f$ is termed the **angular frequency**.*

One can also rewrite the sinusoids in different ways:

1. With $A = R\cos(\Phi)$ and $B = -R\sin(\Phi)$ one can rewrite (45) as

$$\{g(t) = A\cos(2\pi ft) + B\sin(2\pi ft) \ : \ A, B \in \mathbb{R}, \ f \in \mathbb{R}_+\}. \tag{46}$$

2. Note that

$$\exp(2\pi ift) = \cos(2\pi ft) + i\sin(2\pi ft)$$

and

$$\cos(2\pi ft) = \frac{\exp(2\pi ift) + \exp(-2\pi ift)}{2} \text{ and } \sin(2\pi ft) = \frac{\exp(2\pi ift) - \exp(-2\pi ift)}{2i} \tag{47}$$

Thus, one can rewrite (45) with $C = A/2 + B/(2i)$ and its complex conjugate $\overline{C} = A/2 - B/(2i)$ as

$$\{g(t) = C\exp(2\pi ift) + \overline{C}\exp(-2\pi ift) \ : \ C \in \mathbb{C}, \ f \in \mathbb{R}_+\}. \tag{48}$$

Based on these sinusoids, in particular the representation in (48) we now define a transformation of data, which expresses the data in terms of its sinusoidal wave of different frequencies.

Definition 6.2 (Discrete Fourier Transform)**.**
For data $x_0, \ldots, x_{n-1} \in \mathbb{C}$ the discrete Fourier transform (DFT) is given by $b_0, \ldots, b_{n-1} \in \mathbb{C}$, where

$$b_j = \sum_{t=0}^{n-1} x_t \exp\left(-\frac{2\pi ijt}{n}\right) \qquad \text{for } j = 0, \ldots, n-1. \tag{49}$$

In R, the DFT is calculated by the function `fft()`. Note that it always holds that $b_0 = \sum x_t$. Moreover, when $x_0, \dots, x_{n-1} \in \mathbb{R}$ are real numbers (in general, they might also be complex valued), then

$$b_{n-j} = \sum_t x_t \exp\left(-\frac{2\pi i(n-j)t}{n}\right) = \sum_t x_t \exp\left(\frac{2\pi ijt}{n}\right) \exp\left(-2\pi it\right) = \bar{b}_j.$$

For example, for $n = 11$, the DFT can be written as:

$$b_0, b_1, b_2, b_3, b_4, b_5, \bar{b}_5, \bar{b}_4, \bar{b}_3, \bar{b}_2, \bar{b}_1.$$

and for $n = 12$, it is

$$b_0, b_1, b_2, b_3, b_4, b_5, b_6 = \bar{b}_6, \bar{b}_5, \bar{b}_4, \bar{b}_3, \bar{b}_2, \bar{b}_1.$$

Note that b_6 is necessarily real because $b_6 = \bar{b}_6$.
It turns out that the DFT $b_0, \dots, b_{n-1}$ is in one-to-one correspondence with the data $x_0, \dots, x_{n-1}$, because the original data can be uniquely recovered by its DFT, as the following theorem shows. Hence, the DFT $b_0, \dots, b_{n-1}$ and the data $x_0, \dots, x_{n-1}$ contain equivalent information.

Theorem 6.3 (Inverse Fourier Transform)**.**
For data $x_0, \dots, x_{n-1}$ and its DFT $b_0, \dots, b_{n-1}$ as in Definition 6.2 it holds that

$$x_t = \frac{1}{n} \sum_{j=0}^{n-1} b_j \exp\left(\frac{2\pi ijt}{n}\right) \qquad \text{for } t = 0, \dots, n-1. \tag{50}$$

Proof. Start with the right hand side of (50). Using the formula (49), we obtain

$$\begin{aligned} \frac{1}{n} \sum_{j=0}^{n-1} b_j \exp\left(\frac{2\pi ijt}{n}\right) &= \frac{1}{n} \sum_{j=0}^{n-1} \left\{ \sum_{s=0}^{n-1} x_s \exp\left(-\frac{2\pi ijs}{n}\right) \right\} \exp\left(\frac{2\pi ijt}{n}\right) \\ &= \frac{1}{n} \sum_{s=0}^{n-1} x_s \sum_{j=0}^{n-1} \exp\left(\frac{2\pi ij(t-s)}{n}\right) \end{aligned}$$

Check now that the inner sum equals n when $s = t$. For $s \neq t$ we have that $\exp\left(\frac{2\pi i(t-s)}{n}\right) \neq 1$ and hence, the geometric series formula gives

$$\sum_{j=0}^{n-1} \exp\left(\frac{2\pi ij(t-s)}{n}\right) = \sum_{j=0}^{n-1} \exp\left(\frac{2\pi i(t-s)}{n}\right)^j = \frac{1 - \exp\left(2\pi i(t-s)\right)}{1 - \exp\left(\frac{2\pi i(t-s)}{n}\right)} = 0. \tag{51}$$

□

In order to see, why the DFT, indeed, expresses the data in terms of its sinusoidal wave components, note that for $x = (x_0, \dots, x_{n-1})$ one can write

$$x = \frac{1}{n} \sum_{j=0}^{n-1} b_j u^j. \tag{52}$$

with vectors

$$u^j = (1, \exp(2\pi ij/n), \exp(2\pi i 2j/n), \dots, \exp(2\pi i(n-1)j/n)) \quad \text{for } j = 0, \dots, n-1,$$

that is, the sinusoid with frequency j/n evaluated at the time points $t = 0, 1, \dots, (n-1)$ (recall (48)). The frequencies j/n for $j = 0, \dots, n-1$ as called **Fourier frequencies**. Note that for the vectors $u^j \in \mathbb{C}^n$ it holds that for $l \neq k$ (recall (51))

$$\langle u^l, u^k \rangle = \sum_{j=1}^{n} u_j^l \overline{u}_j^k = \sum_{j=1}^{n} \exp(2\pi ijl/n) \exp(-2\pi ijk/n) = \sum_{j=1}^{n} \exp(2\pi ij(l-k)/n) = 0$$

and hence, the vectors $u^0, \dots, u^{n-1}$ form an orthogonal basis of the vector space $\mathbb{C}^n$, where the single basis vectors u^j correspond to the sinusoidal components of frequency j/n. Thus, the DFT $b_0, \dots, b_{n-j}$ is just the representation of the data $x_0, \dots, x_{n-1}$ in the bases $u^0, \dots, u^{n-1}$.
Note that the DFT $b_0, \dots, b_{n-1}$ of real valued data $x_0, \dots, x_{n-1}$ can be complex valued. Thus, in general, to visualize the DFT, one rather plots its absolute value. Note that b_0 is always just the sum of the data, which does not capture much information. Further because $b_{n-j} = \bar{b}_j$, it is enough to look at $|b_j|, 1 \leq j \leq n/2$.

Definition 6.4 (Periodogram)**.**
*For real values data $x_0, \dots, x_{n-1}$ with DFT $b_0, \dots, b_{n-1}$ the **periodogram** is defined as*

$$I(j/n) = \frac{|b_j|^2}{n} \quad \text{for } j = 1, \dots, \lfloor n/2 \rfloor$$

We have seen that b_j gives the jth coefficient of the data $x = (x_0, \dots, x_{n-1})$ in the basis $u^0, \dots, u^{n-1}$, which corresponds to the sinusoids of Fourier frequency j/n, thus we observe the following:

1. If the periodogram shows a single spike for $I(j/n)$ we are sure that the data is a single sinusoid with Fourier frequency j/n.

2. If it shows two spikes, say at $I(j_1/n)$ and $I(j_2/n)$, then the data is is a linear combination of two sinusoids at Fourier frequencies j_1/n and j_2/n with the strengths of these sinusoids depending on the size of the spikes.

3. Multiple spikes indicate that the data is made up of many sinusoids at Fourier frequencies.

4. Sometimes one can see multiple spikes in the DFT even when the structure of the data is not very complicated. A typical example is *leakage* due to the presence of a sinusoid at a non-Fourier frequency.

The following theorem shows an important relation between periodogram $I(j/n)$ and the sample ACVF $\hat{\gamma}(h)$ of some data $x_0, \dots, x_{n-1}$.

Theorem 6.5 (Connection between periodogram and Sample ACVF)**.**
For some data $x_0, \dots, x_{n-1}$ let $\hat{\gamma}(h)$ for $h = 0, \dots, n-1$ be its sample ACVF. Then

$$I(j/n) = \sum_{h=-(n-1)}^{n-1} \hat{\gamma}(h) \exp\left(-\frac{2\pi ijh}{n}\right) \qquad \text{for } j = 1, \dots, \lfloor n/2 \rfloor. \tag{53}$$

Proof. To prove (53), observe first, by the formula for the sum of a geometric series, that

$$\sum_{t=0}^{n-1} \exp\left(-\frac{2\pi i j t}{n}\right) = 0 \qquad \text{for } j = 1, \ldots, \lfloor n/2 \rfloor.$$

In other words, if the data is constant i.e., $x_0 = \cdots = x_{n-1}$, then b_0 equals nx_0 and b_j equals 0 for all other j. Because of this, we can write:

$$b_j = \sum_{t=0}^{n-1} (x_t - \bar{x}) \exp\left(-\frac{2\pi i j t}{n}\right) \qquad \text{for } j = 1, \ldots, \lfloor n/2 \rfloor.$$

Therefore, for $j = 1, \ldots, \lfloor n/2 \rfloor$, we write

$$\begin{aligned}
|b_j|^2 = b_j \bar{b}_j &= \sum_{t=0}^{n-1} \sum_{s=0}^{n-1} (x_t - \bar{x})(x_s - \bar{x}) \exp\left(-\frac{2\pi i j t}{n}\right) \exp\left(\frac{2\pi i j s}{n}\right) \\
&= \sum_{t=0}^{n-1} \sum_{s=0}^{n-1} (x_t - \bar{x})(x_s - \bar{x}) \exp\left(-\frac{2\pi i j (t-s)}{n}\right) \\
&= \sum_{h=-(n-1)}^{n-1} \sum_{t,s: t-s=h} (x_t - \bar{x})(x_{t-h} - \bar{x}) \exp\left(-\frac{2\pi i j h}{n}\right) \\
&= n \sum_{h=-(n-1)}^{n-1} \hat{\gamma}(h) \exp\left(-\frac{2\pi i j h}{n}\right).
\end{aligned}$$

□

6.2 Spectral density

We have just seen that every dataset can be written in terms of sinusoids. The magnitude of the sinusoid component with frequency j/n is given by the respective periodogram $I(j/n)$. In the following we want to extend these definitions to the process $\{X_t\}$ itself. Note that Theorem 6.5 leads to the following natural process-analog of the periodogram.

Definition 6.6 (Spectral density)**.**
For a stationary process with ACVF $\gamma_X(h)$ with $\sum_{h=-\infty}^{\infty} |\gamma_X(h)| < \infty$ we define the spectral density as

$$f(\lambda) := \sum_{h=-\infty}^{\infty} \gamma_X(h) \exp(-2\pi i \lambda h) \qquad \text{for } -1/2 \leq \lambda \leq 1/2. \tag{54}$$

It is easy to check that f is symmetric, that is, $f(-\lambda) = f(\lambda)$. Further, one can show that f is always positive, that is, $f(\lambda) \geq 0$. Analog to the periodogram, the spectral density gives the strengths of sinusoids at various frequencies contributing to a stationary stochastic process. The following theorem shows that the spectral density and the ACVF provide equivalent information.

Theorem 6.7.
For a stationary process with spectral density $f(\lambda)$, $-1/2 \leq \lambda \leq 1/2$, it holds for its ACVF that

$$\gamma_X(h) = \int_{-1/2}^{1/2} e^{2\pi i \lambda h} f(\lambda) d\lambda = \int_{-1/2}^{1/2} \cos(2\pi\lambda h) f(\lambda) d\lambda. \tag{55}$$

Proof. Using the definition of the spectral density we get

$$\int_{-1/2}^{1/2} e^{2\pi i\lambda h} f(\lambda) d\lambda = \int_{-1/2}^{1/2} e^{2\pi i\lambda k} \sum_{k=-\infty}^{\infty} \gamma_X(k) e^{-2\pi i \lambda h} d\lambda = \sum_{k=-\infty}^{\infty} \gamma_X(k) \int_{-1/2}^{1/2} e^{2\pi i \lambda(k-h)} d\lambda$$

note that

$$\int_{-1/2}^{1/2} e^{2\pi i\lambda(k-h)} d\lambda \neq 0 \quad \Leftrightarrow \quad k = j$$

and thus

$$\int_{-1/2}^{1/2} e^{2\pi i\lambda h} f(\lambda) d\lambda = \gamma_X(h).$$

For the second equality note that by symmetry $f(\lambda) = f(-\lambda)$ and the identity (47) it follows that

$$\begin{aligned}
\int_{-1/2}^{1/2} e^{2\pi i\lambda h} f(\lambda) d\lambda &= \int_{-1/2}^{0} e^{2\pi i\lambda h} f(-\lambda) d\lambda + \int_{0}^{1/2} e^{2\pi i\lambda h} f(\lambda) d\lambda \\
&= \int_{0}^{1/2} e^{-2\pi i\lambda h} f(\lambda) d\lambda + \int_{0}^{1/2} e^{2\pi i\lambda h} f(\lambda) d\lambda \\
&= \int_{0}^{1/2} \left(e^{-2\pi i\lambda h} + e^{2\pi i\lambda h}\right) f(\lambda) d\lambda \\
&= \int_{0}^{1/2} 2\cos(2\pi\lambda h) f(\lambda) d\lambda \\
&= \int_{-1/2}^{1/2} \cos(2\pi\lambda h) f(\lambda) d\lambda
\end{aligned}$$

□

Indeed, the identity (55) characterizes the spectral density in the sense that the only function f which satisfies (55) is the spectral density.

Example 6.8 (White Noise)**.**
Suppose $\{X_t\}$ is white noise with mean zero and variance σ^2. Then it is obvious that $\gamma(h) = 0$ for $h \neq 0$ and $\gamma(h) = \sigma^2$ for $h = 0$. In this case, the formula (54) *simply gives*

$$f(\lambda) = \sigma^2 \qquad \text{for every } -1/2 \leq \lambda \leq 1/2.$$

This means that the spectral density of white noise is flat (all frequencies are combined equally).

Example 6.9 (MA(1))**.**
Consider the MA(1) process $X_t = Z_t + \theta Z_{t-1}$. The autocovariance function is given by $\gamma(0) = \sigma_Z^2(1+\theta^2)$, $\gamma(\pm 1) = \theta\sigma_Z^2$ and $\gamma(h)$ equals zero for every other h.
The formula (54) *then immediately gives*

$$\begin{aligned} f(\lambda) &= \gamma(-1)\exp(2\pi i\lambda) + \gamma(0)\exp(0) + \gamma(1)\exp(-2\pi i\lambda) \\ &= \gamma(0) + \gamma(1)\left(\exp(2\pi i\lambda) + \exp(-2\pi i\lambda)\right) \\ &= \gamma(0) + 2\gamma(1)\cos(2\pi\lambda) \\ &= \sigma_Z^2\left(1 + \theta^2 + 2\theta\cos(2\pi\lambda)\right) \qquad \text{for } -1/2 \leq \lambda \leq 1/2. \end{aligned}$$

This function is increasing when $\theta < 0$ and decreasing when $\theta > 0$ (does this make sense?).

Process Representation In the following we want to consider a particular stationary processes, which naturally represents a linear combination of different sinusoidas components. We consider the doubly infinite time series $\{X_t\}$, $t = \ldots, -1, 0, 1, \ldots$, with

$$X_t = \sum_{j=1}^{m} \left(A_j \cos(2\pi\lambda_j t) + B_j \sin(2\pi\lambda_j t)\right) \tag{56}$$

where $0 \leq \lambda_j \leq 1/2$ are fixed constants and $A_1, B_1, A_2, B_2, \ldots, A_m, B_m$ are all uncorrelated random variables with mean zero and

$$\mathrm{var}(A_j) = \mathrm{var}(B_j) = \sigma_j^2.$$

Let $\sum_{j=1}^m \sigma_j^2 = \sigma^2$ so that the variance of the process $\{X_t\}$ equals σ^2. It turns out that the model (56) can approximate any stationary model provided m is large enough and $\lambda_1, \ldots, \lambda_m$ and $\sigma_1^2, \ldots, \sigma_m^2$ are chosen appropriately. The general theory on this is captures by the so called **spectral representation of a process**. However, this requires the notation of a stochastic integral, which is beyond the scope of this course.
However, in the following we want to give some insight on how to chose the the variances $\sigma_1^2, \ldots, \sigma_m^2$ (for some given m), when we wish to approximate a stationary process $\{Y_t\}$ with spectral density $f(\lambda)$ by the time series $\{X_t\}$. For the frequencies λ_j we consider equally spacing in $[0, 1/2]$, that is, $\lambda_j = j/(2m)$.

It is easy to check that for a time series $\{X_t\}$ as in (56) the ACVF is given by

$$\gamma_X(h) = \sum_{i=1}^{m} \sigma_i^2 \cos(2\pi\lambda_j h)$$

Moreover, by Theorem 6.7 it follows for the process $\{Y_t\}$ that

$$\begin{aligned}\gamma_Y(h) &= \int_{-1/2}^{1/2} \cos(2\pi\lambda h) f(\lambda) d\lambda \\ &= 2\int_0^{1/2} \cos(2\pi\lambda h) f(\lambda) d\lambda = 2\sum_{j=1}^{m} \int_{\lambda_{j-1}}^{\lambda_j} \cos(2\pi\lambda h) f(\lambda) d\lambda.\end{aligned}$$

When m is chosen large enough and the spectral density f is sufficiently regular, than it follows that

$$\int_{\lambda_{j-1}}^{\lambda_j} \cos(2\pi\lambda h) f(\lambda) d\lambda \approx f(\lambda_j)\cos(2\pi\lambda_j h)(\lambda_{j-1} - \lambda_j) = \frac{1}{2m} f(\lambda_j)\cos(2\pi\lambda_j h).$$

Thus, if we want $\{X_t\}$ to be a stationary process which has a similar ACVF as $\{Y_t\}$ for any lag h, we require that

$$\sum_{i=1}^{m} \sigma_i^2 \cos(2\pi\lambda_j h) \approx \sum_{i=1}^{m} \frac{1}{m} f(\lambda_j)\cos(2\pi\lambda_j h) \text{ for all } h = 0, 1, \ldots.$$

Thus, we can choose

$$\sigma_j^2 = \frac{f(\lambda_j)}{m} \quad \textbf{and} \quad \lambda_j = \frac{j}{2m}.$$

Thus, the spectral density can be used to approximate any stationary process by sinusoids.

Example 6.10 (MA(1))**.**
For example, we have seen that the spectral density of an MA(1) process is given by

$$f(\lambda) = \sigma_Z^2\left(1 + \theta^2 + 2\theta\cos(2\pi\lambda)\right) \qquad \text{for } -1/2 \le \lambda \le 1/2$$

where σ_Z^2 is the variance of the white noise process. Thus the process (56) *with m large and*

$$\lambda_j = \frac{j}{2m} \quad \text{and} \quad \sigma_j^2 = \frac{\sigma_Z^2}{m}\left(1 + \theta^2 + 2\theta\cos(2\pi\lambda_j)\right)$$

will approximate an MA(1) process. One can use this, for example, for simulating an MA(1) process.

Remark 6.11 (Folding frequency)**.**
Note that in (56) it is indeed sufficient to consider frequencies $0 \le \lambda_j \le 1/2$. To see this, note the following:

1. *If $\lambda < 0$, then we can write $A\cos(2\pi\lambda t) + B\sin(2\pi\lambda t)$ as simply $A\cos(2\pi(-\lambda)t) + (-B)\sin(2\pi(-\lambda)t)$. Clearly $\lambda \ge 0$.*

2. *If* $\lambda \geq 1$*, then (letting* $[\lambda]$ *to be the integer part of* λ*) we can write*

$$\begin{aligned} A\cos(2\pi\lambda t) + B\sin(2\pi\lambda t) &= A\cos(2\pi[\lambda]t + 2\pi(\lambda - [\lambda])t) + B\sin(2\pi[\lambda]t + 2\pi(\lambda - [\lambda])) \\ &= A\cos(2\pi(\lambda - [\lambda])t) + B\sin(2\pi(\lambda - [\lambda])) \end{aligned}$$

   because $\cos(\cdot)$ *and* $\sin(\cdot)$ *are both periodic functions with period* 2π*. Clearly* $0 \leq \lambda - [\lambda] < 1$*.*

3. *If* $\lambda \in [1/2, 1)$*, then*

$$\begin{aligned} A\cos(2\pi\lambda t) + B\sin(2\pi\lambda t) &= A\cos(2\pi t - 2\pi(1-\lambda)t) + B\sin(2\pi t - 2\pi(1-\lambda)t) \\ &= A\cos(2\pi(1-\lambda)t) + (-B)\sin(2\pi(1-\lambda)t). \end{aligned}$$

   because $\cos(2\pi t - x) = \cos x$ *and* $\sin(2\pi t - x) = -\sin x$ *for all integers* t*. Clearly* $0 < 1 - \lambda \leq 1/2$*.*

For $\lambda_j = 1/2$ *the series makes a cycle every two time points. For data that occurs in discrete time points we need at least two points to determine a cycle. Thus, the highest frequency of interest is* 0.5 *cycles per point. This frequency is called the folding frequency. It is the highest frequency that can be seen in discrete sampling. Higher frequencies sampled this way will appear at lower frequencies, called alias. More precisely, given a sinusoid* $A\cos(2\pi\lambda t) + B\sin(2\pi\lambda t)$*, one can always write it as* $A\cos(2\pi\lambda_0 t) + B'\sin(2\pi\lambda_0 t)$ *for some* $\lambda_0 \in [0, 1/2]$ *and* B' *which equals either* B *or* $-B$*, where* λ_0 *is the alias of* λ*.*
An example is the way a camera samples a rotating wheel on a moving automobile in a movie, in which a wheel appears to be rotating at a different rate and sometimes backwards. For example, most movies are recorded at 24 frames per seconds (24 Hertz). If a camera is filming a wheel that is rotating at 24 Hertz, the wheel will appear to stand still.

6.3 Linear Time Invariant Filters

In the following we study the general technique of linear time invariant filters for transforming one time series into another. Linear filters were already introduced in the context of trend estimation in Section 3.1.1. We will see that they are particularly helpful within the frequency domain approaches.

Definition 6.12 (Linear Time invariant Filter)**.**
A linear time-invariant filter with coefficients $\{a_j\}$ for $j = \ldots, -2, -1, 0, 1, 2, 3, \ldots$ transforms an input time series $\{X_t\}$ into an output time series $\{Y_t\}$ via

$$Y_t = \sum_{j=-\infty}^{\infty} a_j X_{t-j}.$$

In the above definition, the coefficients $\{a_j\}$ are often assumed to satisfy $\sum_{j=-\infty}^{\infty} |a_j| < \infty$.

Example 6.13 (Impulse function)**.**
Suppose that the input series $\{X_t\}$ is given by

$$X_t = \begin{cases} 1 & \text{if } t = 0 \\ 0 & \text{otherwise} \end{cases}$$

*Such an $\{X_t\}$ is often called an impulse function. The output of the filter $\{Y_t\}$ can then be easily seen to be $Y_t = a_t$. For this reason the filter coefficients $\{a_j\}$ are often collectively known as the **impulse response function**.*

Example 6.14 (Smoothing and differencing)**.**
We have seen some examples of time invariant filters already in Section 3.1. Smoothing constitutes a particular type of time invariant linear filter with $a_j = 1/(2q+1)$ for $|j| \leq q$ and $a_j = 0$ otherwise. Differencing corresponds to another filter with $a_0 = 1$ and $a_1 = -1$ and all other a_js equal zero.
We have seen that these two filters act very differently; one estimates trend while the other eliminates it.

Suppose that the input time series $\{X_t\}$ is stationary with autocovariance function γ_X. Then for the autocovariance function of $\{Y_t\}$ we observe

$$\begin{aligned} \gamma_Y(h) &:= \operatorname{cov}\left(\sum_j a_j X_{t-j}, \sum_k a_k X_{t+h-k}\right) \\ &= \sum_{j,k} a_j a_k \operatorname{cov}(X_{t-j}, X_{t+h-k}) = \sum_{j,k} a_j a_k \gamma_X(h-k+j). \end{aligned} \tag{57}$$

Note that the above calculation shows also that $\{Y_t\}$ is stationary.
Suppose now that the spectral density of the input stationary series $\{X_t\}$ is f_X. Then for the spectral density f_Y of the output $\{Y_t\}$ we obtain

$$\gamma_X(h) = \int_{-1/2}^{1/2} e^{2\pi i h \lambda} f_X(\lambda) d\lambda.$$

We thus have from (57) that

$$\begin{aligned}\gamma_Y(h) &= \sum_j \sum_k a_j a_k \int e^{2\pi i(h-k+j)\lambda} f_X(\lambda) d\lambda \\ &= \int e^{2\pi i h\lambda} f_X(\lambda) \left(\sum_j \sum_k a_j a_k e^{-2\pi i k\lambda} e^{2\pi i j\lambda} \right) d\lambda\end{aligned} \tag{58}$$

The following definition will simplify notation.

Definition 6.15 (Transfer function)**.**
For a time invariant linear filter with coefficients $\{a_j\}$ define

$$A(\lambda) := \sum_j a_j e^{-2\pi i j\lambda} \qquad \text{for } -1/2 \leq \lambda \leq 1/2. \tag{59}$$

The function

$$\lambda \mapsto A(\lambda)$$

is called the ***transfer function*** *or the* ***frequency response function*** *and the function*

$$\lambda \mapsto |A(\lambda)|^2$$

is called the ***power transfer function****.*

From (58) it clearly follows that

$$\gamma_Y(h) = \int e^{2\pi i\lambda h} f_X(\lambda) A(\lambda) \overline{A(\lambda)} d\lambda,$$

where, of course, $\overline{A(\lambda)}$ denotes the complex conjugate of $A(\lambda)$. As a result, we have

$$\gamma_Y(h) = \int e^{2\pi i\lambda h} f_X(\lambda) \left|A(\lambda)\right|^2 d\lambda.$$

This is clearly of the form $\gamma_Y(h) = \int e^{2\pi i\lambda h} f_Y(\lambda) d\lambda$. We therefore have

$$f_Y(\lambda) = f_X(\lambda) \left|A(\lambda)\right|^2 \qquad \text{for } -1/2 \leq \lambda \leq 1/2. \tag{60}$$

In other words, the action of the filter on the spectrum of the input is very easy to explain. It modifies the spectrum by multiplying it with the power transfer function $|A(\lambda)|^2$. Depending on the value of $|A(\lambda)|^2$, some frequencies may be enhanced in the output while other frequencies will be diminished. Thus, the spectral density is very useful while studying the properties of a filter. While the autocovariance function of the output series γ_Y depends in a complicated way on that of the input series γ_X, the dependence between the two spectral densities is very simple.

Example 6.16 (Power transfer function of the Differencing Filter)**.**
Consider the lag s differencing filter $Y_t = X_t - X_{t-s}$ which corresponds to the weights $a_0 = 1$ and $a_s = -1$ and $a_j = 0$ for all other j. Then the transfer function is clearly given by

$$A(\lambda) = \sum_j a_j e^{-2\pi i j\lambda} = 1 - e^{-2\pi i s\lambda} = 2i\sin(\pi s\lambda)e^{-\pi i s\lambda},$$

where, for the last equality, the formula $1 - e^{i\theta} = -2i\sin(\theta/2)e^{i\theta/2}$ is used. Therefore the power transfer function equals

$$|A(\lambda)|^2 = 4\sin^2(\pi s\lambda) \qquad \text{for } -1/2 \le \lambda \le 1/2.$$

To understand this function, we only need to consider the interval $[0, 1/2]$ because it is symmetric on $[-1/2, 1/2]$. When $s = 1$, the function $\lambda \mapsto |A(\lambda)|^2$ is increasing on $[0, 1/2]$. This means that first order differencing enhances the higher frequencies in the data and diminishes the lower frequencies. Therefore, it will make the data more wiggly. For higher values of s, the function $A(\lambda)$ goes up and down and takes the value zero for $\lambda = 0, 1/s, 2/s, \ldots$. In other words, it eliminates all components of period s.

Example 6.17 (Power transfer function of smoothing filter)**.**
Now consider the smoothing filter which corresponds to the coefficients $a_j = 1/(2q+1)$ for $|j| \le q$. The transfer function is

$$A(\lambda) = \frac{1}{2q+1}\sum_{j=-q}^{q} e^{-2\pi i j\lambda} = \frac{S_{q+1}(\lambda) + S_{q+1}(-\lambda) - 1}{2q+1}, \quad \text{for } -1/2 \le \lambda \le 1/2,$$

where

$$S_q(\lambda) := \sum_{t=0}^{q-1} \exp(2\pi i\lambda t). \tag{61}$$

When $\lambda = 0$ it is easy to see that $S_q(\lambda) = q$ and $A(0) = 1$. When $\lambda \neq 0$ then $\exp(2\pi i\lambda) \neq 1$ and this function can be easily evaluated by the geometric series formula to get

$$S_q(\lambda) = \frac{e^{2\pi i\lambda q} - 1}{e^{2\pi i\lambda} - 1}.$$

Then, because

$$e^{i\theta} - 1 = \cos\theta + i\sin\theta - 1 = 2e^{i\theta/2}\sin(\theta/2)$$

we get

$$S_q(\lambda) = \frac{\sin\pi q\lambda}{\sin\pi\lambda}e^{i\pi\lambda(q-1)}.$$

and thus

$$S_q(\lambda) + S_q(-\lambda) = 2\frac{\sin(\pi q\lambda)}{\sin(\pi\lambda)}\cos(\pi\lambda(q-1)),$$

which implies that the transfer function is given by

$$A(\lambda) = \frac{1}{2q+1}\left(2\frac{\sin(\pi(q+1)\lambda)}{\sin(\pi\lambda)}\cos(\pi q\lambda) - 1\right).$$

This function only depends on q and can be plotted for various values of q. For q large, it drops to zero very quickly. The interpretation is that the filter kills the high frequency components in the input process.

Spectral density of ARMA process As a direct consequence of (60) we can compute the spectral density of the unique stationary solution of a causal ARMA process.

Theorem 6.18 (Spectral density of ARMA process)**.**
Let $\{X_t\}$ be a stationary causal ARMA process $\phi(B)X_t = \theta(B)Z_t$ with ϕ and θ having no common roots. Then, for the spectral density f_X of $\{X_t\}$ as in Definition 6.6 it holds that

$$f_X(\lambda) = \sigma_Z^2 \frac{|\theta(e^{-2\pi i\lambda})|^2}{|\phi(e^{-2\pi i\lambda})|^2} \qquad \text{for } -1/2 \leq \lambda \leq 1/2. \tag{62}$$

Proof. Let $U_t = \phi(B)X_t = \theta(B)Z_t$. Let us first write down the spectral density of $U_t = \phi(B)X_t$ in terms of that of $\{X_t\}$. Clearly, U_t can be viewed as the output of a filter applied to X_t. The filter is given by $a_0 = 1$ and $a_j = -\phi_j$ for $1 \leq j \leq p$ and $a_j = 0$ for all other j. Let $A_\phi(\lambda)$ denote the transfer function of this filter. Then we have

$$f_U(\lambda) = |A_\phi(\lambda)|^2 f_X(\lambda). \tag{63}$$

Similarly, using the fact that $U_t = \theta(B)Z_t$, we can write

$$f_U(\lambda) = |A_\theta(\lambda)|^2 f_Z(\lambda) = \sigma_Z^2 |A_\theta(\lambda)|^2 \tag{64}$$

where $A_\theta(\lambda)$ is the tranfer function of the filter with coefficients $a_0 = 1$ and $a_j = \theta_j$ for $1 \leq j \leq q$ and $a_j = 0$ for all other j. Equating (63) and (64), we obtain

$$f_X(\lambda) = \frac{|A_\theta(\lambda)|^2}{|A_\phi(\lambda)|^2}\sigma_Z^2 \qquad \text{for } -1/2 \leq \lambda \leq 1/2.$$

Now

$$A_\phi(\lambda) = 1 - \phi_1 e^{-2\pi i\lambda} - \phi_2 e^{-2\pi i(2\lambda)} - \cdots - \phi_p e^{-2\pi i(p\lambda)} = \phi(e^{-2\pi i\lambda}).$$

Note that the denominator $A_\phi(\lambda)$ is non-zero for all λ because of stationarity. Similarly $A_\theta(\lambda) = \theta(e^{-2\pi i\lambda})$, which shows the assertion. □

Example 6.19 (MA(1))**.**
For the MA(1) process: $X_t = Z_t + \theta Z_{t-1}$, we have $\phi(z) = 1$ and $\theta(z) = 1 + \theta z$. Therefore

$$\begin{aligned} f_X(\lambda) &= \sigma_Z^2 \left|1 + \theta e^{2\pi i\lambda}\right|^2 \\ &= \sigma_Z^2 \left|1 + \theta\cos 2\pi\lambda + i\theta \sin 2\pi\lambda\right|^2 \\ &= \sigma_Z^2 \left[(1 + \theta\cos 2\pi\lambda)^2 + \theta^2 \sin^2 2\pi\lambda\right] \\ &= \sigma_Z^2 \left[1 + \theta^2 + 2\theta\cos 2\pi\lambda\right] \qquad \text{for } -1/2 \leq \lambda \leq 1/2. \end{aligned}$$

Check that for $\theta = -1$, the quantity $1 + \theta^2 + 2\theta\cos(2\pi\lambda)$ equals the power transfer function of the first differencing filter.

Example 6.20 (AR(1))**.**
For AR(1): $X_t - \phi X_{t-1} = Z_t$, we have $\phi(z) = 1 - \phi z$ and $\theta(z) = 1$. Thus

$$f_X(\lambda) = \sigma_Z^2 \frac{1}{|1 - \phi e^{2\pi i\lambda}|^2} = \frac{\sigma_Z^2}{1 + \phi^2 - 2\phi\cos 2\pi\lambda} \qquad \text{for } -1/2 \leq \lambda \leq 1/2.$$

Example 6.21 (AR(2))**.**
For the AR(2) model: $X_t - \phi_1 X_{t-1} - \phi_2 X_{t-2} = Z_t$, we have $\phi(z) = 1 - \phi_1 z - \phi_2 z^2$ and $\theta(z) = 1$. Here it can be shown that

$$f_X(\lambda) = \frac{\sigma_Z^2}{1 + \phi_1^2 + \phi_2^2 - 2\phi_1(1 - \phi_2)\cos 2\pi\lambda - 2\phi_2 \cos 4\pi\lambda} \qquad \text{for } -1/2 \leq \lambda \leq 1/2.$$

Parametric spectral density estimation When we want to estimate the spectral density of a stationary process, one approach is to consider a parametric ARMA model for the process $\phi(B)X_t = \theta(B)Z_t$, estimate its parameters $\phi_1, \dots, \phi_p, \theta_1, \dots, \theta_q$ as discussed in Section 5.8 and then plug in these estimates $\hat{\phi}_1, \dots, \hat{\phi}_p, \hat{\theta}_1, \dots, \hat{\theta}_q$ into the equation (62). For convenience, usually a parametric spectral estimator is obtained by fitting an AR(p) model, where the order p is determined by model selection such as AIC or BIC (recall Section 5.10).
The following theorem shows that any spectral density can be approximated arbitrary close by the spectrum of an AR process, see (Shumway and Stoffer, 2006, Property 4.7)

Theorem 6.22 (AR Spectal Approximation)**.**
Let $g(\lambda)$ be the spectral density of a stationary process. Then, given $\epsilon > 0$, there is a time series with the representation

$$\phi(B)X_t = Z_t,$$

for some finite order p polynomial ϕ and some white noise Z_t with variance σ^2, such that

$$|f_X(\lambda) - g(\lambda)| < \epsilon \quad \text{for all } \lambda \in [-1/2, 1/2].$$

Moreover, the roots of ϕ outside the unit circle.

Unfortunately, Theorem 6.22 does not tell us how large p, it might be very large in some cases.
In R, one case use the function `spec.ar` to fit the best model via AIC and plot the resulting spectrum.
For further reading on parametric density estimation see (Shumway and Stoffer, 2006, Chapter 4.5)

7 State space models

The following is meant just as a quick overview of state space models[4]

7.1 General state space models/ Hidden Markov models

General state space models (or Hidden Markov models - HMM) consist of

(i) An unobserved (latent) state process (X_t) with Markovian dependence

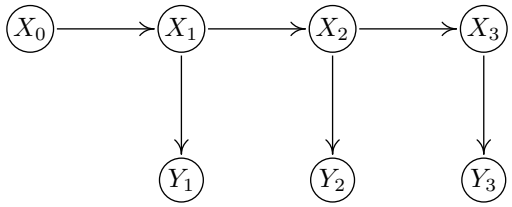
(ii) Observations (Y_t) which are derived from X_t.

Concretely, this means

(i) $X_0, X_1, X_2, \ldots$ is a Markov chain

(ii) Conditionally on (X_t), all Y_t are independent and depend only on X_t.

As a graphical model, can use a directed acyclic graph:



The graphical model is a convenient way of writing down assumptions (i) and (ii). For a general joint distribution of $(X_0, X_1, \ldots, X_T, Y_1, \ldots Y_T)$ we can always write the density as

$$f_0(X_0) \prod_{t=1}^{T} f_t(X_t|X_0, X_1, \ldots, X_{t-1}, Y_1, \ldots, Y_{t-1}) g_t(Y_t|X_0, X_1, \ldots, X_{t-1}, X_t, Y_1, \ldots, Y_{t-1})$$

for appropriate conditional densities f_t, $t = 0, \ldots, T$ and g_t, $t = 1, \ldots, T$. The model above is equivalent to the joint density having a simpler factorization

$$f_0(X_0) \prod_{t=1}^{T} f_t(X_t|X_{t-1}) g_t(Y_t|X_t).$$

Note that (X_t) is a Markov chain and also (Z_t) is a Markov chain if we concatenate $Z_t = (X_t, Y_t)$. However, the observations (Y_t) on their own are not a Markov chain and exhibit more complex time-dependencies. The HMM allows, in other words, to model such more complex time-dependencies in the observations by a simple Markov model.
Goals of HMM analysis include

(a) Given observations $y_1, \ldots, y_T$ and **known** transition densities/probabilities $f_t(X_t|X_{t-1})$ and $g_t(Y_t|X_t)$, provide inference for the underlying state vector $x_0, \ldots, x_T$ (there are several forms of different inference which we return to).

[4] most parts here are in analogy (but simplified) from a book chapter "State Space and Hidden Markov Models" by Prof. Hansruedi Künsch

(b) Given observations $y_1, \dots, y_T$ and **unknown** transition densities/probabilities $f_t(X_t|X_{t-1})$ and $g_t(Y_t|X_t)$, provide inference for the underlying state vector $x_0, \dots, x_T$ and for f_t and g_t simultaneously, either in a Bayesian or frequentist form

In the following, will first assume we are in setting (a), that is the transition densities are assumed to be known.
Examples for state space models:

(i) **Linear state space model** for $X_t \in \mathbb{R}^p$ and $Y_t \in \mathbb{R}^q$:

$$X_t = \mathbf{G}_t X_{t-1} + V_t$$
$$Y_t = \mathbf{H}_t X_t + W_t,$$

where $\mathbf{G}_t \in \mathbb{R}^{p\times p}$ and $\mathbf{H}_t \in \mathbb{R}^{q\times p}$. The state vector X_t could for example be position and velocity of a moving object and Y_t are noisy measurements of the objects location. Goal is to infer X_t as accurately as possible. In case of Gaussian error terms there is an explicit solution (Kalman filter).

(ii) **ARMA models.** Let (Y_t) be a causal and invertible ARMA(p,q) process. Can be written as an HMM. Take as an example the previously discussed case of an AR(p) model:

$$Y_t = \sum_{j=1}^{p} \phi_j Y_{t-j} + W_t.$$

Define X_t as the collection of the past p observations $X_t := (Y_t, Y_{t-1}, Y_{t-2}, \dots, Y_{t-p+1})^t$. Then

$$X_t = \mathbf{\Phi}_t X_{t-1} + \eta_t$$
$$Y_t = \mathbf{H}_t X_t + \varepsilon_t,$$

where

$$\mathbf{\Phi} = \begin{pmatrix} \phi_1 & \dots & \phi_{p-1} & \phi_p \\ & & & 0 \\ & I_{p-1} & & \vdots \\ & & & 0 \end{pmatrix}, \ \eta_t = \begin{pmatrix} W_t \\ 0 \\ 0 \\ \dots \\ 0 \end{pmatrix} \ \mathbf{H_t} = (1, 0, 0, \dots, 0), \text{ and } \varepsilon_t \equiv 0.$$

Advantages of writing an ARMA process as a state space model include the ability to deal easily with missing data (for example by setting $H_t = (0, 0, \dots, 0)$ and $Y - t = 0$ at times t where we have missing data) and the ability to introduce a different type of outlier in the noise distribution (using η_t and ε_t respectively).

(iii) **Speech recognition.** The sequence (X_t) can be seen as the hidden sequence of words a speaker is trying to say and Y_t are the observed sound measurements.

(iv) **Biological examples** include ion-channel analysis (determining whether an ion gate is on or off–the underlying state $X_t \in \{0, 1\}$) based on noisy measurements Y_t of the current flowing through the gate. Includes also DNA analysis where the index of time is taken by the index of position along the chromosome and we can try to infer for example regions with heightened copynumbers of so-called CG-islands (areas where the acids C and G appear more often than A and T).

(v) **Physics/meteorology.** State X_t includes all relevant atmospheric variables at time t. Transition dynamics of X_t are given by a underlying physics (and approximated by meteorological models). The observations Y_t can be satellite measurements, wind and rainfall sensors etc. that help to infer the true underlying state.

7.2 Discrete state space models

Inference is easier (also notationally) for discrete state space models, where without limitation of generality

$$X_t \in \{1, \dots, \ell\}$$
$$Y_t \in \{1, \dots, m\}.$$

Joint density factorizes again as

$$P(X_0 = x_0, X_1 = x_1, \dots, X_t = x_T, Y_1 = y_1, Y_2 = y_2, \dots Y_T = y_T) =$$
$$P(X_0 = x_0) \cdot \prod_{t=1}^{T} \Big[P(X_t = x_t | X_{t-1} = x_{t-1}) \cdot P(Y_t = y_t | X_t = x_t) \Big]$$

Or, taking the logarithm,

$$\log P(X_0 = x_0, X_1 = x_1, \dots, X_t = x_T, Y_1 = y_1, Y_2 = y_2, \dots Y_T = y_T) =$$
$$\log P(X_0 = x_0) + \sum_{t=1}^{T} \Big[\log P(X_t = x_t | X_{t-1} = x_{t-1}) + \log P(Y_t = y_t | X_t = x_t) \Big] \tag{65}$$

Let matrices $\mathbf{A} \in \mathbb{R}^{\ell \times \ell}$ and $\mathbf{B} \in \mathbb{R}^{m \times \ell}$ describe the transition probabilities $X_{t-1} \to X_t$ and $X_t \to Y_t$ in the sense that

$$P(X_t = j' | X_{t-1} = j) = \mathbf{A}_{j',j}$$
$$P(Y_t = o | X_t = j) = \mathbf{B}_{o,j}$$

(Note that often people work with $\mathbf{A}^t$ instead of $\mathbf{A}$ and $\mathbf{B}^t$ instead of $\mathbf{B}$ but for our purposes it is more convenient to define the matrices as above.) We have that (using the fact that conditional probabilities have to be positive and sum to 1):

$$\sum_{j'=1}^{\ell} \mathbf{A}_{j',j} = 1 \text{ for all } j \in \{1, \dots, \ell\} \text{ (matrix } \mathbf{A} \text{ is column-normalised)}$$
$$\text{and } \sum_{o=1}^{m} \mathbf{B}_{o,j} = 1 \text{ for all } j \in \{1, \dots, \ell\} \text{ (matrix } \mathbf{B} \text{ is column-normalised)}$$
$$\text{and } \mathbf{A}_{j',j} \geq 0 \text{ and } \mathbf{B}_{o,j} \geq 0 \text{ for all } j, j' \in \{1, \dots, \ell\} \text{ and } o \in \{1, \dots, m\}.$$

7.3 Filtering, smoothing and prediction

Let $\pi^0 \in \mathbb{R}^\ell$ be the initial/prior distribution of X_0:

$$\pi^0_j := P(X_0 = j) \text{ for all } j = 1, \dots, \ell.$$

Let $y_s^t = (y_s, \dots, y_t)$ be a vector of observations.
Goal: find conditional distribution $P(X_{t+k} | y_s^t)$. This is called

i) **Prediction** if $k > 0$

ii) **Filtering** if $k = 0$

iii) **Smoothing** if $k < 0$.

Will here look mostly at filtering and prediction.

Prediction. The prediction problem can be solved iteratively and hence reduced to filtering.
Let $\pi^{t+k|t}$ be the conditional distribution of X_{t+k}, given y_1^t in the sense that for all $j \in \{1, \dots, \ell\}$,

$$\pi_j^{t+k|t} := P(X_{t+k} = j | y_1^t).$$

We can now get a recursion for $\pi^{t+k|t}$ (a recursion in k) by conditioning on X_{t+k-1} and using the conditional independence between X_{t+k} and Y_1^t given X_{t+k-1}:

$$\begin{aligned}
\pi_j^{t+k|t} &= P(X_{t+k} = j | y_1^t) \\
&= \sum_{j'=1}^{\ell} P(X_{t+k} = j | X_{t+k-1} = j', y_1^t) \cdot P(X_{t+k-1} = j' | y_1^t) \\
&= \sum_{j'=1}^{\ell} P(X_{t+k} = j | X_{t+k-1} = j') \cdot P(X_{t+k-1} = j' | y_1^t) \\
&= \sum_{j'=1}^{\ell} \mathbf{A}_{j,j'} \cdot \pi_{j'}^{t+k-1|t}.
\end{aligned}$$

Or, in vector form,

$$\pi^{t+k|t} = \mathbf{A}\, \pi^{t+k-1|t}.$$

Reiterating back to time t, we get

$$\pi^{t+k|t} = \mathbf{A}^k\, \pi^{t|t}.$$

and we have thus reduced it to a filtering problem since $\pi^{t|t}$ is the conditional distribution of X_t, given $(y_1, \dots, y_t)$.
The distribution of Y_{t+k}, given y_1^t, follows by conditioning on X_{t+k} in similar form. If we set

$$p_o^{t+k|t} := P(Y_{t+k} = o | y_1^t),$$

then, by conditioning,

$$\begin{aligned}
p_o^{t+k|t} &= P(Y_{t+k} = o | y_1^t) \\
&= \sum_{j=1}^{\ell} P(Y_{t+k} = o | X_{t+k} = j, y_1^t) \cdot P(X_{t+k} = j | y_1^t) \\
&= \sum_{j=1}^{\ell} P(Y_{t+k} = o | X_{t+k} = j) \cdot P(X_{t+k} = j | y_1^t)
\end{aligned}$$

In vector form,

$$p^{t+k|t} = \mathbf{B} \cdot \pi^{t+k|t}.$$

Substituting from the result above , we can also write it as

$$p^{t+k|t} = \mathbf{B} \cdot \mathbf{A}^k \cdot \pi^{t|t}.$$

and we are going to look at the filtering distribution $\pi^{t|t}$ next.

Filtering. We want a recursion for the filtering density $\pi_j^{t|t} = P(X_t = j|y_1^t)$, which is also used in the prediction tasks. We can again use conditional independence of Y_1^t and Y_{t+1}, given X_{t+1}, and Bayes formula to get the desired recursion.
Recall that for two events A, B Bayes formula derives from

$$P(A, B) = P(A|B)P(B) = P(B|A)P(A) = P(A, B),$$

and can be written as

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}.$$

We can furthermore condition on yet another event C:

$$P(A|B, C) = \frac{P(B|A, C)P(A|C)}{P(B|C)}.$$

Setting

$$\begin{aligned} A &= \{X_{t+1} = j\} \\ B &= \{Y_{t+1} = y_{t+1}\} \\ C &= \{Y_1^t = y_1^t\}, \end{aligned}$$

we get for the desired filtering density

$$\begin{aligned} \pi_j^{t+1|t+1} &= P(X_{t+1} = j|y_1^{t+1}) \\ &= \frac{P(Y_{t+1} = y_{t+1}|X_{t+1} = j, Y_1^t = y_1^t)P(X_{t+1} = j|Y_1^t = y_1^t)}{P(Y_{t+1} = y_{t+1}|Y_1^t = y_1^t)}. \end{aligned}$$

Using the conditional independencies at this point, we can simplify to

$$\begin{aligned} \pi_j^{t+1|t+1} &= \frac{P(Y_{t+1} = y_{t+1}|X_{t+1} = j)P(X_{t+1} = j|Y_1^t = y_1^t)}{P(Y_{t+1} = y_{t+1}|Y_1^t = y_1^t)} \\ &= \frac{P(Y_{t+1} = y_{t+1}|X_{t+1} = j)P(X_{t+1} = j|Y_1^t = y_1^t)}{\sum_{j'=1}^{\ell} P(Y_{t+1} = y_{t+1}|X_{t+1} = j')P(X_{t+1} = j'|Y_1^t = y_1^t)} \\ &= \frac{\pi_j^{t+1|t}\mathbf{B}_{y_{t+1},j}}{\sum_{j'=1}^{\ell} \pi_{j'}^{t+1|t}\mathbf{B}_{y_{t+1},j'}} \end{aligned}$$

The recursion works thus schematically in computation as

$$\pi^{t|t} \to \pi^{t+1|t} \to \pi^{t+1|t+1} \to \dots,$$

where the second step $\pi^{t+1|t} \to \pi^{t+1|t+1}$ requires the new observation y_{t+1} at time $t+1$. Using from the prediction task the recursion for $\pi^{t+1|t}$, we can directly write the recursion for $\pi^{t|t}$ without going via the prediction density as

$$\pi_j^{t+1|t+1} = \frac{(\mathbf{A}\pi^{t|t})_j \mathbf{B}_{y_{t+1},j}}{\sum_{j'=1}^{\ell} (\mathbf{A}\pi^{t|t})_{j'} \mathbf{B}_{y_{t+1},j'}}.$$

The denominator can be seen as a normalisation that ensures

$$\sum_{j=1}^{\ell} \pi_j^{t|t} = 1 \text{ for all times } t$$

(and can conveniently by implemented by such a normalisation without having to compute the denominator explicitely).

7.4 Posterior mode, viterbi and forward-backward algorithms and dynamic programming

The filtering approach yields posterior densities of, say, X_T, given $y_1, \dots, y_T$. It does not and cannot answer questions about the most likely sequence $x_0^T = (x_0, \dots, x_T)$ under the made observations, which is given by

$$\hat{x}_0^T = \text{argmax}_{x_0^T} \, P(X_0^T = x_0^T | y_1^T).$$

The most likely sequence can be computed with dynamic programming in a forward and backwards recursion.
Taking the log and using the previous decomposition (65),

$$\begin{aligned}
\log P(X_0^T = x_0^T | Y_1^T = y_1^T) &\propto \log P(X_0^T = x_0^T, Y_1^T = y_1^T) \\
&= \log P(X_0 = x_0, X_1 = x_1, \dots, X_T = x_T, Y_1 = y_1, Y_2 = y_2, \dots Y_T = y_T) \\
&= \log P(X_0 = x_0) + \sum_{t=1}^{T} \Big[\log P(X_t = x_t | X_{t-1} = x_{t-1}) + \log P(Y_t = y_t | X_t = x_t) \Big] \\
&= \log \pi^0(x_0) + \sum_{t=1}^{T} \Big[\log \mathbf{A}_{x_t, x_{t-1}} + \log \mathbf{B}_{y_t, x_t} \Big]
\end{aligned}$$

The optimization problem can be seen as one of minimizing the cost of traversing time from 0 to T and passing through $x_0, x_1, \dots, x_T$ along the way where we incur

(i) Cost for passing through the initial state x_0 which depends on the prior distribution π^0 for X_0:

$$-\log \pi^0(x_0)$$

(ii) Cost for passing from state x_{t-1} to state x_t at every time $t = 1, \dots, T$:

$$-\log(\mathbf{A}_{x_t, x_{t-1}})$$

(iii) Cost for state x_t at every time $t = 1, \dots, T$ (depending on the observation y_t at this point).

$$-\log(\mathbf{B}_{y_t, x_t})$$

We can first make a forward recursion going thorugh $t = 1, \dots, T$, and record in $\psi_t(x)$ the lowest cost (negative log-likelihood) achievable up to this point t in time if we end up in position x at time t, that is

$$\psi_t(x) = \min_{(x_0,\dots,x_{t-1}),x_t=x} \Big(-\log\pi^0(x_0) + \sum_{t'=1}^{t} \big[-\log \mathbf{A}_{x_{t'},x_{t'-1}} + -\log \mathbf{B}_{y_{t'},x_{t'}} \big] \Big).$$

Note that

$$\hat{x} = \text{argmin}_{(x_0,\dots,x_T)} \Big(-\log\pi^0(x_0) + \sum_{t'=1}^{T} \big[-\log \mathbf{A}_{x_{t'},x_{t'-1}} + -\log \mathbf{B}_{y_{t'},x_{t'}} \big] \Big)$$

and hence

$$\hat{x}_T = \text{argmax}_x \ \psi_T(x)$$

The function ψ_t can be calculated now for $t = 1, 2, \dots$ in a forward recursion as

$$\begin{aligned} \psi_0(x) &= -\log\pi^0(x) \\ \psi_t(x) &= \min_{x_{t-1}} \big(\psi_{t-1}(x_{t-1}) - \log(\mathbf{A}_{x,x_{t-1}}) - \log(\mathbf{B}_{y_t,x}) \big) \qquad \text{for } t = 1, \dots, T. \end{aligned}$$

We also record the value of x_{t-1} (the back-pointer) for which the minimum was achieved at time $t-1$ if we pass through x at time t as

$$\begin{aligned} \xi_{t-1}(x) &= \text{argmin}_{x_{t-1}} \big(\psi_{t-1}(x_{t-1}) - \log(\mathbf{A}_{x,x_{t-1}}) - \log(\mathbf{B}_{y_t,x}) \big) \\ &= \text{argmin}_{x_{t-1}} \big(\psi_{t-1}(x_{t-1}) - \log(\mathbf{A}_{x,x_{t-1}}) \big). \end{aligned}$$

The optimal path $(\hat{x}_0, \dots, \hat{x}_T)$ is then calculated in a backwards recursion as

$$\begin{aligned} \hat{x}_T &= \text{argmin}_x \psi_T(x) \\ \hat{x}_{t-1} &= \xi_{t-1}(\hat{x}_t) \qquad \text{for } t = T-1, T-2, \dots, 0. \end{aligned}$$

This is sometimes called the Viterbi algorithm.

7.5 Parameter estimation via the EM-algorithm

Assume we have a distribution with discrete observed variables Y and latent X and unknown parameter θ (same ideas work for continuous variables). We would like to get the Maximum-likelihood estimate of θ as

$$\hat{\theta} = \text{argmax}_\theta \ \ell(\theta),$$

where the log-likelihood $\ell(\theta)$ is given by $\log P_\theta(Y = y)$ if the observations of Y are y.
The problem is that $P_\theta(Y)$ is not easily available in tractable form. What is available is the likelihood $P_\theta(Y, X)$ if we could observe the latents X as well as. The EM (Expectation-Maximization; also called Baum-Welch for HMMs) algorithm greedily optimizes the likelihood by alternating between

(i) estimating the latent variables X, given the observed variables Y and the current parameter estimate, and then

(ii) updating the parameter estimates in a second step.

Starting from some initial estimate $\theta^{(1)}$, the steps are for an iteration $t = 1, \ldots,$

E-step (Expectation): Compute the conditional distribution of the latent variables X, given the observed variables and the current parameter estimates $\theta^{(t)}$:

$$P_{\hat{\theta}^{(t)}}(X|Y = y).$$

This is similar to type of inference we discussed. Define the expected log-likelihood under the distribution of X implied by the current parameter estimate as

$$Q_t(\theta) := E_{\hat{\theta}^{(t)}}\Big[\log P_\theta(Y = y, X)|Y = y\Big],$$

where the expectation is with respect to the random X, conditional on $Y = y$ and the current parameter estimate $\hat{\theta}^{(t)}$.

M-step (Maximization): update the parameters as

$$\hat{\theta}^{(t+1)} = \text{argmax}_\theta \; Q_t(\theta).$$

We get monotonically increasing likelihood

$$\ell(\hat{\theta}^{(t)}) \geq \ell(\hat{\theta}^{(t-1)}),$$

that is the parameter estimates $\hat{\theta}^{(t)}$ will converge to a **local** maximum of the likelihood for $t \to \infty$. Depending on the starting value we might reach the global optimum but this is not guaranteed.
The EM algorithm appears very often in practice. Compare also the applications in eg clustering we discuss. Sometimes the expectation is replaced with just computing the most-likely state of x', given $Y = y$ and $\hat{\theta}^{(t)}$ and setting $Q_t(\theta) := \log P_\theta(Y = y, x')$. This is what happens in the K-means clustering algorithm, for example. For HMMs it corresponds to computing the most likely sequence with the Viterbi algorithm, as discussed. The approach is sometimes referred to as hard-EM. The result will again depend on the chosen starting values in general.
But why do we get monotonically increasing likelihood? The proof sheds some more light on EM. To start with, it holds for all functions f over the space $\mathcal{X}$ of the hidden variables with $\sum_{x\in\mathcal{X}} f(x) = 1$ and $f(x) \geq 0$ for all $x \in \mathcal{X}$,

$$\begin{aligned}
\ell(\theta) &= \log P_\theta(Y = y) \\
&= \log \sum_{x\in\mathcal{X}} P_\theta(Y = y, X = x) \\
&= \log \sum_{x\in\mathcal{X}} f(x) \frac{P_\theta(Y = y, X = x)}{f(x)} \\
&\geq \sum_{x\in\mathcal{X}} f(x) \log \frac{P_\theta(Y = y, X = x)}{f(x)},
\end{aligned}$$

where the last inequality uses Jensens inequality and the fact that log is a concave function. Note that equality holds iff

$$\frac{P_\theta(Y = y, X = x)}{f(x)}$$

does not depend on x, for example if $f(x) = P_\theta(X = x|Y = y)$.
Hence there exists a constant $c > 0$ at each time-step[5] such that

$$\ell(\theta') \geq Q_t(\theta') + c \text{ for all } \theta' \qquad \text{and } \ell(\hat{\theta}^{(t)}) = Q_t(\hat{\theta}^{(t)}) + c.$$

Hence

$$\ell(\hat{\theta}^{(t+1)}) \geq Q_t(\hat{\theta}^{(t+1)}) + c \geq Q_t(\hat{\theta}^{(t)}) + c = \ell(\hat{\theta}^{(t)}),$$

where the first inequality is due to the argument just above and the second is true as $\hat{\theta}^{(t+1)}$ is, by definition, maximizing Q_t.

7.6 Kalman filter

Suppose we have a linear dynamical system

$$X_t = AX_{t-1} + V_t$$
$$Y_t = BX_t + W_t,$$

where $X_t \in \mathbb{R}^p$ is the latent state, $Y_t \in \mathbb{R}^q$ the made observations, W_t the so-called process noise and V_t the measurement noise.
Under a Gaussian noise assumption

$$W_t \sim \mathcal{N}(0, W), \qquad V_t \sim \mathcal{N}(0, V),$$

the joint vector of latent and observations over $t = 1, \ldots, T$ will have a joint Gaussian distribution. In principle it is thus easy to derive the conditional distribution of, say, $X_t|Y_1^t$. Let $Z \in \mathbb{R}^p$ be a random vector with a Gaussian distribution,

$$Z \sim \mathcal{N}(\mu, \Sigma).$$

Then the conditional distribution of Z_k, conditional on $Z_S = z_s$ for some $S \subseteq \{1, \ldots, p\}$, will again be Gaussian

$$Z_k|Z_s = z_s \quad \sim \mathcal{N}(\mu_{k|S}, \Sigma_{k,S}),$$

with

$$\mu_{k|S} = \mu_k + \Sigma_{k,S}\Sigma_{S,S}^{-1}(z_S - \mu_S)$$
$$\Sigma_{k|S} = \Sigma_{k,k} - \Sigma_{k,S}^t\Sigma_{S,S}^{-1}\Sigma_{S,k}$$

The problem with the direct approach is that the dimensionality of S grows like pT if we condition on the observations $Y_1^T = (Y_1, \ldots, Y_T)$.
Using the structure of the HMM again and the same message-passing as in the discrete case, we can define

$$\hat{X}_{t|t} = E(X_t|Y_1^t)$$
$$\hat{X}_{t+1|t} = E(X_{t+1}|Y_1^t)$$
$$\Sigma_{t|t} = Cov(X_t|Y_1^t)$$
$$\Sigma_{t+1|t} = Cov(X_{t+1}|Y_1^t)$$

[5] namely entropy of $f(x) = P_{\hat{\theta}^{(t)}}(X = x|Y = y)$ as entropy is $-\sum_x f(x)\log f(x)$

The updates are usually split again into two part,s the time- and the measurement update. The time-update concerns

$$\hat{X}_{t|t} \to \hat{X}_{t+1|t}$$
$$\Sigma_{t|t} \to \Sigma_{t+1|t},$$

while the measurement update concerns

$$\hat{X}_{t+1|t} \to \hat{X}_{t+1|t+1}$$
$$\Sigma_{t+1|t} \to \Sigma_{t+1|t+1},$$

taking into account the new observation made at time t.
Conditioning on Y_1^t, we get

$$X_{t+1}|Y_1^t = (AX_t + V_t)|Y_1^t = AX_t|Y_1^t + V_t$$
$$Y_{t+1}|Y_1^t = (BX_{t+1} + W_t)|Y_1^t = BX_{t+1}|Y_1^t + W_t.$$

The first equation yields the so-called **time-update** (updating $t \to t+1$ without using the new observation at time $t+1$)

$$\hat{X}_{t+1|t} = E(X_{t+1}|Y_1^t) = A\hat{X}_{t|t}$$
$$\Sigma_{t+1|t} = A\Sigma_{t|t}A^t + V$$

The second equation yields the measurement update (where the new observation Y_{t+1} is used to update the conditional distribution). Note that the distribution of $(X_{t+1}, Y_{t+1})|Y_1^t$ has a multivariate Gaussian distribution

$$(X_{t+1}, Y_{t+1})|Y_1^t \quad \sim \quad \mathcal{N}(\mu, S),$$

with

$$\mu = \begin{pmatrix} \hat{X}_{t+1|t} \\ B\hat{X}_{t+1|t} \end{pmatrix}, \qquad S = \begin{pmatrix} \Sigma_{t+1|t} & \Sigma_{t+1|t}^t B^t \\ B\Sigma_{t+1|t} & B^t\Sigma_{t+1|t}B + W \end{pmatrix}.$$

Hence we get the **measurement update** as

$$\hat{X}_{t+1|t+1} = \hat{X}_{t+1|t} + \Sigma_{t+1|t}^t B^t (B^t\Sigma_{t+1|t}B + W)^{-1}(Y_t - B\hat{X}_{t+1|t})$$
$$\Sigma_{t+1|t+1} = \Sigma_{t+1|t} - \Sigma_{t+1|t}^t B^t (B^t\Sigma_{t+1|t}B + W)^{-1}B\Sigma_{t+1|t}$$

Perhaps surprisingly, the error covariance can be computed ahead of time (without seeing any observations).

Steady-state Kalman filter If $\Sigma_{t+1|t}$ converges to a Σ^* (which it will in general), then Σ^* is the solution of a Ricatti-type equation

$$\Sigma^* = A\Sigma^* A^t + V - A(\Sigma^*)^t B^t (B^t\Sigma^* B + W)^{-1}B\Sigma^* A^t.$$

The estimated means follow then the recursion

$$\hat{X}_{t+1|t} = A\hat{X}_{t|t-1} + L(Y_t - B\hat{X}_{t|t-1}),$$

where $L = A(\Sigma^*)^t B^t (B^t\Sigma^* B + W)^{-1}$ is the so-called Kalman gain. The first term updates the guess according to the dynamics of the system while the second corrects it by using the newly available information via the new observation.

References

Brockwell, P. J. and Davis, R. A. (1991). *Time Series: Theory and Methods*. Springer series in statistics. Springer, New York, NY, 2. ed., reprint of the 1991 ed edition. OCLC: 819807707.

Shumway, R. H. and Stoffer, D. S. (2006). *Time Series Analysis and Its Applications - With R Examples*. Springer Texts in Statistics. Springer, 2 edition.