M9121 Time Series I

Štěpán Zapadlo

2024-03-02

Table of contents

Preface

This book follows the course M9121 Time Series I taught by assoc. prof. Kraus on SCI MUNI, Department of Mathematics on Statistics, as it appeared in Fall 2023.

The course offers a comprehensive coverage of selected fundamental methods and models for time series. The course covers theoretical foundations, statistical models and inference, software implementation, application and interpretation.

The students will gain a deeper understanding of the methods and their relations and learn to recognize situations that can be addressed by the models discussed in the course, choose an appropriate model, implement it and interpret the results.

Česká verze:

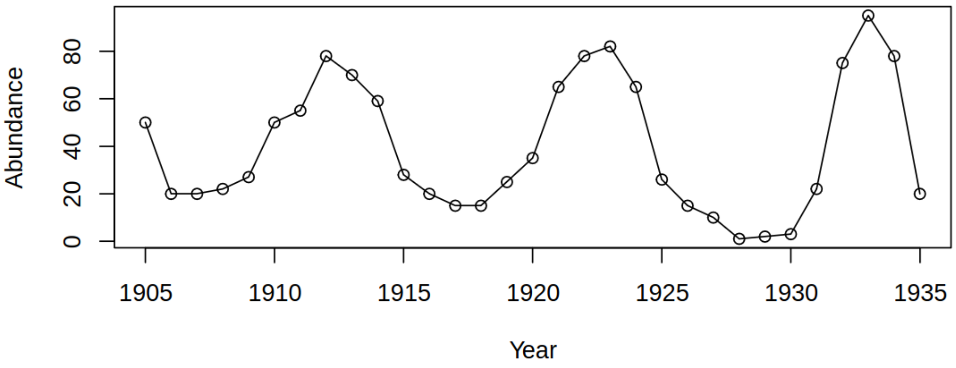
Předmět se věnuje podrobnému výkladu některých základních metod a modelů pro časové řady. Kurs pokrývá teoretické základy, statistické modely a inferenci, softwarovou implementaci, aplikaci a interpretaci.

V kursu studenti získají hlubší pochopení vlastností metod a souvislostí mezi nimi, naučí se rozeznat situace, které lze řešit s pomocí diskutovaných modelů, jsou schopni vybrat vhodný model z této třídy, implementovat jej a interpretovat jeho výsledky.

1 Introduction

Time series data have distinct features – data are collected sequentially over time, order of observations matters and observations *do not arise independently, they are serially dependent*. Also, they serve many purposes, e.g. we can use them to understand or model the stochastic mechanism that gives rise to an observed series or to predict or forecast the future values of a series based on the history, and to quantify the uncertainty of predictions.

1.1 Abundance of Canadian hare

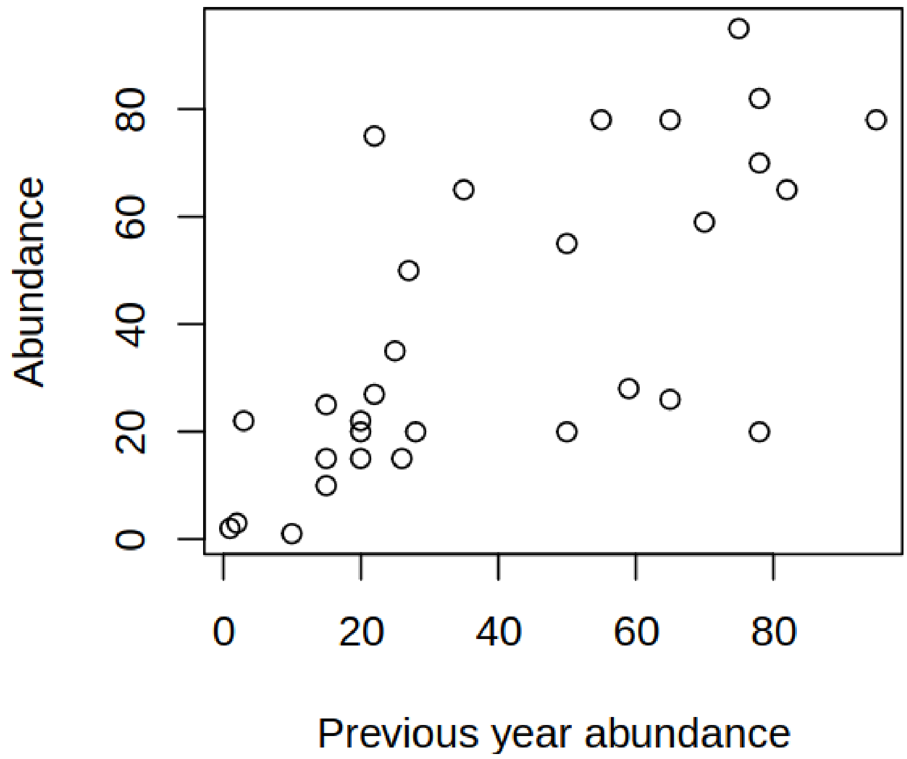


We can surely make the following observations – there is a stable level (oscillating around some stable mean) and no obvious trend. More so, the neighboring values are very closely related, there are no large changes from one year to the next. From even closer analysis a question arises – are consecutive years related? Could be useful for the prediction

Our old approach – Linear regression

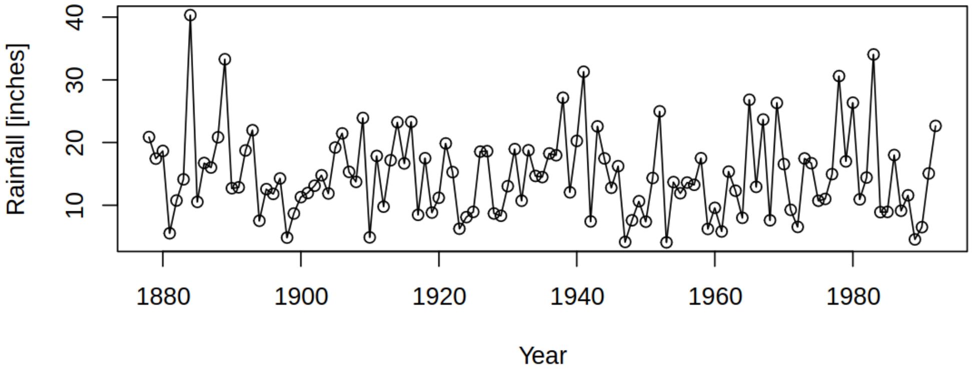
It is easy to see that naive linear regression would not provide good predictive capabilities - it would predict a point on the *mean line*

But we can also deduce from looking at year-to-year changes, that there is an obvious upward trend, low/high values tend to be followed by low/high values and there is a positive serial correlation.

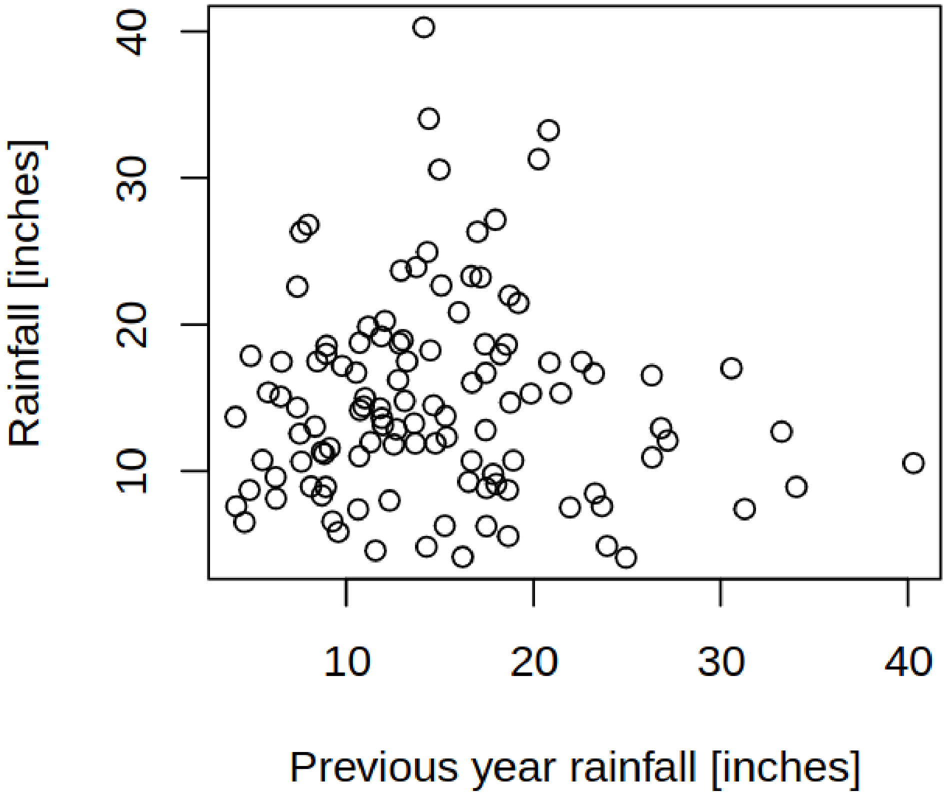


1.2 LA annual rainfall

Consider now the following data about annual rainfall in Los Angeles:

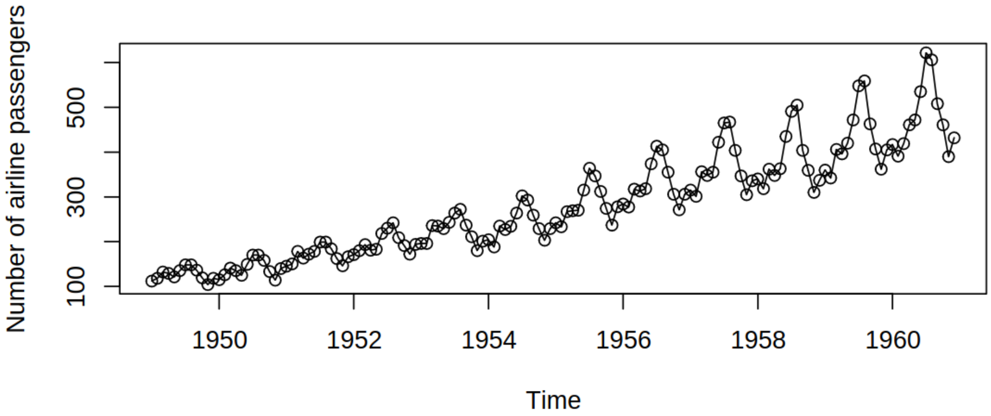


And again we can deduce from the data that there is a considerable variation and no obvious trends (this can be expected a priori).



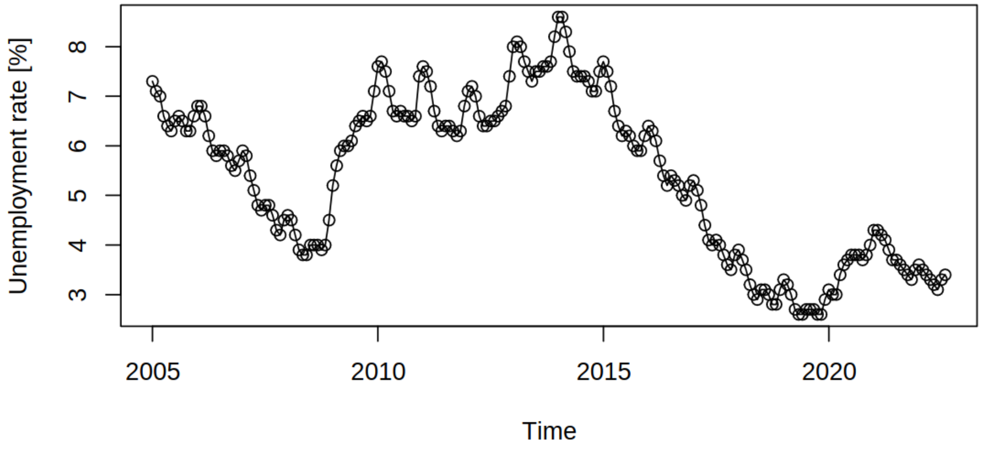
And even when looking at year-to-year changes (so-called a *lag plot*) it shows no general pattern, little correlation between consecutive years and hints at difficult forecasting.

1.3 Airline passengers

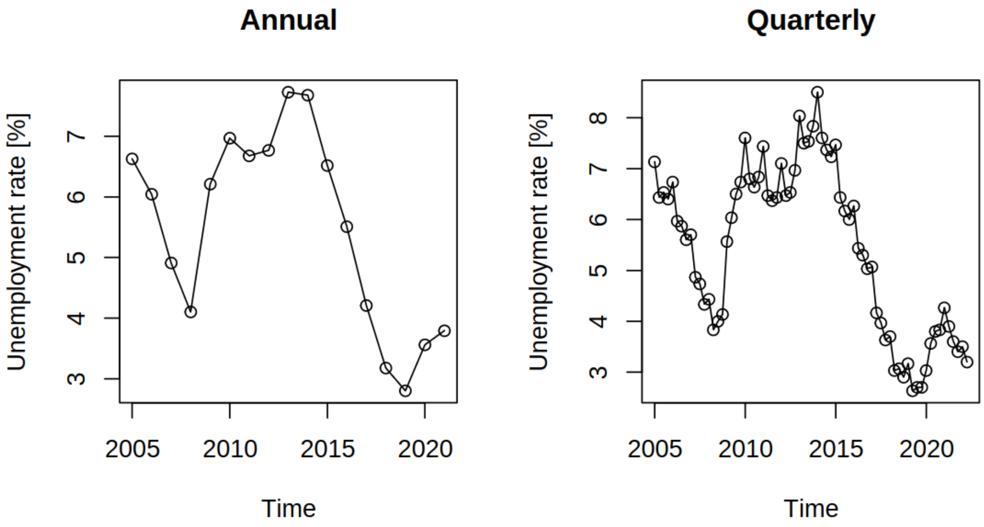


Here is another example of a time series, where this time we can see the following attributes of this data. There is an obvious global increasing trend with a seasonal pattern of behavior and an increasing variance.

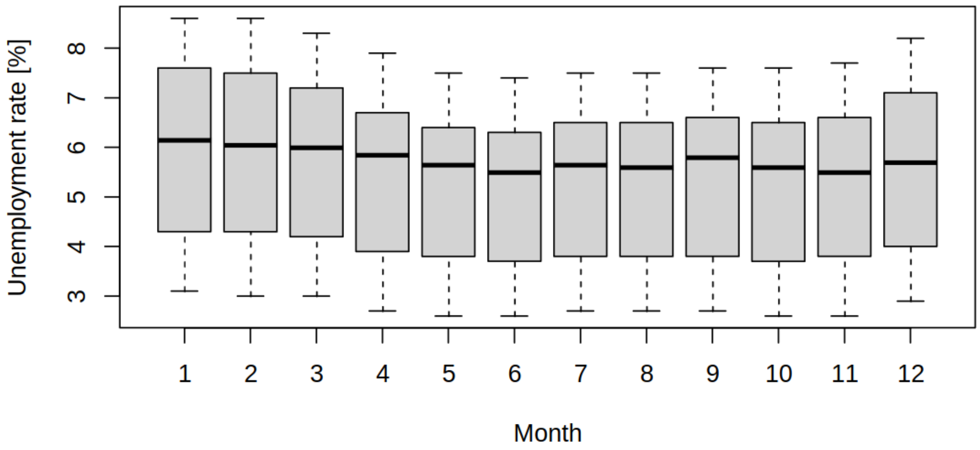
1.4 CZ unemployment rate



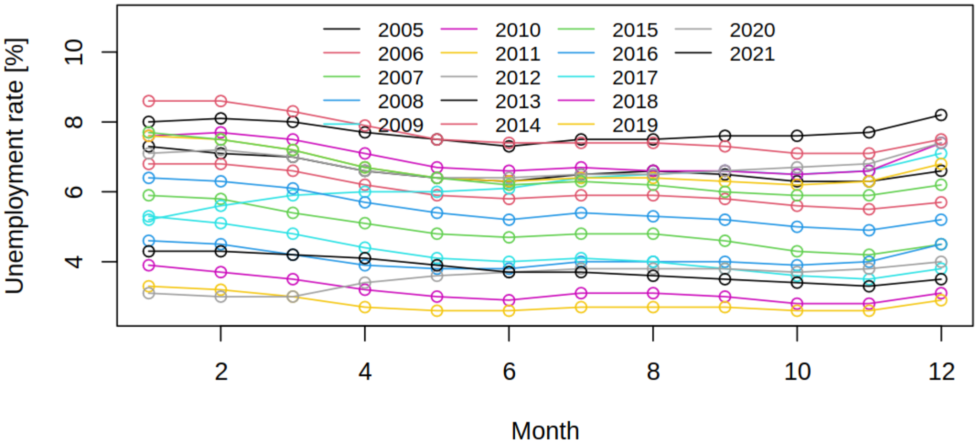
Unlike the previous example, a trend (or a tendency) can be seen here as well, but in a much more complicated manner (in a shorter timeframe, no global trends present). Though it still contains seasonal effects ("oscillations" of sorts)



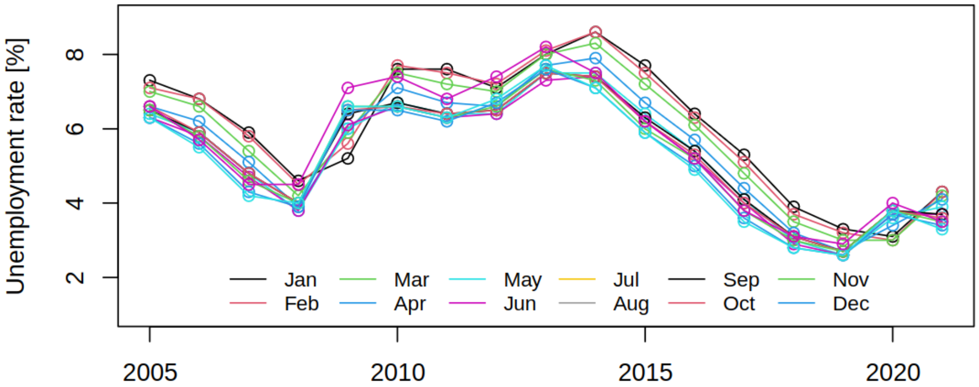
By aggregation of the data, we can preserve the overall trend, but lose the seasonal effects (which we might consider as a noise of sorts). On the other hand, we can also study purely the seasonal trends



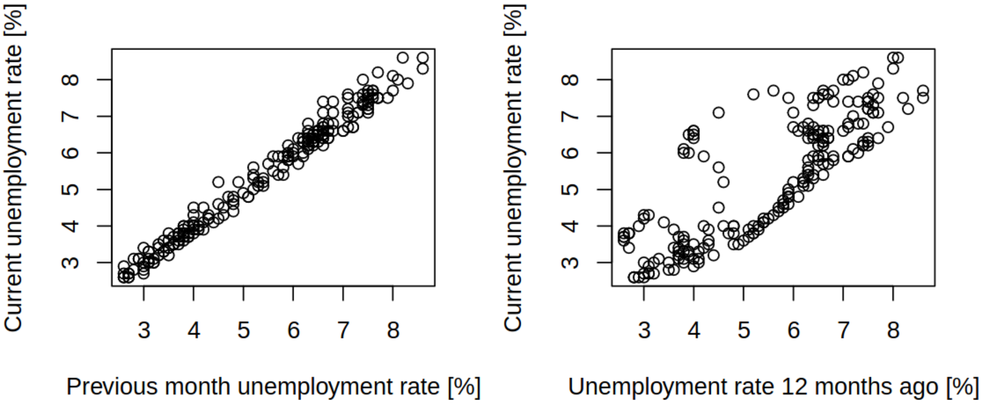
Such visualization shows variability between seasonal values and variability/trends within seasons. Also notice that summer brings a lower unemployment rate, as could be seen in the full data (but less clearly). There are multiple ways of visualizing the same data and each shows a different thing. The following plot shows a correlation between months and the outlier year 2017.



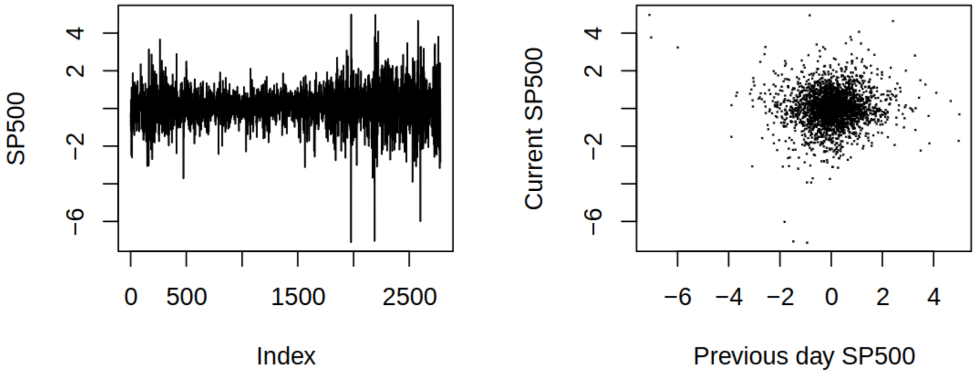
Conversely, next, we can notice the global trends and correlations between years.



Lastly, based on this information we might want to see the *lag plots*, where we will see a strong correlation between consecutive values and a strong correlation between values 12 months apart.



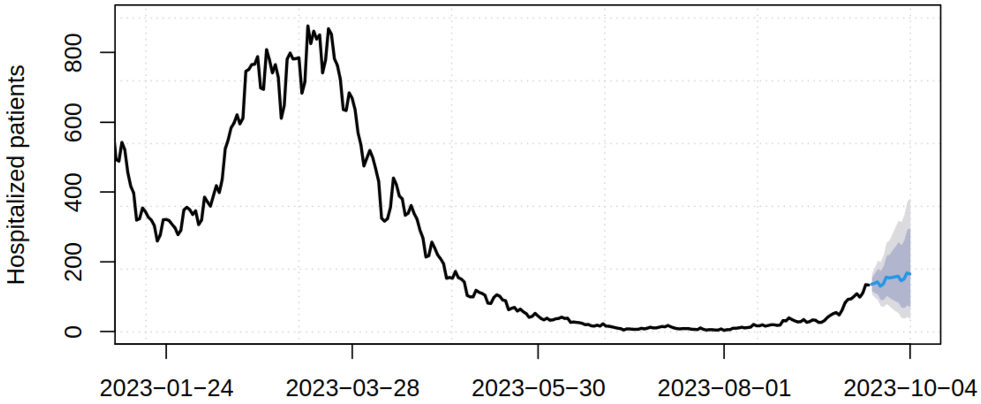
1.5 S&P500 Index series (1990–1999)



Yet another dataset shows other features of time series we can come across. Here that is a changing variance, variability (volatility), which occurs in clusters, and no obvious relationship between consecutive values. These features are typical of financial time series (next semester).

1.6 Covid-19 hospital occupancy

For a Covid-19 disease hospital acceptance (incoming occupancy) rate:



In this case, prediction (and decisions based on that) was the driving force behind this model and as such it needed to include prediction with uncertainty quantification

i Note

Time series analysis provides short-term predictions, rather than long-term extrapolations that require a detailed model of the underlying phenomenon.

From the Covid-19 we can see that ARIMA models (covered in this course) are still useful (and can be among the best models available).

2 Fundamental concepts

2.1 A Stochastic Point of View

The observed time series $X_1, \dots, X_n$ is a sequence of numbers and our aim is to understand the mechanism that generated the series and make use of it. Therefore we need a mathematical model, then capture randomness, by which we mean a model for uncertainty and/or limited knowledge. Partial information also will need to be addressed. The observed data are seen as a sequence of realizations of random variables and as such we need to study them all together, including their relationships.

Definition 2.1. A stochastic process is a family of random variables $\{X_t : t \in T\}$ defined on a probability space $(\Omega, \mathcal{A}, P)$.

Where in the Definition 2.1 the meaning of used symbols is as follows:

- T is the index set;
- $\{X_t : t \in T\}$ can be seen as a function of t and ω, i.e., $\{X(t, \omega) : t \in T, \omega \in \Omega\}$;
- for a fixed $t \in T$, $X_t = X_t(\cdot) = \{X_t(\omega) : \omega \in \Omega\}$ is a random variable defined on Ω, i.e. the process is seen as a collection of random variables indexed by T;
- for a fixed $\omega \in \Omega$, $X = X(\omega) = \{X_t(\omega) : t \in T\}$ is a function on T, i.e. the process is seen as a random function.

As such, we call a realization of $\{X_t : t \in T\}$ a *sample path/trajectory/realization*.

2.1.1 Types of stochastic processes

There are many types of stochastic processes, but most importantly (for us):

- time series (discrete-time processes)
  - $T = \mathbb{Z}$ (or $T \subset \mathbb{Z}$, e.g., $T = \mathbb{N}$);
  - the observed time series is seen as a realization of the stochastic process $\{X_t : t \in \mathbb{Z}\} = \{\dots, X_{-1}, X_0, X_1, \dots\}$;
  - a time series is a random sequence and a sequence of random variables;
- continuous-time processes: $T = \mathbb{R}$ or $T = [a, b]$ (random function);
- spatially indexed processes: $T = \mathbb{R}^d$ or $T \subset \mathbb{R}^d$ (random fields);
- processes on lattices: $T = \mathbb{Z}^d, \dots$;

- spatio-temporal processes: $T = \mathbb{R}^d \times [a,b], ...$;
- etc.

Although we will see other types of index sets in other courses, in our lectures we will stick to time series only (where time is discretized).

2.1.2 Discrete sampling of time series

Time series are random sequences, i.e., $T = \mathbb{Z}$, where the discretization can be due to discrete sampling of a continuous-time process, e.g. closing prices of a share, electrical signal in telecommunications, aggregation of a continuous time process, e.g. daily precipitation, monthly electricity production, or a discrete realization, e.g. regularly repeated medical experiment.

2.2 Distribution of a stochastic process

We can define a finite-dimensional distributions for all $k \in \mathbb{N}, t_1, \dots, t_k \in T$ as

$$F_{t_1,\dots,t_k}(x_1, \dots, x_k) = P(\mathrm{X}_{t_1} \leq x_1, \dots, \mathrm{X}_{t_k} \leq x_k).$$

Here, we took a random vector $\mathrm{X}_{t_1}, \dots, \mathrm{X}_{t_k}$ and looked at its joint distribution.

A system of distribution functions $\{F_{t_1,\dots,t_k} : t_1, \dots, t_k \in T, k \in \mathbb{N}\}$ is called **consistent**, if it has following properties

- $\lim_{x_{k+1}\to\infty} F_{t_1,\dots,t_k,t_{k+1}}(x_1, \dots, x_k, x_{k+1}) = F_{t_1,\dots,t_k}(x_1, \dots, x_k)$;
- $F_{t_1,\dots,t_k}(x_1, \dots, x_k) = F_{t_{i_1},\dots,t_{i_k}}(x_{i_1}, \dots, x_{i_k})$ for all permutations $(i_1, \dots, i_k)$ of $(1, \dots, k)$.

A stochastic process has always a consistent system of distributions.

Theorem 2.1 (Daniell–Kolmogorov). *Let $\{F_{t_1,\dots,t_k} : t_1, \dots, t_k \in T, k \in \mathbb{N}\}$ be a consistent system of distribution functions. Then there exists a stochastic process $\{\mathrm{X}_t : t \in T\}$ such that for all $k \in \mathbb{N}, t_1, \dots, t_k \in T$ the joint distribution function of $(X_{t_1}, \dots, X_{t_k})$ is $F_{t_1,\dots,t_k}$.*

For the distribution of a stochastic process, it holds that a process whose finite-dimensional distributions are all multivariate normal is called *Gaussian*. Often, much information is contained in means, variances and covariances and thus, focusing on the first and second moments is often sufficient. Moreover, if the joint distributions are multivariate normal, the first and second moments completely determine the joint distribution

Definition 2.2 (Mean, autocovariance, autocorrelation). For a stochastic process, we define:

- mean function
$$\mu_t = \mathbb{E}\,\mathrm{X}_t, t \in \mathbb{Z};$$

- autocovariance function (*acov*)

$$\gamma(s,t) = \operatorname{cov}(X_s, X_t) = \mathbb{E}\left((X_s - \mu_s)(X_t - \mu_t)\right),\ s,t \in \mathbb{Z};$$

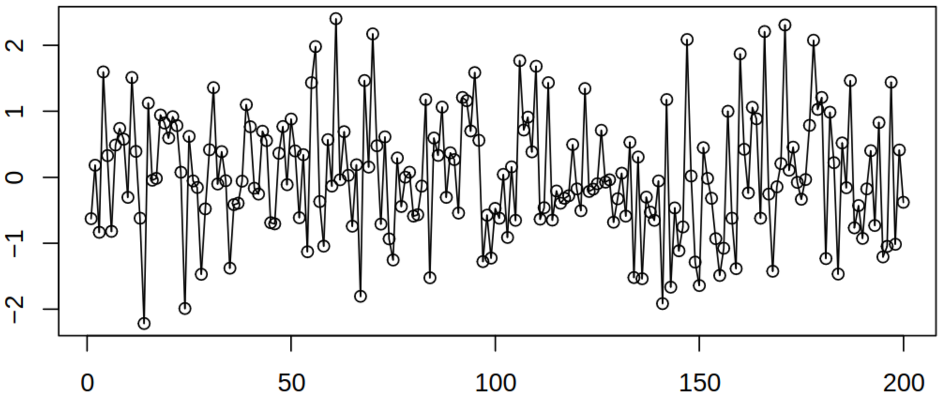
- autocorrelation function (*acf*)

$$\rho(s,t) = \operatorname{cor}(X_s, X_t) = \frac{\operatorname{cov}(X_s, X_t)}{\sqrt{\operatorname{var} X_s \operatorname{var} X_t}},\ s,t \in \mathbb{Z}.$$

2.3 Examples

2.3.1 White noise process

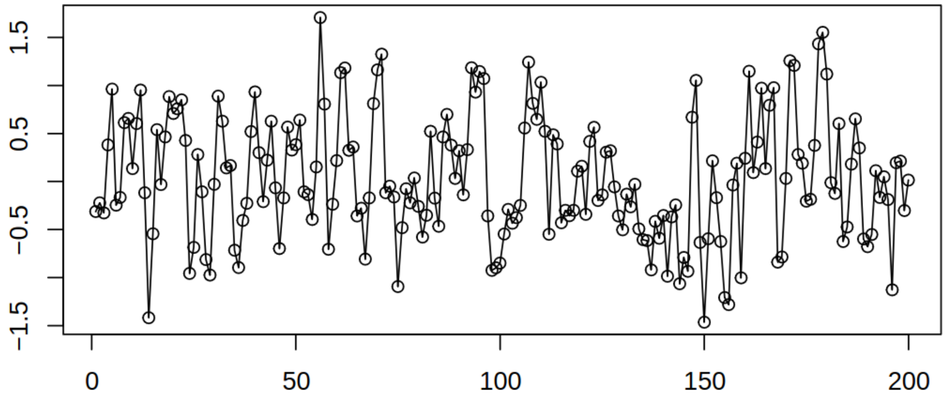
Let $\{\varepsilon_t : t \in \mathbb{Z}\}$ be a sequence of uncorrelated random variables with mean 0 and variance σ^2. That is, $\mathbb{E}\, X_t = 0, \operatorname{var} X_t = \sigma^2, \operatorname{cov}(X_s, X_t) = 0, s \neq t$ and we introduce notation $\{\varepsilon_t\} \sim \mathrm{WN}\left(0, \sigma^2\right)$.



2.3.2 Moving average

Let $\{\varepsilon_t : t \in \mathbb{Z}\}$ be $\mathrm{WN}\left(0, \sigma^2\right)$ and define

$$X_t = (\varepsilon_t + \varepsilon_{t-1})/2.$$



Then the mean function is given by

$$\mu_t = \mathbb{E}\left(\varepsilon_t + \varepsilon_{t-1}\right)/2 = 0,$$

and the variance by

$$\gamma(t,t) = \operatorname{var}\left((\varepsilon_t + \varepsilon_{t-1})/2\right) = \frac{\sigma^2}{2}.$$

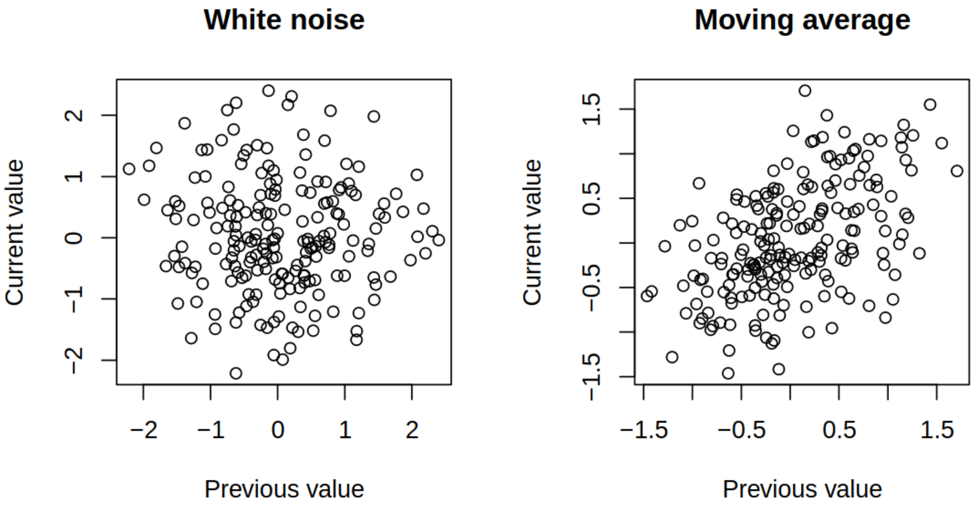
Then autocovariance of the process is

$$\gamma(t,t+1) = \operatorname{cov}\left((\varepsilon_t + \varepsilon_{t-1})/2, (\varepsilon_{t+1} + \varepsilon_t)/2\right) = \frac{\sigma^2}{2},$$
$$\gamma(t,t+h) = \operatorname{cov}\left((\varepsilon_t + \varepsilon_{t-1})/2, (\varepsilon_{t+h} + \varepsilon_{t+h-1})/2\right) = 0, \quad |h| > 1$$

and lastly, the autocorrelation can be expressed as

$$\rho(s,t) = \begin{cases} 1, & s = t, \\ 0.5, & |s-t| = 1, \\ 0, & |s-t| > 1, \end{cases}$$

so in other words, we have a correlation between consecutive values, which is constant in time.



2.3.3 Random walk

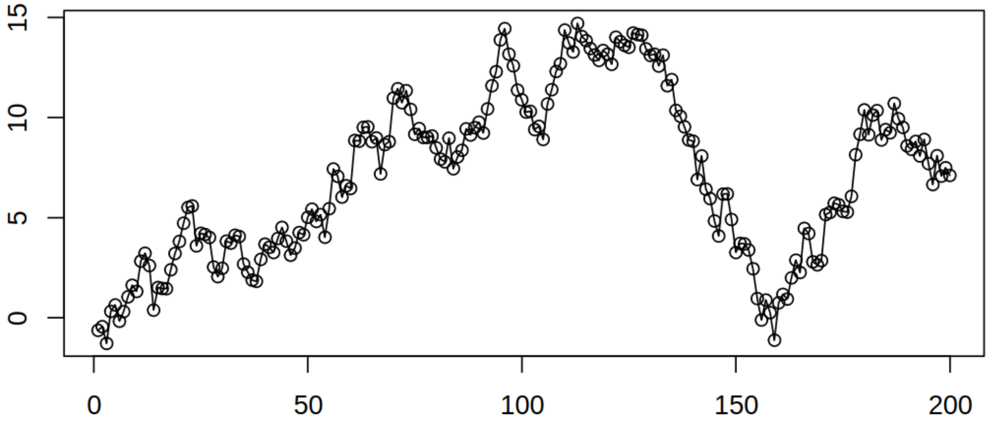
Random walk is a stochastic process that emerges from the cumulative sum of white noise. Let $\varepsilon_1, \varepsilon_2, \ldots$ be a sequence of independent, identically distributed random variables with mean 0 and variance σ^2. Then define

$$X_t = \sum_{j=1}^{t} \varepsilon_j, \ t = 1, 2, \ldots$$

or in other words

$$X_1 = \varepsilon_1, \quad X_{t+1} = X_t + \varepsilon_t, \ t = 2, 3, \ldots,$$

where ε_t are the steps taken by the "random walker", X_t is his position at time t.



3 Autocorrelation and stationarity

⚠ I was absent

These notes are just a summary of the presentation given to us. As such, it is hard to comment on the true intent or intuition behind some of the presented results or sentences.
Thank you for your understanding.

As a starter, it is good to realize, that while correlation is useful, one should be mindful of what it really means.

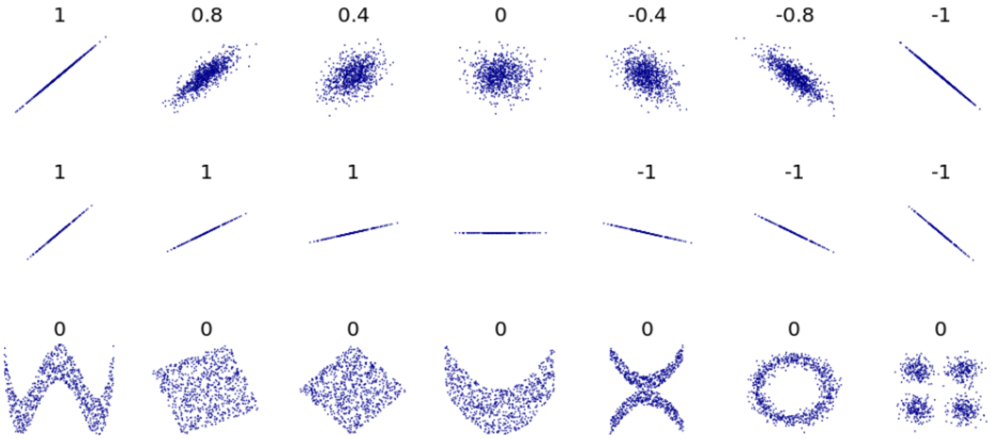


Figure 3.1: Demonstration of correlation

3.1 Properties of autocovariance

As we've seen earlier in the Definition 2.2, the autocovariance function is defined as

$$\gamma(s,t) = \operatorname{cov}(X_s, X_t), \quad s,t \in T \subset \mathbb{R}.$$

We may notice that for any $k \in \mathbb{N}$, $t_1, \ldots, t_k \in T$, $c_1, \ldots, c_k \in \mathbb{R}$

$$0 \leq \operatorname{var}\left(\sum_{j=1}^{k} c_j X_{t_j}\right) = \sum_{j=1}^{k}\sum_{l=1}^{k} c_j c_l \gamma(t_j, t_l).$$

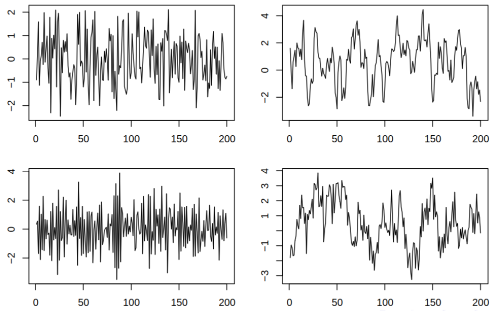
Hence the autocovariance function of a process with finite second moments is *non-negative definite*. The converse holds true as well – for any non-negative definite function g on $T \times T$ there exists a stochastic process such that g is its autocovariance function. Because one can take matrices of the form

$$\boldsymbol{V}_{t_1,\dots,t_k} = \left(g(t_j, t_l)\right)_{j,l=1}^{k}$$

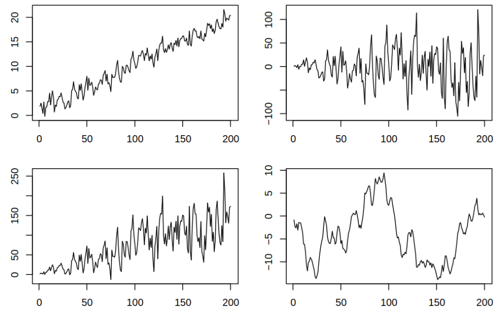
and consider the multivariate Gaussian distributions $\mathcal{N}_k\left(0, \boldsymbol{V}_{t_1,\dots,t_k}\right)$, then by the Daniell-Kolmogorov Theorem 2.1 there exists a Gaussian process with these finite-dimensional distributions.

3.2 Stationarity

Stationarity is a concept best introduced by illustrations.



(a) Simulated **stationary** processes



(b) Simulated **non-stationary** processes

Definition 3.1 (Strict stationarity)**.** A process $\{X_t : t \in \mathbb{Z}\}$ is said to be **strictly stationary** if the joint distribution of $X_{t_1}, \dots, X_{t_k}$ is the same as the joint distribution of $X_{t_1+h}, \dots, X_{t_k+h}$ for all $k \in \mathbb{N}, t_1, \dots, t_k \in \mathbb{Z}, h \in Z$, that is

$$F_{t_1,\dots,t_k}(x_1, \dots, x_k) = F_{t_1+h,\dots,t_k+h}(x_1, \dots, x_k).$$

Definition 3.2 (Weak stationarity)**.** A process $\{X_t : t \in \mathbb{Z}\}$ is said to be **weakly (weak-sense, second-order) stationary** if the μ_t is **constant** in time and $\gamma(s, t)$ depends only on the difference $s - t$.

Any strictly stationary defined in Definition 3.1 process that has a finite mean and a covariance is also stationary in the weak sense. Also, if all joint distributions are Gaussian, then weak and strong stationarity are equivalent.

Tip

By convention, by simply "stationarity" we mean weak-sense stationarity – Definition 3.2.

Example 3.1. Let us define

$$X_t = a\cos\left(2\pi(ft + \Phi)\right), \quad t \in \mathbb{Z}, \tag{3.1}$$

where a is the amplitude, f frequency (e.g. $f = 1/12$) and $\Phi \in [0, 1]$ is the phase. Consider now a random phase given by $\Phi \sim \mathrm{Unif}([0, 1])$. It can be easily computed that

$$\mathbb{E}\, X_t = \int_0^1 a\cos\left(2\pi(ft + \Phi)\right) \phi = 0.$$

From the Definition 2.2 and (3.1), it can be shown the autocovariance function γ of this process has the form

$$\begin{aligned}\gamma(s,t) &= \int_0^1 a\cos\left(2\pi(fs + \Phi)\right) a\cos\left(2\pi(ft + \Phi)\right) \Phi \\ &= \frac{a^2}{2}\cos\left(2\pi(s - t)\right)\end{aligned} \tag{3.2}$$

and as such, the autocorrelation is $\rho(h) = \cos(2\pi f h)$, $h \in \mathbb{Z}$. Hence this stationary process fulfills the criteria for WSS, see Definition 3.2.

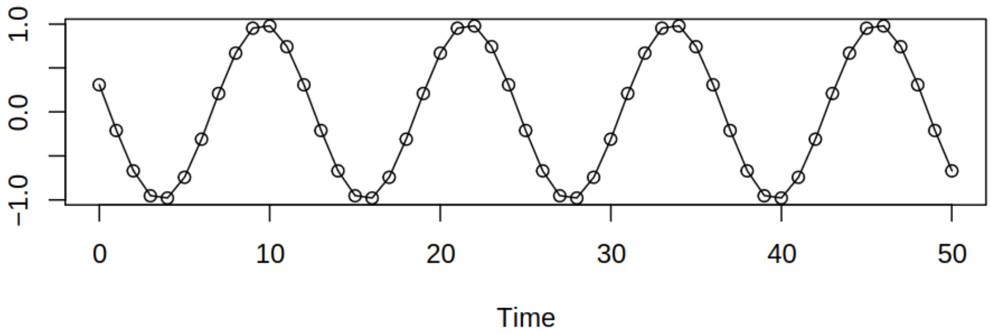


Figure 3.3: "Non-stationary looking" trajectory

It should be noted that stationarity is the property **of the distribution**, not of the realization. Therefore *one trajectory may not look stationary*, but the process as a whole may be stationary, see Figure 3.3. In this example, it is precisely the random phase Φ, that makes the process stationary – independent random amplitude and phase *would not be stationary*.

i Stationarity is good for inference

Stationarity implies that properties of a given process are stable (aka *do not change*) in time. So with more and more data, we collect more and more information about the same, invariant structure. As such, averaging makes sense – without the stationarity averaging wouldn't be meaningful, unless we know how the properties like the mean and the autocovariance function evolve in time.

Let now $\{X_t : t \in \mathbb{Z}\}$ be a *stationary* time series. Let us assume we want to estimate the mean $\mu = \mathbb{E}\, X_t$, autocovariance function $\gamma(h) = \mathrm{cov}(X_t, X_{t+h})$ and autocorrelation function

$\rho(h) = \operatorname{cor}(X_t, X_{t+h})$ given the observed data $X_1, \dots, X_n$. Therefore our goals are to find estimators $\hat{\mu}, \hat{\gamma}(h), \hat{\rho}(h)$, understand their properties and use them in statistical inference and graphics.

3.2.1 Estimation of the mean

As was stated before, we assume that $\mathbb{E}\, X_t = \mu$ for all t. An obvious choice of the estimator would be the sample mean

$$\hat{\mu} = \overline{\mathbf{X}} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Its expected value is given by

$$\mathbb{E}\, \hat{\mu} = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \mu,$$

so $\hat{\mu}$ is *unbiased* (regardless of the covariance structure of the process). To judge the quality of the estimator, let us look at its variance (see (3.2) for reference for derivation)

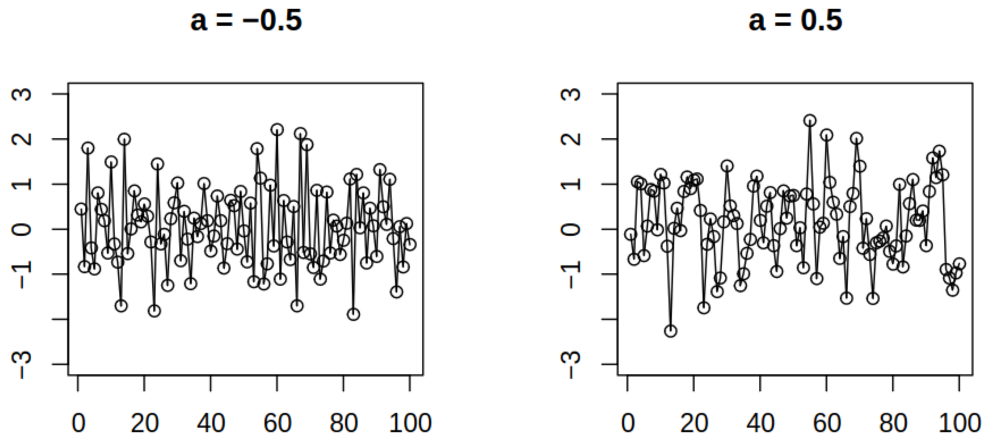
$$\begin{aligned}
\operatorname{var}\hat{\mu} &= \operatorname{var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\operatorname{cov}(X_i, X_j) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\gamma(i-j) \qquad (3.3)\\
&= \frac{1}{n^2}\left(n\gamma(0) + 2(n-1)\gamma(1) + 2(n-2)\gamma(2) + \cdots + 2\gamma(n-1)\right) \\
&= \frac{\gamma(0)}{n}\left(1 + 2\sum_{i=1}^{n-1}\frac{n-i}{n}\rho(i)\right).
\end{aligned}$$

When we perform estimation given by (3.3) on white noise, where $\rho(h) = 0$ for $h \neq 0$, we get $\operatorname{var}\hat{\mu} = \frac{\gamma(0)}{n}$, so it follows the term $2\sum_{i=1}^{n-1}\frac{n-i}{n}\rho(i)$ is a correction reflecting the impact of autocorrelation.

Example 3.2. Let a be a parameter, consider $\{\varepsilon_t\} \sim \mathrm{WN}\left(0, \sigma^2/(1+a^2)\right)$ and define

$$X_t = \varepsilon_t + a\varepsilon_{t-1}.$$

We can compute $\gamma(0) = \sigma^2$, $\gamma(1) = a\sigma^2/(1+a^2)$ and $\gamma(h) = 0$ for $|h| > 1$, thus $\rho(1) = a/(1+a^2)$ and $\rho(h) = 0$ for $|h| > 1$. Therefore this process has the same variance for all a but the correlation depends on a.



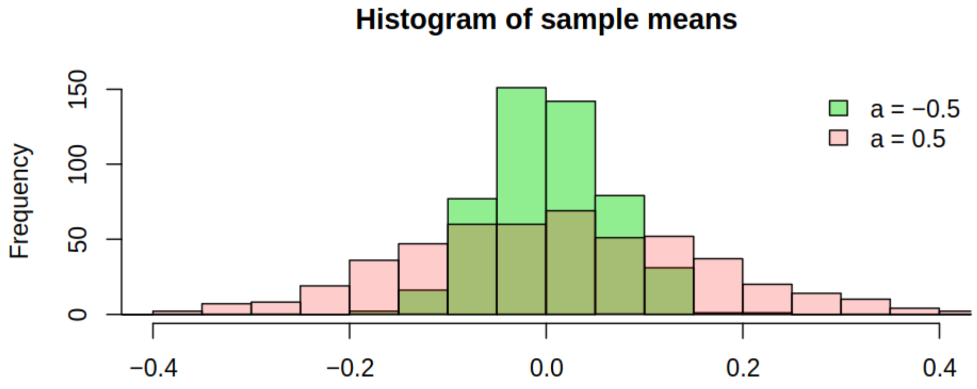
When we now use the (3.3) to compute var $\hat{\mu}$, we get

$$\operatorname{var}\hat{\mu} = \frac{\sigma^2}{n}\left(1 + 2\frac{n-1}{n}\frac{a}{1+a^2}\right) \approx \frac{\sigma^2}{n}\left(1 + 2\frac{a}{1+a^2}\right)$$

for very large n. Thus var $\hat{\mu} \to 0$ as $n \to \infty$ – we can read this as meaning the more data we have, the better. For example

- for $a = -0.5$, var $\hat{\mu} \approx 0.2\sigma^2/n$,
- for $a = 0.5$, var $\hat{\mu} \approx 1.8\sigma^2/n$.

We can make the observation, that *negative correlation improves* the estimation of μ – there will be more oscillations back and forth across the mean. On the other hand, a positive correlation reduces the accuracy of the estimate. Just as an illustration, we can simulate 500 realizations of length $n = 100$ with $\mu = 0$ and compute $\hat{\mu}$ for each of them



Tip

In the slides, there are two more examples:

- Random walk
- IID with additional noise

3.2.2 Consistency of the sample mean

We may recall *the consistency* of an estimator from previous courses.

Definition 3.3 (Consistency of an estimator)**.** If $\hat{\theta}_n$ is an estimator of some parameter θ from observations $X_1, \dots, X_n$, it is called *consistent* when $\hat{\theta}_n \to \theta$ in *some appropriate way*, as $n \to \infty$, for all values of θ.

The possible ways of convergence are:

- **Almost surely** $\hat{\theta}_n \overset{\text{a.s.}}{\to} \theta : \hat{\theta}_n(\omega) \to \theta$ for all $\omega \in A, P(A) = 1$ (*strong consistency*)
- **In mean square** $\hat{\theta}_n \overset{L^2}{\to} \theta : \mathbb{E}\left((\hat{\theta}_n - \theta)^2\right) \to 0$ (for unbiased estimators, it holds if $\operatorname{var}\hat{\theta}_n \to 0$)
- **In probability** $\hat{\theta}_n \overset{P}{\to} \theta : P\left(\left|\hat{\theta}_n - \theta\right| > \varepsilon\right) \to 0$ for all $\varepsilon > 0$ (*weak consistency*; is true if convergence is almost surely or if convergence in mean square holds)

Recall now that when $\{X_t\}$ is stationary, we have computed in (3.3), after some changes, that

$$\operatorname{var}\hat{\mu} = \operatorname{var}\overline{\mathbf{X}}_n = \frac{1}{n}\sum_{|i|<n}\left(1 - \frac{|i|}{n}\right)\gamma(i)$$

Theorem 3.1. *If $\{X_t\}$ is stationary with mean μ and autocovariance function $\gamma(h)$ such that $\sum_{i=-\infty}^{\infty}|\gamma(i)| < \infty$, then as $n \to \infty$*

$$n\operatorname{var}\overline{\mathbf{X}}_n \to \sum_{i=-\infty}^{\infty}\gamma(i).$$

In particular, $\operatorname{var}\overline{\mathbf{X}}_n \to 0$ *and hence* $\overline{\mathbf{X}}_n \overset{L^2}{\to} \mu$.

The Theorem 3.1 follows from the expression

$$n\operatorname{var}\overline{\mathbf{X}}_n = \sum_{|i|<n}\left(1 - \frac{|i|}{n}\right)\gamma(i)$$

by dominated convergence theorem (where the sum is thought of as an abstract integral with respect to an appropriate measure).

3.2.3 Approximate distribution

Consider a situation, where we want to say something about the true μ that the estimator $\hat{\mu} = \overline{\mathbf{X}}_n$ fluctuates around. We can observe the magnitude of this fluctuation – the variance – but we want to know (or approximate) the distribution.

Tip

Histograms suggest approximate normality of $\hat{\mu} = \overline{\mathbf{X}}_n$ – they are symmetric and centered at the true mean. This can be also deducted from the central limit theorem.

Theorem 3.2. *If $\{X_t\}$ is stationary, then under certain conditions $\overline{\mathbf{X}}_n$ is approximately (for large n) normal with mean μ and variance $n^{-1}v = n^{-1}\sum_{i=-\infty}^{\infty}\gamma(i)$, i.e.,*

$$n^{1/2}(\overline{\mathbf{X}}_n - \mu) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, v).$$

Examples of *certain conditions* mentioned in Theorem 3.2 can be found in the slides. One such condition is satisfied if

$$X_t = \mu + \sum_{j=-\infty}^{\infty} \psi_j Z_{t-j},$$

where $\{Z_t\}$ are iid with mean 0 and variance $\sigma^2 \in (0, \infty)$, $\sum_{j=-\infty}^{\infty} \left|\psi_j\right| < \infty$ and $\sum_{j=-\infty}^{\infty} \psi_j \neq 0$, we have $v = \sigma^2 \left(\sum_{j=-\infty}^{\infty} \psi_j\right)^2$.

3.2.4 Confidence intervals for the mean

What's more, we can construct intervals that cover the true mean with approximate probability ($1 - \alpha$, say). Naturally, we may consider symmetric intervals of the form

$$I = (\hat{\mu} - d, \hat{\mu} + d)$$

and we want to determine d such that

$$P(\mu \in I) = 1 - \alpha.$$

If we recall that $\hat{\mu}$ is approximately $\mathcal{N}(\mu, v/n)$, then

$$I = \left(\hat{\mu} - \frac{u_{1-\alpha/2}v^{1/2}}{n^{1/2}}, \hat{\mu} + \frac{u_{1-\alpha/2}v^{1/2}}{n^{1/2}}\right)$$

so practically, v must be estimated. Ignoring autocorrelation, one would use the true (if known) or sample variance in place of v (can lead to too short (for positively correlated data) or too long (for negatively correlated data) intervals).

From the previous consideration, the need for the estimation of limiting (long-run) variance arises

$$v = \sum_{h=-\infty}^{\infty} \gamma(h),$$

but using the obvious estimator

$$\hat{\gamma}(0) \cdot \left(1 + 2 \sum_{i=1}^{n-1} \frac{n-i}{n} \hat{\rho}(i) \right)$$

here does **not** work – the estimation of $\hat{\rho}(i)$ is unstable for high i – they correspond to high-delay-autocorrelations on which we have little data. As such, the *Newey-West* weighted estimator of the form

$$\hat{\gamma}(0) \cdot \left(1 + 2 \sum_{i=1}^{K_n} w_{i,n} \hat{\rho}(i) \right)$$

works with an appropriate truncation point K_n and weights $w_{i,n}$, e.g. $K_n = 4(n/100)^{2/9}$, $w_{i,n} = 1 - \frac{i}{K_n+1}$.

💡 Tip

In R we can use package `sandwich` (because these estimates are called *sandwich estimates*) and function `lrvar`.

4 Estimation of the autocorrelation function

There are multiple approaches when trying to determine the autocorrelation function ρ of given data. The first obvious way is to determine ρ from the used model. While this works, their autocorrelation ρ is often very complicated. It should be also said that direct modeling of the autocorrelation ρ is difficult. The second way is determining the form of acf ρ directly from data without any underlying model assumptions. These estimates can be used for inference on the relationships between the variables and for identifying a good model.

Therefore we are in a situation where we have observations $X_1, \dots, X_n$ of a *stationary* series and let our goal be to estimate

$$\gamma(h) = \operatorname{cov}(X_t, X_{t+h}) = \mathbb{E}\left((X_t - \mu)(X_{t+h} - \mu)\right).$$

Naturally, we can use the empirical counterpart

$$\hat{\gamma}(h) = \frac{\sum_{i=1}^{n-h}(X_i - \overline{\mathbf{X}})(X_{i+h} - \overline{\mathbf{X}})}{n},$$

where alternatively the denominator can be $n - h$ or something similar, but n is typically used in the context of time series (for positive semi-definiteness and simplification of acf ρ). Now we can estimate the autocorrelation function $\rho(h) = \frac{\gamma(h)}{\gamma(0)}$ by

$$\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)} = \frac{\sum_{i=1}^{n-h}(X_i - \overline{\mathbf{X}})(X_{i+h} - \overline{\mathbf{X}})}{\sum_{i=1}^{n}(X_i - \overline{\mathbf{X}})^2}.$$

We can picture this with simulated white noise and its correlogram (plot of the estimated autocorrelation function ρ), see Figure 4.1.

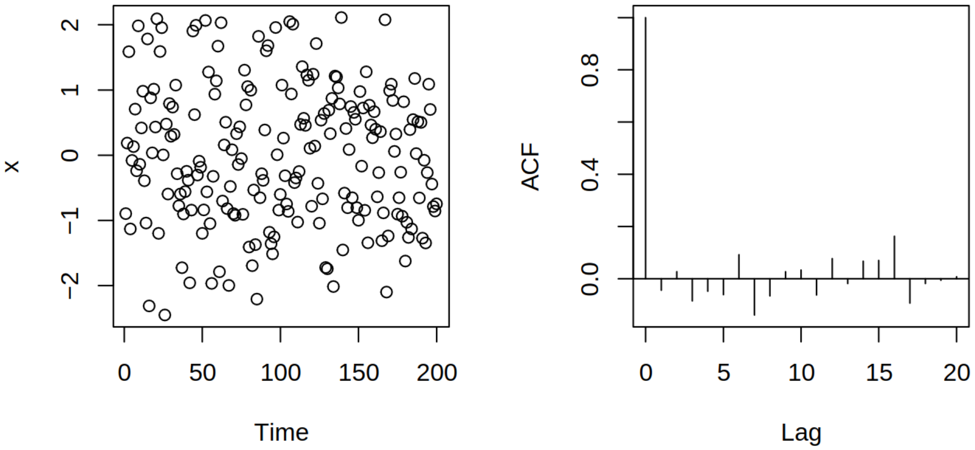


Figure 4.1: Simulated white noise and its correlogram

4.1 Properties of the empirical autocovariance and autocorrelation

Under the non-restrictive assumptions, $\hat{\gamma}(h)$ and $\hat{\rho}(h)$ consistently estimate $\gamma(h), \rho(h)$ respectively. Furthermore, we would like to see whether the autocorrelations (bars in the correlogram) are important or not. Therefore we will study the asymptotic distribution of

$$\hat{\mathbf{c}}(h) = (\hat{\gamma}(0), \hat{\gamma}(1), \ldots, \hat{\gamma}(h))^\top$$

and

$$\hat{\mathbf{r}}(h) = (\hat{\rho}(1), \ldots, \hat{\rho}(h))^\top$$

for $n \to \infty$.

Theorem 4.1. *If $\{X_t\}$ is a stationary process, then, under* certain assumptions, *for any fixed h the vector of estimated autocovariances $\hat{\mathbf{c}}(h)$ is approximately (for $n \to \infty$) normally distributed, i.e.,*

$$n^{1/2}(\hat{\mathbf{c}}(h) - \boldsymbol{c}(h)) \xrightarrow{d} \mathcal{N}_{h+1}(0, \boldsymbol{V}),$$

where $\boldsymbol{c}(h) = (\gamma(0), \ldots, \gamma(h))^\top$.

Note that the matrix $\boldsymbol{V}$ can be given explicitly. In particular, if the process consists of uncorrelated variables ($\gamma(j) = 0, j \neq 0$), then $\boldsymbol{V}$ is diagonal with $V_{jj} = \gamma(0)^2$ for $j > 0$ (hence the components of $\hat{\mathbf{c}}(h)$ are asymptotically independent). Examples of certain conditions are again given in the slides, but one can remember that ARMA processes satisfy them

4.2 Asymptotic qualities of the sample autocorrelation

Most often, we wish to make inferences about the dependence structure regardless of scale and we would like to have the asymptotic distribution for autocorrelations rather than autocovariances. The sample autocovariances are (more or less) sums of variables (some central limit theorem is behind the proof). The sample autocorrelation

$$\hat{\rho}(j) = \frac{\hat{\gamma}(j)}{\hat{\gamma}(0)}$$

is a non-linear function (ratio) of the autocovariances. So the question is: Can we deduce the asymptotic distribution of a function of an estimator if we have it for the estimator itself?

4.3 Limit theorems

Theorem 4.2 (Continuous mapping). *Let g be a continuous function. Then*

- *if* $\mathrm{R}_n \xrightarrow[n\to\infty]{P} \mathrm{R}$, *then* $g(\mathrm{R}_n) \xrightarrow[n\to\infty]{P} g(\mathrm{R})$
- *or if* $\mathrm{R}_n \xrightarrow[n\to\infty]{d} \mathrm{R}$, *then* $g(\mathrm{R}_n) \xrightarrow[n\to\infty]{d} g(\mathrm{R})$,

where the g actually needs to only be continuous on a set X such that $P(\mathrm{R} \in X) = 1$.

Clearly, we can use the continuous mapping Theorem 4.2 when we have a random variable $\hat{\mathrm{v}}_n$, which consistently estimates a variance of interest. Then $\hat{\mathrm{v}}_n^{1/2}$ consistently estimates the corresponding standard deviation.

Theorem 4.3 (Slutsky's). *Let* $\mathrm{R}_n \xrightarrow[n\to\infty]{d} \mathrm{R}$ *and* $\mathrm{S}_n \xrightarrow[n\to\infty]{P} s$, *where s is a non-random constant. Then*

$$\mathrm{R}_n\mathrm{S}_n \xrightarrow[n\to\infty]{d} s\mathrm{R} \quad \& \quad \mathrm{R}_n + \mathrm{S}_n \xrightarrow[n\to\infty]{d} \mathrm{R} + s.$$

Again, a familiar use of Slutsky's Theorem 4.3 might be if we have

$$n^{1/2}(\hat{\theta}_n - \theta) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, v)\,, \qquad \hat{\mathrm{v}}_n \xrightarrow[n\to\infty]{P} v,$$

then

$$\hat{\mathrm{v}}_n^{-1/2} n^{1/2}(\hat{\theta}_n - \theta) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, 1)\,.$$

4.4 Delta method

Let $\hat{\boldsymbol{\theta}}_n \in \mathbb{R}^p$ be a random vector and assume it is asymptotically normal with mean $\boldsymbol{\theta}$ and variance matrix $a_n^{-2}\boldsymbol{V}$, i.e.

$$a_n(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow[n\to\infty]{d} \mathcal{N}_p\left(\mathbf{0}, \boldsymbol{V}\right).$$

Our goal is now to find the asymptotic distribution of $\boldsymbol{g}(\hat{\boldsymbol{\theta}}_n)$, where $\boldsymbol{g} : \mathbb{R}^p \to \mathbb{R}^q$. If $\boldsymbol{g}$ is continuously differentiable at $\boldsymbol{\theta}$ with $\nabla \boldsymbol{g}(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{g}(\boldsymbol{\theta}) \in \mathbb{R}^{p\times q}$, then by Taylor series, Theorem 4.2 and Theorem 4.3, we get

$$a_n\left(\boldsymbol{g}(\hat{\boldsymbol{\theta}}_n) - \boldsymbol{g}(\boldsymbol{\theta})\right) = \nabla \boldsymbol{g}(\boldsymbol{\theta}^*) a_n(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow[n\to\infty]{d} \nabla \boldsymbol{g}(\boldsymbol{\theta}) \mathcal{N}_p(\mathbf{0}, \boldsymbol{V}) = \mathcal{N}_q\left(\mathbf{0}, \nabla \boldsymbol{g}(\boldsymbol{\theta}) \cdot \boldsymbol{V} \cdot \nabla \boldsymbol{g}(\boldsymbol{\theta})^\top\right).$$

Here the "*delta*" stands for differentiation.

Tip

The delta method is a very useful general tool (bear in mind it is not limited to this particular context).

4.5 Asymptotic qualities of the sample autocorrelation – cont.

Consider $\boldsymbol{g}(x_0, \ldots, x_h) = (x_1/x_0, \ldots, x_h/x_0)^\top$, then

$$\hat{\boldsymbol{\rho}}(h) = \boldsymbol{g}(\hat{\mathbf{c}}(h)), \quad \boldsymbol{\rho}(h) = \boldsymbol{g}(\boldsymbol{c}(h)),$$

and also $\boldsymbol{g}$ has the $h \times (h+1)$ Jacobi matrix of partial derivatives in form

$$\nabla \boldsymbol{g}(x_0, \ldots, x_h) = \frac{1}{x_0^2}\begin{pmatrix} -x_1 & x_0 & 0 & \cdots & 0 \\ -x_2 & 0 & x_0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -x_h & 0 & 0 & \cdots & x_0 \end{pmatrix}.$$

At $\boldsymbol{x} = \boldsymbol{c}(h)$, this is $\nabla \boldsymbol{g}(\boldsymbol{c}(h)) = \frac{1}{\gamma(0)}\begin{pmatrix} -\boldsymbol{r}(h) & \boldsymbol{I}_h \end{pmatrix}$ with $\boldsymbol{I}_h$ denoting $h \times h$ identity matrix. Then $\hat{\mathbf{r}}(h)$ is approximately normal with mean $\boldsymbol{r}(h)$ and covariance matrix $n^{-1}\nabla \boldsymbol{g}(\boldsymbol{c}(h))\boldsymbol{V}\nabla \boldsymbol{g}(\boldsymbol{c}(h))^\top$.

Theorem 4.4. *If $\{X_t\}$ is a stationary process, then, under certain assumptions, for any fixed h the vector of estimated autocorrelations $\hat{\mathbf{r}}(h)$ is approximately (for $n \to \infty$) normally distributed, i.e.*

$$n^{1/2}(\hat{\mathbf{r}}(h) - \boldsymbol{r}(h)) \xrightarrow{d} \mathcal{N}_h(\mathbf{0}, \boldsymbol{W}),$$

where $\mathbf{W}$ *is the* $h \times h$ *matrix with entries given by the* ***Barlett formula***

$$W_{ij} = \sum_{k=1}^{\infty} \left(\rho(k+i) + \rho(k-i) + 2\rho(i)\rho(k)\right) \cdot \left(\rho(k+j) + \rho(k-j) + 2\rho(j)\rho(k)\right).$$

For white noise (that is, $\rho(t) = 0, t \neq 0$) we get $W_{ij} = 1$ for $i = j$, $W_{ij} = 0$ otherwise.

4.5.1 Sample autocorrelation of white noise

Theorem 4.5. *If* $\{X_t\}$ *is a sequence of iid random variables, then, under certain assumptions, for any h the vector of estimated autocorrelations* $\hat{\mathbf{r}}(h)$ *is approximately (for* $n \to \infty$*) normally distributed with mean zero, variances* $1/n$ *and independent components, i.e.*

$$n^{1/2}\hat{\mathbf{r}}(h) \xrightarrow[n\to\infty]{d} \mathcal{N}_h(0, \boldsymbol{I}_h).$$

It should be clear that Theorem 4.5 follows from Theorem 4.4. This allows us to test certain hypotheses because $\hat{\rho}(j) \dot{\sim} \mathcal{N}(0, 1/n)$ under independence. Hence $\hat{\rho}(j)$ should be between $-1.96/\sqrt{n}$ and $1.96/\sqrt{n}$ with approximate probability 95 % under independence

4.6 Inference about the autocorrelation structure

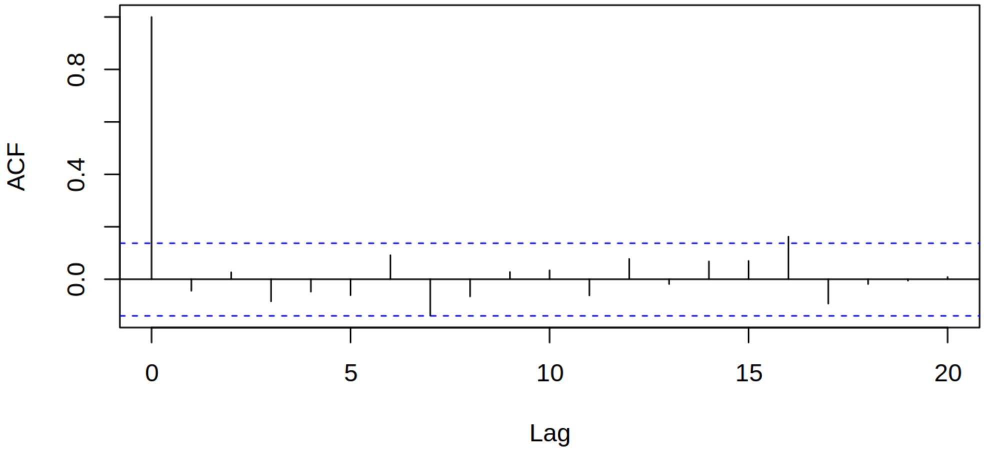


Figure 4.2: Example correlogram

Here in Figure 4.2, the limits are at $\pm 1.96/\sqrt{n}$. At lag j, they indicate the rejection regions for testing the null hypothesis of no autocorrelation against the alternative that $\rho(j) \neq 0$. When the estimated acf is between the blue dotted lines, i.e. in the confined region, we do **not** reject the null hypothesis. Due to the approximate independence, one can expect $\alpha \times 100\%$ false rejections on average (e.g., 1 out of 20).

Caution

Although R calls it confidence intervals, it is more correctly a region of rejection/validity of sorts.

4.6.1 Asymptotic distribution of estimated acf of a moving average process

Consider white noise $\{\varepsilon_t\}$ and define

$$X_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q},$$

for which we have already shown that $\rho(j) = 0$ for $j > q$. Thus we can test

$$H_0 : \rho(j) = 0,\ j \geq q+1 \quad \text{vs} \quad H_1 : \rho(q+1) \neq 0,\ j \geq q+2$$

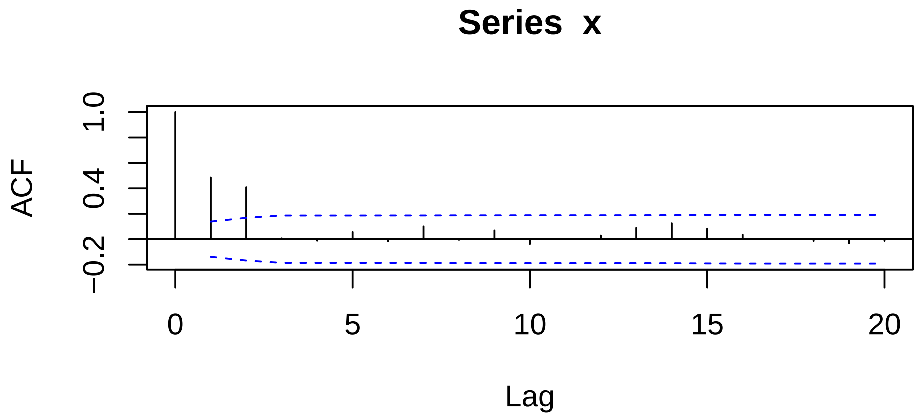
and such we will reject $\hat{\rho}(q+1)$ outside the limits $\pm 1.96 n^{-1/2}(1+2\sum_{k=1}^{q} \hat{\rho}(k)^2)^{1/2}$. Note that there are wider limits because we put no restrictions on $\rho(j), j \leq q$.

Example 4.1. Given a specific example

$$X_t = \varepsilon_t + 0.6\varepsilon_{t-1} + 0.9\varepsilon_{t-2},$$

we get, with the following code, these results:

```r
set.seed(1)
x = filter(rnorm(202),
    sides=1,
    filter=c(1,.6,.9),
    method="convolution"
)[-(1:2)]
acf(x,ci.type="ma",lag.max=20)
```



4.6.2 The Ljung-Box test

Let us assume our goal is to test the global hypothesis of no autocorrelation. Begin with a null hypothesis $H_0 : \rho(h) = 0$ for all $h = 1, 2, \dots, L$ for $L \in \mathbb{N}$ and alternative $H_1 : \exists h \in \{1, \dots, L\}$, such that $\rho(h) \neq 0$. We use the *Box-Pierce* test statistic

$$\mathrm{Q} = n(\hat{\rho}(1)^2 + \cdots + \hat{\rho}(L)^2).$$

Now under H_0, Q is approximately χ^2_L distributed – as such we will reject H_0 for $\mathrm{Q} > q_L(1-\alpha)$ where $q_L(1-\alpha)$ is $(1-\alpha)$-quantile of χ^2_L. In truth, the Ljung-Box test statistic reads (more precise for smaller values)

$$\mathrm{Q}_* = n(n+2)\left(\frac{\hat{\rho}(1)^2}{n-1} + \cdots + \frac{\hat{\rho}(L)^2}{n-L}\right),$$

which has the same approximate distribution (Q_* is much closer to chi-square than Q)

Example 4.2. Consider the example (as an illustration) where the maximum delay of 10 was used:

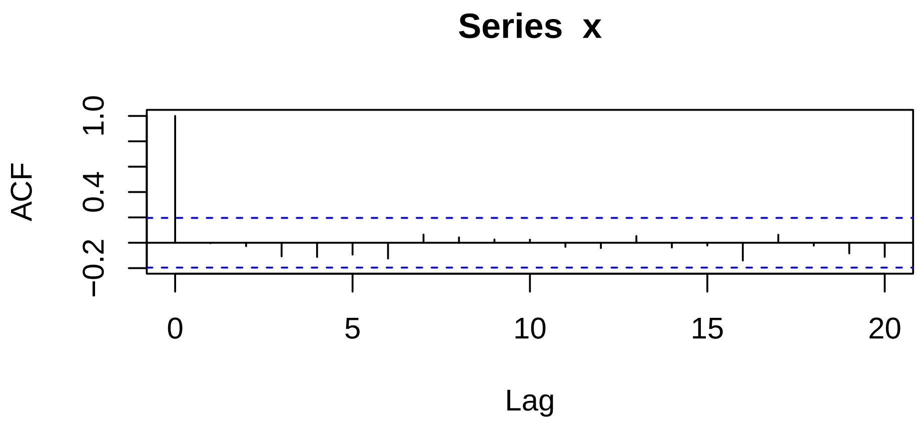
```
set.seed(1)
x = rnorm(100)
Box.test(x,lag=10,type="Ljung-Box")
```

```

    Box-Ljung test

data:  x
X-squared = 6.0721, df = 10, p-value = 0.8092
```

```
acf(x)
```



Tip

We can't choose too big a lag, because we would have too many parameters. As a heuristics, lag should be approximately $< \frac{n}{10}$

4.6.3 Maximum correlation test

We shall now aim to test whether any ρ at lags up to L is significant. We might have an idea to use single tests and combine them in

$$\mathrm{T} = \max\left(|\hat{\rho}(1)|, \dots, |\hat{\rho}(L)|\right).$$

We can then look for c such that $P(\mathrm{T} > c) \doteq \alpha$ under H_0 and as such, $\hat{\rho}(1), \dots, \hat{\rho}(L)$ are approximately *iid* under H_0. Thus, also under H_0,

$$\begin{aligned} 1 - \alpha \doteq P(\mathrm{T} \leq c) &= P\left(\max\left(|\hat{\rho}(1)|, \dots, |\hat{\rho}(L)|\right) \leq c\right) \\ &= \prod_{j=1}^{L} P\left(|\hat{\rho}(j)| \leq c\right) \\ &= \prod_{j=1}^{L} \left(\Phi(n^{1/2}c) - \Phi(-n^{1/2}c)\right) \\ &= \left(2\Phi(n^{1/2}c) - 1\right)^{L}. \end{aligned}$$

Hence the critical value is $c = n^{-1/2}\Phi^{-1}\left((1-\alpha)^{1/L}\right)$. Regions for testing multiple autocorrelations at once have boundaries at $\pm c$ (whereas usual rejection regions in the correlogram are for tests about single autocorrelations).

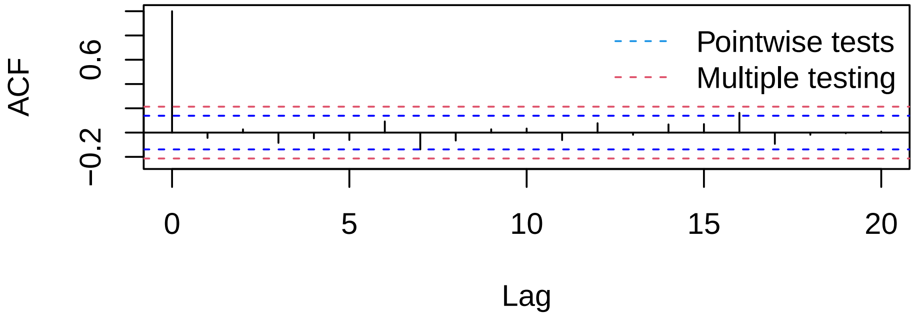
```r
set.seed(2); n = 200; x = rnorm(n); L = 20
# standard rejection limit for single tests
qnorm(.975)/sqrt(n)
```

```
[1] 0.1385904
```

```r
# rejection regions corrected for multiple testing
(max.lim = qnorm(.975^(1/L))/sqrt(n))
```

```
[1] 0.2135257
```

```
acf(x,
    lag.max=L,
    main="",
    ylim=c(-.25,1)
)
abline(h=c(-1,1)*max.lim,
    col=2,
    lty=2
)
legend("topright",
    legend=c("Pointwise tests","Multiple testing"),
    col=c(4,2),
    lty=2,
    bty="n"
)
```



4.6.4 Problems with acf estimation

Surely, the ρ estimation and the correlogram make sense for stationary data. Now consider iid white noise $\{\varepsilon_t\}$ and define $X_t = t + \varepsilon_t$. Then X_t are *independent* variables, hence uncorrelated. If we directly apply the correlogram, we get misleading results, see Figure 4.3.

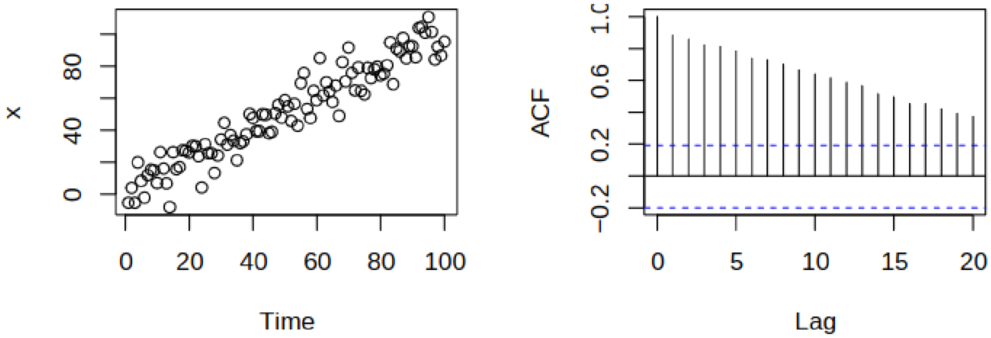


Figure 4.3: Misleading correlogram

As this example shows, trends/deterministic components should be accounted for: e.g., model them by regression. Now consider the random walk $X_t = \sum_{j=1}^{t} \varepsilon_j$ (this time series is non-stationary). Directly applying the correlogram to X_t indicates a complicated autocorrelation structure, see Figure 4.4.

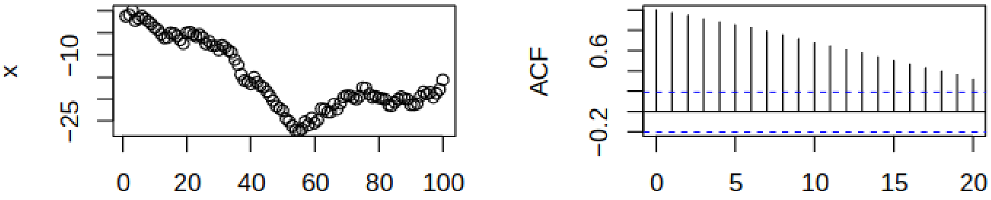


Figure 4.4: Complicated correlation structure for random walk

On the other hand, applying the correlogram to the differentiated series $\{X_t - X_{t-1}\}$ suggests a much simpler structure, see Figure 4.5.

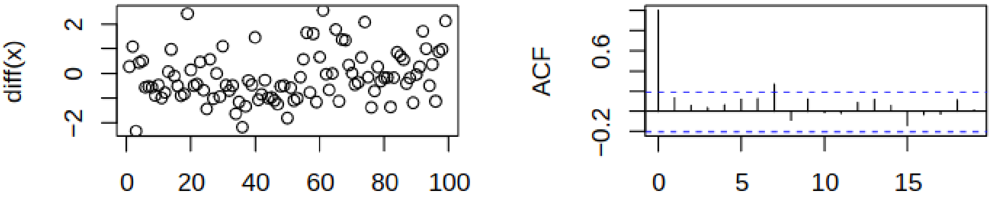


Figure 4.5: Simple correlation structure from differences

Therefore transient trends should be accounted for, e.g., by differencing.

5 Regression Methods for Deterministic Components

5.1 Linear regression models

Observable processes follow some deterministic patterns but randomly deviate from them. From these assumptions, we may write our time series as a model plus noise decomposition

$$X_t = \text{Deterministic}_t + \text{Stochastic}_t.$$

Furthermore, the deterministic part can be modeled as follows

$$\text{Deterministic}_t = \text{Trend}_t + \text{Seasonality}_t \ (+ \text{OtherVariablesEffects}_t)$$

and the stochastic component as (we will see models of this kind, e.g. ARMA, later)

$$\text{Stochastic}_t = \text{PredictableVariation}_t + \text{WhiteNoise}_t.$$

Note

The above decompositions are schematic, and not always applicable. (non-stationarity, non-linearity, changing variability etc.)

We shall now look at *parametric* modeling of the deterministic part.

5.1.1 Examples

Example 5.1 (Annual global temperature series)**.** With the following code, we can plot Figure 5.1, where a possible monotonic trend (linear, maybe quadratic) can be seen.

```
data(gtemp,package="astsa")
plot(gtemp,type="o",ylab="Global temperature")
```

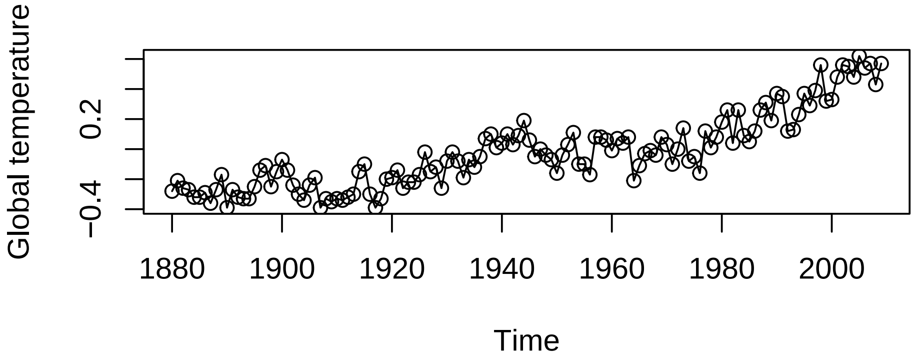


Figure 5.1: Annual global temperature

Example 5.2 (Monthly carbon dioxide levels)**.** Now, consider the following time series, in which one can see a clear monotonic trend (linear) and seasonality.

```
data(co2,package="TSA")
plot(co2,type="o",ylab="CO2")
```

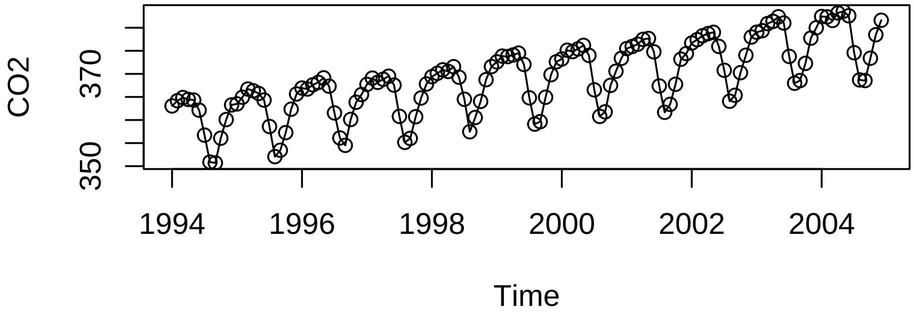
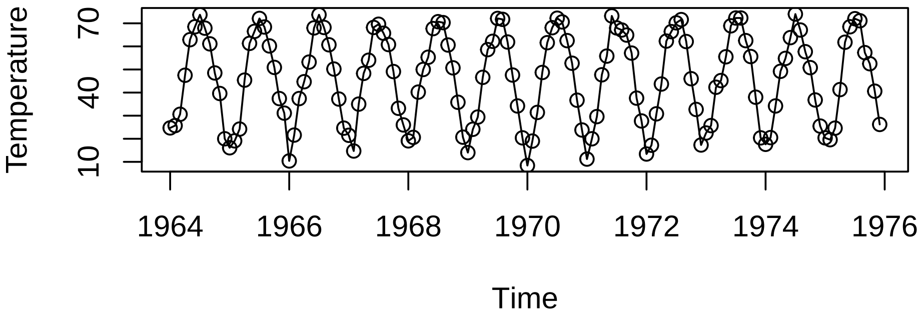


Figure 5.2: CO_2 levels

Example 5.3 (Monthly temperatures)**.** Consider the following time series – this time, no obvious trend is visible, but seasonality plays a big role.

```
data(tempdub,package="TSA")
plot(tempdub,type="o",ylab="Temperature")
```



5.1.2 Linear models for trends in time

Consider a model

$$X_t = \mu_t + U_t,$$

where $\mu_t = \mathbb{E}\,X_t$ is the mean, U_t are residuals $\mathbb{E}\,U_t = 0$. Models for the trend should be simple, model only obvious trends (such as monotonic) and seasonality. Unclear/varying/transient features (stochastic trends) are often handled by differencing, for example

- *linear trend:* $\mu_t = \beta_0 + \beta_1 t$;
- *quadratic trend:* $\mu_t = \beta_0 + \beta_1 t + \beta_2 t^2$.

Tip

We usually do not use higher-order polynomials (a complicated graph of the whole process may likely be due to stochastic trends).

Example 5.4. As introduced earlier, we can use the following code to model with deterministic components the Example 5.1 data:

```
gtemp.m1 = lm(gtemp~time(gtemp))
gtemp.m2 = lm(gtemp~time(gtemp)+I(time(gtemp)^2))
par(mfrow=c(1,2))
plot(gtemp,ylab="Global temperature")
```

```
abline(gtemp.m1)
plot(gtemp,ylab="Global temperature")
lines(as.vector(time(gtemp)),fitted(gtemp.m2))
```

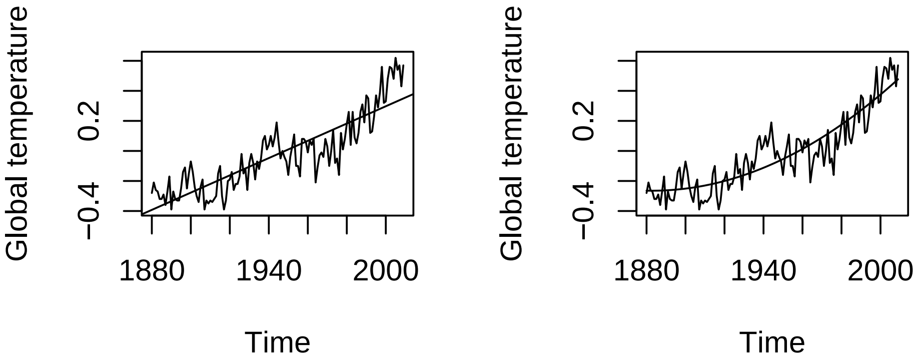


Figure 5.3: Models with deterministic parts for global temperatures

Now, we shall look at the summary statistics of the used models.

```
summary(gtemp.m2)
```

```
Call:
lm(formula = gtemp ~ time(gtemp) + I(time(gtemp)^2))

Residuals:
      Min         1Q    Median        3Q       Max 
-0.300105 -0.080650  0.004134  0.074619  0.280003 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)       1.647e+02  2.916e+01   5.647 1.02e-07 ***
time(gtemp)      -1.752e-01  3.000e-02  -5.840 4.12e-08 ***
I(time(gtemp)^2)  4.653e-05  7.714e-06   6.032 1.65e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1108 on 127 degrees of freedom
Multiple R-squared:  0.8065,    Adjusted R-squared:  0.8035 
F-statistic: 264.7 on 2 and 127 DF,  p-value: < 2.2e-16
```

But can we trust these standard errors, confidence intervals and p-values? In the theory of linear regression models, we mostly assumed independence. We can look at the residuals

```
par(mfrow=c(1,2))
plot(as.vector(time(gtemp)),
    resid(gtemp.m2),
    type="l",
    xlab="Time",
    ylab=""
)
acf(resid(gtemp.m2),
    main=""
)
```

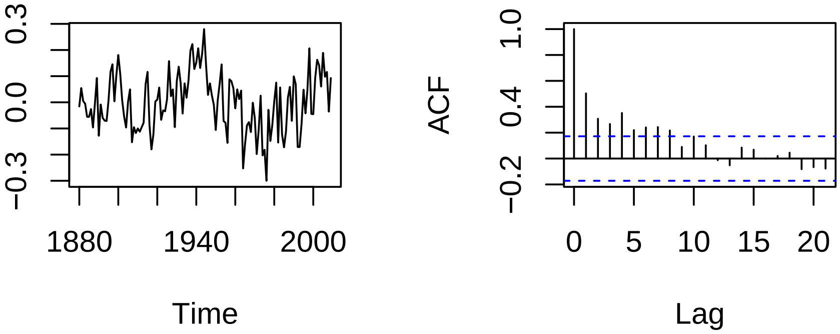


Figure 5.4: Residuals and it's correlogram

As can be seen in Figure 5.4, the residuals are strongly autocorrelated and the standard errors, p-values are probably incorrect (they are computed under the assumption of iid errors).

5.1.3 Linear models for trends and seasonality

Again, consider a model $X_t = \mu_t + U_t$, where $\mu_t = \mathbb{E}\,X_t$ is the mean, U_t are residuals $\mathbb{E}\,U_t = 0$. Now we can include the seasonal effect. Consider s seasons (e.g. $s = 12$), each having its own level. Then the model with linear trend and seasonal indicators is of form

$$\begin{aligned}\mu_t &= \beta_0 + \beta_1 t + \sum_{j=2}^{s} \alpha_j \mathbb{1}_{t=ks+j} \\ &= \begin{cases} \beta_0 + \beta_1 t, & t = 1, s+1, 2s+1, \ldots \\ \beta_0 + \beta_1 t + \alpha_2, & t = 2, s+2, 2s+2, \ldots \\ \vdots & \\ \beta_0 + \beta_1 t + \alpha_s, & t = s, 2s, 3s, \ldots, \end{cases}\end{aligned}$$

where season 1 is the reference level and $\alpha_2, \ldots, \alpha_s$ are the differences against it.

Using the following code, we get the model seen in Figure 5.5.

```
mth = factor(cycle(co2))
co2.m1 = lm(co2~time(co2)+mth)
plot(co2,type="p")
lines(as.vector(time(co2)),fitted(co2.m1))
```

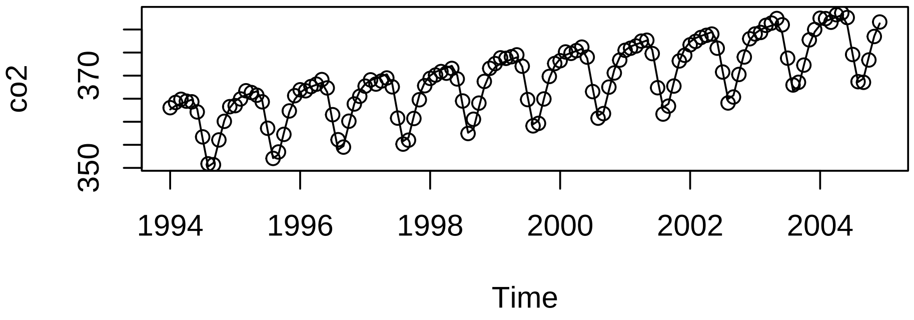


Figure 5.5: Time series CO_2 model

Or we can consider other equivalent models.

```
lm(co2~time(co2)+mth)
```

```
Call:
lm(formula = co2 ~ time(co2) + mth)
```

```
Coefficients:
(Intercept)    time(co2)         mth2         mth3         mth4         mth5
 -3290.5412       1.8321       0.6682       0.9637       1.2311       1.5275
       mth6         mth7         mth8         mth9        mth10        mth11
    -0.6761      -7.2851     -13.4415     -12.8205      -8.2604      -3.9277
      mth12
    -1.3367
```

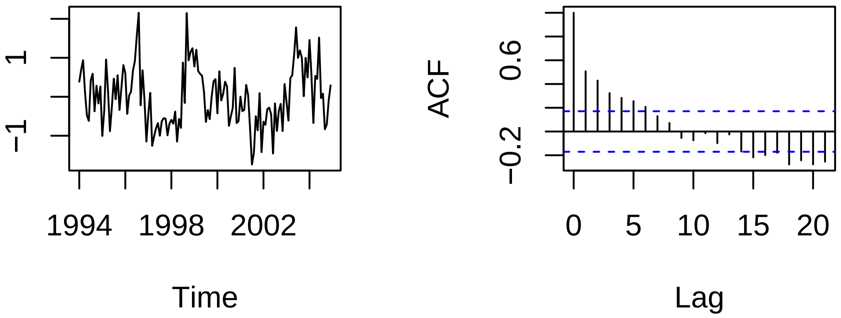
```
lm(co2~time(co2)-1+mth)
```

```
Call:
lm(formula = co2 ~ time(co2) - 1 + mth)

Coefficients:
time(co2)       mth1       mth2       mth3       mth4       mth5       mth6
    1.832  -3290.541  -3289.873  -3289.577  -3289.310  -3289.014  -3291.217
     mth7       mth8       mth9      mth10      mth11      mth12
-3297.826  -3303.983  -3303.362  -3298.802  -3294.469  -3291.878
```

Also, we can look at the residuals and last, but not least, we can plot fitted values of our CO_2 model.

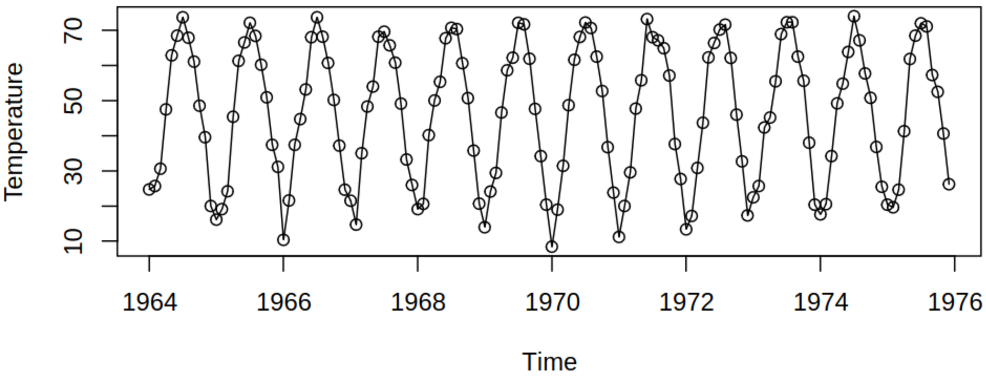
```
par(mfrow=c(1,2))
plot(as.vector(time(co2)),
    resid(co2.m1),
    type="l",
    xlab="Time",
    ylab=""
)
acf(resid(co2.m1),
    main=""
)
```



Notice that there are parallel monthly lines with different distances and parallel (almost equidistant) profiles within years

5.1.4 Seasonality via harmonic components

Consider the following time series from Example 5.3.



Here, waves look like cosines (unlike in the CO_2 series), so our idea might be to model the deterministic component with a cosine wave with known frequency (e.g., $f = \frac{1}{12}$) and unknown amplitude and phase

$$\mu_t = \beta_0 + \beta_1 t + a \cos\left(2\pi f t + \varphi\right).$$

Now, we have to estimate intercept β_0, slope β_1, amplitude a, phase φ, but the model is **non-linear** in φ, which would be difficult to estimate. From trigonometry we get

$$a \cos\left(2\pi f t + \varphi\right) = \alpha_1 \cos(2\pi f t) + \alpha_2 \sin(2\pi f t),$$

where $\alpha_1 = a\cos(\varphi), \alpha_2 = -\sin(\varphi)$. This transforms our model to

$$\mu_t = \beta_0 + \beta_1 t + \alpha_1 \cos(2\pi f t) + \alpha_2 \sin(2\pi f t)$$

and this model is **linear** in parameters (and we have 4 parameters to estimate: $\beta_1, \beta_2, \alpha_1, \alpha_2$). It is good to use known, logical frequencies for deterministic modeling (e.g. $f = 1/12$). Compared with seasonal indicators we have less flexibility, more restrictive, but also fewer parameters.

Related topics

More components, continuum of components, spectral analysis, relative importance of components, periodogram, FFT, random cosine wave stationary model, ...

Example 5.5. The derived model yields Figure 5.6

```
tempdub.harm = lm(tempdub~time(tempdub)
    +cos(2*pi*time(tempdub))
    +sin(2*pi*time(tempdub))
)
plot(tempdub,type="p")
lines(as.vector(time(tempdub)),
    fitted(tempdub.harm)
)
```

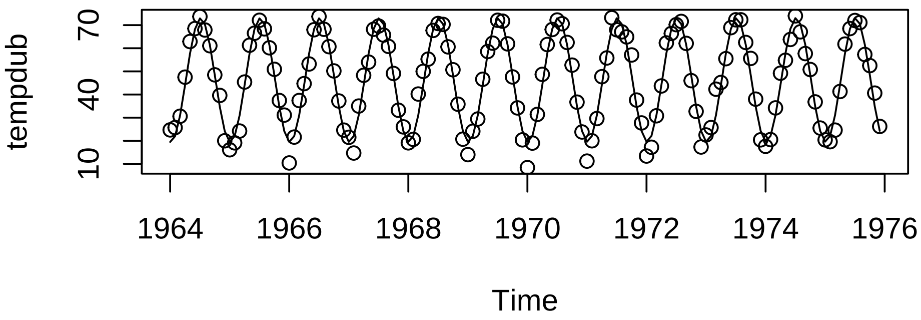


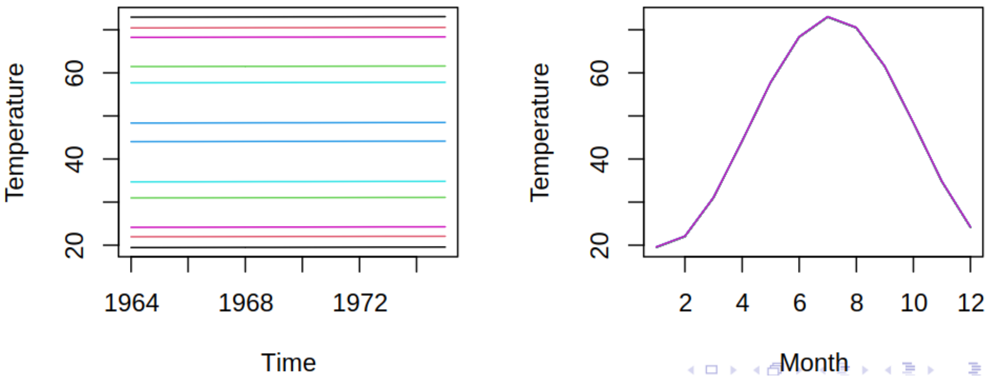
Figure 5.6: Harmonic model

We can now look at the fitted values in the harmonic model for monthly temperatures.

```
tempdub.harm
```

```
Call:
lm(formula = tempdub ~ time(tempdub) + cos(2 * pi * time(tempdub)) +
    sin(2 * pi * time(tempdub)))

Coefficients:
                (Intercept)                time(tempdub)
                   23.85687                      0.01138
cos(2 * pi * time(tempdub))  sin(2 * pi * time(tempdub))
                  -26.70699                     -2.16621
```



5.2 Regression estimators and their properties

Consider a model $X_t = \mu_t + U_t$ with

$$\mu_t = \sum_{j=1}^{p} \beta_j Z_{tj}, \quad t = 1, 2, \ldots$$

Presume we have observations at $t = 1, \ldots, n$. We write $\mu = \boldsymbol{Z\beta}$ in matrix notation ($\boldsymbol{Z}$ is $n \times p$, assume full rank p). Typically, columns of $\boldsymbol{Z}$ are $\mathbf{1}$ (intercept), t (linear trend), possibly t^2, seasonal indicators or sines and cosines, and maybe other explanatory variables (**assume non-random**). Now our goal is to estimate $\boldsymbol{\beta}$ by (ordinary) least squares (OLS or LS): minimize

$$\|\mathbf{X} - \boldsymbol{Z\beta}\|_2^2 = \sum_{t=1}^{n} (X_t - \boldsymbol{Z}_t^\top \boldsymbol{\beta})^2.$$

Although the textbook solution is

$$\hat{\boldsymbol{\beta}} = (\mathbf{Z}^\top \mathbf{Z})^{-1}\mathbf{Z}^\top \mathbf{X},$$

we typically use QR-decomposition, i.e. we factorize $\boldsymbol{Z} = \boldsymbol{Q}\boldsymbol{R}$, where $\boldsymbol{Q}$ $(n \times p)$ has orthonormal columns and $\boldsymbol{R}$ is $(p \times p)$ upper triangular, so the objective becomes

$$\|\mathbf{X} - \boldsymbol{Z}\boldsymbol{\beta}\|_2^2 = \|\mathbf{X} - \boldsymbol{Q}\boldsymbol{R}\boldsymbol{\beta}\|_2^2 = \|\mathbf{X} - \boldsymbol{Q}\boldsymbol{\gamma}\|_2^2\,,$$

where $\boldsymbol{\gamma} = \boldsymbol{R}\boldsymbol{\beta}$. Since the transformation $\boldsymbol{\gamma} = \boldsymbol{R}\boldsymbol{\beta}$ is one-to-one ($\boldsymbol{R}$ has full rank), we can first minimize $\|\mathbf{X} - \boldsymbol{Q}\boldsymbol{\gamma}\|_2^2$ over $\boldsymbol{\gamma} \in \mathbb{R}^p$ to get $\hat{\boldsymbol{\gamma}}$ and the solve $\boldsymbol{R}\boldsymbol{\beta} = \hat{\boldsymbol{\gamma}}$ to get $\hat{\boldsymbol{\beta}}$.

Minimizing $\|\mathbf{X} - \boldsymbol{Q}\boldsymbol{\gamma}\|_2^2$ is easy because $\boldsymbol{Q}$ has orthonormal columns, thus $\hat{\gamma}_j = \left\langle \mathbf{X}, \boldsymbol{Q}_{\cdot,j} \right\rangle$. Also, solving $\boldsymbol{R}\boldsymbol{\beta} = \hat{\boldsymbol{\gamma}}$ is cheap because $\boldsymbol{R}$ is triangular (back-substitution can be used).

5.2.1 Properties of the OLS estimator

Surely $\hat{\boldsymbol{\beta}}$ is unbiased (in spite of ignoring dependence)

$$\mathbb{E}\,\hat{\boldsymbol{\beta}} = (\mathbf{Z}^\top \mathbf{Z})^{-1}\mathbf{Z}^\top(\mathbb{E}\,\mathbf{X}) = \boldsymbol{\beta}.$$

Moreover, if $\mathbf{U}$ is stationary with $\operatorname{cov}\mathbf{U} = \boldsymbol{\Gamma} = \sigma^2\mathbf{R}$, the variance is

$$\begin{aligned} \operatorname{var}\hat{\boldsymbol{\beta}} &= (\mathbf{Z}^\top \mathbf{Z})^{-1}\mathbf{Z}^\top \boldsymbol{\Gamma}\mathbf{Z}(\mathbf{Z}^\top \mathbf{Z})^{-1} \\ &= \sigma^2(\mathbf{Z}^\top \mathbf{Z})^{-1}\mathbf{Z}^\top \mathbf{R}\mathbf{Z}(\mathbf{Z}^\top \mathbf{Z})^{-1}. \end{aligned} \tag{5.1}$$

ℹ Note

Recall the form $\operatorname{var}\hat{\boldsymbol{\beta}} = \sigma^2(\mathbf{Z}^\top \mathbf{Z})^{-1}$ for $\mathbf{R} = \boldsymbol{I}$ for the iid case in (5.1).

Thus the usual standard errors, confidence intervals, and p-values will be incorrect under autocorrelation. Often, the inference about the deterministic components is not the goal, we just want to remove the trend and focus on the stochastic part, then move on to prediction.

One possible remedy could be the sandwich estimators

$$\operatorname{var}\hat{\boldsymbol{\beta}} = (\mathbf{Z}^\top \mathbf{Z})^{-1}\mathbf{Z}^\top \hat{\boldsymbol{\Gamma}}\mathbf{Z}(\mathbf{Z}^\top \mathbf{Z})^{-1}$$

where $\hat{\boldsymbol{\Gamma}}$ is an estimator such that $\boldsymbol{Z}^\top \hat{\boldsymbol{\Gamma}}\boldsymbol{Z}/n$ consistently estimates the limit of $\boldsymbol{Z}^\top \boldsymbol{\Gamma}\boldsymbol{Z}/n$ (similar to long-run variance in mean estimation), see `sandwich::vcovHAC`.

5.2.2 Generalized least squares

Surely, we can use the (assumed) correct covariance matrix derived earlier for estimation. Then we get **generalized least squares** (GLS), where we solve

$$\min_{\boldsymbol{\beta}} (\mathbf{X} - \boldsymbol{Z\beta})^\top \boldsymbol{\Gamma}^{-1} (\mathbf{X} - \boldsymbol{Z\beta}),$$

which is motivated by

- *Gaussian likelihood:* maximize the likelihood for $\mathbf{X} \sim \mathcal{N}_n(\boldsymbol{Z\beta}, \boldsymbol{\Gamma})$
$$(2\pi)^{-k/2}(\det \boldsymbol{\Gamma})^{-1/2} \exp\left(-(\mathbf{X} - \boldsymbol{Z\beta})^\top \boldsymbol{\Gamma}^{-1} (\mathbf{X} - \boldsymbol{Z\beta})/2\right);$$
- *Standardization and OLS:* surely $\boldsymbol{\Gamma}^{-1/2}\mathbf{X}$ has mean $\boldsymbol{\Gamma}^{-1/2}\boldsymbol{Z\beta}$ and variance $\boldsymbol{\Gamma}^{-1/2}\boldsymbol{\Gamma}\boldsymbol{\Gamma}^{-1/2} = \boldsymbol{I}$, hence $\boldsymbol{\Gamma}^{-1/2}\mathbf{X}$ are uncorrelated. As such we can solve the OLS minimization.
$$\begin{aligned}\left(\boldsymbol{\Gamma}^{-1/2}\mathbf{X} - \boldsymbol{\Gamma}^{-1/2}\boldsymbol{Z\beta}\right)^\top & \left(\boldsymbol{\Gamma}^{-1/2}\mathbf{X} - \boldsymbol{\Gamma}^{-1/2}\boldsymbol{Z\beta}\right) \\ &= (\mathbf{X} - \boldsymbol{Z\beta})^\top \boldsymbol{\Gamma}^{-1/2} (\mathbf{X} - \boldsymbol{Z\beta}).\end{aligned}$$

All in all, the textbook solution is now

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{Z}^\top \boldsymbol{\Gamma}^{-1} \boldsymbol{Z}\right)^{-1} \boldsymbol{Z}^\top \boldsymbol{\Gamma}^{-1} \mathbf{X}.$$

Note that the difficulty with GLS is the need to know $\boldsymbol{\Gamma}$, but we can specify a model for $\boldsymbol{\Gamma}$ (or $\mathbf{R}$) up to some parameters and estimate both $\boldsymbol{\beta}$ and the parameters of $\boldsymbol{\Gamma}$ by maximum likelihood. As an example, for moving average errors, one can use a banded matrix.

How do we know that the correlation model was correctly specified? The residuals should look like white noise (in particular, no correlation). The function `gls` in the `R` package `nlme` can do this estimation but it is primarily designed for something else (grouped data), we will see tools that are more appropriate for time series

5.2.3 Transformations

Some of the common problems – think *heteroskedasticity* (e.g. variance increases with mean), *non-linearity*, *non-stationarity* or *non-normality* (skewness) – can be addressed using **transformations** to make residuals look stationary and normal. After transforming our data, we then fit the regression models to the transformed series. Most commonly, a log or square root transformations are used, e.g.

$$\mathrm{X}_t = m_t \mathrm{U}_t \mapsto \log \mathrm{X}_t = \log(m_t) + \log(\mathrm{U}_t)$$

where a non-stationary, heteroskedastic series is transformed to a trend part plus stationary part (if U_t is stationary). In general, the log transformation isn't the only possible one in this case (that gives us linearity and stationarity). Consider a transformation g such that

$$\mathrm{X}_t \mapsto g(\mathrm{X}_t) = m_t^* + \mathrm{U}_t^*,$$

where g may be, for example, the **Box-Cox** transformation, see Definition 5.1.

Definition 5.1 (Box-Cox Transformation)**.** Let $\mathrm{X}_t = m_t \mathrm{U}_t$ be a time series and g the **Box-Cox** transformation given by

$$g(x) = \frac{x^\lambda - 1}{\lambda}$$

for $\lambda > 0$ and $g(x) = \log x$ for $\lambda = 0$.

Tip

Here $\lambda = 0$ corresponds to a multiplicative transformation and $\lambda = 1$ to an additive one.

But now a question arises how and when do Box-Cox transformations work? Suppose that $\mathrm{X}_t > 0$ and that

$$\mathbb{E}\,\mathrm{X}_t = \mu_t \quad \& \quad \operatorname{var} \mathrm{X}_t = \sigma^2 v(\mu_t).$$

Now consider the Taylor expansion

$$g(\mathrm{X}_t) \approx g(\mu_t) + g'(\mu_t)(\mathrm{X}_t - \mu_t)$$

and compute expectations and variances on both sides to get

$$\mathbb{E}\, g(\mathrm{X}_t) \approx g(\mu_t), \quad \operatorname{var} g(\mathrm{X}_t) \approx g'(\mu_t)^2 \operatorname{var} \mathrm{X}_t = g'(\mu_t)^2 \sigma^2 v(\mu_t).$$

To further stabilize the variance, i.e. make it independent of the mean, use g such that $g'(x) = v^{-\frac{1}{2}}(x)$. As can be shown, the Box-Cox transformations do this for $v(x) = cx^{2(1-\lambda)}$ – in other words when the standard deviation of X_t is proportional to $\mu_t^{1-\lambda}$.

Note

Using this we get a log transformation for the variance *quadratic* in the mean and *square root* for the variance linear in the mean.

5.3 Nonparametric regression techniques

Consider a time series describing global temperature

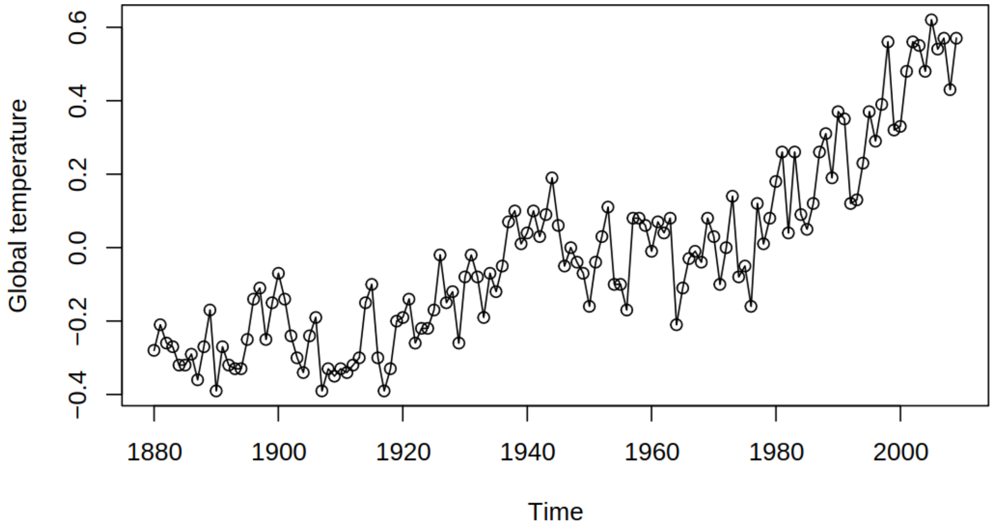


Figure 5.7: Global temperature series

Here the "trend" does not really look either quadratic or linear (or simple in any way), but we would still like to estimate the central trajectory and remove the fluctuations. And for this we may use *nonparametric smoothing*.

Our goal now is to take a time series and "extract" its main shape while removing the noise. Thus, suppose we can decompose the series into

$$X_t = T_t + E_t, \tag{5.2}$$

where T_t is the trend component and E_t are random, irregular fluctuations (noise) and we want to estimate T_t. Previously we used linear regression, e.g. we set $T_t = a + bt$, but we can allow T_t to vary more flexibly. Instead of assuming a parametric (e.g., linear) model for the trend, we will estimate it nonparametrically as a general function of t.

ℹ Note

This process has many applications, e.g. exploratory analysis, graphics, detrending or simply the preparation for the analysis of the random part.

5.3.1 Moving averages

As mentioned before, our goal is to remove fluctuations and preserve the "central" values of a given time series. Naturally, we can assume that locally the series values fluctuate around a similar level. The fluctuations have a zero mean and as such should cancel out when averaged (locally). Thus we consider a **moving average smoother**, see Definition 5.2.

Definition 5.2 (Moving average Smoother)**. Moving average smoother** of order $m = 2k + 1$ is a transformed series

$$\hat{T}_t = \frac{1}{m} \sum_{j=-k}^{k} X_{t+j},$$

or for $m = 2k$ even, we can use

$$\hat{T}_t = \frac{1}{2m} X_{t-k} + \frac{1}{m} X_{t-k+1} + \cdots + \frac{1}{m} X_{t+k-1} + \frac{1}{2m} X_{t+k}.$$

As a terminology side note, "moving average" is used for the above estimator of the trend as well as for the model based on the moving average of white noise. Recall the global temperature time series from Figure 5.7, on which we can demonstrate the moving average smoother, see Figure 5.8.

```
library(forecast)
```

```
Registered S3 method overwritten by 'quantmod':
  method            from
  as.zoo.data.frame zoo
```

```
plot(gtemp,type="o",ylab="Global temperature")
lines(ma(gtemp,15),col=2)
```

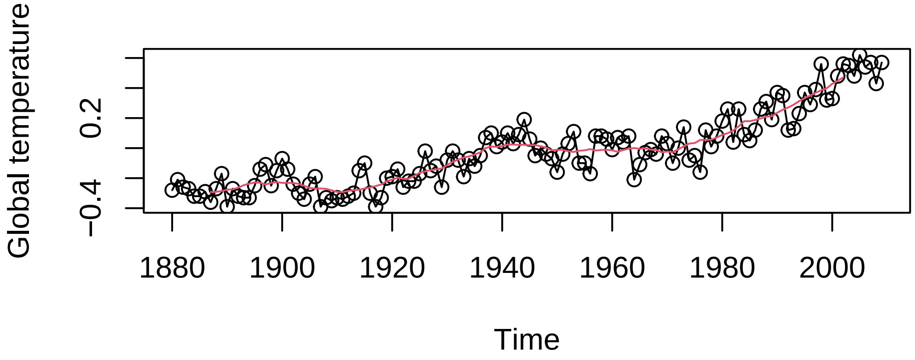
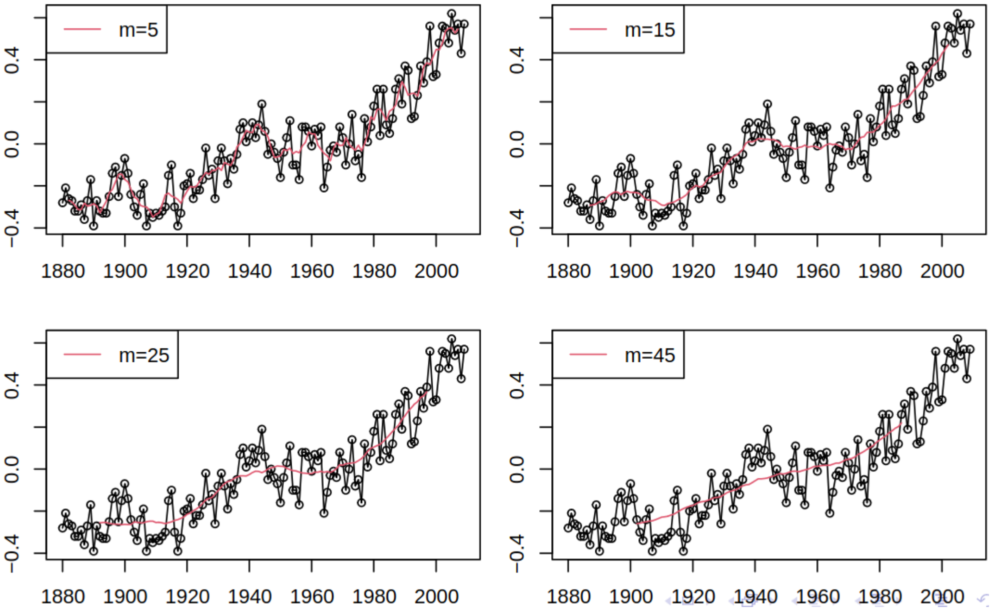


Figure 5.8: Relatively smooth trajectory estimating the global tendency

Furthermore, we can compare the effect of different orders of the moving average.



5.3.2 Kernel smoothing

The aforementioned moving average smoother, see Definition 5.2, locally averages observations, but all are given the same weight. We could generalize this concept – consider moving *weighted* averages with weights depending on the proximity to the time point of interest.

Definition 5.3 (Kernel Smoothing (Nadaraya-Watson estimator))**.** The Kernel smoothing uses

$$\hat{T}_t = \sum_{i=1}^{n} w_i(t) X_i,$$

where the weights w_i are defined as

$$w_i(t) = \frac{K\left(\frac{i-t}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{j-t}{h}\right)}.$$

Here K is a **kernel function**, e.g. $K(z) = (2\pi)^{-\frac{1}{2}} \exp -\frac{z^2}{2}$, and h is the **bandwidth** that controls the smoothness.

Tip

For large h, $w_i(t)$ is large even when i is far from t, and thus $w_i(t)$ is "flat", averaging relies on distant observations and there is a possibility of over-smoothing.

> On the other hand, for small h, $w_i(t)$ is large only when i is close to t, and thus $w_i(t)$ drops quickly. Averaging then depends mainly on nearby observations but under-smoothing is possible.

If we again apply this method to our example, see Figure 5.7, we get Figure 5.9.

```
plot(gtemp,type="o",ylab="Global temperature")
lines(ksmooth(
        time(gtemp),
        gtemp,
        kernel="normal",
        bandwidth=10
    ),
    col=2
)
```

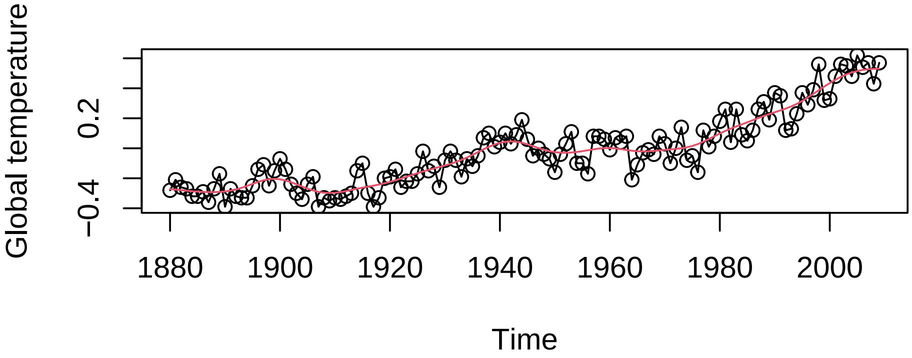


Figure 5.9: Trend approximated with kernel smoothing

What's more, we can also compare the effect of bandwidth on the estimate.

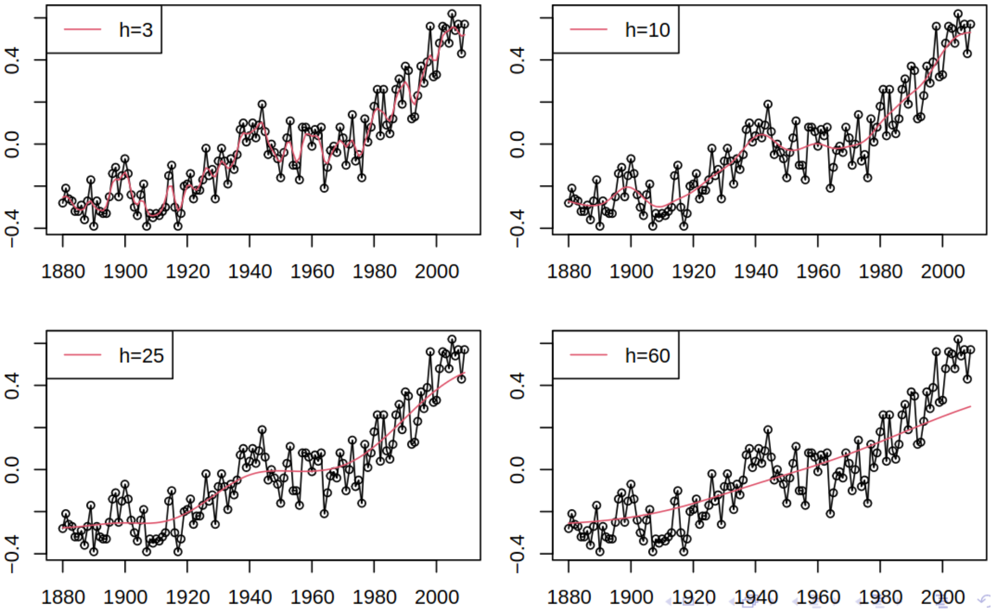


Figure 5.10: Effect of bandwidth on kernel smoothing

As for the properties of the kernel smoother (see Definition 5.3), $\hat{T}_t$ in fact estimates the smoothed version of T_t taken from (5.2) (also see Definition 5.2) and

$$\mathbb{E}\,\hat{T}_t = \sum_{i=1}^{n} w_i(t) T_i.$$

As $\hat{T}_t$ is a linear function of data, the variance of the kernel smoother is easy to find, if the autocovariance γ of the series is known (e.g. in iid case). Then

$$\operatorname{var}\hat{T}_t = \sum_{i=1}^{n}\sum_{j=1}^{n} w_i(t) w_j(t)\gamma(i,j).$$

In time series analysis, we typically use kernel smoothing as an exploratory technique, when we are not much interested in the standard errors, confidence intervals etc. There is also a *bias-variance tradeoff*:

- large h gives us *high bias* but *low variance*;
- small h gives us *low bias* but *high variance*.

And while automatic procedures to find the best compromise do exist, they often assume iid errors and risk possible overfitting with positive autocorrelation, i.e., series of similar values.

Note

There are, of course, more advanced kernel smoothing techniques:

- locally polynomial fitting (Nadaraya–Watson is locally constant),
- locally adaptive choice of h (h depending on t),
- etc.

5.3.3 Other smoothing techniques

Although we don't have the time to explain and show the following methods in this course, one can also use

- Splines
  - Express T_t as a linear combination of spline functions;
  - Estimate by least squares – possibly with penalization for "roughness" (to avoid fitting the data instead of the trend);
- Lowess (locally weighted scatterplot smoothing);
- Loess;
- Nearest neighbors.

6 Decomposition Techniques and Simple Prediction Methods

6.1 Decomposition

Consider a series with different components, see Figure 6.1 for examples.

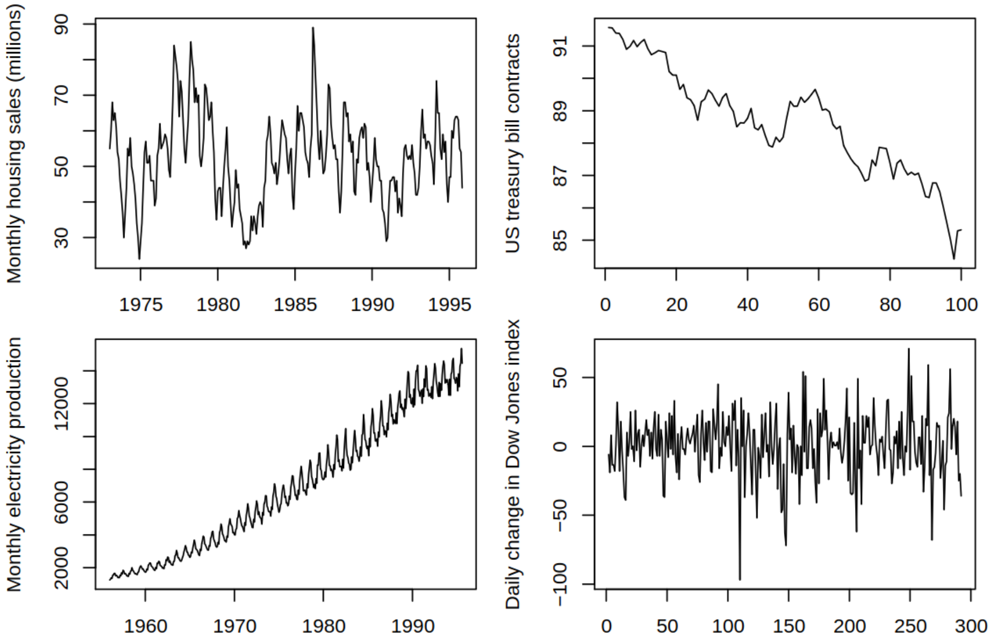


Figure 6.1: Examples of series with different components

As we did earlier, our goal will be to decompose X_t as

$$X_t = f(T_t, S_t, E_t)$$

where

- T_t is the trend,
- S_t is the seasonal component (length of season s, e.g. $s = 12$),
- E_t is the random component (remainder, irregular, error term).

All in all, our approach is similar to what we've used before, but now we allow the components to vary in a flexible, non-parametric way. As examples of this decomposition, consider

- Additive decomposition
$$\mathrm{X}_t = T_t + S_t + \mathrm{E}_t;$$
- Multiplicative decomposition
$$\mathrm{X}_t = T_t S_t \mathrm{E}_t.$$

In general, the goal of the decomposition is to aid us in exploratory analysis or it can serve as a preliminary step before the analysis of the random component.

6.1.1 Classical Decomposition

Consider a "classical" *additive decomposition* of X_t as

$$\mathrm{X}_t = T_t + S_t + \mathrm{E}_t,$$

where T_t is allowed to change in time flexibly. Moreover, S_t is periodic with period s with the seasonal indices presumed to be constant in time. So our modeling procedure goes:

1. estimate T_t by the moving average of order s to obtain $\hat{\mathrm{T}}_t$;
2. calculate the de-trended $\mathrm{X}_t - \hat{\mathrm{T}}_t$;
3. estimate the seasonal component for each season (e.g. month) by averaging the de-trended values for that season. Then we adjust them so that they add up to zero, by which we produce $\hat{\mathrm{S}}_t$;
4. calculate the remainder component $\hat{\mathrm{E}}_t = \mathrm{X}_t - \hat{\mathrm{T}}_t - \hat{\mathrm{S}}_t$.

Example 6.1. As an example, consider housing sales and the following code:

```
data(hsales, package="fma")
plot(decompose(hsales,type="additive"))
```

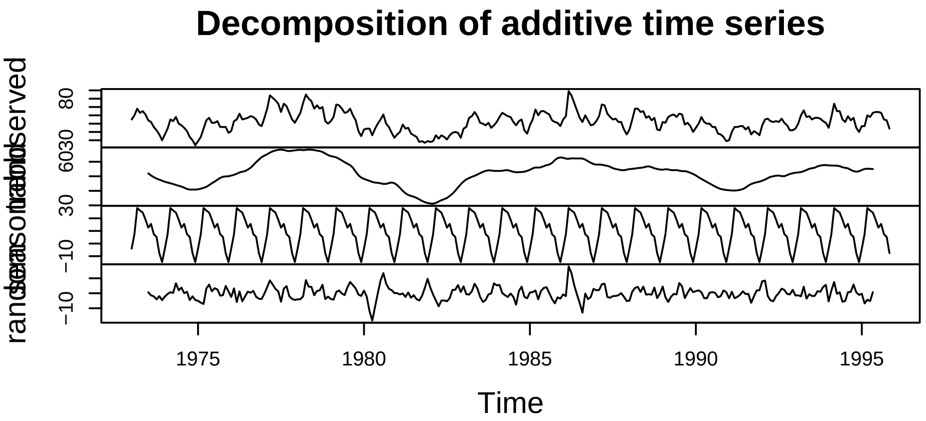


Figure 6.2: Classical decomposition of housing sales

6.1.2 Multiplicative Classical Decomposition

This time we decompose X_t as

$$X_t = T_t S_t E_t$$

with the same interpretation of the used symbols. Our modeling process stays almost the same as well:

1. estimate T_t by the moving average of order s to obtain $\hat{T}_t$;
2. calculate the de-trended $X_t/\hat{T}_t$;
3. estimate the seasonal component for each season (e.g. month) by averaging the de-trended values for that season. Then we adjust them so that they add up to zero, by which we produce $\hat{S}_t$;
4. calculate the remainder component $\hat{E}_t = \frac{X_t}{\hat{T}_t \hat{S}_t}$.

Example 6.2. As an example, this time consider electricity production data.

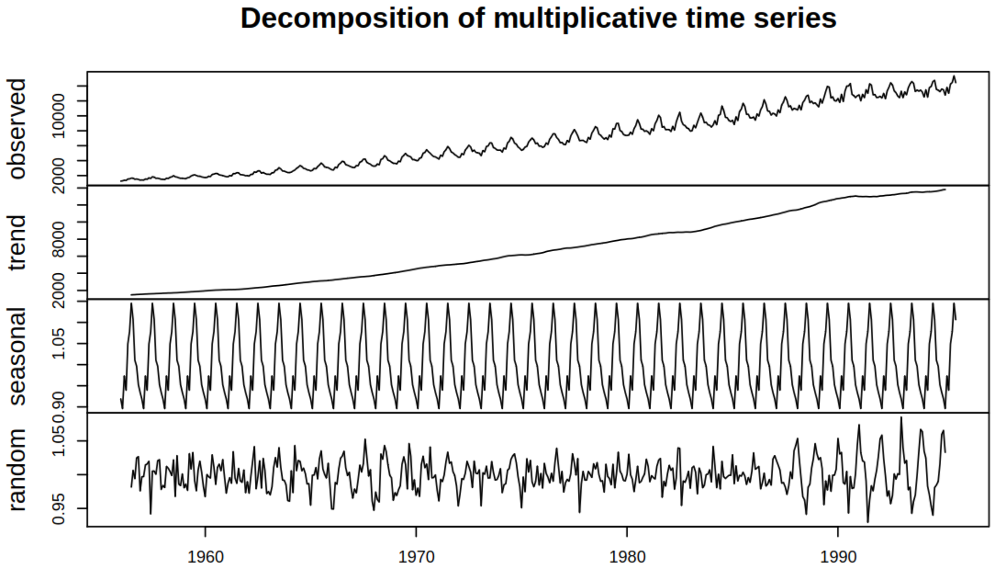


Figure 6.3: Multiplicative classical decomposition of electricity production

6.1.3 Problems with classical decomposition

As we might have noticed, there are certain problems with the classical decomposition. For one, due to the moving average estimation, the trend is not available at the beginning and end of our data (roughly at the first and last $s/2$ time points). Also, we assumed constant seasonal effects, which may be OK for some series but not for all, see Figure 6.4.

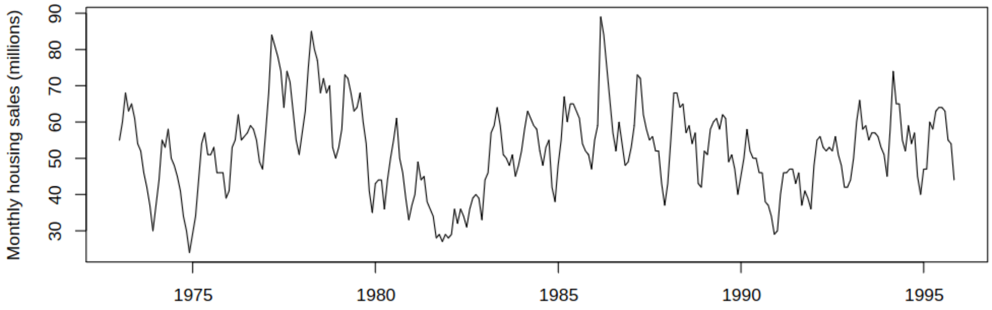


Figure 6.4: Non-consant seasonal effect

Basically, the latter problem boils down to the question of if the *trend* is allowed to vary flexibly for exploratory analysis, why not the seasonals?

6.1.4 STL Decomposition

Once again, consider *additive* decomposition of X_t as

$$X_t = T_t + S_t + E_t,$$

though this time we estimate T_t, S_t, E_t using **STL** (*Seasonal and Trend decomposition using Loess*). Hence now, T_t is allowed to change in time in a flexible way, and the smoothness of the trend component can be controlled. Also, S_t is now allowed to change in time flexibly (not only periodically), and the rate of change can also be controlled. Both time-varying functions are estimated nonparametrically by loess, which is a nonparametric regression technique similar to kernel smoothing.

❗ Important

STL decomposition is **only additive**, so one needs to use log or Box-Cox (see Definition 5.1) transformations for different types of decompositions.

Recall the classical decomposition of housing sales, see Example 6.1. We now decompose the same data using STL with the following code:

```
plot(stl(hsales,s.window=13))
```

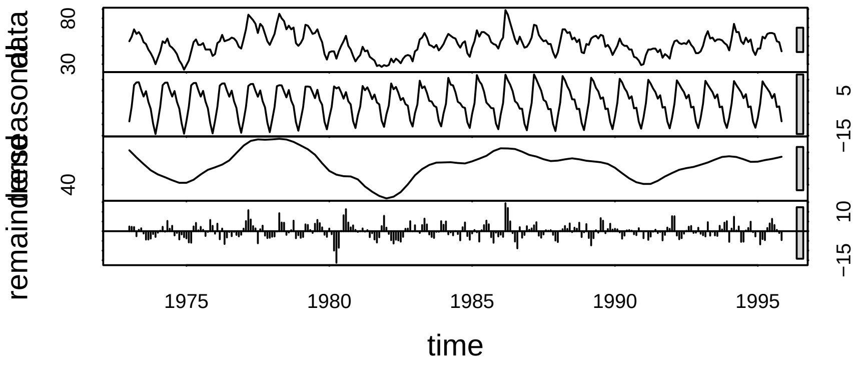
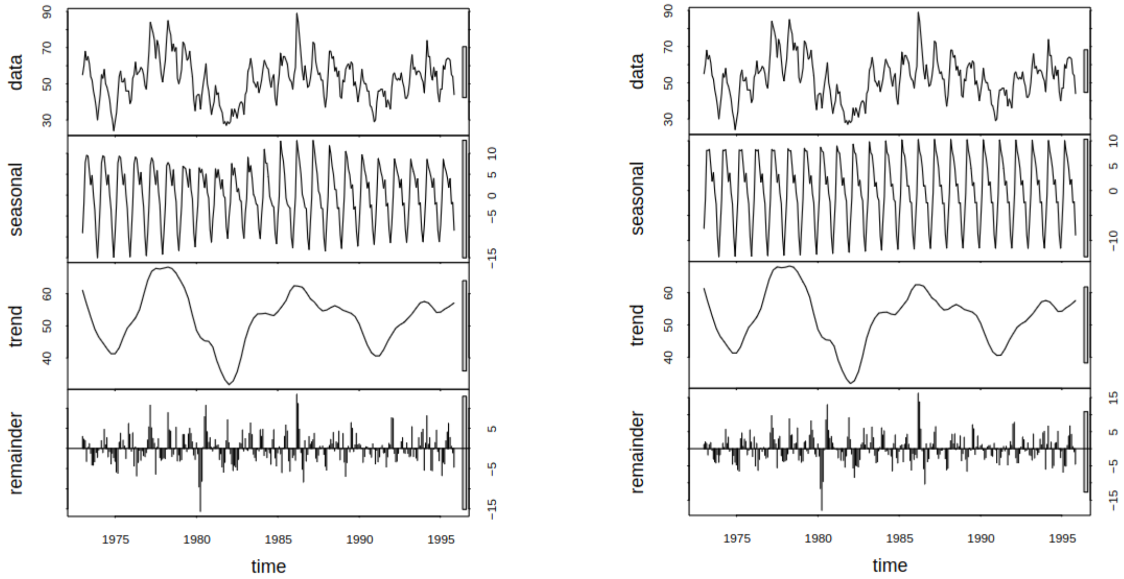


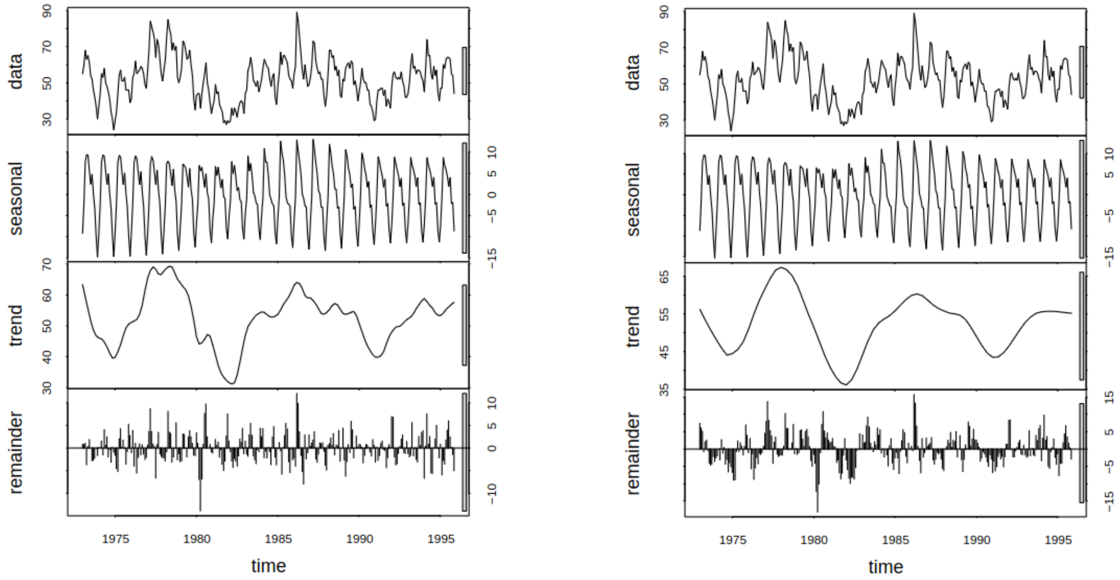
Figure 6.5: Housing sales decomposed with STL

In general, STL decomposition first estimates the trend by loess. Then it de-trends the series and applies loess smoothing to each seasonal sub-series. It follows by removing the seasonal component. All of these steps are iterated until a satisfying result is achieved (or another stopping condition is met).



(a) `s.window = 11`   (b) `s.window = 19`

Figure 6.6: Comparison of different *seasonal* smoothing parameters for STL

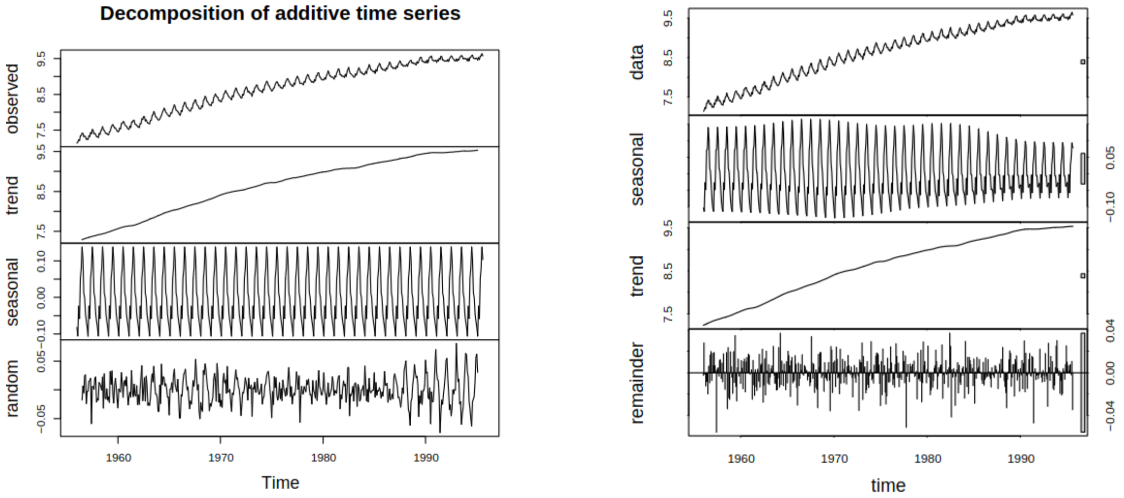


(a) `t.window = 15`   (b) `t.window = 40`

Figure 6.7: Comparison of different *trend* smoothing parameters for STL

6.1.5 Comparison of Classical and STL Decompositions

When comparing these two methods on the log-transformed electricity data, see Example 6.2, we see that the classical decomposition fails to capture the decrease of seasonal effects at the end and as such, they propagate to the remainder component. What's more, the remainder from STL looks far more random and irregular, compared to the one produced by the classical decomposition.



(a) Classical decomposition

(b) STL decomposition

Figure 6.8: Comparison of classical and STL decompositions of log-transformed electricity production data

6.2 Exponential Smoothing

Let us now assume our goal is to predict (forecast) future values as a function of past observations. Therefore we need to incorporate changing the level, trend and seasonal patterns.

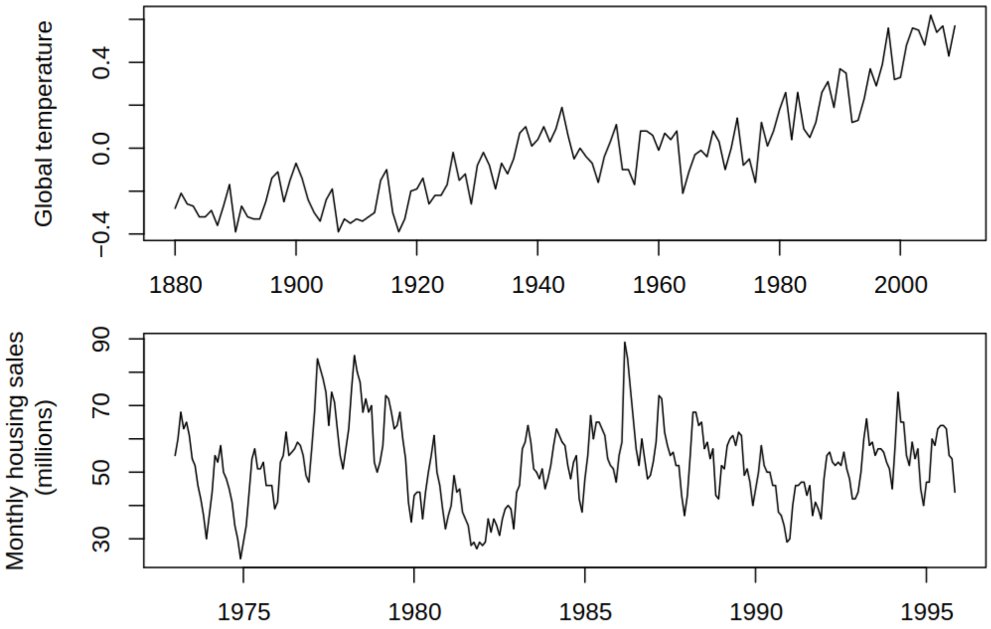


Figure 6.9: Possible time series for prediction

Let us presume we have observed data $X_1, \ldots, X_n$ and we want to predict the future values X_{n+h} for $h = 1, 2, \ldots$. Firstly, let us introduce a new notation $\hat{X}_{n+h|n}$ to denote the prediction of X_{n+h} from observed $X_1, \ldots, X_n$. Moreover, we assume there is *no systematic trend or seasonal effects*. This means that the level of the process *can* vary with time but we have no information about the likely direction of these changes.

As such, we model our data with

$$X_t = \mu_t + e_t,$$

where μ_t is the varying mean of the process and e_t are independent random inputs with mean 0 and standard deviation σ. Also, let a_t be an estimate of μ_t.

Since there is no systematic trend, X_{n+h} can be simply predicted by the **forecating equation**:

$$\hat{X}_{n+h|n} = a_n.$$

The estimate a_t can be obtained using the same way – no systematic trend implies that it is reasonable to estimate μ_t by a weighted average of our estimate at time $t-1$ and the new observation at time t, i.e. by the **smoothing equation**

$$a_t = \alpha X_t + (1-\alpha) a_{t-1} \tag{6.1}$$

for some $\alpha \in (0, 1)$. Here we call α a *smoothing parameter*. For α close to 1, we put more weight on the new observation X_t, which produces fast changes. On the other hand, for α close to 0, we put more weight on the previous estimate a_{t-1}, which forces changes to be rather slow. We can repeat the process behind (6.1) with the back-substitution method to get

$$a_t = \alpha X_t + \alpha(1-\alpha)X_{t-1} + \alpha(1-\alpha)^2 X_{t-2} + \dots$$

> **Note**
>
> From this arises the name *exponential smoothing* or *exponentially weighted moving average*.

As an example, consider exponential smoothing applied to global temperatures with the following code

```r
data(gtemp,package="astsa")
 # fit the model
m = HoltWinters(gtemp,
    alpha=.3,
    beta=F,
    gamma=F
)
# compute predictions 10 years ahead
p = predict(m,10)
plot(m,p,main="")
```

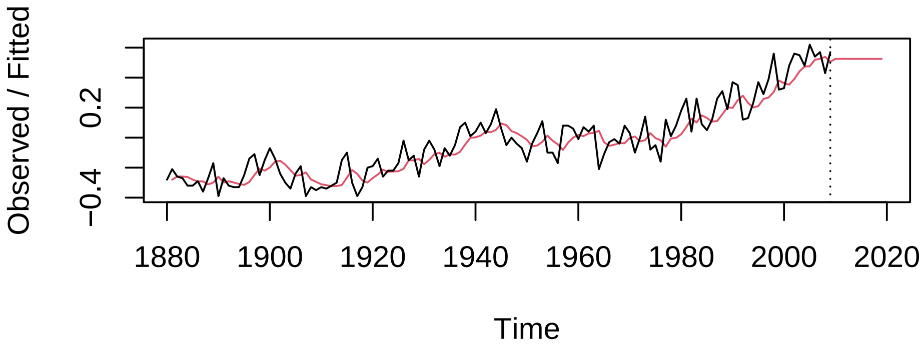


Figure 6.10: Exponential smoothing on global temperatures

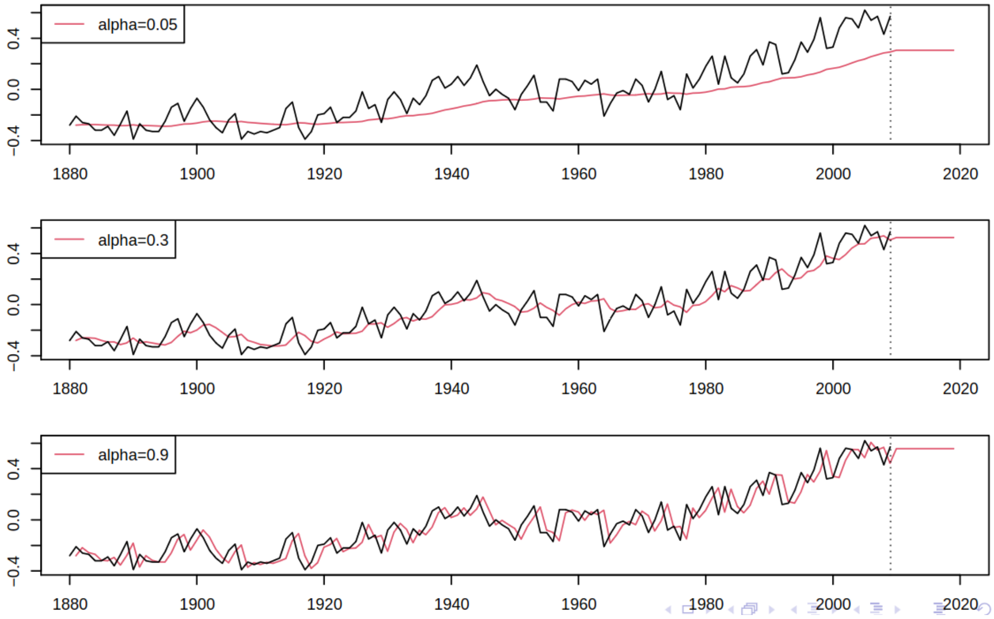


Figure 6.11: The effect of the smoothing α parameter

6.2.1 Selection of the smoothing parameter

As always, the exact selection of the smoothing parameter is a tricky matter. In general, small α leads to over-smoothing and the smoothed level lacks flexibility, while big α promotes under-smoothing, and the smooth level follows the data too closely. As heuristics, values around 0.2 to 0.3 are often recommended.

In linear regression we estimate parameters by minimizing the sum of squared errors, so here we can use the same idea. Surely, we can estimate the unknown parameters and the initial values for exponential smoothing by minimizing the sum of squared errors. In this case, we consider the prediction errors $X_t - \hat{X}_{t|t-1}$. Putting all of this together we get that the criterion to minimize is

$$\sum_{t=1}^{n} \left(X_t - \hat{X}_{t|t-1}\right)^2 .$$

Unlike with ordinary least squares, this problem is non-linear and numerical methods (like Newton's method) must be used.

Tip

In R one can use `HoltWinters` to estimate the smoothing parameter α.

What's more, we can construct prediction intervals that will contain the future observation X_{n+h} with a high probability, e.g. 0.95, as in Figure 6.12 using the following code.

```
m = HoltWinters(gtemp,
    alpha=.3,
    beta=F,
    gamma=F
)
p = predict(m,10,
    prediction.interval=TRUE
)
plot(m,p,main="")
```

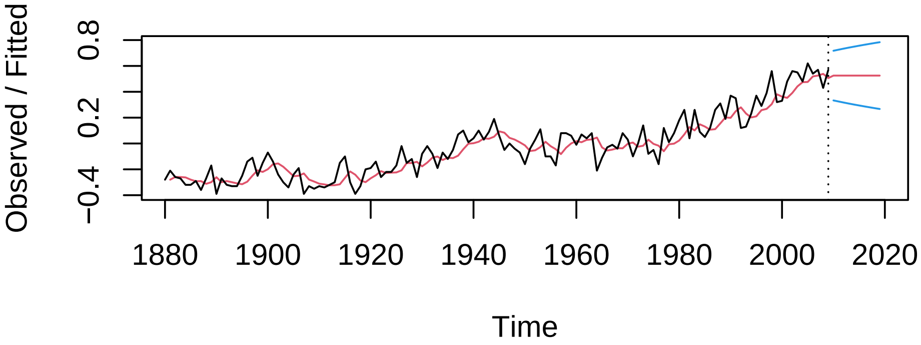


Figure 6.12: Prediction intervals for future values of global temperatures

6.3 Holt's linear method

Exponential smoothing assumes no trend of the level, predictions are constant, but this may be insufficient, e.g., global temperatures exhibit temporary trends. As such, we use Holt's method which incorporates trends. Therefore we predict X_{n+h} from data up to time n by

$$\hat{X}_{n+h|n} = a_n + hb_n,$$

for $h = 1, 2, \ldots$, where a_n is an estimate of level, and b_n is an estimate of slope. So now we estimate the level at t by a combination of the new observation and linearly extrapolated previous estimate

$$a_t = \alpha X_t + (1 - \alpha)(a_{t-1} + b_{t-1}).$$

Similarly, we estimate the slope by a combination of the slope of the new level change and the previous estimate

$$b_t = \beta(a_t - a_{t-1}) + (1 - \beta)b_{t-1}.$$

Same as before, $\alpha, \beta \in (0, 1)$ are smoothing parameters controlling the flexibility of the estimates a_t, b_t – how quickly they adapt to new data.

To demonstrate this method, consider the following code, which plots Figure 6.13.

```
# fit the model
m = HoltWinters(gtemp,
    alpha=.4,
    beta=.1,
    gamma=F
)
# point and interval predictions
p = predict(m,10,
    prediction.interval=TRUE
)
plot(m,p,main="")
```

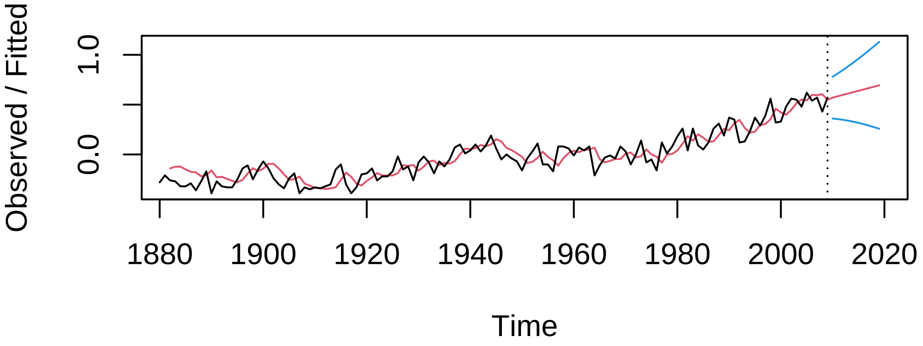


Figure 6.13: Holt's linear method

To get a better understanding of the effect of change of the smoothing parameters α, β, we may plot smoothed levels for a fixed α and different β (we've seen the effect of α already, see Figure 6.11).

Note

Again, we can optimize for the value of parameters against some criterion using, for example, least squares.

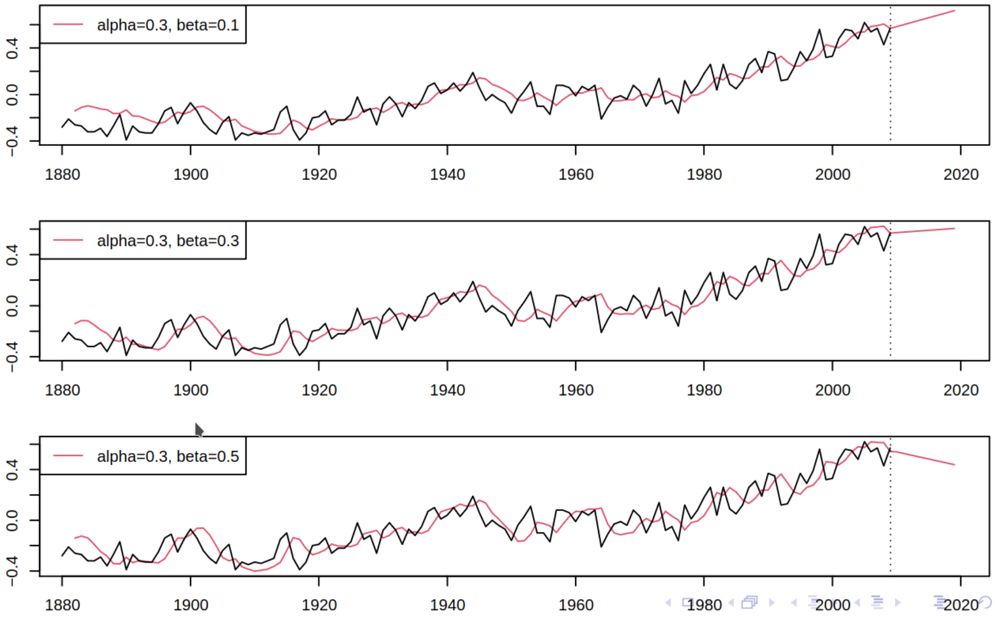


Figure 6.14: Effect of β on smoothed level using Holt's linear method

6.4 Holt-Winters method

While Holt's linear method is more general than pure exponential smoothing, it still lacks any seasonality. An extension of Holt's linear method to include the aforementioned seasonality is called a *Holt-Winters method*. Consider a season with length p, e.g. $p = 12$. We now predict X_{n+h} from data up to time n by

$$\hat{X}_{n+h|n} = a_n + hb_n + s_{n+h-\lceil h/p \rceil p},$$

for $h = 1, 2, \ldots$, where

- a_n is an estimate of the level;
- b_n is an estimate of the slope;
- $s_{n+h-\lceil h/p \rceil p}$ is an estimate of the seasonal effect at time $n + h - \lceil h/p \rceil p$ (i.e., the corresponding seasonal effect within the last p observation times, e.g., the seasonal of the last observed April if $n + h$ is in April).

The updating equations now are

$$\begin{aligned} a_t &= \alpha(X_t - s_{t-p}) + (1 - \alpha)(a_{t-1} + b_{t-1}), \\ b_t &= \beta(a_t - a_{t-1}) + (1 - \beta)b_{t-1}, \\ s_t &= \gamma(X_t - a_t) + (1 - \gamma)s_{t-p}, \end{aligned}$$

where $\alpha, \beta, \gamma \in (0, 1)$ are the smoothing parameters. Again, to demonstrate consider the housing sales data and the following code

```
# let alpha, beta, gamma be selected automatically
m = HoltWinters(hsales)
# compute predictions 10 years ahead
p = predict(m,24,
    prediction.interval=TRUE
)
plot(m,p,main="")
```

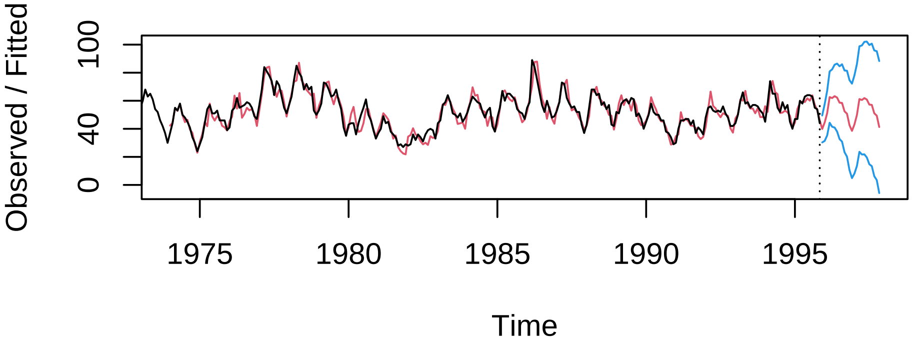


Figure 6.15: Smoothing and prediction using Holt-Winters method

7 ARMA Processes

7.1 From differences equations to ARMA models

Autoregressive models work on the idea that we may want to model or describe real-world phenomena by recurrence relations

$$x_t = g(x_{t-1}, \dots, x_{t-p}),$$

for example $x_t = \varphi x_{t-1}$ or $\Delta x_t \equiv x_t - x_{t-1} = \varphi x_{t-1}$. Often it is sufficient to take a linear relation g, thus we get *homogeneous linear difference equation of order p*

$$x_t = \varphi_1 x_{t-1} + \cdots + \varphi_p x_{t-p}.$$

We will find these models as a motivation for a class of models for time series or as tools for deriving their properties.

Tip

ARMA model stands for *AutoRegressive Moving Average model*.

7.1.1 Backshift operator

Definition 7.1 (Backshift operator)**.** The backshift (or lag) operator B is the operator that maps a sequence $\{x_t;\ t \in \mathbb{Z}\}$ to the sequence $\{x_{t-1};\ t \in \mathbb{Z}\}$, i.e., element-wise $\mathsf{B}x_t = x_{t-1}$.

By a k-power of B we denote a repeated application

$$\mathsf{B}^k x_t = \overbrace{\mathsf{B} \dots \mathsf{B}}^{k} x_t = x_{t-k}.$$

Surely $\mathsf{B}^0 x_t = 1x_t = x_t$. We may also notice that this operator is linear

$$a\mathsf{B}^k x_t + b = ax_{t-k} + b,$$

and its difference is defined as $(1-\mathsf{B})x_t = x_t - x_{t-1}$. Considering polynomials, for $P(z) = \sum_{j=0}^{p} \alpha_j z^j$, we get $P(\mathsf{B})x_t = \sum_{j=0}^{p} \alpha_j x_{t-j}$. Together, the difference equation

$$x_t - \varphi_1 x_{t-1} - \cdots - \varphi_p x_{t-p} = 0$$

can be written as

$$\Phi\left(\mathsf{B}\right) x_t = 0,$$

where $\Phi\left(z\right) = 1 - \sum_{j=1}^{p} \varphi_j z^j$.

7.1.2 Solving difference equations

Consider a difference equation

$$x_t - \varphi_1 x_{t-1} - \cdots - \varphi_p x_{t-p} = 0,$$

i.e. $\Phi\left(\mathsf{B}\right) x_t = 0$, and let our goal be to find solution(s), i.e. find $\{x_t, t \in \mathbb{N}\}$ that satisfies this equations. For $p = 1$, straightforward substitution gives

$$x_t = \varphi_1 x_{t-1} = \cdots = \varphi_1^t x_0.$$

Notice that φ_1 satisfies $\Phi\left(\varphi_1^{-1}\right) = 0$. Thus, the solution is of the form $c\xi_0^{-t}$, where ξ_0 is the root of Φ. For $p > 1$, if ξ_0 is a simple root of Φ, we can factorize Φ as

$$\Phi\left(z\right) = \Phi^*(z)(1 - \xi_0^{-1} z).$$

Hence $c\xi_0^{-t}$ solves the equation $\Phi\left(\mathsf{B}\right) x_t = \Phi^*(\mathsf{B})(1 - \xi_0^{-1}\mathsf{B})x_t = 0$, because it solves the first order equation $(1 - \xi_0^{-1}\mathsf{B})x_t = 0$. This procedure can be repeated with all simple roots. On the other hand, if ξ_0 is a root with multiplicity m, then Φ can be factorized as

$$\Phi\left(z\right) = \Phi^*(z)(1 - \xi_0^{-1} z)^m,$$

where the equation $(1 - \xi_0^{-1} z)^m$ is then solved by $\xi_0^{-t}, t\xi_0^{-t}, \dots, t^{m-1}\xi_0^{-t}$. Because it can be shown that any non-trivial linear combination of solutions is a solution, then the general solution is given by the following expression

$$\sum_{j=1}^{q} \sum_{k=0}^{m_j - 1} c_{jk} t^k \xi_j^{-t},$$

where $\xi_1, \dots, \xi_q$ are distinct roots with multiplicities $m_1, \dots, m_q$, i.e. $\sum_{j=1}^{q} m_j = p$. For a pair of complex conjugate roots $\left(\xi_j, \overline{\xi_j}\right)$, the solution becomes

$$\sum_{k=0}^{m-1} a_{jk} t^k \left|\xi_j\right|^{-t} \cos(\arg(\xi_j)t + b_{jk}).$$

Note that using initial conditions (values of x_t at p time points), we can determine c_{jk} (and a_{jk}, b_{jk}).

7.1.3 Stochastic difference equations

In real data, difference equations may not hold exactly as observed data are less regular. But we could theoretically exploit the asymptotic behavior of difference equations: when all roots are outside the unit circle ($\left|\xi_j\right| > 1$), the solution converges to 0 but in real data, we see stabilization rather than extinction.

To overcome these issues, we may add some randomness to the equation

$$x_t - \varphi_1 x_{t-1} - \cdots - \varphi_p x_{t-p} = 0.$$

Let $\{\varepsilon_t, t \in \mathbb{Z}\}$ be a sequence of random variables with mean 0, variance $\sigma^2 > 0$ and covariance 0 (white noise $\mathrm{WN}\left(0, \sigma^2\right)$). Now consider the stochastic difference equation

$$X_t - \varphi_1 X_{t-1} - \cdots - \varphi_p X_{t-p} = \varepsilon_t,$$

i.e. $\Phi\left(\mathsf{B}\right) X_t = \varepsilon_t$. Here ε_t are *random* errors/disturbances/perturbations of the model/*random* inputs.

Definition 7.2 (Autoregressive (AR) model)**.** The process $\{X_t, t \in \mathbb{Z}\}$ satisfying

$$X_t - \varphi_1 X_{t-1} - \cdots - \varphi_p X_{t-p} = \varepsilon_t,$$

where $\{\varepsilon_t\} \in \mathrm{WN}\left(0, \sigma^2\right)$, is called an **autoregressive process** of order p.

The Definition 7.2 coins the so-called **autoregression**

$$X_t = \varphi_1 X_{t-1} + \cdots + \varphi_p X_{t-p} + \varepsilon_t,$$

which means that past values are predictors of future values. It is convenient for forecasting and it is constructive, as it describes the underlying physical process and the random data generating process. Moreover, random errors influence future outcomes as they accumulate, i.e. it features the propagation of errors.

Tip

Compare, e.g., with linear regression $Y_i = \beta_0 + \beta_1 Z_i + e_i$ – the outcome is also model (straight line) plus random error, but the errors do not enter as input in the model.

In time series, we have dependent (correlated) sequences of variables. Correlation often becomes small or disappears for observations far apart.

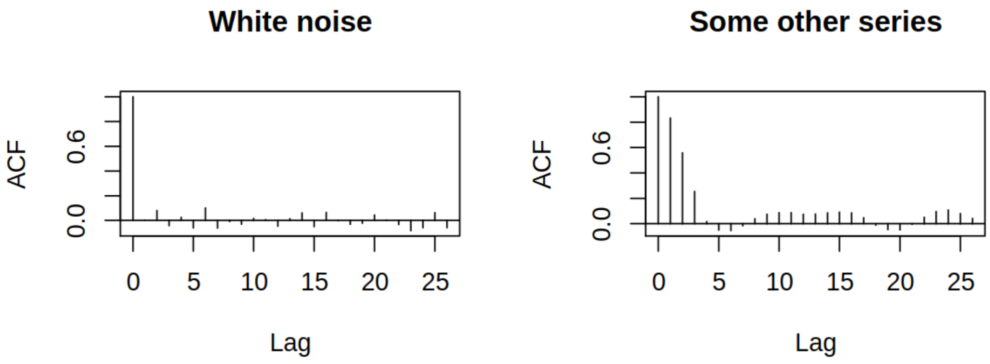


Figure 7.1: ACF from simulated series

This suggests that we may attempt to construct variables that are correlated up to some lag, then uncorrelated, e.g. consider linear combinations of white noise.

7.1.4 Moving average model

Definition 7.3. The process $\{X_t, t \in \mathbb{Z}\}$ satisfying

$$X_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q},$$

for $t \in \mathbb{Z}$, where $\{\varepsilon_t\} \sim \text{WN}\left(0, \sigma^2\right)$, is called a **moving average process** of order q.

Again, we can write $X_t = \Theta(\mathsf{B})\varepsilon_t$ for $\Theta(z) = 1 + \theta_1 z + \ldots \theta_q z^q$. Obviously, $\text{cov}\left(X_t, X_{t+h}\right) = 0$ for $|h| > q$, and surely it is stationary.

7.2 Towards ARMA models

We have 2 different concepts:

- *Deterministic difference equation*: the ideal data generating process

$$x_t - \varphi_1 x_{t-1} - \cdots - \varphi_p x_{t-p} = 0;$$

- *Stochastic difference equation*: ideal process with random perturbation

$$X_t - \varphi_1 X_{t-1} - \cdots - \varphi_p X_{t-p} = \varepsilon_t.$$

Because random inputs (disturbances, errors, ...), i.e. uncorrelated variables (white noise) may sometimes be too restrictive, we may want to allow correlated perturbations of the difference equation. Thus we may use an MA process instead of a pure white noise.

Definition 7.4 (ARMA model)**.** The process $\{X_t, t \in \mathbb{Z}\}$ satisfying

$$X_t - \varphi_1 X_{t-1} - \cdots - \varphi_p X_{t-p} = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}, \tag{7.1}$$

for $t \in \mathbb{Z}$, where $\{\varepsilon_t\} \sim \text{WN}\left(0, \sigma^2\right)$, is called an **autoregressive moving average process** of order (p, q).

Clearly, the equation (7.1) from Definition 7.4 can be re-written as

$$\Phi\left(\mathsf{B}\right) X_t = \Theta\left(\mathsf{B}\right) \varepsilon_t, \quad t \in \mathbb{Z}$$

for

$$\Phi(z) = 1 - \sum_{j=1}^{p} \varphi_j z^j, \quad \Theta(z) = 1 + \sum_{k=1}^{q} \theta_k z^k.$$

ℹ Note

Trivially, $\text{ARMA}(p, 0) = \text{AR}(p)$ and $\text{ARMA}(0, q) = \text{MA}(q)$.

7.2.1 Parameter redundancy

Consider now $\text{ARMA}(0, 0)$ process given by $X_t = \varepsilon_t$. Equivalently, $\alpha X_{t-1} = \alpha\varepsilon_{t-1}$. By subtracting these two equations, we get

$$X_t - \alpha X_{t-1} = \varepsilon_t - \alpha\varepsilon_{t-1}, \tag{7.2}$$

which looks like an $\text{ARMA}(1, 1)$ process but X_t is still white noise (i.e. $X_t = \varepsilon_t$ solves the equation). Thus we got parameter redundancy (or over-parametrization).

Hence we can rewrite (7.2) in form $\Phi(\mathsf{B}) X_t = \Theta(\mathsf{B})\varepsilon_t$, then

$$(1 - \alpha\mathsf{B})X_t = (1 - \alpha\mathsf{B})\varepsilon_t$$

and we can apply $(1 - \alpha\mathsf{B})^{-1}$ to get $X_t = \varepsilon_t$. In general, we can do this with all common roots (factors) of Φ and Θ to reduce the complexity of the parametrization.

7.3 Basic properties of ARMA processes

7.3.1 Infinite moving average

Recall the moving average $\text{MA}(q)$

$$X_t = \varepsilon_t + \theta_1\varepsilon_{t-1} + \cdots + \theta_q\varepsilon_{t-q}$$

defined in Definition 7.3. We can extend it to $\text{MA}(\infty)$ as follows

$$X_t = \varepsilon_t + \psi_1\varepsilon_{t-1} + \cdots = \sum_{j=0}^{\infty} \psi_j\varepsilon_{t-j}.$$

Remember now the already presented $\text{AR}(1)$ has the form

$$X_t = \varphi_1 X_{t-1} + \varepsilon_t = \varphi_1^2 X_{t-2} + \varphi_1\varepsilon_{t-1} + \varepsilon_t = \cdots = \varphi_1^k X_{t-k} + \sum_{j=0}^{k-1} \varphi_1^j \varepsilon_{t-j},$$

then maybe

$$X_t = \sum_{j=0}^{\infty} \varphi_1^j \varepsilon_{t-j}.$$

Much care is needed though – the series above should be the limit of random variables $\sum_{j=0}^{n} \psi_j\varepsilon_{t-j}$ as $n \to \infty$, but we haven't specified in which sense we view this limit and it is not trivial to determine whether it exists.

7.4 Intermezzo on L^2

Consider $L^2(\Omega, \mathcal{A}, P)$ s the set of all random variables on $(\Omega, \mathcal{A}, P)$ with finite second moments ($\mathbb{E}\, X^2 < \infty$). Then $L^2(\Omega, \mathcal{A}, P)$ is a linear space. We define $\langle X, Y\rangle = \mathbb{E}\,(XY)$, where $\langle X, Y\rangle$ is an inner (scalar/dot) product – meaning it is symmetric, linear, $\langle X, X\rangle \geq 0$ and $\langle X, X\rangle = 0 \iff X = 0$. Also we define the norm $\|X\| = \langle X, X\rangle^{\frac{1}{2}} = (\mathbb{E}\, X^2)^{1/2}$. Two important properties then hold

- $\|X + Y\| \leq \|X\| + \|Y\|$ (*triangle inequality*);
- $|\langle X, Y\rangle| \leq \|X\|\,\|Y\|$ (*Cauchy-Schwarz inequality*).

Convergence in this space is the convergence in the norm $\|\cdot\|$

$$X_n \xrightarrow[n\to\infty]{L^2} X \iff \|X - X_n\|^2 = \mathbb{E}\left((X_n - X)^2\right) \underset{n\to\infty}{\longrightarrow} 0,$$

which is called *convergence in* L^2 (or *convergence in the mean/convergence in mean-square*). Now recall that a sequence is called **Cauchy** if $\|x_n - x_k\| \to 0$ as $n, k \to \infty$ and that a metric space is called **complete** if every Cauchy sequence of elements of the space has a limit in the space. Hence it can be deduced that L^2 is a **Hilbert space**.

> **i** Note
>
> A space with inner product, that is complete, i.e. every Cauchy sequence has the limit in the space, is called Hilbert space.

7.5 Existence of linear processes

Theorem 7.1. *If* $\{\varepsilon_t, t \in \mathbb{Z}\} \sim \mathrm{WN}\left(0, \sigma^2\right)$ *and the sequence* $\left\{\psi_j, j \in \mathbb{Z}\right\}$ *is such that* $\sum_{j=1}^{\infty} \psi_j^2 < \infty$, *then the random series* $\sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}$ *converges in* L^2, *i.e., there is a variable* $Y \in L^2(\Omega, \mathcal{A}, P)$ *such that*

$$\mathbb{E}\left(Y - \sum_{j=0}^{n} \psi_j \varepsilon_{t-j}\right) \to 0$$

as $n \to \infty$.

Proof. Since L^2 is a Hilbert space, it is enough to verify the Cauchy criterion, which is implied by the following:

$$\mathbb{E}\left(\sum_{j=0}^{n+m} \psi_j \varepsilon_{t-j} - \sum_{j=0}^{n} \psi_j \varepsilon_{t-j}\right)^2 = \sum_{j,k=n+1}^{n+m} \psi_j \psi_k\, \mathbb{E}\left(\varepsilon_j \varepsilon_k\right) = \sum_{j=n+1}^{n+m} \psi_j^2 \sigma^2 \to 0.$$

□

Similarly, we can formulate a more general theorem.

Theorem 7.2. *If $\{Y_t, t \in \mathbb{Z}\}$ is a* ***mean zero stationary sequence*** *and the sequence $\left\{\psi_j, j \in \mathbb{Z}\right\}$ is such that $\sum_{j=1}^{\infty} \left|\psi_j\right| < \infty$, then the random series $\sum_{j=0}^{\infty} \psi_j Y_{t-j}$ converges in L^2 (in mean square) and also absolutely with probability* 1.

This way we assumed all general stationary processes (not only white noise) and got almost sure convergence (not only in L^2). Thus given a linear process

$$X_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}$$

with mean

$$\mathbb{E}\, X_t = \sum_{j=0}^{\infty} \psi_j \,\mathbb{E}\, \varepsilon_{t-j} = 0$$

and covariance

$$\begin{aligned}\operatorname{cov}(X_s, X_t) &= \mathbb{E}\left(\left(\sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}\right)\left(\sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}\right)\right) \\ &= \sum_{j,k=1}^{\infty} \psi_j \psi_k \,\mathbb{E}\left(\varepsilon_{s-j}\varepsilon_{t-k}\right) = \sum_{j=0}^{\infty} \psi_j \psi_{j+|t-s|} \sigma^2,\end{aligned}$$

it follows that $\{X_t\}$ is *weakly stationary*, see Definition 3.2.

7.6 Causality

Consider AR (1) and iteratively substitute

$$X_t = \varphi_1 X_{t-1} + \varepsilon_t = \varphi_1^2 X_{t-2} + \varphi_1 \varepsilon_{t-1} + \varepsilon_t = \cdots = \varphi_1^k X_{t-k} + \sum_{j=0}^{k-1} \varphi_1^j \varepsilon_{t-j},$$

then if $|\varphi_1| < 1$ and $\{X_t\}$ is stationary, the the remainder term $\varphi_1^k X_{t-k}$ goes to 0 in L^2 as $k \to \infty$, because

$$\mathbb{E}\left((\varphi_1^k X_{t-k})^2\right) = \varphi_1^{2k}\, \mathbb{E}\left(X_{t-k}^2\right) \to 0.$$

Also the series $\sum_{j=0}^{\infty} \varphi_1^j \varepsilon_{t-j}$ converges because $\sum_{j=0}^{\infty} (\varphi_1^j)^2 < \infty$. Hence we have the **causal representation**

$$X_t = \sum_{j=0}^{\infty} \varphi_1^j \varepsilon_{t-j},$$

where *causal* means the current state depends on some history. We can also iterate in the opposite direction

$$\begin{aligned} X_t &= \varphi_1^{-1} X_{t+1} - \varphi_1^{-1} \varepsilon_{t+1} \\ &= \varphi_1^{-2} X_{t+2} - \varphi_1^{-2} \varepsilon_{t+2} - \varphi_1^{-1} \varepsilon_{t+1} \\ &= \dots \\ &= \varphi_1^{-k} X_{t+k} - \sum_{j=1}^{k-1} \varphi_1^{-j} \varepsilon_{t+j}, \end{aligned}$$

so if $|\varphi_1| > 1$ and $\{X_t\}$ is **stationary**, we get the future dependent representation

$$X_t = -\sum_{j=1}^{\infty} \varphi_1^{-j} \varepsilon_{t+j}.$$

Hence a stationary solution exists, but is *practically useless*, because it requires the knowledge of the future for the prediction of the future. Also note that for $|\varphi_1| = 1$, the process is not stationary (we get a random walk).

Definition 7.5 (Causality)**.** An ARMA (p, q) process is said to be **causal** if there exists a sequence $\left\{\psi_j, j \in \mathbb{N}_0\right\}$ such that $\sum_{j=0}^{\infty} \left|\psi_j\right| < \infty$ and

$$X_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}, \quad t \in \mathbb{Z},$$

where $\boldsymbol{\psi}$ is called a **filter** and this representation is called a **causal representation** of the process.

Since

$$\sum_{j=0}^{\infty} \psi_j^2 \leq \left(\sum_{j=0}^{\infty} \left|\psi_j\right| \right)^2 < \infty,$$

a causal process is a linear process. Also since linear processes are stationary, causal ARMA processes are stationary.

7.6.1 Causality of an autoregressive process of order p

Consider an AR (p) process of the form $\Phi(\mathsf{B}) X_t = \varepsilon_t$. Let us try to find its causal representation. First, we factorize

$$\Phi(z) = \prod_{j=1}^{r} \left(1 - \xi_j^{-1} z\right)^{m_j},$$

where ξ_j are the distinct roots with multiplicities m_j, $\sum_{j=1}^{r} m_j = p$. Thus

$$\left(1 - \xi_1^{-1}\right)^{m_1} \cdot \dots \cdot \left(1 - \xi_r^{-1}\right)^{m_r} X_t = \varepsilon_t.$$

So if $\left|\xi_j^{-1}\right| < 1$ (i.e. $\left|\xi_j\right| > 1$), we have $\left(1 - \xi_j^{-1}\right)^{-1} = \sum_{h=0}^{\infty} \xi_j^{-h} \mathsf{B}^h$. Then we successively apply $\left(1 - \xi_1^{-1}\right)^{-1}$ to both sides to work out the coefficients. The key assumption here is that all roots of $\Phi(z)$ lie outside the unit circle.

Theorem 7.3. *Let $\{\mathsf{X}_t\}$ be an* ARMA (p, q) *process for which the polynomials Φ and Θ have no common roots (otherwise we can cancel them out). Then $\{\mathsf{X}_t\}$ is* ***causal*** *if and only if all roots of Φ lie outside the unit circle, i.e., $\Phi(z) \neq 0$ for all $z \in \mathbb{C}$ such that $|z| \leq 1$. Then coefficients $\left\{\psi_j\right\}$ can be determined by the relation*

$$\Psi(z) = \sum_{j=0}^{\infty} \psi_j z^j = \Theta(z)/\Phi(z), \quad |z| \leq 1.$$

7.6.2 Non-uniqueness of MA models and inverting them

Consider two MA (1) models

$$\begin{aligned} \mathsf{X}_t &= (1 + \theta_1 \mathsf{B})\varepsilon_t, \quad \varepsilon_t \sim \mathrm{WN}\left(0, \sigma^2\right), \\ \mathsf{X}_t^* &= (1 + \theta_1^{-1} \mathsf{B})\varepsilon_t^*, \quad \varepsilon_t^* \sim \mathrm{WN}\left(0, \theta_1^2 \sigma^2\right), \end{aligned}$$

then both X_t and X_t^* have the same autocovariance fuction

$$\gamma(0) = (1 + \theta_1^2)\sigma^2, \quad \gamma(1) = \theta_1 \sigma^2, \quad \gamma(h) = 0$$

for $|h| > 1$. Hence we have two representations for the same covariance structure of the observed process (if $\theta_1 \neq 1$). Thus we choose one of them – by mimicking the idea of causality for AR, we choose the **invertible** one, that is, X_t if $|\theta_1| < 1$ and X_t^* if $|\theta_1| > 1$.

Therefore, WLOG, let us assume that $|\theta_1| < 1$. Then we can invert MA (1) to get the its AR (∞) representation

$$\varepsilon_t = (1 + \theta_1 \mathsf{B})^{-1} \mathsf{X}_t = \sum_{j=0}^{k} (-\theta_1)^j \mathsf{X}_{t-j}.$$

Consider now a moving average process MA (q)

$$\mathsf{X}_t = \Theta(\mathsf{B})\varepsilon_t, \quad t \in \mathbb{Z}.$$

Then we factorize Θ to get

$$\mathsf{X}_t = \left(1 - \xi_1^{-1} \mathsf{B}\right)^{m_1} \cdot \ldots \cdot \left(1 - \xi_r^{-1} \mathsf{B}\right)^{m_r} \varepsilon_t,$$

where ξ_j are the distinct roots with multiplicities m_j, $\sum_{j=1}^{r} m_j = p$. Thus, WLOG, we assume that $\left|\xi_j^{-1}\right| < 1$, that is, all roots of $\Theta(z)$ are outside the unit circle (possibly after flipping to the reciprocal in each factor and multiplying the error variance as shown before).

Analogously to the AR (p) case, we can invert the MA (q) representation and successively apply $\left(1-\xi_j^{-1}\mathsf{B}\right)^{-m_j}$ to both sides to get the infinite AR representation (or **invertible** representation)

$$\varepsilon_t = \sum_{j=0}^{\infty} \pi_j X_{t-j}.$$

ℹ Note

Notice that $X_t = \sum_{j=1}^{\infty} X_{t-j} + \varepsilon_t$, which is AR (∞) process.

Definition 7.6. An ARMA (p, q) process is said to be **invertible** if there exists a sequence $\left\{\pi_j, j \in \mathbb{N}_0\right\}$ such that $\sum_{j=0}^{\infty} \left|\pi_j\right| < \infty$ and

$$\varepsilon_t = \sum_{j=0}^{\infty} \pi_j X_{t-j}, \quad t \in \mathbb{Z}.$$

Theorem 7.4. *Let $\{X_t\}$ be an ARMA (p, q) process for which the polynomials Φ and Θ have no common roots. Then $\{X_t\}$ is **invertible** if and only if all roots of Θ lie outside of the unit circle, i.e., $\Theta(z) \neq 0$ for all $z \in \mathbb{C}$ such that $|z| \leq 1$. The coefficients $\left\{\pi_j\right\}$ can be determined by relation*

$$\Pi(z) = \sum_{j=0}^{\infty} \pi_j z^j = \Phi(z)/\Theta(z), \quad |z| \leq 1.$$

8 Correlation Structure in ARMA Processes

Our motivation for the last lecture was to describe autocorrelated time series. Thus we should study the autocorrelation structure of ARMA models. Consider now an MA (q) process

$$X_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q},$$

where $\{\varepsilon_t\} \sim \mathrm{WN}\left(0, \sigma^2\right)$. Recall that $\gamma(h) = \gamma(-h)$ and compute for $h \geq 0$

$$\begin{aligned} \gamma(h) &= \mathrm{cov}(X_t, X_t + h) \\ &= \mathrm{cov}\left(\sum_{j=0}^{q} \theta_j \varepsilon_{t-j}, \sum_{j=0}^{q} \theta_j \varepsilon_{t+h-j}\right) \\ &= \begin{cases} \sigma^2 \sum_{j=0}^{q-h} \theta_j \theta_{j+h}, & 0 \leq h \leq q, \\ 0, & h > q \end{cases} \end{aligned}$$

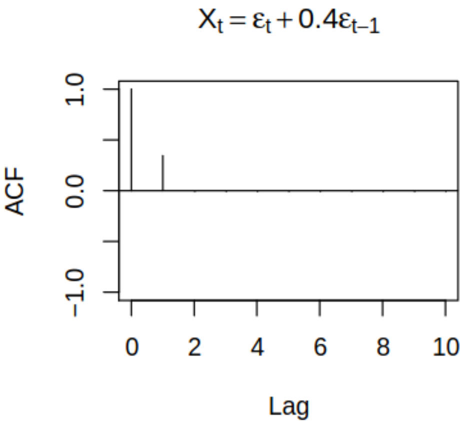
and that the autocorrelation function ρ has, in this case, the form

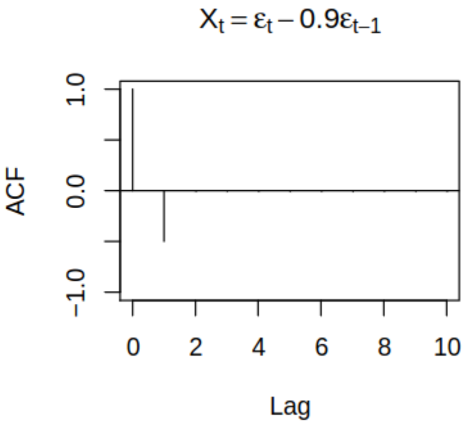
$$\rho(h) = \begin{cases} \dfrac{\sum_{j=0}^{q-h} \theta_j \theta_{j+h}}{\sum_{j=0}^{q} \theta_j^2}, & 0 \leq h \leq q, \\ 0, & h > q. \end{cases}$$

To illustrate the preceding derivations, consider the following MA (1) processes and observe

$$\rho(1) = \frac{\theta_1}{1 + \theta_1^2}, \quad \rho(h) = 0$$

for $h > 1$.


$X_t = \varepsilon_t + 0.4\varepsilon_{t-1}$


$X_t = \varepsilon_t - 0.9\varepsilon_{t-1}$

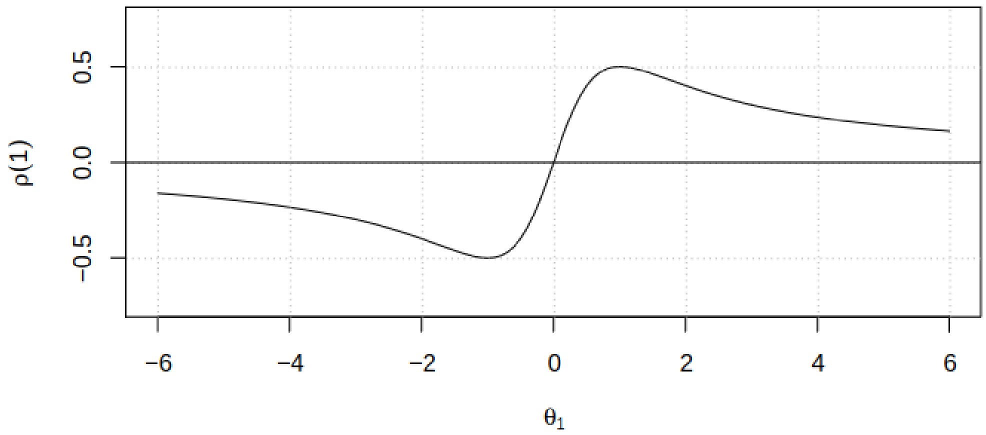


Figure 8.1: Dependence of $\rho(1)$ on θ_1

Notice that $|\rho(1)| \leq 0.5$ and it has maximum at $\theta_1 = \pm 1$. Also, observe that for most values of $\rho(1)$ there exist 2 choices for θ_1. From the fact that $|\rho(1)| \leq 0.5$ follows that we cannot use MA (1) to model data with acf $(1, 0.7, 0, 0, \dots)$. Consider now MA (2) models and look at their autocorrelations.

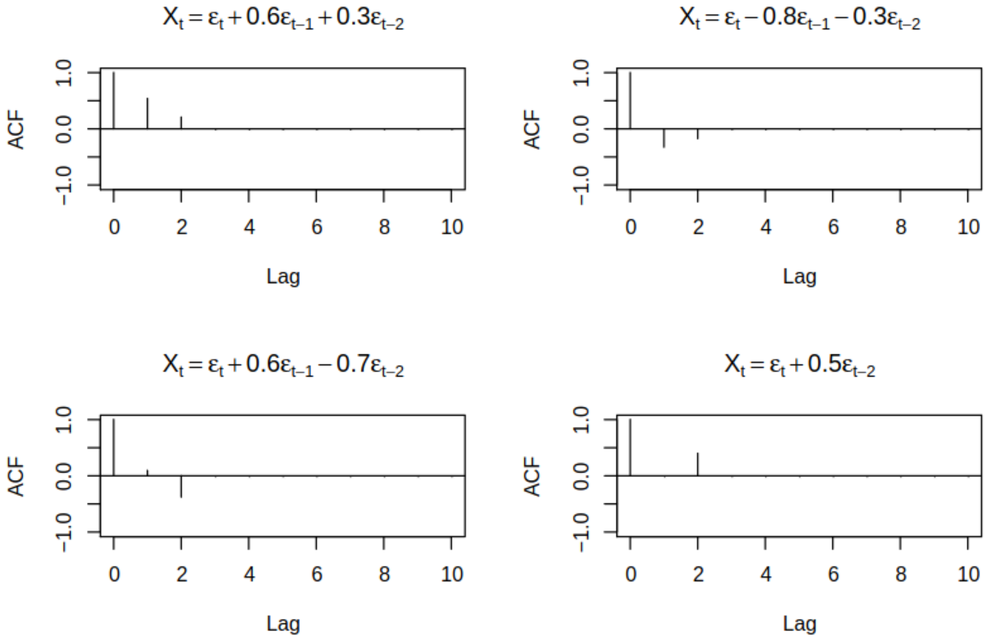


Figure 8.2: Autocorrelations of MA (2) models

It can be computed that

$$\rho(1) = \frac{\theta_1 + \theta_1\theta_2}{1 + \theta_1^2 + \theta_2^2}, \quad \rho(2) = \frac{\theta_2}{1 + \theta_1^2 + \theta_2^2}, \quad \rho(h) = 0$$

for $h > 1$, which can be seen in Figure 8.3. It can also be shown easily that $|\rho(2)| \leq 0.5$ with maximum at $\theta_1 = 0$ and $\theta_2 = \pm 1$.

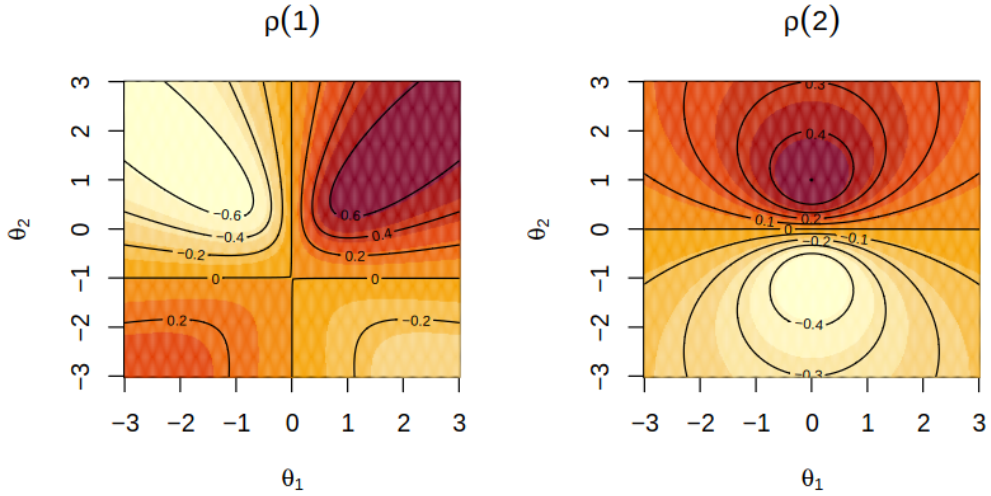


Figure 8.3: Depedence of ρ on θ_1, θ_2

8.1 Autocorrelation of the AR process

Let X_t be a **causal** AR (p) process

$$X_t = \varphi_1 X_{t-1} + \cdots + \varphi_p X_{t-p} + \varepsilon_t,$$

where $\{\varepsilon_t\} \sim \mathrm{WN}\left(0, \sigma^2\right)$. Since $\{X_t\}$ is **causal**, we see that $X_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}$ and thus $\mathbb{E}\, X_t = 0$. Now compute for $h \geq 0$ that

$$\begin{aligned}\gamma(h) &= \mathbb{E}\left(X_t X_{t-h}\right) \\ &= \mathbb{E}\left(\left(\varphi_1 X_{t-1} + \cdots + \varphi_p X_{t-p} + \varepsilon_t\right) X_{t-h}\right) \\ &= \varphi_1 \gamma(h-1) + \cdots + \varphi_p \gamma(h-p) + \sigma^2 \mathbb{1}_{h=0}.\end{aligned}$$

For $h \geq p$ we get

$$\gamma(h) - \varphi_1 \gamma(h-1) - \cdots - \varphi_p \gamma(h-p) = 0$$

or by dividing through by $\gamma(0)$ it yields

$$\rho(h) - \varphi_1 \rho(h-1) - \cdots - \varphi_p \rho(h-p) = 0,$$

which are called the **Yule-Walker** equations.

Caution

Here we can see that we started with some difference equation, then we added a stochastic part, but on the autocorrelation level, we still have the same difference equation. For polynomial decay time series, we have to use different methods/models.

Thus we recognize that $\gamma(h)$ (or $\rho(h)$), $h = 0, 1, \ldots$, satisfies the (deterministic) p-th order difference equation (for $h \geq p$)

$$\rho(h) - \varphi_1 \rho(h-1) - \cdots - \varphi_p \rho(h-p) = 0,$$

i.e. $\Phi(\mathsf{B})\rho(h) = 0$. The general solution then is

$$\rho(h) = \sum_{j=1}^{r} \sum_{k=0}^{m_j - 1} c_{jk} h^k \xi_j^{-h}$$

where ξ_j are the distinct roots of Φ with multiplicities m_j, $\sum_{j=1}^{r} m_j = p$. The initial conditions then become

$$\rho(0) = 1, \quad \rho(h) - \varphi_1 \rho(h-1) - \cdots - \varphi_p \rho(h-p) = 0,$$

where $h = 0, 1, \ldots, p-1$. To compute $\gamma(0)$ (and thus $\gamma(h)$), divide the equation for $\gamma(h)$ at $h = 0$ by $\gamma(0)$ and solve to get

$$\gamma(0) = \sigma^2 / \left(1 - \varphi_1 \rho(1) - \cdots - \varphi_p \rho(p)\right).$$

Tip

For AR processes, expect very fast (exponential) decay in parameter values – this can also be seen in their autocorrelations.

Example 8.1. Together, the autocorrelation of AR (1) process

$$\mathrm{X}_t - \varphi_1 \mathrm{X}_{t-1} = \varepsilon_t$$

has the following corresponding difference equation

$$\rho(h) - \varphi_1 \rho(h-1) = 0, \quad h \geq 1,$$

which is solved by

$$\rho(h) = \varphi_1 \rho(h-1), \quad h \geq 1.$$

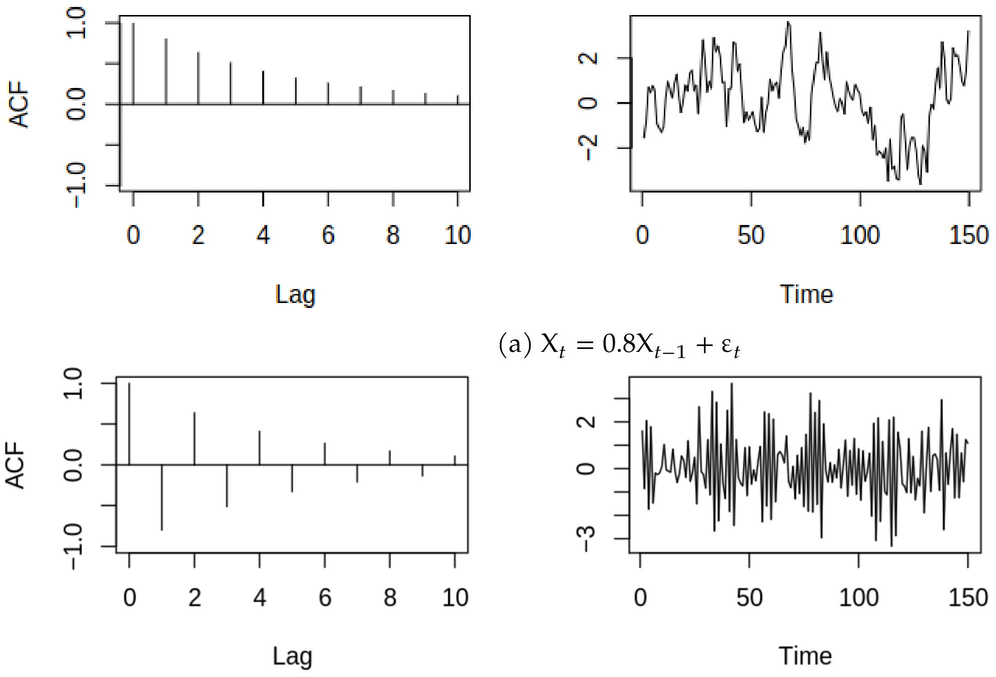
Given the initial condition $\rho(0) = 1$, this transforms to

$$\rho(h) = \varphi_1^h, \quad h \geq 0$$

and then $\gamma(0) = \sigma^2/(1 - \varphi_1^2)$.

Note

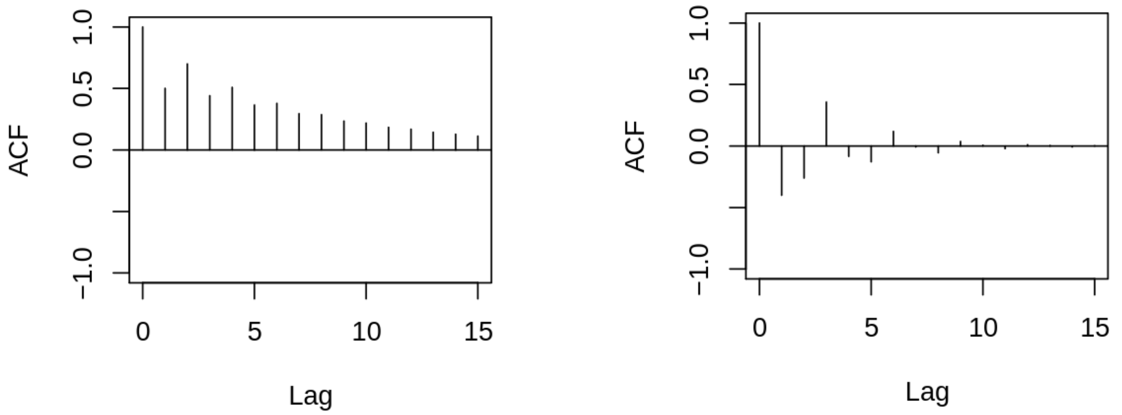
This can also be solved using the causal representation.



(a) $X_t = 0.8X_{t-1} + \varepsilon_t$

(b) $X_t = -0.8X_{t-1} + \varepsilon_t$

Figure 8.4: Examples of acf for AR (1)



(a) $X_t = 0.2X_{t-1} + 0.6X_{t-2} + \varepsilon_t$

(b) $X_t = -0.6X_{t-1} - 0.5X_{t-2} + \varepsilon_t$

Figure 8.5: Examples of acf for AR (2)

8.1.1 Autocorrelation of an ARMA process

Consider a **causal** ARMA (p, q) process

$$X_t = \varphi_1 X_{t-1} + \cdots + \varphi_p X_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q},$$

for which surely $\mathbb{E}\, X_t = 0$, which follows from the causal representation $X_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}$. We can then compute for $h \geq 0$ that

$$\begin{aligned}\gamma(h) = \mathbb{E}\left(X_t X_{t-h}\right) &= \mathbb{E}\left(\left(\sum_{j=1}^{p} \varphi_j X_{t-j} + \sum_{j=0}^{q} \theta_j \varepsilon_{t-j}\right) X_{t-h}\right) \\ &= \sum_{j=1}^{p} \varphi_j \gamma(h-j) + \sigma^2 \sum_{j=0}^{q} \theta_j \psi_{j-h},\end{aligned}$$

because

$$\mathbb{E}\left(\varepsilon_{t-j} X_{t-h}\right) = \mathbb{E}\left(\varepsilon_{t-j} \sum_{k=0}^{\infty} \psi_k \varepsilon_{t-h-k}\right) = \sigma^2 \psi_{j-h}.$$

For $h \geq \max\{p, q+1\}$, we get homogeneous difference equations

$$\gamma(h) - \varphi_1 \gamma(h-1) - \cdots - \varphi_p \gamma(h-p) = 0$$

or

$$\rho(h) - \varphi_1 \rho(h-1) - \cdots - \varphi_p \rho(h-p) = 0.$$

Given the initial conditions

$$\gamma(h) - \varphi_1 \gamma(h-1) - \cdots - \varphi_p \gamma(h-p) = \sigma^2 \sum_{j=h}^{q} \theta_j \psi_{j-h}$$

for $h = 0, \ldots, \max\{p, q+1\} - 1$. The general solution with the initial conditions stays the same as in the homogeneous case. Lastly, from the fact that X_t is causal, we get that $\left|\xi_j\right| > 1$ and thus $\rho(h)$ converges **exponentially fast to zero** as $h \to \infty$ (in sinusoidal fashion if some of the roots are complex).

8.2 Partial autocorrelation

We have seen that for an MA (q) process $\rho(h) = 0$ for $h > q$ and $\rho(q) \neq 0$, thus ρ provides information about q. On the other hand, for AR (p), no such cut-off exists, as $\rho(h)$ exponentially dampens. Therefore we might look for an analog of the autocorrelation that would provide information about the order of autoregression.

Consider now AR (1), $X_t = \varphi_1 X_{t-1} + \varepsilon_t$. Then $\operatorname{cor}(X_t, X_{t-2}) \neq 0$, because X_t is dependent on X_{t-2} through X_{t-1} and the dependencies are linear. Thus we need to break the dependence chain and remove the linear effect of everything in-between.

Definition 8.1 (Partial correlation)**.** The partial correlation between *mean zero* variables Y_1 and Y_2 given $\mathbf{Z} = (Z_1, \dots, Z_m)^\top$ is defined as

$$\rho_{Y_1 Y_2 | \mathbf{Z}} = \operatorname{cor}(Y_1 - \hat{Y}_1, Y_2 - \hat{Y}_2),$$

where $\hat{Y}_j$ is the best linear prediction (*BLP*) of Y_j from $\mathbf{Z}$.

Here the BLP takes the form $\hat{Y}_j = \hat{\beta}_1^{(j)} Z_1 + \cdots + \hat{\beta}_m^{(j)} Z_m$ where $(\hat{\beta}_1^{(j)}, \dots, \hat{\beta}_m^{(j)})$ is the minimizer of $\mathbb{E}\left(Y_j - \hat{\beta}_1^{(j)} Z_1 - \cdots - \hat{\beta}_m^{(j)} Z_m\right)$. Thus $\rho_{Y_1 Y_2 | \mathbf{Z}}$ is the correlation between Y_1 and Y_2 when the linear effect of $\mathbf{Z}$ is removed. Clearly, for a multivariate Gaussian distribution, the BLP equals the conditional expectation and hence $\rho_{Y_1 Y_2 | \mathbf{Z}} = \operatorname{cor}(Y_1, Y_2 \mid \mathbf{Z})$.

> **Note**
>
> This concept relates to the precision (concentration) matrix
>
> $$\boldsymbol{P} = \operatorname{var}(Y_1, Y_2, \mathbf{Z})^{-1}: \quad \rho_{Y_1 Y_2 | \mathbf{Z}} = -P_{12} / \sqrt{P_{11} P_{22}}.$$

Definition 8.2. The **partial autocorrelation function (PACF)** of a stationary process, $\{X_t\}$, denoted as $\alpha(h)$, is defined for $h = 1, 2, \dots$ as

$$\alpha(1) = \operatorname{cor}(X_t, X_{t+1}) = \rho(1)$$

and

$$\begin{aligned} \alpha(h) &= \operatorname{cor}(X_t - \hat{X}_t, X_{t+h} - \hat{X}_{t+h}) \\ &= \rho_{X_t X_{t+h} | (X_{t+1}, \dots, X_{t+h-1})^\top}, \quad h = 2, 3, \dots \end{aligned}$$

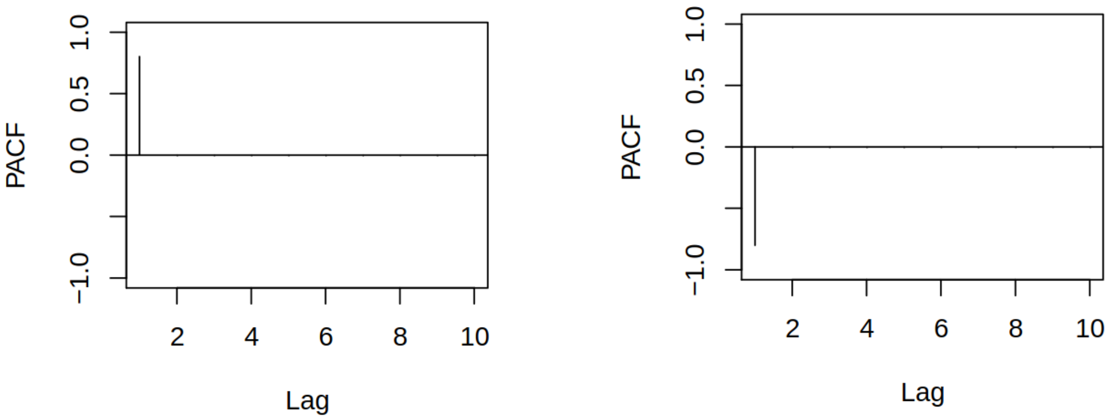
where $\hat{X}_t$ and $\hat{X}_{t+h}$ are the *best linear predictions* of X_t and X_{t+h} from $X_{t+1}, \dots, X_{t+h-1}$.

Let now X_t be a causal AR(p) process, $X_t = \varphi_1 X_{t-1} + \cdots + \varphi_p X_{t-p} + \varepsilon_t$. For $h > p$, compute the mean squared prediction error for prediction of X_{t+h} from $X_{t+1}, \dots, X_{t+h-1}$, which we want to minimize,

$$\begin{aligned} &\mathbb{E}\left(X_{t+h} - \beta_1 X_{t+h-1} - \cdots - \beta_{h-1} X_{t+1}\right)^2 \\ &= \mathbb{E}\left(\varepsilon_{t+h} + \sum_{j=1}^{p} (\varphi_j - \beta_j) X_{t+h-j} - \sum_{j=p+1}^{h-1} \beta_j X_{t+h-j}\right)^2 \\ &= \mathbb{E}\,\varepsilon_{t+h}^2 + \mathbb{E}\left(\sum_{j=1}^{p} (\varphi_j - \beta_j) X_{t+h-j} - \sum_{j=p+1}^{h-1} \beta_j X_{t+h-j}\right)^2 \geq \mathbb{E}\,\varepsilon_{t+h}^2 \end{aligned}$$

with the assumption that there is no correlation between ε_{t+h} and the past. The minimum is attained at $\beta_j = \varphi_j$ for $j = 1, \dots, p$ and $\beta_j = 0$ otherwise. So the BLP is $\hat{X}_{t+h} = \varphi_1 X_{t+h-1} + \cdots + \varphi_p X_{t+h-p}$. Hence for $h > p$, we get

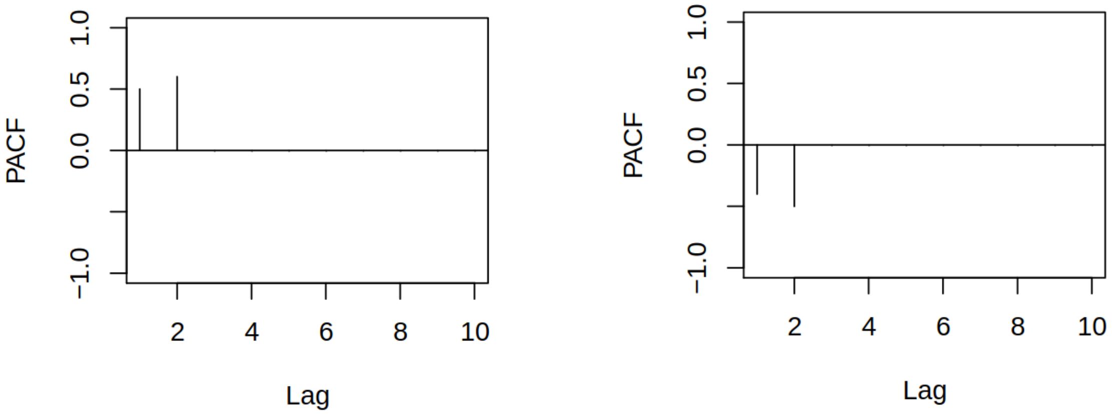
$$\alpha(h) = \operatorname{cor}(X_t - \hat{X}_t, X_{t+h} - \hat{X}_{t+h}) = \operatorname{cor}(\varepsilon_{t+h}, X_t - \hat{X}_t) = 0.$$



(a) $X_t = 0.8X_{t-1} + \varepsilon_t$

(b) $X_t = -0.8X_{t-1} + \varepsilon_t$

Figure 8.6: Examples of pacf for AR (1)



(a) $X_t = 0.2X_{t-1} + 0.6X_{t-2} + \varepsilon_t$

(b) $X_t = -0.6X_{t-1} - 0.5X_{t-2} + \varepsilon_t$

Figure 8.7: Examples of pacf for AR (2)

Let X_t be now a causal invertible ARMA (p, q) process

$$\Phi(\mathsf{B})X_t = \Theta(\mathsf{B})\varepsilon_t,$$

then by invertibility, we have the AR (∞) representation

$$X_t = -\sum_{j=1}^{\infty} \pi_j X_{t-j} + \varepsilon_t.$$

Thus no finite AR representation exists, hence the PACF does not cut off.

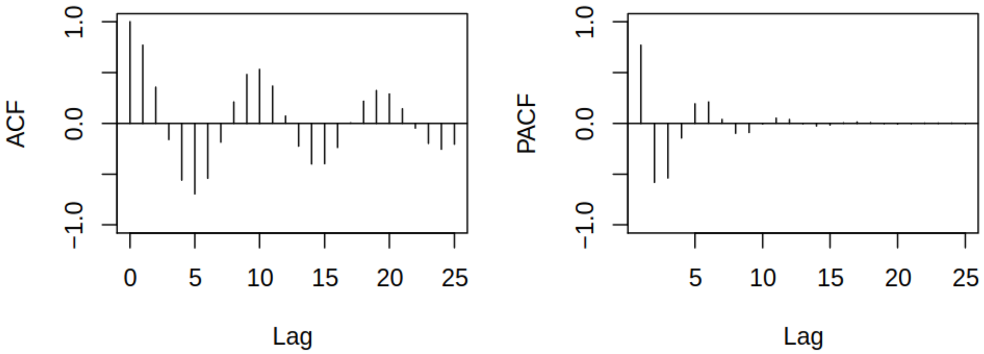


Figure 8.8: Examples of acf and pacf for ARMA process $X_t = 1.5X_{t-1} - 0.9X_{t-2} + \varepsilon_t - 0.7\varepsilon_{t-1} + 0.6\varepsilon_{t-2}$

We can also check the invertibility and the roots by:

```
polyroot(c(1,-1.5,.9)); abs(polyroot(c(1,-1.5,.9)))
```

```
[1] 0.8333333+0.6454972i 0.8333333-0.6454972i
```

```
[1] 1.054093 1.054093
```

```
polyroot(c(1,-.7,.6)); abs(polyroot(c(1,-.7,.6)))
```

```
[1] 0.583333+1.15169i 0.583333-1.15169i
```

```
[1] 1.290994 1.290994
```

Altogether, we get the following table Table 8.1.

Table 8.1: Behavior of ρ and α for all possible ARMA models

	ACF	PACF
AR (p)	Exponential decay	Cuts off after lag p
MA (q)	Cuts off after lag q	Exponential decay
ARMA (p,q)	Exponential decay	Exponential decay

9 Prediction of ARMA Processes

9.1 Linear prediction

Let our goal now be to predict future values X_{n+h} based on the data up to time n, that is $\{X_n, \dots, X_1\}$. Specifically, we want to

- predict the future value (point prediction);
- quantify uncertainty (prediction error, prediction intervals).

Thus we need to specify the criteria of quality of predictions, try to find good (optimal) predictions, focus on predictions that are easy to compute, and provide good algorithms.

9.1.1 Mean squared error criterion

Hence we want to predict Y given $Z_1, \dots, Z_n$ (mean zero variables with finite second moments). Given a measurable function $g : \mathbb{R}^n \to \mathbb{R}$ such that $g(\mathbf{Z})$ is on average close the unobserved value of Y, we shall minimize the mean squared error

$$\begin{aligned}\mathbb{E}\left(Y - g(\mathbf{Z})\right)^2 &= \mathbb{E}\left(Y - \mathbb{E}\left(Y \mid \mathbf{Z}\right)\right)^2 + \mathbb{E}\left(\mathbb{E}\left(Y \mid \mathbf{Z}\right) - g(\mathbf{Z})\right)^2 \\ &\quad + 2\,\mathbb{E}\left(\left(Y - \mathbb{E}\left(Y \mid \mathbf{Z}\right)\right)\left(\mathbb{E}\left(Y \mid \mathbf{Z}\right) - g(\mathbf{Z})\right)\right).\end{aligned}$$

Clearly, the last term is zero and $\mathbb{E}\left(Y - \mathbb{E}\left(Y \mid \mathbf{Z}\right)\right)^2 \geq 0$. Thus for all functions g

$$\mathbb{E}\left(Y - g(\mathbf{Z})\right)^2 \geq \mathbb{E}\left(Y - \mathbb{E}\left(Y \mid \mathbf{Z}\right)\right)^2$$

and the minimum is then attained for $g(\mathbf{Z}) = \mathbb{E}\left(Y \mid \mathbf{Z}\right)$, i.e. the best prediction is the conditional expectation. Unfortunately, this is often complicated, as it depends on the joint distribution, but in the Gaussian case, the conditional expectation is linear. As such, we shall focus on linear predictions.

9.2 Intermezzo

9.2.1 The Hilbert space setting for prediction

As we've mentioned earlier, we decided to focus on linear projections, which naturally work in the linear space of mean zero second-order variables. Now consider a Hilbert space with inner product $\langle \mathbf{X}, \mathbf{Y} \rangle = \mathbb{E}\,\mathbf{X}\mathbf{Y}$ and distance $\|\mathbf{X} - \mathbf{Y}\|^2 = \mathbb{E}\,(\mathbf{X} - \mathbf{Y})^2$. Now, given $\mathbf{Z}_1, \dots, \mathbf{Z}_n$, candidate linear predictions $c_1\mathbf{Z}_1 + \cdots + c_n\mathbf{Z}_n$ form the linear span of $\mathbf{Z}_1, \dots, \mathbf{Z}_n$. Unfortunately, we will also need to consider the infinite sets $\mathbf{Z}_1, \mathbf{Z}_2, \dots$ We define the linear span as the set of all **finite** linear combinations

$$\operatorname{lin}\{\mathbf{Z}_1, \mathbf{Z}_2, \dots\} = \left\{c_1\mathbf{Z}_{i_1} + \cdots + c_m\mathbf{Z}_{i_m} : m \in \mathbb{N}, c_1, \dots, c_m \in \mathbb{R}\right\}.$$

Surely, its closure consists of all limits (in the mean square) of convergent sequences and it is a closed subspace of the Hilbert space. It can be shown that the best approximation of an element of the Hilbert space by an element in a subspace is found by orthogonal projection.

Theorem 9.1. *Let M be a closed subspace of a Hilbert space H. Then every $y \in H$ can be uniquely decomposed as $y = \hat{y} + u$ where $\hat{y} \in M$ and u is orthogonal to M (i.e., $\langle u, z \rangle = 0$ for all $z \in M$). Furthermore,*

$$\|y - \hat{y}\| = \min_{z \in M} \|y - z\|$$

and

$$\|y\|^2 = \|\hat{y}\|^2 + \|u\|^2.$$

9.3 Linear prediction - cont.

Consider the linear space $\operatorname{lin}\{\mathbf{Z}_1, \dots, \mathbf{Z}_n\}$ (surely it has finite dimension and is automatically closed) and the linear predictions of form $\hat{\mathbf{Y}} = \sum_{j=1}^n c_j\mathbf{Z}_j$. Now our task is to find constants $c_1, \dots, c_n$ such that $\mathbb{E}\left(\mathbf{Y} - \hat{\mathbf{Y}}\right)^2$ is minimal, which means to find the orthogonal projection on the $\operatorname{lin}\{\mathbf{Z}_1, \dots, \mathbf{Z}_n\}$, which satisfies

$$\mathbf{Y} - \hat{\mathbf{Y}} \perp \operatorname{lin}\{\mathbf{Z}_1, \dots, \mathbf{Z}_n\}.$$

In other words,

$$\mathbb{E}\left(\mathbf{Z}(\mathbf{Y} - \mathbf{Z}^\top \boldsymbol{c})\right) = 0,$$

that is, $\operatorname{var}(\mathbf{Z})\boldsymbol{c} = \operatorname{cov}(\mathbf{Z}, \mathbf{Y})$ or in the case of invertibility $\boldsymbol{c} = \operatorname{var}(\mathbf{Z})^{-1}\operatorname{cov}(\mathbf{Z}, \mathbf{Y})$.

Tip

In the linear models, we had $(\mathbf{X}^\top\mathbf{X})$, which corresponds to $\operatorname{var}(\mathbf{Z})$, and $(\mathbf{X}^\top\mathbf{Y})$ in the role of $\operatorname{cov}(\mathbf{Z}, \mathbf{Y})$.

The prediction error is then given by

$$\mathbb{E}\left(\mathbf{Y} - \mathbf{Z}^\top \boldsymbol{c}\right)^2 = \operatorname{var}\mathbf{Y} - \operatorname{cov}(\mathbf{Y}, \mathbf{Z}) \operatorname{var}(\mathbf{Z})^{-1} \operatorname{cov}(\mathbf{Z}, \mathbf{Y}).$$

If $\operatorname{cov}(\mathbf{Y}, \mathbf{Z}) = 0$, the prediction error is $\operatorname{var}\mathbf{Y}$, which can be interpreted as *we have the same amount of uncertainty about* $\mathbf{Y}$ *as if we did not observe* $\mathbf{Z}$ *at all and as such* $\mathbf{Z}$ *provides no additional information about* $\mathbf{Y}$. Hence correlation is good for prediction (which is intuitively true). In regression with iid errors, prediction can be done only by extrapolating the fitted mean while in time series we can exploit the association (similarity, dissimilarity) between variables. Also notice that the prediction error is positive (unless $\mathbf{Y}$ is a linear function of $\mathbf{Z}$) and there will always be some uncertainty, and some additional randomness in $\mathbf{Y}$ that is not contained in $\mathbf{Z}$.

9.3.1 Time-series forecasting

Now, let's focus on zero-mean stationary time series with finite second moments based on $X_1, \dots, X_n$, we attempt to predict X_{n+1} by $\hat{X}_{n+1}$ (one step ahead) or X_{n+h} by $\hat{X}_{n+h}(n)$ (h steps ahead). Thus we look for $\hat{X}_{n+1} = \sum_{j=1}^{n} \varphi_{nj} X_{n+1-j}$. With stationarity, the estimating equation for $\hat{X}_{n+1}$ becomes

$$\boldsymbol{\Gamma}^{(n)} \boldsymbol{\varphi}^{(n)} = \gamma^{(n)},$$

where $\boldsymbol{\Gamma}^{(n)}$ is the $(n \times n)$ matrix with entries $\Gamma_{ij}^{(n)} = \gamma(i - j)$ (autocovariance), $\gamma^{(n)} = (\gamma(1), \dots, \gamma(n))^\top$ and $\boldsymbol{\varphi}^{(n)} = (\varphi_{n1}, \dots, \varphi_{nn})^\top$. The non-stationary case can be solved analogously. Also, a similar procedure could be performed for h-steps predictions ahead. Surely, if $\boldsymbol{\Gamma}^{(n)}$ is invertible, one can compute

$$\boldsymbol{\varphi}^{(n)} = \boldsymbol{\Gamma}^{(n)^{-1}} \gamma^{(n)}.$$

Theorem 9.2. *For a mean zero stationary* L^2 *sequence with* $\gamma(0) > 0$ *and* $\gamma(k) \to 0$ *as* $k \to \infty$*, the matrix* $\boldsymbol{\Gamma}^{(n)}$ *is non-singular for every* n.

9.4 Recursive algorithms for computing predictions

9.4.1 Difficulties in the computation of predictions

Realize that solving

$$\boldsymbol{\Gamma}^{(n)} \boldsymbol{\varphi}^{(n)} = \gamma^{(n)},$$

may be difficult. While, a special structure of $\boldsymbol{\Gamma}^{(n)}$ (e.g., banded for MA models) may help, it still would be no small feat. If the series is long (e.g., $n = 10000$), we need to solve a large linear system, i.e. invert a large matrix (problems with numerical errors, computing time, memory). Hence we would like to avoid inverting matrices. Also when a new observation arrives, the prediction needs to be recomputed from scratch, thus we would like to update previous predictions, without needing to recalculate everything. Combined, this suggests using recursive procedures that use simple operations to update previous results

9.4.2 Innovations

The linear prediction

$$\hat{X}_{n+1} = \sum_{j=1}^{n} \varphi_{nj} X_{n+1-j}$$

is an element of the linear span $\operatorname{lin}\{X_1, \dots, X_n\}$. The principal difficulty in solving $\boldsymbol{\varphi}^{(n)} = \boldsymbol{\Gamma}_n^{-1} \gamma^{(n)}$ is that a non-trivial matrix must be inverted. Thus we try to re-express (change the basis) the linear span to get a simpler matrix. Now notice that

$$\operatorname{lin}\{X_1, \dots, X_n\} = \operatorname{lin}\left\{X_1 - \hat{X}_1, \dots, X_n - \hat{X}_n\right\},$$

where we set $\hat{X}_1 = 0$, as there is no history to predict from. **Innovations** $X_k - \hat{X}_k$ are differences of the actual observation and its one-step-ahead prediction based on the past. So instead of looking for

$$\hat{X}_{n+1} = \sum_{j=1}^{n} \varphi_{nj} X_{n+1-j},$$

we equivalently search for

$$\hat{X}_{n+1} = \sum_{j=1}^{n} \theta_{nj} \left(X_{n+1-j} - \hat{X}_{n+h-j} \right).$$

Notice that the innovations are uncorrelated, i.e.

$$\mathbb{E} \left(X_j - \hat{X}_j \right) \left(X_k - \hat{X}_k \right) = 0 \tag{9.1}$$

for all $j < k$, because

- $X_j - \hat{X}_j \in \operatorname{lin}\left\{X_1, \dots X_j\right\}$;
- $X_k - \hat{X}_k \perp \operatorname{lin}\left\{X_1, \dots, X_j\right\}$ by orthogonality of the projection.

Hence $X_1 - \hat{X}_1, \dots, X_n - \hat{X}_n$ are truly uncorrelated, as we claimed, and orthogonal. From the fact, that projections on orthogonal elements are particularly easy, and the fact that the orthogonal projection of Y on $\mathbf{Z}^{(n)} = (Z_n, \dots, Z_1)^\top$ is given by

$$\hat{Y} = \sum_{j=1}^{n} \theta_{nj} Z_{n+j-j} = \boldsymbol{\theta}^{(n)^{-1}} \mathbf{Z}^{(n)},$$

where

$$\boldsymbol{\theta}^{(n)} = \operatorname{cov}\left(\mathbf{Z}^{(n)}\right)^{-1} \operatorname{cov}(\mathbf{Z}^{(n)}, Y).$$

Now if $\mathbf{Z}$ are orthogonal, then $\operatorname{cov}(\mathbf{Z}^{(n)})$ is diagonal. Thus it is easy to invert and

$$\theta_{nj} = \frac{\operatorname{cov}(Z_{n+1-j}, Y)}{\operatorname{var} Z_{n+1-j}}.$$

As innovations $Z_k = X_{k+1} - \hat{X}_{k+1}$, $k = 1, \dots, n$ are orthogonal, denote

$$v_k = \mathbb{E}\, Z_k^2 = \mathbb{E}\left(X_{k+1} - \hat{X}_{k+1}\right)^2. \tag{9.2}$$

The covariance matrix of the innovations is $\operatorname{diag}\{v_{n-1}, v_{n-2}, \dots, v_0\}$. Prediction coefficients then become

$$\theta_{n,n-k} = v_k^{-1}\, \mathbb{E}\, X_{n+1} Z_k = v_k^{-1}\, \mathbb{E}\, X_{n+1}(X_{k+1} - \hat{X}_{k+1})$$

for $k = 1, \dots, n$, then

$$\begin{aligned}\theta_{n,n-k} &= v_k^{-1}\, \mathbb{E}\left(X_{n+1}\left(X_{k+1} - \hat{X}_{k+1}\right)\right)\\ &= v_k^{-1}\left(\gamma(n+1, k+1) - \mathbb{E}\left(X_{n+1}\hat{X}_{k+1}\right)\right).\end{aligned}$$

To compute $\mathbb{E}\, X_{n+1}\hat{X}_{k+1}$ recall that

$$\hat{X}_{k+1} = \sum_{j=0}^{k-1} \theta_{k,k-j}(X_{j+1} - \hat{X}_{j+1}), \tag{9.3}$$

hence

$$\mathbb{E}\, X_{n+1}\hat{X}_{k+1} = \sum_{j=0}^{k-1} \theta_{k,k-j}\, \mathbb{E}\, X_{n+1}(X_{j+1} - \hat{X}_{j+1}),$$

where for $j < k$ we get by (9.1)

$$\begin{aligned}\mathbb{E}\left(X_{n+1}(X_{j+1} - \hat{X}_{j+1})\right) &= \mathbb{E}\left((X_{n+1} - \hat{X}_{j+1})(X_{j+1} - \hat{X}_{j+1})\right)\\ &\quad + \mathbb{E}\left(\hat{X}_{n+1}(X_{j+1} - \hat{X}_{j+1})\right)\\ &= \mathbb{E}\left(\sum_{h=0}^{n-1} \theta_{n,n-h}(X_{h+1} - \hat{X}_{h+1})(X_{j+1} - \hat{X}_{j+1})\right)\\ &= \theta_{n,n-j} v_j.\end{aligned}$$

Put together, we get

$$\theta_{n,n-k} = v_k^{-1}\left(\gamma(n+1, k+1) - \sum_{j=0}^{k-1} \theta_{k,k-j}\theta_{n,n-j} v_j\right), \tag{9.4}$$

and when we then compute $v_n = \mathbb{E}\left(X_{n+1} - \hat{X}_{n+1}\right)^2$, using orthogonality

$$\mathbb{E}\, X_{n+1}^2 = \mathbb{E}\left(X_{n+1} - \hat{X}_{n+1}\right)^2 + \mathbb{E}\, \hat{X}_{n+1}^2,$$

thus by Definition 2.2 and the assumption $\{X_t\}$ has mean zero, together with (9.3) and (9.4), it finally yields

$$\begin{aligned}v_n &= \mathbb{E}\, X_{n+1}^2 - \mathbb{E}\, \hat{X}_{n+1}^2\\ &= \gamma(n+1, n+1) - \mathbb{E}\left(\sum_{j=0}^{n-1} \theta_{n,n-j}(X_{j+1} - \hat{X}_{j+1})\right)^2\\ &= \gamma(n+1, n+1) - \sum_{j=0}^{n-1} \theta_{n,n-j}^2 v_j,\end{aligned}$$

which can be summarized in the following Theorem 9.3.

Theorem 9.3 (Innovations)**.** *If $\{X_t\}$ has mean zero and $\mathbb{E}\, X_i X_j = \gamma(i,j)$, where the matrix with entries $\gamma(i,j)$, $i,j = 1,\dots,n$ is non-singular for each $n = 1,2,\dots$, then the **one-step predictors** $\hat{X}_{n+1}$, $n \geq 0$, and their **mean squared errors** v_n, see (9.2), are given by*

$$\hat{X}_{n+1} = \begin{cases} 0, & n = 0 \\ \sum_{j=1}^{n} \theta_{n,j}(X_{n+1-j} - \hat{X}_{n+1-j}), & n \geq 1 \end{cases}$$

and

$$\begin{aligned} v_0 &= \gamma(1,1), \\ \theta_{n,n-k} &= v_k^{-1}\left(\gamma(n+1,k+1) - \sum_{j=0}^{k-1} \theta_{k,k-j}\theta_{n,n-j}v_j\right), \quad k = 0,\dots,n-1, \\ v_n &= \gamma(n+1,n+1) - \sum_{j=0}^{n-1} \theta_{n,n-j}^2 v_j. \end{aligned}$$

! Important

Thus the one-step prediction can be easily computed (fast) without any matrix operations (but only using basic operations).

Thus this algorithm relies on the recursive computation of $\theta_{n,j}$ and v_n using previously computed values, and it features no requirements of matrix inversion, just basic operations, and easy updating. Also, stationarity is not necessary. Essentially, this is the Gram–Schmidt orthogonalization procedure and it is useful in maximum likelihood estimation.

Tip

A similar (but better) algorithm is a Kalman filter, which we will discuss in the next course.

We can also simplify the innovations algorithm for MA (q) processes, as the coefficients $\theta_{n,q+1}, \theta_{n,q+2}, \dots$ are zero, because $\mathbb{E}\, X_{n+1}(X_{k+1} - \hat{X}_{k+1})$ for $n - k > q$

9.4.3 Recursive prediction for ARMA model

For MA (q), the *innovations algorithm* simplifies due to the simple autocorrelation structure. On the other hand, ARMA autocorrelation is complicated, but a question arises whether we could transform it into an MA model. Thus consider ARMA (p,q)

$$\Phi(\mathsf{B})\, X_t = \Theta(\mathsf{B})\varepsilon_t, \quad \{\varepsilon_t\} \sim \mathrm{WN}\left(0, \sigma^2\right).$$

For $m = \max(p,q)$, define the transformed process

$$\begin{aligned} W_t &= \sigma^{-1} X_t, \quad t = 1,\dots,m \\ W_t &= \sigma^{-1}\Phi(\mathsf{B})\, X_t, \quad t > m. \end{aligned}$$

The linear spans of $X_1, \dots, X_n$ and $W_1, \dots, W_n$ are the same, so now we can apply the algorithm to W_t to get

$$\hat{W}_{n+1} = \sum_{j=1}^{\min(p,q)} \theta_{nj}(W_{n+1-j} - \hat{W}_{n+1-j}),$$

where the coefficients θ_{nj} and mean square error r_n are found by the recursive procedures. So W_t is MA (q) and thus $\theta_{nj} = 0$ for $j > q$. Now we would like to obtain $\hat{X}_t$ from $\hat{W}_t$. We can project each side of the defining equation for W_t on $X_1, \dots, X_{t-1}$ to get

$$\begin{aligned} \hat{W}_t &= \sigma^{-1}\hat{X}_t, \quad t = 1, \dots, m, \\ \hat{W}_t &= \sigma^{-1}\left(\hat{X}_t - \varphi_1 X_{t-1} - \cdots - \varphi_p X_{t-p}\right), \quad t > m, \end{aligned}$$

and thus $X_t - \hat{X}_t = \sigma(W_t - \hat{W}_t)$, which gives

$$\begin{aligned} \hat{X}_{n+1} &= \sum_{j=1}^{n} \theta_{nj}(X_{n+1-j} - \hat{X}_{n+1-j}), \quad 1 \le n \le m \\ \hat{X}_{n+1} &= \varphi_1 X_{n-1} + \cdots + \varphi_p X_{n-p} + \sum_{j=1}^{n} \theta_{nj}(X_{n+1-j} - \hat{X}_{n+1-j}), \quad n \ge m \end{aligned}$$

and finally $v_n = \mathbb{E}\left(X_{n+1} - \hat{X}_{n+1}\right)^2 = \sigma^2 r_n$.

> Tip
>
> To summarize, we derived a way to make a one-step prediction for MA (q) and then used this in ARMA model, where the AR part is easy to predict.

9.4.4 h-step prediction

Given data $X_1, \dots, X_n$, we want to predict X_{n+h} using $X_1, \dots, X_n$ – via best linear prediction – which leads us to solve

$$\boldsymbol{\Gamma}\boldsymbol{\varphi}^{(h)} = \boldsymbol{\gamma}^{(h)}.$$

For this task, we can continue the iteration of the innovations algorithm to h-step prediction

$$\hat{X}_{n+h}(n) = \sum_{j=h}^{n+h-1} \theta_{n+h-1,j}(X_{n+h-j} - \hat{X}_{n+h-j}(n+h-j-1))$$

with squared error

$$v_{n+h-1} = \gamma(n+h, n+h) - \sum_{j=h-1}^{n+h-2} \theta^2_{n+h-1,n-j} v_j.$$

Values of $\boldsymbol{\theta}^{(n)}$ are finally obtained by continued iteration.

9.4.4.1 Limiting behavior of the h-step best linear prediction

Now let $h \to \infty$. What is then the behavior of the prediction and its error? We have that

$$\boldsymbol{\varphi}^{(h)} = \boldsymbol{\Gamma}^{-1}\gamma^{(h)},$$

and often $\gamma_{(h)}$ as $h \to \infty$ (e.g., for stationary ARMA). Hence $\varphi_{(h)} \to 0$ and $\hat{X}_{n+h}(n) \to 0$ as $h \to \infty$. This implies the convergence to the mean of the series. Under stationarity, the prediction error is

$$\gamma(0) - \boldsymbol{\gamma}^{(h)\top}\boldsymbol{\Gamma}^{-1}\gamma^{(h)} \to \gamma(0)$$

as $h \to \infty$.

> ⚠ Warning
>
> Thus long-term forecasts have approximately the same degree of uncertainty as if **no data were observed**.

9.4.5 Prediction with trend

Now consider a model with a deterministic trend

$$Y_t = \mu_t + X_t,$$

where X_t is stationary and, for example, $\mu_t = \beta_0 + \beta_1 t + \beta_2 s(t)$. Then we have data $Y_1, \dots, Y_n$ and our goal is to predict Y_{n+h} by $\hat{Y}_{n+h}(n)$, so we extrapolate the trend (which can be either known or estimated) and then predict by

$$\hat{Y}_{n+h}(n) = \mu_{n+h} + \hat{X}_{n+h}(n)$$

where $\hat{X}_{n+h}(n)$ is constructed as the h-step prediction from $X_t = Y_t - \mu_t$, $t = 1, \dots, n$.

9.4.6 Prediction errors and intervals

We can derive the following asymptotic behavior

$$\hat{Y}_{n+h}(n) - \mu_{n+h} = \hat{X}_{n+h}(n) \to 0, \quad h \to \infty$$

and keep in mind, that **in the limit** we predict by **extrapolating the trend**. Error is then given by

$$\hat{Y}_{n+h}(n) - Y_{n+h} = \hat{X}_{n+h}(n) - X_{n+h},$$

thus the *mean squared prediction error* is the same as for $\hat{X}_{n+h}(n)$ and it converges to the series variance $\gamma(0)$. The trend will be estimated with a parametric estimation error of order $n - 1$, comparatively smaller than the prediction error.

Note

Our prediction was simply constructed as a best linear prediction and as such we made **no assumptions** on the distribution of our data.

If the data are Gaussian, then

$$X_{n+h} \mid (X_1, \dots, X_n) \sim \mathcal{N}\left(\hat{X}_{n+h}(n), \mathbb{E}\left(X_{n+h} - \hat{X}_{n+h}(n)\right)^2\right),$$

where the variance is v_{n+h-1}, or equivalently $\gamma(0) - {\gamma^{(h)}}^\top \boldsymbol{\Gamma}^{-1}\gamma^{(h)}$. Gaussian prediction interval with coverage $1-\alpha$ is then given by

$$\left(\hat{Y}_{n+h}(n) - c_{1-\alpha/2}v_{n+h-1}^{1/2}, \hat{Y}_{n+h}(n) + c_{1-\alpha/2}v_{n+h-1}^{1/2}\right)$$

with $c_{1-\alpha/2}$ being the $(1-\alpha/2)$-quantile of $\mathcal{N}(0,1)$.

Tip

It is a good practice to use wider prediction intervals (80% or 90%) – using too strict intervals makes (a bit) less sense.

9.5 Prediction of ARMA processes with infinite past

Recall a AR (p) model given by

$$X_t = \varphi_1 X_{t-1} + \dots + \varphi_p X_{t-p} + \varepsilon_t$$

and we want to predict X_{n+1} based on $X_1, \dots, X_n$, $n \geq p$. Best linear prediction is then given trivially by the autoregressive formulation

$$\hat{X}_{n+1} = \varphi_1 X_n + \dots + \varphi_p X_{n-p+1}.$$

9.5.1 h-step prediction in AR process

Now we want to predict X_{n+h} given $X_1, \dots, X_n$, $n \geq p$. We can obtain the projection $\hat{X}_{n+h}(n) = \mathsf{P}_n X_{n+h}$ of X_{n+h} on $X_1, \dots, X_n$ by projecting first on $X_1, \dots, X_{n+h-1}$, then we take the projection $\mathsf{P}_{n+h-1}X_{n+h}$ and again project it on $X_1, \dots, X_{n+h-2}$, etc., and finally on $X_1, \dots, X_n$, from which we obtain

$$\hat{X}_{n+h}(n) = \mathsf{P}_n X_{n+h} = \dots = \mathsf{P}_n \mathsf{P}_{n+1} \dots \mathsf{P}_{n+h-1} X_{n+h}.$$

As an example, the two-steps-ahead best prediction is

$$\hat{X}_{n+2}(n) = \varphi_1 \hat{X}_{n+1} + \varphi_2 X_n + \dots + \varphi_p X_{n-p+2}$$

and analogously, the h-steps-ahead best prediction is

$$\hat{X}_{n+h}(n) = \varphi_1 \hat{X}_{n+h-1}(n) + \cdots + \varphi_p \hat{X}_{n+h-p},$$

where we set $\hat{X}_t = X_t$ for $t = 1, \ldots, n$. Hence we have a straightforward recursive computation.

9.5.2 Prediction with infinite past

Recall that prediction in AR (p) is autoregression on p preceding series values. Now also remember that for an invertible MA (q) process we may write

$$\varepsilon_t = \Theta(B)^{-1} X_t = \sum_{j=0}^{\infty} \pi_j X_{t-j},$$

which is an autoregression on the infinite past (the process must be invertible). Thus assume now the infinite past

$$X_n, X_{n-1}, \ldots, X_1, X_0, X_{-1}, \ldots$$

is available and investigate $\tilde{X}_{n+h}(n) = \tilde{P}_n X_{n+h}$ the projection of X_{n+h} on the infinite past. Consider a full ARMA (p, q) model

$$\Phi\left(B\right) X_t = \Theta\left(B\right) \varepsilon_t$$

and let it be **causal** and **invertible**, i.e. we can write

$$X_{n+h} = \sum_{j=0}^{\infty} \psi_j \varepsilon_{n+h-j}, \quad \varepsilon_{n+h} = \sum_{j=0}^{\infty} \pi_j X_{n+h-j}.$$

By projecting both sides in the AR (∞) presentation on the past we get

$$0 = \hat{X}_{n+h}(n) + \sum_{j=1}^{\infty} \pi \hat{X}_{n+h-j}(n),$$

therefore

$$\hat{X}_{n+h}(n) = -\sum_{j=1}^{h-1} \pi_j \hat{X}_{n+h-j} - \sum_{j=h}^{\infty} \pi_j X_{n+h-j},$$

which gives us recursive computation of the prediction. The mean squared prediction error is again $\mathbb{E}\left(X_{n+h} - \hat{X}_{n+h}(n)\right)^2$. Also, the past

$$X_n, X_{n-1}, \ldots, X_1, X_0, X_{-1}, \ldots$$

is equivalent to

$$\varepsilon_n, \varepsilon_{n-1}, \ldots, \varepsilon_1, \varepsilon_0, \varepsilon_{-1}, \ldots$$

as the *subspaces generated by both sequences coincide* due to **causality** and **invertibility**. By projecting both sides in the MA (∞) representation on the past we get (recall the orthogonality)

$$\hat{X}_{n+h}(n) = \sum_{j=0}^{\infty} \psi_j \tilde{P}_n \varepsilon_{n+h-j} = \sum_{j=h}^{\infty} \psi_j \varepsilon_{n+h-j}.$$

From this, we may obtain that the mean square prediction error is

$$\mathbb{E}\left(\mathrm{X}_{n+h} - \hat{\mathrm{X}}_{n+h}(n)\right)^2 = \mathbb{E}\left(\sum_{j=0}^{h-1} \psi_j \varepsilon_{n+h-j}\right) = \sigma^2 \sum_{j=0}^{h-1} \psi_j^2.$$

As we consider the prediction horizon $h \to \infty$, from

$$\hat{\mathrm{X}}_{n+h}(n) = \sum_{j=h}^{\infty} \psi_j \varepsilon_{n+h-j}$$

and the fact that the ψ-weights decay exponentially, we get that $\hat{\mathrm{X}}_{n+h}(n)$ quickly converges to the mean (in the L^2 sense). As for the MSPE, we see that

$$\mathbb{E}\left(\mathrm{X}_{n+h} - \hat{\mathrm{X}}_{n+h}(n)\right)^2 = \sigma^2 \sum_{j=0}^{h-1} \psi_j^2 \to \sigma^2 \sum_{j=0}^{\infty} \psi_j^2 = \gamma(0).$$

Caution

Hence the mean squared prediction error quickly converges to the variance of the series. This implies that short-term predictions are good, but long-term predictions behave like without any observed data.

9.5.3 Truncated prediction in ARMA models with finite past

Lastly, consider finite observed past $\mathrm{X}_1, \ldots, \mathrm{X}_n$ instead of infinite past and assume n sufficiently large. We use the same relation as with infinite past

$$\hat{\mathrm{X}}_{n+h}(n) = -\sum_{j=1}^{h-1} \pi_j \hat{\mathrm{X}}_{n+h-j} - \sum_{j=h}^{\infty} \pi_k \mathrm{X}_{n+h-j},$$

but set $\mathrm{X}_t = 0$ for $t \leq 0$ (also, their weights are small if n is large), that is we use the **truncated forecast**

$$\hat{\mathrm{X}}_{n+h}(n) = -\sum_{j=1}^{h-1} \pi_j \hat{\mathrm{X}}_{n+h-j} - \sum_{j=h}^{n+h-1} \pi_k \mathrm{X}_{n+h-j},$$

which again leads to recursive computation.

10 ARIMA, Seasonal ARMA and SARIMA Models

10.1 ARIMA models

One ARIMA model that we can see is the random walk AR (1), i.e. $X_t = \varphi_1 X_{t-1} + \varepsilon_t$, with $\varphi_1 = 1$, thus

$$X_t = X_{t-1} + \varepsilon_t = \sum_{j=1}^{t} \varepsilon_t, \quad t = 1, 2, \ldots$$

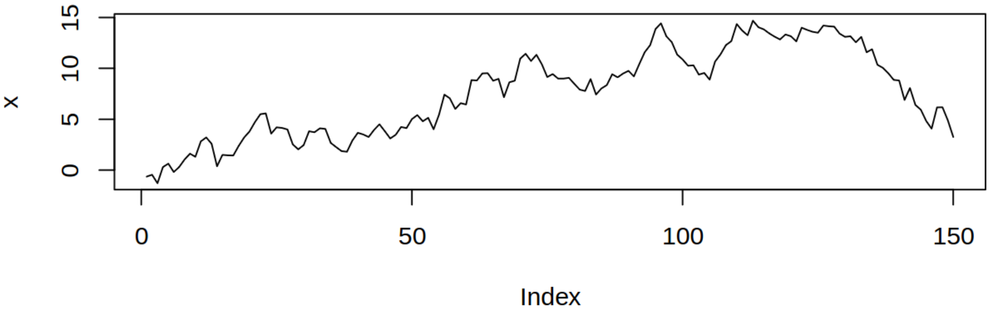


Figure 10.1: Realization of random walk

What's more, we can compare different AR models and notice the differences in their scales, stationarity vs increasing variance and stable behavior vs temporary, transient trends.

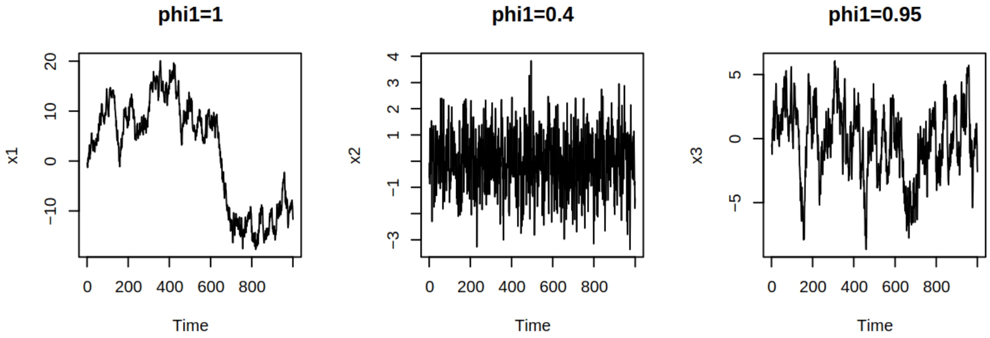


Figure 10.2: Different AR models (some stationary and some not)

As we have discussed before, the correlogram of the random hints (falsely) at a complicated autocorrelation structure (remember that it is misused in this case).

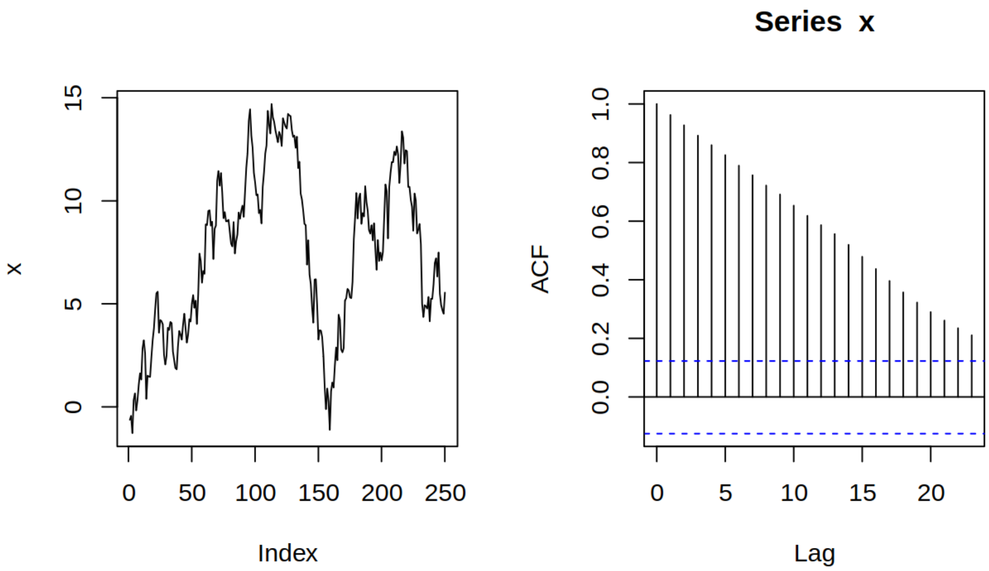


Figure 10.3: ACF of random walk

We can see the same behavior in global temperature data, which features similar temporary stochastic trends and even the acf is a look-alike.

```
data(gtemp,package="astsa")
par(mfrow=c(1,2))
plot(gtemp,ylab="Global temperature")
acf(gtemp)
```

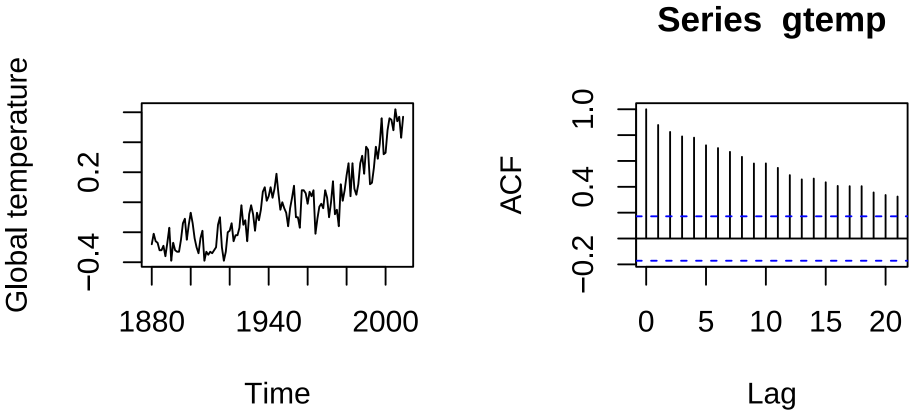


Figure 10.4: Global temperature data

10.1.1 Non-stationarity of random walk

Recall that the random walk

$$X_t = X_{t-1} + \varepsilon_t = \sum_{j=1}^{t} \varepsilon_t, \quad t = 1, 2, \ldots$$

has mean $\mathbb{E}\, X_t = \mathbb{E} \sum_{j=1}^{n} \varepsilon_j = 0$ and

$$\gamma(t,t) = \operatorname{var}\left(\sum_{j=1}^{t} \varepsilon_j\right) = t\sigma^2 \tag{10.1}$$

$$\gamma(s,t) = \mathbb{E}\left(\sum_{j=1}^{s} \varepsilon_j \sum_{j=1}^{t} \varepsilon_j\right) = \min(s,t)\sigma^2. \tag{10.2}$$

Notice, that the random walk is not stationary but the difference

$$X_t - X_{t_1} = \varepsilon_t$$

is a white noise sequence, hence it **is** stationary. This suggests that differencing could lead to stationarity.

10.1.2 Differencing time series

Let us define the **first difference** as

$$\Delta X_t = X_t - X_{t-1} = X_t - \mathsf{B}X_t = (1 - \mathsf{B})X_t$$

and d-th difference

$$\Delta^d X_t = \Delta \dots \Delta X_t = (1 - \mathsf{B})^d X_t.$$

One can realize that differencing

- *preserves stationarity*: if X_t is stationary, then $(1-\mathsf{B})^d X_t$ is also stationary;
- *introduces stationarity*:
  - with deterministic trends: if μ_t is a polynomial of order d and X_t is stationary, then $(1-\mathsf{B})^d(\mu_t + X_t)$ is stationary (with constant mean);
  - with stochastic trend: random walk becomes stationary after differencing.

Now although differencing may lead to stationarity, it can also give rise to complicated covariance structure if overused. One should consider the *principle of parsimony* which says that *models should be as simple as possible* (but not any simpler).

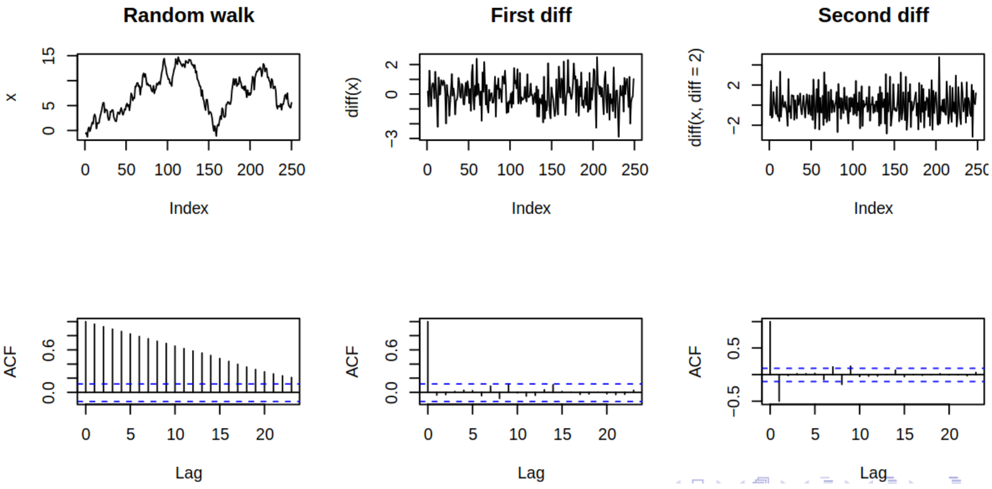


Figure 10.5: The effect of differencing

10.1.3 ARIMA model

Definition 10.1. A process X_t is said to be ARIMA (p, d, q) (**integrated ARMA**) if $\Delta^d X_t = (1 - \mathsf{B})^d X_t$ is ARMA (p, q).

We shall again use the notation

$$\Phi(\mathsf{B})\,(1-\mathsf{B})^d X_t = \Theta(\mathsf{B})\varepsilon_t,$$

then one can realize that random walk is ARIMA $(0,1,0)$, or I(1). If the underlying model is ARMA plus mean (instead of just ARMA), i.e., $\mathbb{E}\left(\Delta^d X_t\right) = \mu$, then $\Phi\mathsf{B}(\Delta^d X_t - \mu) = \Theta(\mathsf{B})\,\varepsilon_t$, hence

$$\Phi(\mathsf{B})\,(1-\mathsf{B})^d X_t = \delta + \Theta(\mathsf{B})\varepsilon_t,$$

where $\delta = \mu(1-\varphi_1-\cdots-\varphi_p)$. Also, in this case $\Delta^d X_t - \mu = \Delta^d(X_t - \mu t^d/d!)$, i.e., X_t has a polynomial trend. Now consider the following example showcasing the first difference on global temperature data.

```r
require(forecast)
```

```
Loading required package: forecast
```

```
Registered S3 method overwritten by 'quantmod':
  method            from
  as.zoo.data.frame zoo
```

```r
ggtsdisplay(diff(gtemp))
```

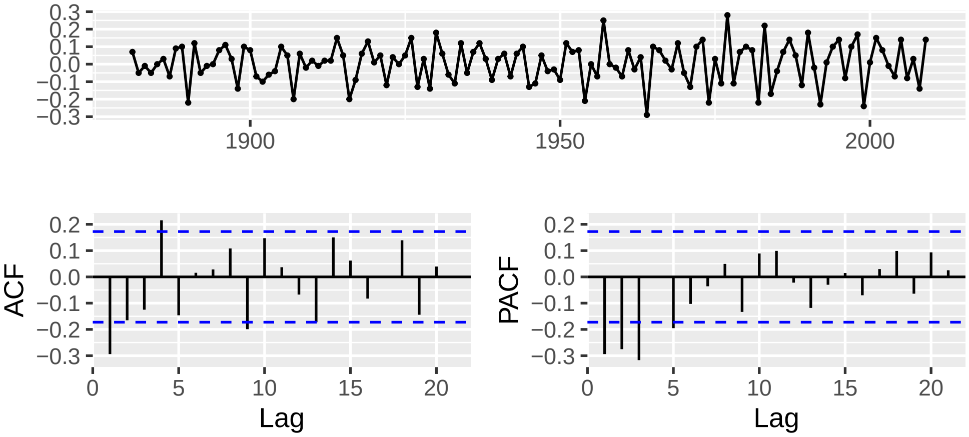


Figure 10.6: Stationary-looking difference of global temperatures

10.1.4 Estimating ARIMA models

Clearly, after differencing d times, we obtain ARMA (p,q) from ARIMA (p,d,q). Then we can estimate the coefficients, mean and white noise variance of the differenced process by methods for ARMA models (maximum likelihood estimation, conditional least squares etc.). Also, we do not need to center the ARIMA process, as differencing will remove additive constants, see

$$\Delta(X_t - a) = \Delta X_t.$$

Let now $Y_t = \Delta^d X_t$ be an ARMA model. We can then obtain forecasts of Y_t by methods for ARMA processes. Forecasts of X_t can then be obtained by "anti-differencing" – consider $d = 1$, then

- we have observed data until time n;
- we predict Y_{n+1} by $\hat{Y}_{n+1}$;
- since $X_{n+1} = X_n + Y_{n+1}$, predict X_{n+1} by $\hat{X}_{n+1} = X_n + \hat{Y}_{n+1}$;
- Predict Y_{n+2} by $\hat{Y}_{n+2}(n)$;
- Since $X_{n+2} = X_{n+1} + Y_{n+2}$, predict X_{n+2} by
$$\hat{X}_{n+2}(n) = \hat{X}_{n+1} + \hat{Y}_{n+2}(n) = X_n + \hat{Y}_{n+1} + \hat{Y}_{n+2}(n);$$
- etc.

Example 10.1 (Random walk with drift)**.** Consider random walk

$$X_t = \delta + X_{t-1} + \varepsilon_t, \quad \text{i.e.} \quad Y_t = \Delta X_t = \delta + \varepsilon_t$$

with $X_0 = 0$. We shall also have observed data $X_1, \dots, X_n$ and compute one-step-ahead forecast (which is a projection on $\operatorname{lin}\{1, X_1, \dots, X_n\}$)

$$\hat{X}_{n+1} = \mathsf{P}_n X_{n+1} = \mathsf{P}_n(\delta + X_n + \varepsilon_{n+1}) = \delta + X_n.$$

Two-step-ahead forecast has form $\hat{X}_{n+2}(n) = \delta + \hat{X}_{n+1} = 2\delta + X_n$ and similarly m-step-ahead forecast goes like $\hat{X}_{n+m}(n) = m\delta + X_n$. Compared with stationary AR (1), $X_t = \delta + \varphi_1 X_{t-1} + \varepsilon_t$ with $|\varphi_1| < 1$, we see (by successively projecting on the past until $n+m-1, n+m-2, \dots, n$ and using $\mu = \delta/(1-\varphi_1)$) that

$$\hat{X}_{n+m}(n) = \delta \frac{1-\varphi_1^m}{1-\varphi_1} + \varphi_1^m X_m = \mu + \varphi_1^m (X_n - \mu),$$

from which we get different asymptotic behavior (as $m \to \infty$):

- straight line vs constant;
- persistent vs vanishing effect on X_n.

Now to obtain prediction errors, recall that

$$X_n = n\delta + \sum_{j=1}^{n} \varepsilon_j \quad \& \quad X_{n+m} = m\delta + X_n + \sum_{j=n+1}^{n+m} \varepsilon_j$$

and the prediction error is

$$\mathbb{E}\left(X_{n+m} - \hat{X}_{n+m}(n)\right)^2 = \mathbb{E}\left(\sum_{j=n+1}^{n+m} \varepsilon_j\right)^2 = m\sigma^2.$$

Also remember that for a stationary AR (1) model, we have seen that

$$\mathbb{E}\left(X_{n+m} - \hat{X}_{n+m}(n)\right)^2 = \sigma^2 \frac{1-\varphi_1^{2m}}{1-\varphi_1^2},$$

which also shows different asymptotic behavior (as $m \to \infty$) of growing (or diverging) vs converging prediction errors.

Example 10.2 (IMA$(1,1)$ and exponential smoothing)**.** Consider an ARIMA $(0,1,1)$, or IMA$(1,1)$ model,

$$X_t = X_{t-1} + \varepsilon_t - \lambda\varepsilon_{t-1}$$

with $|\lambda| < 1$, for $t = 1, 2, \dots$ and $X_0 = 0$. Now

$$X_t - X_{t-1} = \varepsilon_t - \lambda\varepsilon_{t-1}$$

has an **invertible** representation $\sum_{j=0}^{\infty} \lambda^j \mathsf{B}^j (X_t - X_{t-1}) = \varepsilon_t$. Then for large t we can approximate (set $X_t = 0, t \leq 0$)

$$X_t = \sum_{j=1}^{\infty} (1-\lambda)\lambda^{j-1} X_{t-j} + \varepsilon_t.$$

Now approximate one-step prediction by

$$\tilde{X}_{n+1} = \sum_{j=1}^{\infty} (1-\lambda)\lambda^{j-1} X_{n+1-j} = (1-\lambda)X_n + \lambda\tilde{X}_n.$$

Thus the forecast is a linear combination of the old forecast and the new observation, which is also called *exponentially weighted moving average* (*EWMA*).

Tip

This is then equivalent to Holt's linear method, as we've seen before.

10.2 Seasonal ARMA models

Seasons are important in applications: economics, environment, etc. We have so far seen

- Classical decomposition;
- Seasonal indicators, harmonic functions;
- STL;
- Holt–Winters,

but often, there is a strong dependence on the past at multiples of some seasonal lag s (e.g., $s = 12$ months). At the same time, there is some variation (beyond fixed seasonal effects) Hence our goal becomes to incorporate this into ARMA models

One can realize that seasonal lags correspond to powers of the backshift operator B^{js}, e.g. $\mathsf{B}^{12}X_t = X_{t-12}$ is the value one year ago for monthly time series. We can then define the **seasonal autoregressive operator**

$$\Phi^*(\mathsf{B}^s) = 1 - \varphi_1^*\mathsf{B}^s - \varphi_2^*\mathsf{B}^{2s} - \cdots - \varphi_P^*\mathsf{B}^{Ps}$$

and **seasonal moving average operator**

$$\Theta^*(\mathsf{B}^s) = 1 + \theta_1^*\mathsf{B}^s + \cdots + \theta_Q^*\mathsf{B}^{Qs},$$

from which we obtain **pure seasonal ARMA models** by combining

$$\Phi^*(\mathsf{B}^s)X_t = \Theta^*(\mathsf{B}^s)\varepsilon_t$$

Note that pure seasonal models only include powers of B^s, but the dependence at other than seasonal lags is also needed (e.g., previous month). Thus we combine both seasonal and non-seasonal (typically shorter lag) dependence to get **multiplicative seasonal ARMA model**

$$\Phi(\mathsf{B})\Phi^*(\mathsf{B}^s)X_t = \Theta(\mathsf{B})\Theta^*(\mathsf{B}^s)\varepsilon_t.$$

As an example, we can take a look at ARMA $(0,1)\ (0,1)_{12}$

$$X_t = 0.8X_{t-12} + \varepsilon_t - 0.5\varepsilon_{t-1}.$$

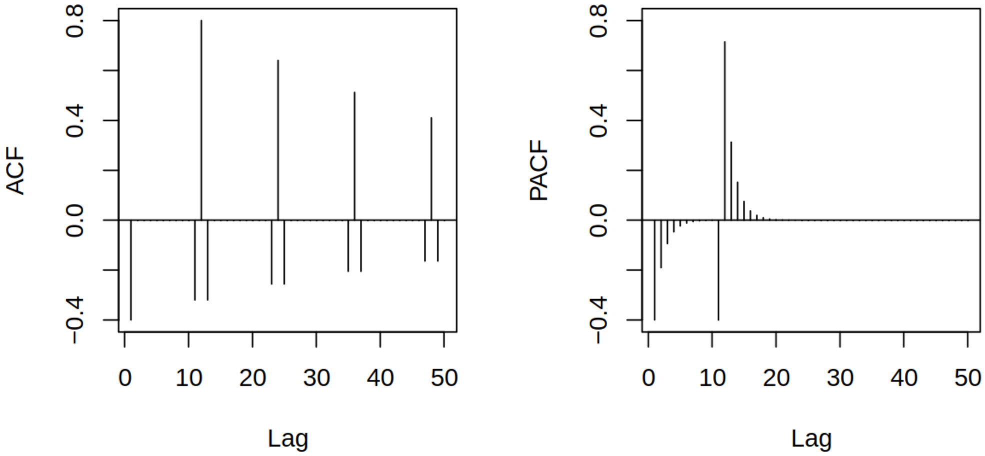


Figure 10.7: Seasonal ARMA model ARMA $(0,1)\ (0,1)_{12}$

10.2.1 Estimation and prediction

Clearly, the seasonal ARMA model is an ARMA model with special AR and MA polynomials, which puts constraints on the coefficients (thus reduces free parameters), e.g.

$$(1-\varphi_1^*\mathsf{B}^{12})(1-\varphi_1\mathsf{B})X_t = \varepsilon_t$$

is an AR (13) model but has only two parameters. Estimation procedures are then similar to non-seasonal ARMA (MLE etc.) but they exploit these parameter constraints (which gives us fewer parameters to estimate). Lastly, prediction is similar as with the usual ARMA model.

10.3 Seasonal ARIMA models

Recall the CO_2 series, see Example 5.2.

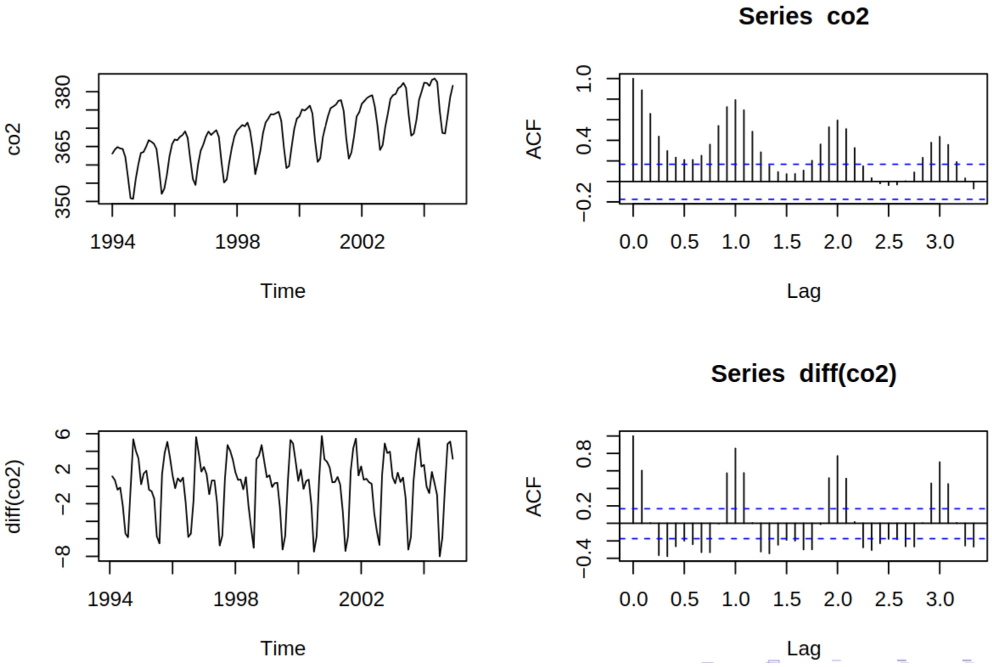


Figure 10.8: Time series of CO_2 concentration

Now by multiple differencing, we get Figure 10.9, where we can see that while autocorrelation ρ at seasonal lags decreases slowly, with the seasonal differencing we get a fast decrease.

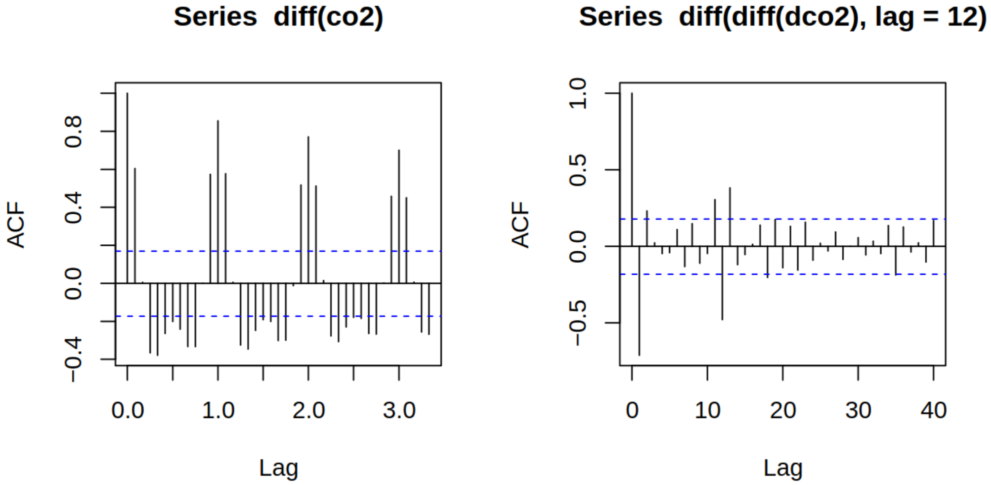


Figure 10.9: Seasonal differencing of CO_2 series

10.3.1 Multiplicative seasonal ARIMA models

Seasonal differencing is defined by

$$(1-\mathsf{B}^s)X_t = X_t - X_{t-s},$$

e.g. $(1-\mathsf{B}^{12})X_t = X_t - X_{t-12}$ (beware, it is **not** $(1-\mathsf{B})^{12} = \Delta^{12}$). Higher-order seasonal differences are then $(1-\mathsf{B}^s)^D X_t$. Now we can combine seasonal differencing, ordinary differencing, seasonal and ordinary AR and MA modeling.

Definition 10.2 (SARIMA model)**.** The multiplicative seasonal autoregressive integrated moving average (SARIMA) model is given by

$$\Phi^*(\mathsf{B}^s)\Phi(\mathsf{B})(1-\mathsf{B}^s)^D(1-\mathsf{B})^d X_t = \delta + \Theta^*(\mathsf{B}^s)\Theta(\mathsf{B})\varepsilon_t,$$

where ε_t is white noise and $\Phi, \Phi^*, \Theta, \Theta^*$ are polynomials of order p, P, q, Q respectively, which we denote by ARIMA (p, d, q) $(P, D, Q)_s$.

Example 10.3 (ARIMA $(0,1,1)$ $(0,1,1)_{12}$)**.** Consider a model given by

$$(1-\mathsf{B}^{12})(1-\mathsf{B})X_t = \Theta^*(\mathsf{B}^{12})\Theta(\mathsf{B})\varepsilon_t,$$

where $\Theta^*(\mathsf{B}^{12}) = 1 + \theta_1^*\mathsf{B}^{12}$ and $\Theta\mathsf{B} = 1 + \theta_1\mathsf{B}$. By expanding and rearranging we get

$$X_t = X_{t-1} + X_{t-12} - X_{t-13} + \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_1^*\varepsilon_{t-12} + \theta_1\theta_1^*\varepsilon_{t-13}.$$

Due to the multiplicative nature of the model, the coefficient at ε_{t-13} is $\theta_1\theta_1^*$ rather than a free parameter.

10.3.2 Estimation and forecasting or SARIMA models

First, one should difference the series according to the model (take d ordinary and D seasonal differences), then estimate the corresponding SARMA model (i.e., ARMA with a special form of the polynomials) and lastly forecast the differenced series and anti-difference it.

10.4 Further reading

10.4.1 Long seasonal periods

SAR(I)MA models are suitable for short periods, e.g., 12 (or 4) for monthly (or quarterly) data within years, 7 for daily data within weeks, 24 for hourly data within days etc. Long periods, e.g., 365 for daily data within years, are not meaningful practically and not supported by standard software. In such cases, one should use harmonic (or other periodic) functions instead, see forecasting with long seasonal periods.

10.4.2 Constants, trends,... in integrated models

In non-differenced models, one can include the intercept (constant vertical shift of the series). In the case of first-order differencing, the vertical shift is not estimable (differencing removes it) but the slope of the linear trend is. Similarly, under differencing of order d, the vertical shift of the series by a polynomial of degree $< d$ is not estimable; shift by t^d is estimable but not permitted for $d > 1$ by `forecast::Arima` (dangerous extrapolation), for more information see constants and ARIMA models in R.

Also, the regression matrix in `xreg` (for regression plus AR(I)MA errors) must not be collinear with the vertical shift and must not contain non-estimable terms (those whose d-th difference is zero)

11 SARIMA Model Building and Diagnostics

11.1 Model building and selection

The richness of the class of (S)AR(I)MA models is a double-edged sword – they are flexible but also difficult to use:

- ARMA has two orders: AR order p, MA order q;
- ARIMA has one more order: order of difference d;
- seasonal ARMA has another two orders: seasonal AR order P, seasonal MA order Q;
- SARIMA has one more order: order of seasonal difference D.

Now we need sensible model-building strategies and model adequacy criteria. Thus we will honor the *principle of parsimony*: models should be as simple as possible but not simpler.

> *"All models are wrong but some are useful"*

Also, recall that over-fitting and/or over-differencing leads to overly complicated models and increased variance of estimates. Thus we should first plot the data. Then we may transform it to make the variance constant if necessary and also possibly remove deterministic components (trends, seasonality) if appropriate. Now we need to model the stochastic component, which can be achieved with

- differencing appropriately (until stationary, not too much, use unit root tests if uncertain);
- starting with a simple model (white noise);
- performing diagnostics of residuals (acf, pacf, Box test);
- identifying orders and modifying the model (increase model complexity) to remedy;
- continuing until approximately white noise;
- possibly checking the model by slightly overfitting.

Last, but not least, we use the model for prediction. But a question might arise whether we should differentiate or not:

- if the series is non-stationary, differencing may make it stationary;
- if the series is stationary, differencing will preserve stationarity but complicate autocorrelation (non-invertibility).

Recall that the random walk is an AR sequence with $\Phi(z) = 1 - z$, hence Φ has a unit root ($\Phi(1) = 0$). The presence of a unit root among the roots of the AR polynomial implies non-stationarity and thus the need to differentiate. Methods exist for hypothesis testing about unit roots with hypotheses:

- null hypothesis H_0: 1 is a root;
- alternative H_1: all roots our outside the unit circle.

One can use, for example, the Dickey-Fuller test or Phillips-Perron test, which use non-standard distributions of the test statistics (non-stationary data under the null hypothesis) and in `R` they can be used with `PP.test`, or `adf.test` and `pp.test` in the package `tseries.`

11.2 Modal diagnostics

For model diagnostics, we consider the standardized residuals

$$\hat{e}_t = \frac{X_t - \hat{X}_t(\hat{\beta})}{v_{t-1}(\hat{\beta})^{1/2}},$$

where

- $\hat{\beta}$ is an estimator (e.g. maximum likelihood) of all model parameters (AR, MA, coefficients, white noise variance, possibly mean and covariates);
- $v_{t-1}(\hat{\beta})$ is the prediction error for X_t from $X_1, \dots, X_{t-1}$.

If the right model is used, then for large n the residuals are approximately white noise, so we can plot ACF to check for no correlation or use portmanteau tests (Ljung–Box test) to test the significance of autocorrelations up to some lag (in `R` use function `tsdiag`). Also, check for approximate normality of the residuals using the QQ plot, histogram,... (needed for prediction intervals).

11.3 Model order selection

We need an objective, which would quantitatively measure model adequacy because surely bigger models provide a better fit – likelihood or least squares are always better in more complex models. However, there is a price to pay as a more complex model increases the variance of estimates. Also, we need to find a good compromise between fit and variance and penalize for complexity – that is to maximize

$$\text{(model fit)} - \text{(model complexity penalty)},$$

or minimize

$$\text{(model error)} + \text{(model complexity penalty)}.$$

To perform this optimization, we search in a set of candidate models to optimize the selection criterion. Now it is needed to choose selection criteria – one of the most widespread is the **Akaike's Information Criterion** (AIC), which is defined as

$$\text{AIC} = -2\ell(\hat{\beta}) + 2r,$$

where

- $\hat{\beta}$ denotes collectively all parameters (AR, MA coefficients, white noise variance, possibly mean and regression coefficients);
- $r = \dim(\beta)$ is the number of parameters (e.g. $r = 1 + 1 + p + q$ for an ARMA (p, q) model with mean);
- $\ell(\beta) = \log f(X_1, \dots, X_n; \beta)$ is the log-likelihood.

Our goal is then to select the model that has the smallest value of AIC (within a set of candidate models, e.g., with orders up to some limits).

🔥 Caution

Note that we use selection criteria for models estimated from the same data, in particular, do not compare models with different orders of differencing!

The use of pure AIC is discouraged as it does not penalize large models enough. Hence the corrected AIC, denoted as AIC$_c$, is recommended

$$\text{AIC}_c = -2\ell(\hat{\beta}) + 2r\frac{n}{n-r-1}.$$

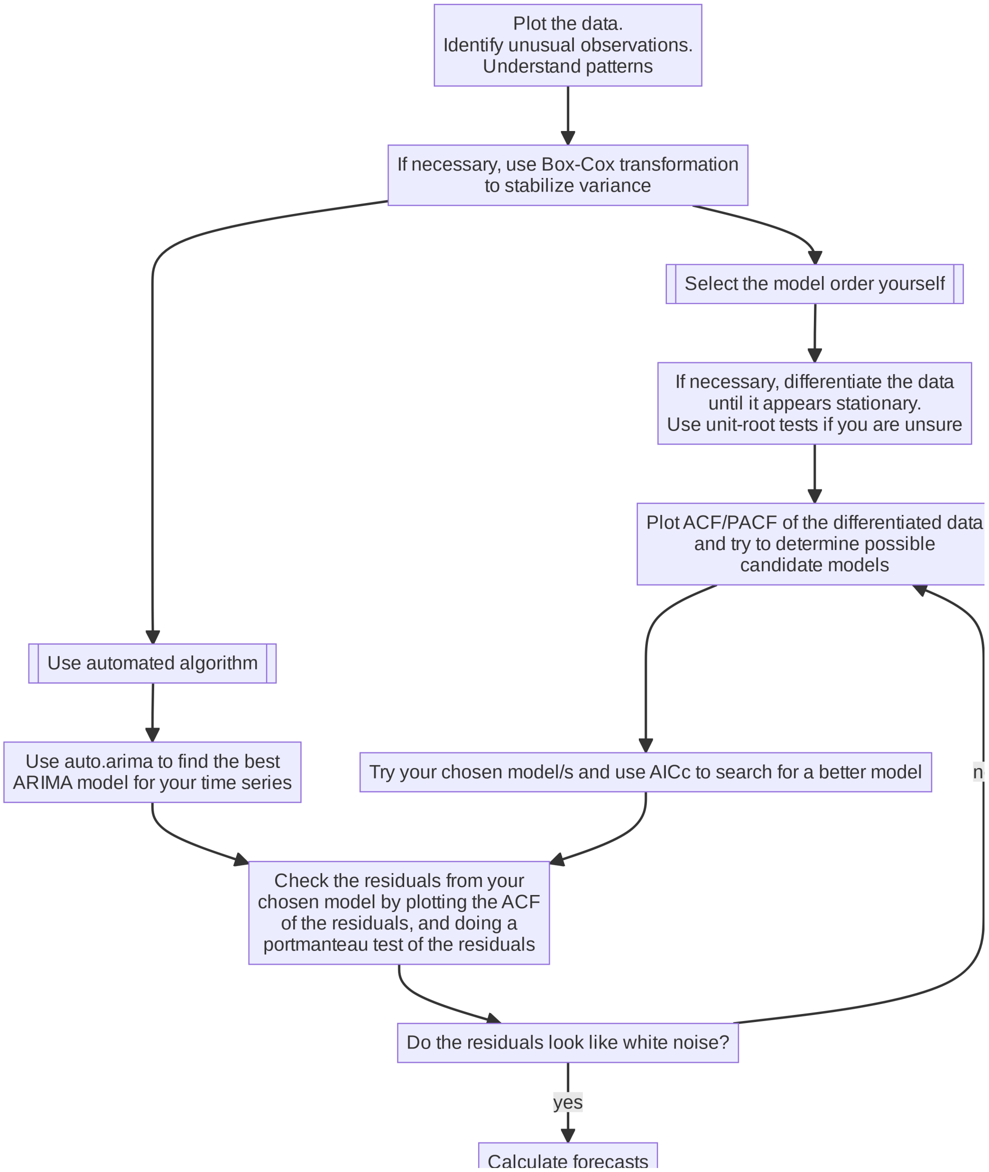
Also one can use BIC (*Bayesian Information Criterion*, or Schwarz's selection rule), which is defined as

$$\text{BIC} = -2\ell(\hat{\beta}) + r\log n,$$

which again strictly penalizes large models and as such tends to select smaller models.

💡 Tip

As a rule of thumb, AIC or AIC$_c$ is recommended to use for **forecasting**, whereas BIC is better used for **model estimation**.


Plot the data.
Identify unusual observations.
Understand patterns
If necessary, use Box-Cox transformation to stabilize variance
Select the model order yourself
If necessary, differentiate the data until it appears stationary.
Use unit-root tests if you are unsure
Plot ACF/PACF of the differentiated data and try to determine possible candidate models
Use automated algorithm
Use auto.arima to find the best ARIMA model for your time series
Try your chosen model/s and use AICc to search for a better model
n
Check the residuals from your chosen model by plotting the ACF of the residuals, and doing a portmanteau test of the residuals
Do the residuals look like white noise?
yes
Calculate forecasts

12 Basic Parameter Estimators For ARMA Models

12.1 Method of moments estimation

Consider an ARMA (p, q) model

$$(X_t - \mu) - \varphi_1(X_{t-1} - \mu) - \cdots - \varphi_p(X_{t-p} - \mu) = \varepsilon_t + \theta_1\varepsilon_{t-1} + \cdots + \theta_+\varepsilon_{t-q},$$

that is

$$\Phi\left(\mathsf{B}\right)\left(X_t - \mu\right) = \Theta(\mathsf{B})\varepsilon_t,$$

where $\{\varepsilon_t\} \sim \text{WN}\left(0, \sigma^2\right)$. Let our goal be to estimate the unknown parameters $\varphi_1, \ldots, \varphi_p, \theta_1, \ldots, \theta_q, \mu, \sigma^2$ and we assume the order p, q are known. Realize that the mean μ is easy to estimate from the data

$$\hat{\mu} = \overline{\mathbf{X}}.$$

Also, see that covariance function γ is also easy to estimate from data

$$\hat{\gamma}(h) = \frac{1}{n-h-1}\sum_{i=1}^{n-h}(X_i - \overline{\mathbf{X}})(X_{i+h} - \overline{\mathbf{X}}).$$

Now we can try to use it for the estimation of the parameters of the model. With the method of moments, we try to choose parameters for which the theoretical moments are equal to the empirical moments.

12.1.1 Yule-Walker estimation for AR model

Consider an AR (p) model with mean 0,

$$\Phi(\mathsf{B})X_t = \varepsilon_t, \quad \text{i.e.} \quad X_t - \varphi_1 X_{t-1} - \cdots - \varphi_p X_{t-p} = \varepsilon_t.$$

Now for $h = 1, \ldots, p$, we get

$$\begin{aligned}\gamma(h) &= \mathbb{E}\left(X_{t-h}X_t\right)\\ &= \mathbb{E}\left(X_{t-h}(\varphi_1 X_{t-1} + \cdots + \varphi_p X_{t-p} + \varepsilon_t)\right)\\ &= \gamma(h-1)\varphi_1 + \cdots + \gamma(h-p)\varphi_p\end{aligned}$$

and for $h = 0$

$$\begin{aligned}\gamma(0) &= \mathbb{E}\left(X_t^2\right)\\ &= \mathbb{E}\left(X_t(\varphi_1 X_{t-1} + \cdots + \varphi_p X_{t-p} + \varepsilon_t)\right)\\ &= \gamma(1)\varphi_1 + \cdots + \gamma(p)\varphi_p + \sigma^2.\end{aligned}$$

The Yule-Walker equations in matrix form can be now written as

$$\boldsymbol{\Gamma}^{(p)}\boldsymbol{\varphi} = \boldsymbol{\gamma}^{(p)}, \quad \gamma(0) - \left\langle \boldsymbol{\gamma}^{(p)}, \boldsymbol{\varphi} \right\rangle = \sigma^2. \tag{12.1}$$

where

$$\boldsymbol{\Gamma}^{(p)} = (\gamma(i-j))_{i,j=1}^{p}, \quad \boldsymbol{\gamma}^{(p)} = (\gamma(1), \dots, \gamma(p))^\top, \quad \boldsymbol{\varphi} = (\varphi_1, \dots, \varphi_p)^\top.$$

We have already seen Yule-Walker equations (12.1) when calculating the autocorrelation function of AR(p), see Section 8.1, where $\varphi_1, \dots, \varphi_p$ were known and we solved for $\gamma(h)$. Now we solve for $\varphi_1, \dots, \varphi_p$ and also replace $\gamma(h)$ by their corresponding estimators $\hat{\gamma}(h)$. Clearly, Yule-Walker estimators $\hat{\boldsymbol{\varphi}}$ of $\varphi_1, \dots, \varphi_p$ solve the first equation of (12.1)

$$\hat{\boldsymbol{\Gamma}}^{(p)}\hat{\boldsymbol{\varphi}} = \hat{\boldsymbol{\gamma}}^{(p)}$$

and the estimator for σ^2 solves

$$\hat{\sigma}^2 = \hat{\gamma}(0) - \left\langle \hat{\boldsymbol{\gamma}}^{(p)}, \hat{\boldsymbol{\varphi}} \right\rangle.$$

Then, we can finally compute

$$\hat{\boldsymbol{\varphi}} = \left(\hat{\boldsymbol{\Gamma}}^{(p)}\right)^{-1} \hat{\boldsymbol{\gamma}}^{(p)},$$

if the matrix is invertible. If p is large or models with many different orders need to be estimated, this may get computationally demanding – instead, the Durbin-Levinson recursive algorithm is typically used.

Theorem 12.1. *For a* ***causal*** AR(p) *process, the Yule-Walker estimators satisfy*

$$n^{1/2}(\hat{\boldsymbol{\varphi}} - \boldsymbol{\varphi}) \xrightarrow[n\to\infty]{d} \mathcal{N}\left(0, \sigma^2 \left(\boldsymbol{\Gamma}^{(p)}\right)^{-1}\right)$$

and $\hat{\sigma}^2 \xrightarrow[n\to\infty]{P} \sigma^2$.

Let X_t be an AR(p) process and n be large. The distribution of $n^{1/2}(\hat{\boldsymbol{\varphi}} - \boldsymbol{\varphi})$ is approximately $\mathcal{N}\left(0, \sigma^2 \left(\hat{\boldsymbol{\Gamma}}^{(p)}\right)^{-1}\right)$. Now with the approximate probability $1-\alpha$, the interval

$$\left(\hat{\varphi}_j - c_{1-\alpha/2} n^{-1/2} \hat{\sigma} \left(\hat{\boldsymbol{\Gamma}}^{(p)}\right)_{jj}^{1/2}, \hat{\varphi}_j + c_{1-\alpha/2} n^{-1/2} \hat{\sigma} \left(\hat{\boldsymbol{\Gamma}}^{(p)}\right)_{jj}^{1/2}\right)$$

covers the true value of φ_j, where $c_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of the standard normal distribution.

12.1.2 Method of moments estimation for general ARMA models

We can again compute the theoretical ACF, then the expressions depend on the parameters and we can we can equate them to the empirical ACF. This means to solve nonlinear equations. For example, consider MA (1):

- we have already seen that
$$\rho(1) = \frac{\theta_1}{1+\theta_1^2};$$
- then we can estimate θ_1 by solving
$$\hat{\rho}(1) = \frac{\hat{\theta}_1}{1+\hat{\theta}_1^2};$$
- and we will use the invertible solution (if it exists).

For higher-order MA or ARMA models, this quickly gets complicated (it is an inefficient estimator, as we will see later).

12.2 Conditional least squares

Consider again the AR (p) model

$$X_t - \mu = \varphi_1(X_{t-1} - \mu) + \cdots + \varphi_p(X_{t-p} - \mu) + \varepsilon_t.$$

By subtracting $\hat{\mu} = \overline{\mathbf{X}}$, we can ignore the constant level for now. Then $X_t - \hat{\mu}$, $t = p+1, \dots, n$ satisfy a linear regression model:

- response is $X_t - \hat{\mu}$;
- covariates are $X_{t-1} - \hat{\mu}, \dots, X_{t-p} - \hat{\mu}$;
- coefficients are $\varphi_1, \dots, \varphi_p$;
- uncorrelated errors ε_t with mean zero and variance σ^2 are assumed.

We then estimate the coefficients $\varphi_1, \dots, \varphi_p$ by ordinary least squares, that is to minimize the following

$$S_0(\boldsymbol{\varphi}) = \sum_{t=p+1}^{n} \left((X_t - \hat{\mu}) - \varphi_1(X_{t-1} - \hat{\mu}) - \cdots - \varphi_p(X_{t-p} - \hat{\mu})\right)^2,$$

which is called *conditional least squares* as the first p values are taken as fixed (and their randomness is ignored). More specifically, in matrix notation we get

- response vector (length $n - p$)
$$\mathbf{Y} = (X_{p+1} - \hat{\mu}, \dots, X_n - \hat{\mu})^\top,$$

- covariate (predictor) matrix (of dimension $(n-p) \times p$)

$$\Xi = (\Xi_{ij}), \quad \Xi_{ij} = X_{p+i-j} - \hat{\mu},$$

- coefficient vector (length p)

$$\boldsymbol{\varphi} = (\varphi_1, \dots, \varphi_p)^\top,$$

- errors (length $n-p$)

$$\mathbf{e} = (\varepsilon_{p+1}, \dots, \varepsilon_n)^\top.$$

Then the regression model has the form

$$\mathbf{Y} = \Xi\boldsymbol{\varphi} + \mathbf{e}$$

and least-squares criterion reads as follows

$$S_0(\boldsymbol{\varphi}) = \|\mathbf{Y} - \Xi\boldsymbol{\varphi}\|^2.$$

12.2.1 Normal equations and relation to Yule-Walker

Recall that we can solve the following optimization problem

$$\min_{\boldsymbol{\varphi}} \|\mathbf{Y} - \Xi\boldsymbol{\varphi}\|^2,$$

which corresponds to "normal equations"

$$\Xi^\top \Xi \boldsymbol{\varphi} = \Xi^\top \mathbf{Y}.$$

For $j \neq k$, the (j,k) entry of $(n-p-1)^{-1}\Xi^\top\Xi$ is

$$\frac{1}{n-p-1} \sum_{i=1}^{n-p} (X_{p+i-j} - \overline{\mathbf{X}})(X_{p+i-k} - \overline{\mathbf{X}}),$$

which is similar to $\hat{\gamma}(j-k)$. Just as well, for the diagonal entries, we get similarity with $\hat{\gamma}(0)$. Thus the OLS estimator is similar to the Yule-Walker estimator (if p is small relative to n).

12.2.2 Conditional least squares for an autoregression model with unknown mean

We can rewrite the model

$$X_t - \mu = \varphi_1(X_{t-1} - \mu) + \cdots + \varphi_p(X_{t-p} - \mu) + \varepsilon_t.$$

as

$$X_t = \alpha + \varphi_1(X_{t-1} - \mu) + \cdots + \varphi_p(X_{t-p} - \mu) + \varepsilon_t,$$

where $\alpha = (1 - \varphi_1 - \cdots - \varphi_p)\mu$. Now we can estimate both $\boldsymbol{\varphi}$ and μ (or α now) with least squares. This gives us a linear model with intercept α and we can estimate $\boldsymbol{\varphi}, \alpha$ by minimizing

$$S_0(\boldsymbol{\varphi}, \alpha) = \sum_{t=p+1}^{n} \left(\mathrm{X}_t - \alpha - \varphi_1 \mathrm{X}_{t-1} - \cdots - \varphi_p \mathrm{X}_{t-p}\right)^2,$$

then estimate σ^2 by $\hat{\sigma}^2 = (n - p - 1)^{-1} S_0(\hat{\boldsymbol{\varphi}}, \hat{\alpha})$. Consider now the least squares criterion for AR (1)

$$S_0(\varphi_1, \mu) = \sum_{t=2}^{n} \left(\mathrm{X}_t - \mu - \varphi_1(\mathrm{X}_{t-1} - \mu)\right)^2,$$

which we can differentiate with respect to μ to get

$$\mu = \frac{1}{(n-1)(1-\varphi_1)} \left(\sum_{t=2}^{n} \mathrm{X}_t - \varphi_1 \sum_{t=2}^{n} \mathrm{X}_{t-1}\right).$$

For large n, we can use an approximation

$$\frac{1}{n-1} \sum_{t=2}^{n} \mathrm{X}_t \approx \frac{1}{n-1} \sum_{t=2}^{n} \mathrm{X}_{t-1} \approx \overline{\mathbf{X}},$$

thus, regardless of φ_1, $\hat{\mu} = \frac{1}{1-\varphi_1}(\overline{\mathbf{X}} - \varphi_1 \overline{\mathbf{X}}) = \overline{\mathbf{X}}$. What's more, we can differentiate $S_0(\varphi_1, \overline{\mathbf{X}})$ w.r.t. φ_1 and solve to get

$$\hat{\varphi}_1 = \frac{\sum_{t=2}^{n} (\mathrm{X}_t - \overline{\mathbf{X}})(\mathrm{X}_{t-1} - \overline{\mathbf{X}})}{\sum_{t=2}^{n} (\mathrm{X}_{t-1} - \overline{\mathbf{X}})^2}.$$

Except for the term $(\mathrm{X}_n - \overline{\mathbf{X}})^2$ that is missing in the denominator, this is the same as $\hat{\rho}(1)$, which is the Yule-Walker estimator for $\rho(1)$. Hence, for large n, OLS and YW are almost identical.

13 Parameter Estimation in ARMA Models

13.1 Maximum likelihood estimation

Consider an ARMA (p, q) model

$$(X_t - \mu) - \varphi_1(X_{t-1} - \mu) - \cdots - \varphi_p(X_{t-p} - \mu) = \varepsilon_t + \theta_1\varepsilon_{t-1} + \cdots + \theta_q\varepsilon_{t-q},$$

that is $\Phi(\mathsf{B})(X_t - \mu) = \Theta(\mathsf{B})\,\varepsilon_t$, where $\{\varepsilon_t\} \sim \mathrm{WN}\left(0, \sigma^2\right)$. Now let us assume we observe data $X_1, \ldots, X_n$ and let our goal be to estimate unknown parameters

$$\boldsymbol{\beta} = (\overbrace{\varphi_1, \ldots, \varphi_p}^{\boldsymbol{\varphi}}, \overbrace{\theta_1, \ldots, \theta_q}^{\boldsymbol{\theta}}, \mu, \sigma^2)$$

given known orders p, q.

ℹ Note

Here, $\boldsymbol{\beta}$ may also include other parameters of the deterministic trend μ_t, e.g. regression coefficients for external covariates, see `xreg` in `R`.

Clearly, there are multiple possible approaches to the estimation:

- *Method of moments:* we can equate the theoretical moments to their empirical estimates and solve for the parameters (e.g. Yule-Walker);
- *Least squares:* we can minimize the sum of squared distances between the observed values and their model-based expectations;
- *Maximum likelihood estimation.*

Thus let the joint density of $X_1, \ldots, X_n$ be $f(x_1, \ldots, x_n; \boldsymbol{\beta}) = f(\boldsymbol{x}; \boldsymbol{\beta})$ and we shall strive to find the value of $\boldsymbol{\beta}$ for which the likelihood $\mathcal{L}(\boldsymbol{\beta}) = f(X_1, \ldots, X_n; \boldsymbol{\beta})$ is maximal.

Advantages & Disadvantages of MLE

This approach of using MLE brings some advantages like efficiency (low variance), optimality, clear justification of our results and a unified framework. Also often, the Gaussian assumption is reasonable! But, on the other hand, it leads to possibly difficult optimization (the solution is given implicitly and as such it requires iterative numerical procedure). Furthermore, we need to choose a good starting point (and often use other estimators for this alone).

13.1.1 Gaussian likelihood function

Let us assume the process is now Gaussian and data $\mathbf{X} = (X_1, \dots, X_n)^\top$ are jointly Gaussian with mean μ and covariance matrix $\boldsymbol{\Gamma}$. The likelihood function is then

$$\mathcal{L}(\boldsymbol{\beta}) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Gamma}|^{1/2}} \exp\left(-\frac{(\mathbf{X}-\mu)^\top \boldsymbol{\Gamma}^{-1}(\mathbf{X}-\mu)}{2}\right),$$

which the MLE approach maximizes.

13.1.1.1 Maximum likelihood estimation for $\mathrm{AR}(1)$ process

Consider now a case where X_t, $t = 1, \dots, n$ follow an $\mathrm{AR}(1)$ model, i.e. $X_t = \mu + \varphi_1(X_{t-1} - \mu) + \varepsilon_t$, and rewrite it again as a linear model

$$X_t = \alpha + \varphi_1 X_{t-1} + \varepsilon_t,$$

where $\alpha = \mu(1 - \varphi_1)$. The covariance matrix of $(X_1, \dots, X_n)^\top$ has then entries

$$\Gamma_{ij} = \frac{\sigma^2}{1-\varphi_1^2}\varphi_1^{|i-j|}.$$

Note that this leads to a difficult matrix inverse and determinant computation in the likelihood function

$$\frac{1}{(2\pi)^{n/2}|\boldsymbol{\Gamma}|^{1/2}} \exp\left(-\frac{(\mathbf{X}-\mu)^\top \boldsymbol{\Gamma}^{-1}(\mathbf{X}-\mu)}{2}\right).$$

> **ℹ Note**
>
> Notice the difference from situations with independent data!

For fixed parameter values, it is still difficult to evaluate the likelihood, hence we use *conditioning*

$$\begin{aligned} f(x_1, \dots, x_n; \boldsymbol{\beta}) &= f(x_n \mid x_{n-1}, \dots, x_1; \boldsymbol{\beta}) f(x_{n-1}, \dots, x_1; \boldsymbol{\beta}) \\ &= \left(\prod_{i=2}^{n} f(x_i \mid x_{i-1}, \dots, x_1; \boldsymbol{\beta})\right) f(x_1; \boldsymbol{\beta}). \end{aligned}$$

For Gaussian data, all conditional distributions are still Gaussian and for $i \geq 2$, the conditional expectation and variance are

$$\mathbb{E}(X_i \mid X_{i-1}, \dots, X_1) = \alpha + \varphi_1 X_{i-1}, \quad \operatorname{var}(X_i \mid X_{i-1}, \dots, X_1) = \sigma^2,$$

hence $f(x_i \mid x_{i-1}, \dots, x_1; \boldsymbol{\beta}) \sim \mathcal{N}\left(\alpha + \varphi_1 x_{i-1}, \sigma^2\right)$. What's more, the (unconditional) distribution of X_1 is, too, Gaussian with mean $\mu = \alpha/(1-\varphi_1)$ and variance (by causality)

$$\mathbb{E}(X_1 - \mu)^2 = \mathbb{E}\left(\sum_{j=0}^{\infty} \varphi_1^j \varepsilon_{1-j}\right)^2 = \frac{\sigma^2}{1-\varphi_1^2}.$$

Hence $f(x_1) \sim \mathcal{N}\left(\alpha/(1-\varphi_1), \sigma^2/(1-\varphi_1^2)\right)$ and put together, the likelihood is

$$\begin{aligned}\mathcal{L}(\boldsymbol{\beta}) = &\left(\prod_{i=2}^{n} \frac{1}{(2\pi)^{1/2}\sigma} \exp\left(-\frac{(\mathrm{X}_i - \alpha - \varphi_1 \mathrm{X}_{i-1})^2}{2\sigma^2}\right)\right) \\ &\times \frac{1}{(2\pi)^{1/2}\sigma/(1-\varphi_1^2)^{1/2}} \exp\left(-\frac{(\mathrm{X}_1 - \alpha/(1-\varphi_1))^2}{2\sigma^2/(1-\varphi_1^2)}\right).\end{aligned}$$

As one can see, now it is easy to compute the likelihood given a set of parameter values. To solve for optimal $\boldsymbol{\beta}$, as usual, we define a log-likelihood (ignoring any terms independent of parameters) to get

$$\begin{aligned}\ell(\boldsymbol{\beta}) = &-\sum_{i=2}^{n} \frac{(\mathrm{X}_i - \alpha - \varphi_1 \mathrm{X}_{i-1})^2}{2\sigma^2} - \frac{(\mathrm{X}_1 - \alpha/(1-\varphi_1))^2}{2\sigma^2/(1-\varphi_1^2)} \\ &- n\log\sigma - \frac{1}{2}\log(1-\varphi_1^2),\end{aligned}$$

which is clearly non-linear in parameters and as such needs to be maximized numerically.

Notice that all terms except the one for X_1 have the same form in the likelihood function $\mathcal{L}(\boldsymbol{\beta})$, thus we can use an alternative approach where we consider X_1 as fixed. From this, we get a conditional likelihood

$$\begin{aligned}\mathcal{L}(\boldsymbol{\beta} \mid \mathrm{X}_1) &= f(\mathrm{X}_2, \dots, \mathrm{X}_n \mid \mathrm{X}_1; \boldsymbol{\beta}) \\ &= \prod_{i=2}^{n} f(\mathrm{X}_i \mid \mathrm{X}_1, \dots, \mathrm{X}_{i-1}; \boldsymbol{\beta}) \\ &= \prod_{i=2}^{n} \frac{1}{(2\pi)^{1/2}\sigma} \exp\left(-\frac{(\mathrm{X}_i - \alpha - \varphi_1 \mathrm{X}_{i-1})^2}{2\sigma^2}\right)\end{aligned}$$

and the conditional log-likelihood similarly becomes

$$\ell(\boldsymbol{\beta} \mid \mathrm{X}_1) = -\sum_{i=2}^{n} \frac{(\mathrm{X}_i - \alpha - \varphi_1 \mathrm{X}_{i-1})^2}{2\sigma^2} - (n-1)\log\sigma.$$

One can notice that this is the conditional least squares problem (or conditional sum) and that we can solve this explicitly (as opposed to the implicit formulation before).

13.1.1.2 Maximum likelihood estimation for an autoregressive process

Now consider an AR (p) process, i.e.

$$\mathrm{X}_t = \alpha\varphi_1 \mathrm{X}_{t-1} + \cdots + \varphi_p \mathrm{X}_{t-p} + \varepsilon_t,$$

for which the likelihood comes in the form

$$\begin{aligned}\mathcal{L}(\boldsymbol{\beta}) &= \prod_{i=p+1}^{n} f(\mathrm{X}_i \mid \mathrm{X}_1, \dots, \mathrm{X}_{i-1}; \boldsymbol{\beta}) f(\mathrm{X}_1, \dots, \mathrm{X}_p; \boldsymbol{\beta}) \\ &= \mathcal{L}\left(\boldsymbol{\beta} \mid \mathrm{X}_1, \dots, \mathrm{X}_p\right) f(\mathrm{X}_1, \dots, \mathrm{X}_p; \boldsymbol{\beta})\end{aligned}$$

with conditional densities

$$f(\mathrm{X}_i \,|\, \mathrm{X}_1, \dots, \mathrm{X}_{i-1}; \boldsymbol{\beta}) \sim \mathcal{N}\left(\alpha + \varphi_1 x_{i-1} + \dots + \varphi_p x_{i-p}, \sigma^2\right).$$

The density $f(\mathrm{X}_1, \dots, \mathrm{X}_p; \boldsymbol{\beta})$ is Gaussian with a complicated covariance matrix. Also, the conditional likelihood $\mathcal{L}\left(\boldsymbol{\beta} \,|\, \mathrm{X}_1, \dots, \mathrm{X}_p\right)$ leads to a least squares problem.

13.1.2 Evaluation of the likelihood function in the general case

In ARMA (p, q) models, the likelihood is much more complicated due to MA terms and it requires a large non-diagonal matrix to be inverted and its determinant to be computed. Similarly to how conditioning on previous values was useful in AR models, even in the general case, the likelihood becomes a product of simpler terms (densities in this product correspond to independence in data).

Tip

Note that we have already seen a transformation of series into uncorrelated variables via **innovations**.

13.2 Recursive likelihood calculation using innovations

13.2.1 Innovations

Recall the innovations $\mathrm{Z}_i = \mathrm{X}_i - \hat{\mathrm{X}}_i$, where $\hat{\mathrm{X}}_i$ is the best linear prediction of X_i from $\mathrm{X}_1, \dots, \mathrm{X}_{i-1}$ (this corresponds to conditional expectation in the Gaussian case). Predictions $\hat{\mathrm{X}}_i$ can be expressed in the terms of previous observations $\mathrm{X}_1, \dots, \mathrm{X}_{i-1}$ or in the terms of previous innovations as

$$\hat{\mathrm{X}}_i = \mu + \sum_{j=1}^{i-1} \varphi_{i-1,j}(\mathrm{X}_{i-j} - \mu) = \mu + \sum_{j=1}^{i-1} \theta_{i-1,j}(\mathrm{X}_{i-j} - \hat{\mathrm{X}}_{i-j}).$$

The innovations have the following properties:

- zero mean;
- variance (or prediction error) $\mathbb{E}\left(\mathrm{X}_i - \hat{\mathrm{X}}_i\right)^2 = v_{i-i} = r_{i-1}\sigma^2$ for appropriate $r_{i-1} > 0$;
- most importantly, they are uncorrelated (even independent in the Gaussian case) because $\mathrm{Z}_j \in \operatorname{lin}\left\{\mathrm{X}_1, \dots, \mathrm{X}_j\right\}$ and $\mathrm{Z}_k \perp \operatorname{lin}\left\{\mathrm{X}_1, \dots, \mathrm{X}_j\right\}$ for $j < k$.

The innovations $\mathbf{Z} = (\mathrm{Z}_1, \dots, \mathrm{Z}_n)^\top$ are obtained from the observations $\mathbf{X} = (\mathrm{X}_1, \dots, \mathrm{X}_n)^\top$ by linear transformation

$$\mathbf{Z} = \boldsymbol{A}(\mathbf{X} - \mu),$$

where the $n \times n$ matrix $\boldsymbol{A}$ is given as

$$\boldsymbol{A} = \begin{pmatrix} 1 & 0 & \cdots & \cdot & \cdot & \cdot \\ -\theta_{1,1} & 1 & 0 & \cdots & \cdot & \cdot \\ -\theta_{2,2} & -\theta_{2,1} & 1 & 0 & \cdots & \cdot \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ -\theta_{n-2,n-2} & -\theta_{n-2,n-3} & \cdots & -\theta_{n-2,1} & 1 & 0 \\ -\theta_{n-1,n-1} & -\theta_{n-1,n-2} & \cdots & -\theta_{n-1,2} & -\theta_{n-1,1} & 1 \end{pmatrix}.$$

Clearly, the matrix $\boldsymbol{A}$ is **regular**, hence the correspondence between the data X_i and the innovations $Z_i = X_i - \hat{X}_i$ is **one-to-one**. Thus they contain the same information and have the same likelihood. Since the linear transformation preserves normality, i.e. if $X_1, \dots, X_n$ are jointly Gaussian, then $Z_1, \dots, Z_N$ are too jointly Gaussian. All in all, Z_i is Gaussian with mean zero and variance $\sigma^2 r_{i-1}$, i.e.

$$f(z_i; \boldsymbol{\beta}) = \frac{1}{(2\pi)^{1/2} \sigma r_{i-1}^{1/2}} \exp\left(- \frac{z_i^2}{2\sigma^2 r_{i-1}} \right),$$

thus Z_i's are **uncorrelated** and by the Gaussian distribution properties they are even **independent**. Hence, the joint density of $\mathbf{Z} = (Z_1, \dots, Z_n)^\top$ is the product of marginal densities

$$\begin{aligned} f(\boldsymbol{z}; \boldsymbol{\beta}) = f(z_1, \dots, z_n; \boldsymbol{\beta}) &= \prod_{i=1}^{n} f(z_i; \boldsymbol{\beta}) \\ &= \frac{1}{(2\pi)^{n/2} \sigma^n \prod_{i=1}^{n} r_{i-1}^{1/2}} \exp\left(- \sum_{i=1}^{n} \frac{z_i^2}{2\sigma^2 r_{i-1}} \right). \end{aligned}$$

The likelihood then becomes

$$\begin{aligned} \mathcal{L}(\boldsymbol{\beta}) &= \frac{1}{(2\pi)^{n/2} \sigma^n \prod_{i=1}^{n} r_{i-1}^{1/2}} \exp\left(- \sum_{i=1}^{n} \frac{Z_i^2}{2\sigma^2 r_{i-1}} \right) \\ &= \frac{1}{(2\pi)^{n/2} \sigma^n \prod_{i=1}^{n} r_{i-1}^{1/2}} \exp\left(- \sum_{i=1}^{n} \frac{(X_i - \hat{X}_i)^2}{2\sigma^2 r_{i-1}} \right), \end{aligned}$$

where $\hat{X}_i$ and r_{i-1} depend on the covariance structure of the series, i.e. on the parameters of the model. For a given set of parameter values, the value of the likelihood function can be easily computed, provided we can obtain $\hat{X}_i$ and r_{i-1}. Then, the **innovations algorithm** can be used to compute the one-step predictions $\hat{X}_i$ and their errors $v_{i-1} = \sigma^2 r_{i-1}$ recursively using simple linear operations.

Tip

This can be used for any Gaussian series, not only ARMA processes. Also, another recursive approach we might use is the *Kálmán* filter.

13.2.2 Conditional least squares

For an ARMA (p, q) model, one can condition on the first p (or possibly more) values of the series and then minimize $\sum_{i=p+1}^{n} \hat{\varepsilon}_i^2(\boldsymbol{\beta})$, where

$$\begin{aligned}\hat{\varepsilon}_i(\boldsymbol{\beta}) =& \mathrm{X}_i - \mu - \varphi_1(\mathrm{X}_{i-1-\mu}) - \cdots - \varphi_p(\mathrm{X}_{i-p} - \mu) \\ &- \theta_1\hat{\varepsilon}_{i-1}(\boldsymbol{\beta}) - \cdots - \theta_q\hat{\varepsilon}_{i-q}(\boldsymbol{\beta})\end{aligned}$$

for $i > p$ and $\hat{\varepsilon}_i(\boldsymbol{\beta}) = 0$ for $i \leq p$. Afterward, we can compare the exact likelihood and condition least squares, since the exact likelihood uses the innovations Z_i (errors of predictions using the complete observed history) and the conditional least squares use the residuals $\hat{\varepsilon}_i$ (residuals obtained from the model formula, fixing the initial values).

13.3 Computation

Although the conditional likelihood for AR models can be maximized directly analytically via least squares, other situations require numerical optimization methods. While grid search might be useful sometimes, the general approach is the Newton-Raphson algorithm.

13.3.1 the Newton-Raphson algorithm

Our objective now shall be to maximize $\ell(\boldsymbol{\beta})$. First, compute the gradient

$$\nabla\ell(\boldsymbol{\beta}) = \left(\frac{\partial\ell(\boldsymbol{\beta})}{\partial\beta_1}, \ldots, \frac{\partial\ell(\boldsymbol{\beta})}{\partial\beta_d}\right)^{\top}$$

and afterward solve for $\nabla\ell(\boldsymbol{\beta}) = \mathbf{0}$, for which we can use a Taylor expansion

$$\mathbf{0} = \nabla\ell(\boldsymbol{\beta}) \approx \nabla\ell\left(\boldsymbol{\beta}_{(0)}\right) - \nabla^2\ell\left(\boldsymbol{\beta}_{(0)}\right)\left(\boldsymbol{\beta} - \boldsymbol{\beta}_{(0)}\right),$$

where $\nabla^2\ell(\boldsymbol{\beta})$ is the Hessian matrix of ℓ. Solving the linearized equation yields the next iteration

$$\boldsymbol{\beta}_{(k+1)} = \boldsymbol{\beta}_{(k)} + \nabla^2\ell\left(\boldsymbol{\beta}_{(k)}\right)^{-1}\nabla\ell\left(\boldsymbol{\beta}_{(k)}\right).$$

Here a good initial estimate is important for the convergence of the iterative process. As such, we typically use the conditional least squares, and then we iterate until convergence under some criterion, for example

- small relative change of the solution, e.g.

$$\frac{\left\|\boldsymbol{\beta}_{(k)} - \boldsymbol{\beta}_{(k-1)}\right\|}{\left\|\boldsymbol{\beta}_{(k-1)}\right\|} < \delta;$$

- small relative change in the value of the objective function, e.g.

$$\frac{|\ell(\boldsymbol{\beta}_{(k)}) - \ell(\boldsymbol{\beta}_{(k-1)})|}{|\ell(\boldsymbol{\beta}_{(k-1)})|} < \delta;$$

- small norm of the gradient vector

$$\|\nabla \ell(\boldsymbol{\beta}_{(k)})\| < \delta.$$

13.4 Properties of estimators

Theorem 13.1. *Under* appropriate technical assumptions, *the asymptotic distribution of the maximum likelihood estimator is given as follows,*

$$n^{1/2}\left((\hat{\boldsymbol{\varphi}}, \hat{\boldsymbol{\theta}})^\top - (\boldsymbol{\varphi}, \boldsymbol{\theta})^\top\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \sigma^2 \boldsymbol{V}^{-1}\right),$$

where the block $\boldsymbol{V}_{\boldsymbol{\varphi},\boldsymbol{\varphi}}$ of the matrix

$$\boldsymbol{V} = \begin{pmatrix} \boldsymbol{V}_{\boldsymbol{\varphi},\boldsymbol{\varphi}} & \boldsymbol{V}_{\boldsymbol{\varphi},\boldsymbol{\theta}} \\ \boldsymbol{V}_{\boldsymbol{\theta},\boldsymbol{\varphi}} & \boldsymbol{V}_{\boldsymbol{\theta},\boldsymbol{\theta}} \end{pmatrix}$$

contains the autocovariances up to lag $p-1$ of the AR process

$$\Phi(\mathsf{B})\, Y_t = \varepsilon_t, \tag{13.1}$$

the block $\boldsymbol{V}_{\boldsymbol{\theta},\boldsymbol{\theta}}$ contains the autocovariances up to lag $q-1$ of the AR process

$$\Theta(\mathsf{B})\, Z_t = \varepsilon_t \tag{13.2}$$

and the block $\boldsymbol{V}_{\boldsymbol{\varphi},\boldsymbol{\theta}}$ the cross-covariances between the processes (13.1) and (13.2).

Example 13.1 (Maximum likelihood estimation asymptotics for ARMA (1, 1))**.** Consider the model

$$X_t = \varphi_1 X_{t-1} + \varepsilon_t + \theta_1 \varepsilon_{t-1},$$

for which the limiting distribution of $n^{1/2}((\hat{\varphi}_1, \hat{\theta}_1)^\top - (\varphi_1, \theta_1)^\top)$ is *bivariate Gaussian* with mean zero and covariance matrix $\sigma^2 \boldsymbol{V}^{-1}$. Here

$$\boldsymbol{V} = \begin{pmatrix} \boldsymbol{V}_{\varphi,\varphi} & \boldsymbol{V}_{\varphi,\theta} \\ \boldsymbol{V}_{\theta,\varphi} & \boldsymbol{V}_{\theta,\theta} \end{pmatrix}$$

and $V_{\varphi,\varphi}$ is given the autocovariance of the AR (1) process $Y_t = \varphi_1 Y_{t-1} + \varepsilon_t$, that is

$$V_{\varphi,\varphi} = \gamma_{\mathbf{Y}}(0) = \frac{\sigma^2}{1 - \varphi_1^2}.$$

Furthermore, $V_{\theta,\theta}$ is set by the autocovariance of the AR (1) process $Z_t = -\theta_1 Z_{t-1} + \varepsilon_t$, that is

$$V_{\theta,\theta} = \gamma_{\mathbf{Z}}(0) = \frac{\sigma^2}{1-\theta_1^2}.$$

Last, but not least, the entry $V_{\varphi,\theta}$ is given by the covariance of **Y** and **Z**, that is

$$\begin{aligned} V_{\varphi,\theta} &= \operatorname{cov}(Y_t, Z_t) = \operatorname{cov}(\varphi_1 Y_{t-1} + \varepsilon_t, -\theta_1 Z_{t-1} + \varepsilon_t) \\ &= -\varphi_1\theta_1 \operatorname{cov}(Y_{t-1}, Z_{t-1}) + \sigma^2. \end{aligned}$$

In the light of **stationarity** we get $V_{\varphi,\theta} = -\varphi_1\theta_1 V_{\varphi,\theta} + \sigma^2$, thus

$$V_{\varphi,\theta} = \frac{\sigma^2}{1+\varphi_1\theta_1} = V_{\theta,\varphi}.$$

Combined, this produces

$$\boldsymbol{V} = \sigma^2 \begin{pmatrix} (1-\varphi_1^2)^{-1} & (1+\varphi_1\theta_1)^{-1} \\ (1+\varphi_1\theta_1)^{-1} & (1-\theta_1^2)^{-1} \end{pmatrix},$$

therefore, $(\hat{\varphi}_1, \hat{\theta}_1)^\top$ is approximately normal with mean $(\varphi_1, \theta_1)^\top$ and covariance matrix

$$\frac{1}{n} \begin{pmatrix} (1-\varphi_1^2)^{-1} & (1+\varphi_1\theta_1)^{-1} \\ (1+\varphi_1\theta_1)^{-1} & (1-\theta_1^2)^{-1} \end{pmatrix}^{-1}.$$

Tip

Notice here the **invariance** with respect to σ^2.