FACULTY
OF INFORMATION
TECHNOLOGY
CTU IN PRAGUE

Assignment of bachelor's thesis

Title:	Modeling and properties of autoregressive processes
Student:	Artem Tokarevskikh
Supervisor:	Ing. Kamil Dedecius, Ph.D.
Study program:	Informatics
Branch / specialization:	Knowledge Engineering
Department:	Department of Applied Mathematics
Validity:	until the end of summer semester 2022/2023

Instructions

Autoregressive (AR) processes where the previous values (plus a noise term) determine the current output variable are omnipresent in nature and human activities. They are one of the most fundamental parts of many other processes like the autoregressive-moving average (ARMA) or generalized autoregressive conditional heteroskedasticity (GARCH) processes and their variants used, e.g., in econometrics, or the processes with exogenous inputs used in the control theory.

The aim of this thesis is to focus on the AR processes and their properties like the order, their stationarity and its testing, means for their modeling and description. In addition, the student should provide a convenient real-life example or a brief study of a particular AR-related problem.

Electronically approved by Ing. Karel Klouda, Ph.D. on 10 February 2021 in Prague.

FACULTY
OF INFORMATION
TECHNOLOGY
CTU IN PRAGUE

Bachelor's thesis

Modeling and properties of autoregressive processes

Artem Tokarevskikh

Department of Applied Mathematics

Supervisor: Ing. Kamil Dedecius, Ph.D.

May 13, 2021

Acknowledgements

First and foremost, I want to thank my supervisor Ing. Kamil Dedecius, Ph.D. Thanks to his sincere and valuable encouragement, I got interested in statistics and topics covered in this thesis. I am grateful for his assistance and dedicated involvement in every step throughout the process.

I also want to express my sincere gratitude to my beloved partner Kristina for her unceasing encouragement, support, and attention.

Last but not least, I would like to thank my family (and especially my dear Mother) and friends for supporting me during my studies. I would never have been able to achieve this without you.

Declaration

I hereby declare that the presented thesis is my own work and that I have cited all sources of information in accordance with the Guideline for adhering to ethical principles when elaborating an academic final thesis.

I acknowledge that my thesis is subject to the rights and obligations stipulated by the Act No. 121/2000 Coll., the Copyright Act, as amended, in particular that the Czech Technical University in Prague has the right to conclude a license agreement on the utilization of this thesis as a school work under the provisions of Article 60 (1) of the Act.

In Prague on May 13, 2021 . . . . . .

Czech Technical University in Prague

Faculty of Information Technology

Citation of this thesis

Tokarevskikh, Artem. *Modeling and properties of autoregressive processes.* Bachelor's thesis. Czech Technical University in Prague, Faculty of Information Technology, 2021.

Abstrakt

Problematika streamovaných dat přitahuje v poslední době mnoho pozornosti v odvětvích jako jsou IoT, sociální sítě či elektronický obchod. Jde o data, která jsou svými zdroji kontinuálně generovaná a je potřeba je zpracovávat v reálném čase. Vzhledem k vysoké frekvenci jejich získávání je potřeba využívat metody, které jsou výpočetně minimálně náročné a zároveň mají minimální požadavky na paměť. Autoregresni (AR) modely jsou jedním ze základních přístupů k modelování časových řad. Myšlenka autoregresních modelů spočívá v tom, že současná hodnota řady je lineárně závislá na hodnotách předchozích. Díky své struktuře AR modely umožňují efektivně zpracovávat streamovaná data i v situacích, kdy složitější modely nemohou být kvůli související výpočetní a paměťové náročnosti využity. Tato práce zkoumá AR procesy a jejich vlastnosti, nastiňuje teorii, která za nimi stojí, a uvádí příklad AR modelování na reálných datech týkajících se průběhu pandemie COVID-19 v České republice.

Klíčová slova Časová řada, Náhodný proces, Autoregrese, Autokorelace, Autokovariance, Modelování

Abstract

Streaming data has recently attracted attention in numerous fields such as IoT, social networks, and e-Commerce. Streaming data is the data that is continuously generated by some sources and there is a need for their real-time processing. Given that the frequency of a stream may be very high, the methods for its processing must be computationally cheap and have low memory requirements. Autoregressive (AR) models are one of the fundamental approaches to time series modelling. The idea behind the autoregressive models is that the current value of the series is linearly dependent on its most recent past values. Thanks to its structure, AR models are able to effectively process streaming data even in situations when more complex models cannot achieve the desired performance for their computational and memory burden. This thesis investigates AR processes and their properties, outlines the theory behind them, and provides an illustrative example of AR modelling on real data related to the COVID-19 pandemic course in Czech Republic.

Keywords Time series, Stochastic process, Autoregression, Autocorrelation, Autocovariance, Modelling

Contents

List of Figures

Introduction

Streaming data – data that is continuously generated by some sources – has recently attracted attention in numerous fields such as IoT, social networks, e-Commerce, etc. Given that the frequency of a stream may be very high (e.g., a click-stream on Facebook), the methods applicable must be as computationally inexpensive as possible and consume a small amount of memory.

Autoregression (AR) is a common approach for time series modeling. Autoregressive models operate under the premise that the past values linearly affect the current values, which makes this technique very popular. Thanks to their structure, the AR models are able to effectively process streaming data even in situations when more complex models cannot achieve the desired performance. In addition, together with the moving-average (MA) model, it is a key component of the more general autoregressive–moving-average (ARMA) and autoregressive integrated moving average (ARIMA) time series models, which have a more complicated structure and, thus, are more computationally costly.

The aim of this thesis is to focus on AR processes and their properties like the order and stationarity, describe the means for their modelling, and provide an example of AR modelling on a real-world time series.

We expect the reader to be familiar with essential concepts of probability theory, e.g., the probability space, random value, mean, variance, correlation, statistical hypothesis testing, etc. Furthermore, in this thesis we will use basic concepts from linear algebra and calculus.

Structure of the Thesis

This thesis is divided into three chapters. Chapter 1 introduces an AR process and focuses on its theoretical properties with regard to the probability theory in order to describe its fundamental characteristics and determine what affects it. In Chapter 2, we focus on the time series modelling with AR model and go through each step of the modelling procedure. An illustrative example of AR modelling on real-world time series data is provided in Chapter 3. In addition, the methods introduced in Chapter 2 are compared.

Chapter 1

Theoretical part

This chapter is dedicated to the analysis of an AR process as a stochastic process. We will start with defining what is a stochastic process and focus on its main properties. Then we will move to AR processes and their description with regard to probability theory. AR processes are an idealization of real processes, however it constructs the basis for the AR modelling. In the last section we will describe the behaviour of the stochastic process using spectral analysis.

1.1 Stochastic processes

Autoregression is a common approach to time series modeling, so before we dive into the AR processes, we must define what the time series is. We assume discrete time instants in this thesis, hence let us introduce the set of time indices $T \subset \mathbb{Z}$.

Definition 1 (Time series) *Let T be a set of time indices. A Time series Y is a sequence of measurements taken at successive equally spaced points $t \in T$, i.e., $Y = \{y_t \,|\, t \in T\}$.*

This definition considers a time series to be a sequence of measurements, whereas our goal is to deduce properties of an underlying random process. Let us rewrite this definition with regard to the probability theory. Every point Y_t of a series might be considered as a realization of a *random variable* from an underlying probability space $(\Omega, \mathcal{F}, P)$. Sequence of random variables is called the **stochastic process**

Definition 2 (Stochastic process) *Let $(\Omega, \mathcal{F}, P)$ be a probability space and T be a set of time indices. A stochastic process Y is a set $\{Y_t \,|\, t \in T\}$, where Y_t is a random variable from $(\Omega, \mathcal{F}, P)$.*

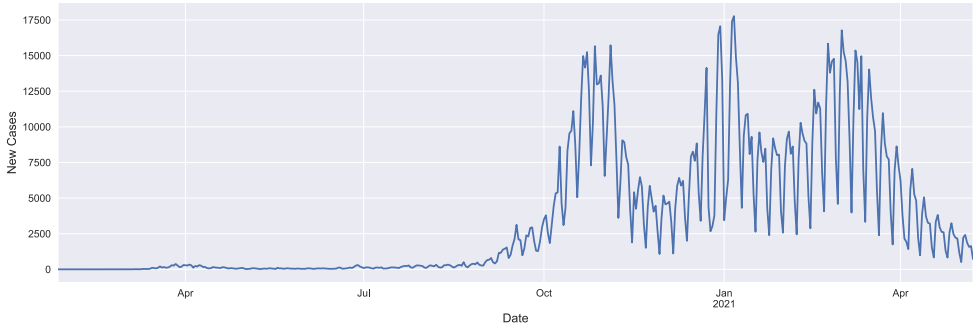


Figure 1.1: Time series with daily confirmed cases of COVID-19 disease.

The probabilistic structure of such a process is determined by its joint probability distribution, whose analysis is very complicated. Fortunately, much of the information in these distributions can be described in terms of the central and non-central moments (mean, variance, covariance, etc.) [1].

For the stochastic process Y, the **mean** function is defined by

$$\mu_t = \mathbb{E}[Y_t], \qquad t \in T. \tag{1.1}$$

Note that μ_t might be different at each time point t.

Similarly, the **variance** function for the stochastic process

$$\sigma_t^2 = var(Y_t), \qquad t \in T. \tag{1.2}$$

Since the covariance of the time series elements describes the inner structure of a single underlying process, the covariance is usually called the **autocovariance** function (ACVF) and it is defined as follows

$$\gamma_{t_1,t_2} = Cov(Y_{t_1}, Y_{t_2}) = \mathbb{E}[(Y_{t_1} - \mu_{t_1})(Y_{t_2} - \mu_{t_2})], \qquad t_1,\, t_2 \in T. \tag{1.3}$$

Finally, let us define **autocorrelation** function (ACF), which is a normalized autocovariance function. The autocorrelation function is defined as

$$\rho_{t_1,t_2} = Corr(Y_{t_1}, Y_{t_2}) = \frac{Cov(Y_{t_1}, Y_{t_2})}{\sigma_{t_1}\sigma_{t_2}}, \qquad t_1,\, t_2 \in T. \tag{1.4}$$

Note that the autocovariance (autocorrelation) function gives the covariance (correlation) of the process with itself at pairs of time points. The prefix *auto* is to convey the notion of self-correlation.

To describe a stochastic process and its structure from the statistical perspective, we usually make some assumptions about its structure. The most important assumption about the stochastic process is **stationarity**. The basic idea of stationarity is that the characteristics of a process (like moments) that affect its behavior do not change over time [1].

Definition 3 (Strict stationarity) *A stochastic process Y is said to be* ***strictly stationary*** *if the joint distribution of $Y_{t_1}, Y_{t_2}, \ldots, Y_{t_n}$ is the same as $Y_{t_1-k}, Y_{t_2-k}, \ldots, Y_{t_n-k}$, for all $t \in T$, $k \in \mathbb{Z}$.*

The definition expresses the idea that any shift by time k does not affect the joint distribution of the process. It only depends on time $t_k \in T$.

Note that this condition is very strict, especially for processes with large n. Therefore, a weaker form of stationarity is preferred in many engineering disciplines – **weak-sense stationarity** (WSS). Whenever the term "stationarity" is used in this thesis, its weak variant is considered.

Definition 4 (Weak-sense stationarity) *A stochastic process Y is said to be* ***weakly stationary*** *if*

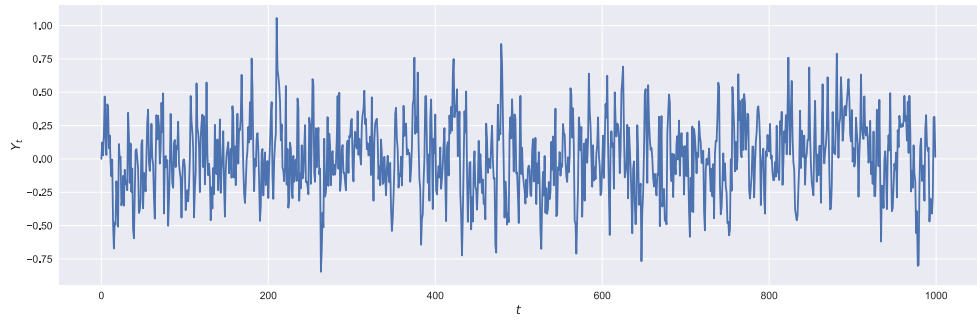
1. *The first moment (mean) function is constant over time.*
2. $\gamma_{t,\,t-k} = \gamma_{0,\,k}$ *for all time $t \in T$ and lag $k \in \mathbb{N}$.*

If Y is a weakly stationary process then the mean μ and the variance σ^2 are time-independent, and further the ACVF function depends only on the lag between t_1 and t_2 [1]. This implies that the autocovariance and autocorrelation functions can be expressed as functions of a time-lag $\tau = t_2 - t_1$ only. Considering this, the autocovariance function (1.3) can be rewritten as

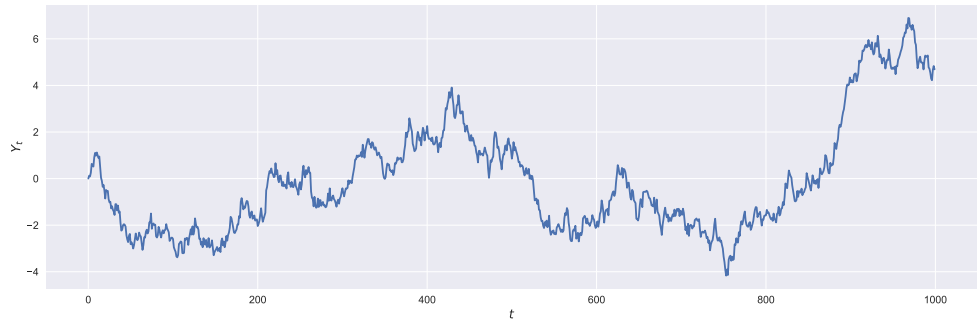
$$\gamma(\tau) = Cov(Y_t, Y_{t+\tau}), \qquad t \in T, \tag{1.5}$$

and, similarly, the autocorrelation function (2.9)

$$\rho(\tau) = Corr(Y_t, Y_{t+\tau}), \qquad t \in T. \tag{1.6}$$



(a) Stationary time series



(b) Non-stationary time series

Figure 1.2: In Figure (a) we can see an example of a stationary process. Values of the series are concentrated around zero and its mean remains the same. In Figure (b) we can see a non-stationary process. In fact, this is an example of a random walk.

Moreover, the weak stationarity of Y implies that ACVF and ACF are *even*.

$$\gamma(\tau) = \mathbb{E}[Y_t Y_{t+\tau}] = \mathbb{E}[Y_{t+\tau} Y_t] = \mathbb{E}[Y_t Y_{t-\tau}] = \mathbb{E}[Y_t Y_{t+(-\tau)}] = \gamma(-\tau).$$

It can be shown that any strictly stationary process which has a finite mean and covariance is also WSS [2].

1.2 Autoregressive processes

The autoregressive process is a representation of a type of stochastic process. Autoregression specifies that the output variable at time t depends linearly on its own p most recent past values, and the noise term. The number p is called the **model order**. Let us define the autoregressive process.

Definition 5 (Autoregressive process of order p) *The autoregressive process of order p denoted by AR(p) is defined as*

$$\begin{aligned} Y_t &= \phi_0 + \phi_1 Y_{t-1} + \ldots + \phi_p + \varepsilon_t \\ &= \phi_0 + \sum_{k=1}^{p} \phi_k Y_{t-k} + \varepsilon_t. \end{aligned} \tag{1.7}$$

Using the matrix notation we can rewrite it as follows

$$\begin{aligned} Y_t &= \begin{bmatrix} \phi_0 \\ \phi_1 \\ \vdots \\ \phi_p \end{bmatrix}^\intercal \begin{bmatrix} 1 \\ Y_{t-1} \\ \vdots \\ Y_{t-p} \end{bmatrix} + \varepsilon_t \\ &= \phi^\intercal \xi_t + \varepsilon_t, \end{aligned} \tag{1.8}$$

where $\phi = [\phi_0, \phi_1, \ldots, \phi_p]$ *is the vector of regression coefficients and* ε_t *stands for the noise term.*

We add the constant ϕ_0 to every Y_t value, so it is an intercept which shifts the series along Y-axis. Later, we will show that when the process Y has the coefficient $\phi_0 = 0$, the mean $\mathbb{E}[Y] = 0$. The noise term ε_t incorporates everything new in the series at time t which cannot be described by the past values [1]. However, the term must satisfy $\mathbb{E}[\varepsilon_t] = 0$.

The assumption that the current value depends on the previous ones made those processes very popular among different disciplines. The original work on autoregressive processes was carried out by George Udny Yule in 1926 [3].

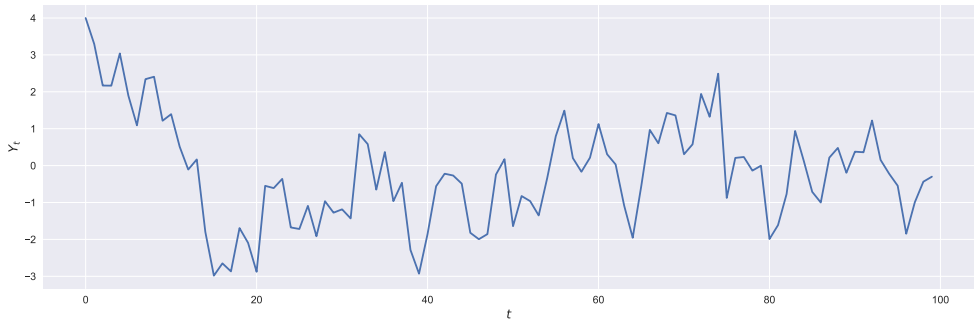


Figure 1.3: An example of AR(2) process with regression coefficients $\phi = [0, -0.3, 0.5]^\intercal$ and noise variance $\sigma^2 = 1$.

Theoretically, the autoregressive process could be of any order p. However, it is reasonable to avoid high-order models, since the impact of past values

is vanishing with an increase of the time lag. We will go deeper into the subtleties of AR modelling in the next chapter.

There is another way to write AR process using **backshift operator** [4], denoted B. It operates on the time index of a series and shifts it one time unit back. Mathematically,

$$BY_t = Y_{t-1}.$$

The backshift operator can be raised to arbitrary integer powers so that

$$\begin{aligned} B^{-1}Y_t &= Y_{t+1} \\ B^k Y_t &= Y_{t-k}. \end{aligned}$$

Thanks to these properties, we can rewrite AR(p) process using the backshift operator. Assume that $\mathbb{E}[Y] = 0$, this condition imposes the zero mean of the process and makes the following calculations easier.

$$\begin{aligned} Y_t &= \sum_{k=1}^{p} \phi_k Y_{t-k} + \varepsilon_t \\ \varepsilon_t &= Y_t - \sum_{k=1}^{p} \phi_k Y_{t-k} \\ &= Y_t - \phi_1 B^1 Y_t - \ldots - \phi_p B^p Y_t \\ &= \Big(1 - \sum_{k=1}^{p} \phi_k B^k\Big) Y_t \\ &= \Phi(B) Y_t, \end{aligned}$$

where $\Phi(B)$ is known as the **characteristic polynomial** of the process. Its roots determine whether the process is stationary or not. Let us introduce **stationarity condition** [5].

Theorem 1 (Stationarity condition) *Assume an autoregressive process Y of order p with regression coefficients $\phi = [\phi_0, \phi_1, \ldots, \phi_p]^\intercal$. The process Y is said to be weakly stationary if the roots of its characteristic polynomial $\Phi(B) = 1 - \sum_{k=1}^{p} \phi_k B^k$ lie outside the unit circle[1], that is for each root B_k, $k \in \hat{p} : |B_k| > 1$. Otherwise, the process is not stationary.*

1.2.1 White noise

The simplest form of autoregressive process is AR(0). The zero-order autoregressive process is given by

$$Y_t = \varepsilon_t. \tag{1.9}$$

[1] The unit circle is the circle of radius 1 centered at the origin $(0, 0)$ in the complex plane.

That is, the current value is not affected by the past values. In fact, this process is the so-called white noise process. This white noise process is defined as a collection of independent, identically distributed (iid) random variables ε_t. This implies the *strict stationarity* of the process, since the multivariate distribution of iids cannot be affected by any shift. Furthermore, the independence implies the absence of correlation, so the autocovariance function for the white noise process is given by

$$\gamma(\tau) = \begin{cases} \sigma^2 & \text{for } \tau = 0, \\ 0 & \text{for } \tau \neq 0. \end{cases} \tag{1.10}$$

Similarly, the autocorrelation function

$$\rho(\tau) = \begin{cases} 1 & \text{for } \tau = 0, \\ 0 & \text{for } \tau \neq 0. \end{cases} \tag{1.11}$$

1.2.2 The First-Order autoregressive process

The first-order autoregressive process AR(1) is given by

$$Y_t = \phi_0 + \phi_1 Y_{t-1} + \varepsilon_t. \tag{1.12}$$

Let us now investigate its moments, autocorrelation structure, and define when the process is stationary. Let us designate that there is no point in deriving non-stationary process moments, since they are not constant. Therefore, we consider the process to be weakly stationary.

Let us start with the mean. Using Equation 1.1 we get

$$\mathbb{E}[Y_t] = \mathbb{E}[\phi_0 + \phi_1 Y_{t-1} + \varepsilon_t].$$

Using the linearity of the expectation, we can rewrite

$$\begin{aligned} \mathbb{E}[Y_t] &= \mathbb{E}[\phi_0] + \mathbb{E}[\phi_1 Y_{t-1}] + \mathbb{E}[\varepsilon_t] \\ &= \phi_0 + \phi_1 \mathbb{E}[Y_{t-1}] + 0. \end{aligned}$$

With the weakly stationarity assumption, the mean is constant over time, according to Definition 4. Thus, $\mathbb{E}[Y_t] = \mathbb{E}[Y_{t-1}] = \mu$ and we get

$$\mu = \phi_0 + \phi_1 \mu.$$

Extracting μ we get a (theoretical) mean of AR(1) process

$$\mu = \frac{\phi_0}{1 - \phi_1}. \tag{1.13}$$

Now it is clear that the process has zero mean if the coefficient $\phi_0 = 0$.

Variance of the AR(1) process is derived the same way. By independence of the errors ε_t and the values Y_t, we get

$$\begin{aligned} var(Y_t) &= var(\phi_0) + var(\phi_1 Y_{t-1}) + var(\varepsilon_t) \\ &= \phi_1^2 var(Y_{t-1}) + \sigma_\varepsilon^2. \end{aligned}$$

By the stationarity assumption, $var(Y_t) = var(Y_{t-1}) = \sigma^2$. So,

$$\sigma^2 = \phi_1^2 \sigma^2 + \sigma_\varepsilon^2.$$

Solving this equation for σ^2 we get the variance of AR(1) process

$$\sigma^2 = \frac{\sigma_\varepsilon^2}{1 - \phi_1^2}. \tag{1.14}$$

Since the variance is non-negative moment, it follows that $1 - \phi_1^2 > 1$ and therefore $|\phi_1| < 1$. We will face this condition later in defining the stationarity regions for AR(1).

Let us now derive the autocovariance function. For simplicity and without loss of generality, let us assume $\phi_0 = 0$. The autocovariance function for a unit time lag reads

$$\begin{aligned} \gamma_1 = Cov(Y_t Y_{t+1}) &= \mathbb{E}[Y_t Y_{t+1}] \\ &= \mathbb{E}[Y_t(\phi_1 Y_t + \varepsilon_{t+1})] \\ &= \mathbb{E}[\phi_1 Y_t^2 + Y_t \varepsilon_{t+1}] \\ &= \phi_1 \mathbb{E}[Y_t^2] + \mathbb{E}[Y_t \varepsilon_{t+1}]. \end{aligned}$$

Since Y_t is independent of ε_t for any t, and $\mathbb{E}[Y_t] = 0$ we get

$$\gamma_1 = \phi_1 \mathbb{E}[Y_t^2] + \mathbb{E}[Y_t]\mathbb{E}[\varepsilon_{t+1}] = \phi_1 \mathbb{E}[(Y_t - 0)^2] = \phi_1 var(Y_t). \tag{1.15}$$

Using Equation 2.9 and assuming that the process is stationary, we get

$$\rho_1 = Corr(Y_t Y_{t+1}) = \frac{Cov(Y_t Y_{t+1})}{\sigma_t \sigma_{t+1}} = \frac{Cov(Y_t Y_{t+1})}{var(Y_t)} = \frac{\phi_1 var(Y_t)}{var(Y_t)} = \phi_1. \tag{1.16}$$

Note that the value of the autocorrelation function at lag 1 is equal to ϕ_1, which is pretty natural taking into account the definition of AR(1) model. Furthermore, ρ_1 is bounded by $\langle -1, 1 \rangle$ by definition, thus ϕ_1 is also bounded and cannot exceed 1 in absolute value. We will describe the nature of this constraint later.

To find the autocovariance γ_h between the observations h time periods apart, we multiply each side of the process for Y_t by Y_{t-h}, then take expectations. Note that $h = |t_2 - t_2| \geq 0$.

$$\begin{aligned} Y_t &= \phi_1 Y_{t-1} + \varepsilon_t \\ Y_{t-h} Y_t &= \phi_1 Y_{t-h} Y_{t-1} + Y_{t-h} \varepsilon_t \\ \mathbb{E}[Y_{t-h} Y_t] &= \mathbb{E}[\phi_1 Y_{t-h} Y_{t-1}] + \mathbb{E}[Y_{t-h} \varepsilon_t] \\ \gamma_h &= \phi_1 \gamma_{h-1}. \end{aligned}$$

We can see the recursive nature of the autocovariance function and if we start at γ_1, and move recursively forward we get

$$\gamma_h = \phi_1^h \gamma_0. \tag{1.17}$$

By definition $\gamma_0 = var(Y_t)$, so this gives us

$$\gamma_h = \phi_1^h var(Y_t). \tag{1.18}$$

Since the autocorrelation function is just normalized autocovariance by variance, we get

$$\rho_h = \frac{\phi_1^h var(Y_t)}{var(Y_t)} = \phi_1^h. \tag{1.19}$$

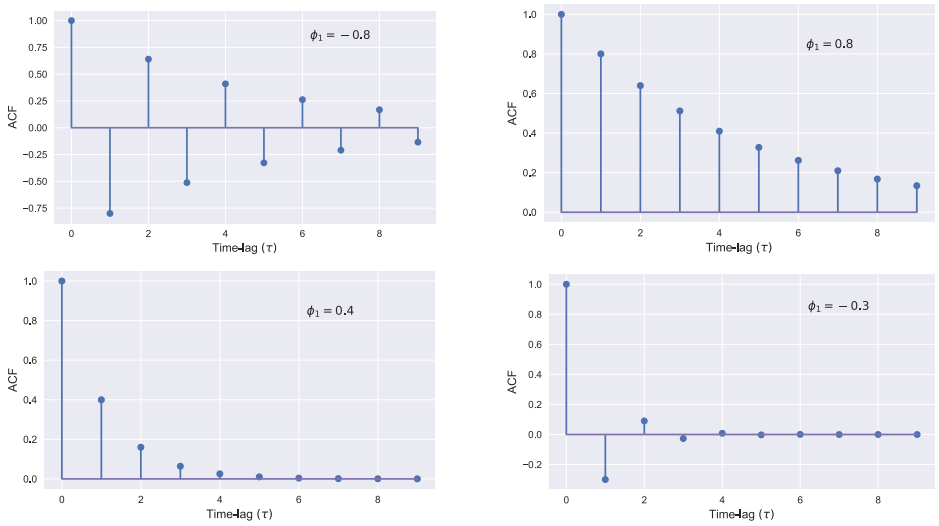


Figure 1.4: Examples of several autocorrelation functions generated by AR(1) processes with different regression coefficients.

In Figure 1.4 examples of ACF for AR(1) processes with different coefficients are displayed. First, note that for processes with $0 > \phi_1 > 1$, all correlations

are positive, whereas the signs of ACF values for processes with $-1 < \phi_1 < 0$ alternate from positive to negative. In both cases, the magnitude of autocorrelations decreases exponentially as the lag τ increases. Second, it is clear that the magnitude of ACF values is proportional to $|\phi_1|$, that is, the higher the value of ϕ_1 in the absolute value, the stronger the correlation [1].

To decide whether the process is stationary or not, we use Theorem 1. The characteristic polynomial for AR(1) process is

$$\Phi(B) = 1 - \phi_1 B, \tag{1.20}$$

the corresponding characteristic equation is

$$0 = 1 - \phi_1 B.$$

This equations has only one root

$$B = \frac{1}{\phi_1},$$

which must lie outside the unit circle, so

$$\left|\frac{1}{\phi_1}\right| > 1 \Rightarrow |\phi_1| < 1 \tag{1.21}$$

Theorem 2 (Stationarity condition for AR(1)) *The AR(1) process is stationary if only if* $|\phi_1| < 1$ *or* $-1 < \phi_1 < 1$.

Recall this condition from the derivation of variance. We assumed that the process is stationary and unintentionally got the correct stationary condition.

1.2.3 The Second-order Autoregressive Process

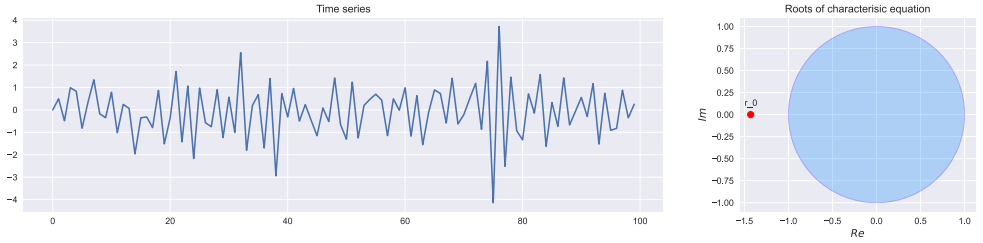
AR(2) process is given by

$$Y_t = \phi_0 + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \varepsilon_t \tag{1.22}$$

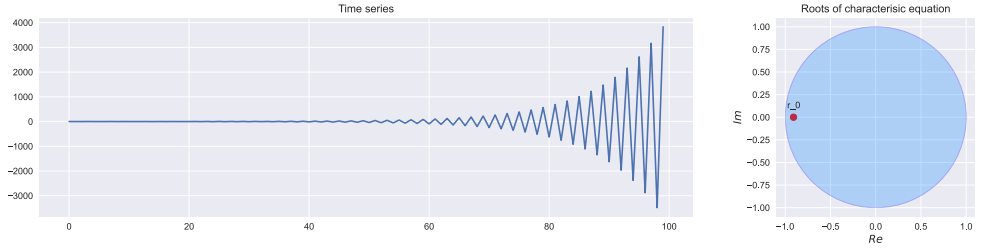
Derivation of moments of the AR(2) process is similar, so let us concentrate on the stationarity of the process. The characteristic equation of the AR(2) process is given by

$$\Phi(B) = 1 - \phi_1 B - \phi_2 B^2. \tag{1.23}$$

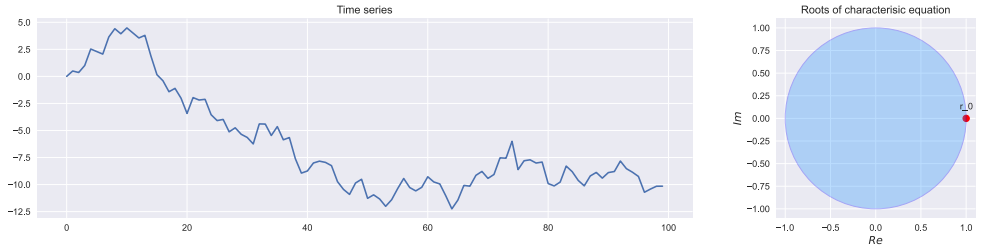
In order to ease the calculations, we introduce **poles**.



(a) $\phi_1 = -0.7$



(b) $\phi_1 = -1.1$



(c) $\phi_1 = 1$

Figure 1.5: Examples of the AR(1) processes with the roots of corresponding characteristic equations. Recall, that the roots of characteristic equation must lie outside the unit circle (blue area) in order the process to be stationary. In Figure (a) there is an example of a stationary process. The process in Figure (b) is called explosive. In Figure (c) the value of the root is 1, thus the corresponding process is not stationary.

Definition 6 (Zeros and Poles) *A **zero** of a function f is a complex number z such that $f(z) = 0$. A **pole** of f is a zero of $\frac{1}{f}$.*

In this situation, use of poles is more convenient, because solving the reduced quadratic equation (when the leading coefficient is one) is easier. According to the Theorem 1, the roots of $\Phi(B)$ must lie outside the unit circle so that the process is stationary, therefore, the poles must lie **inside** the unit circle. So we will be looking for zeros of $\frac{1}{\Phi(B)} = 0$. In fact, the polynomial $\frac{1}{\Phi(B)}$ is usually referred as **reciprocal polynomial**. The reciprocal equation in this case is following

$$0 = \lambda^2 - \phi_1\lambda - \phi_2, \tag{1.24}$$

reciprocal zeros are

$$\lambda_{1,2} = \frac{\phi_1 \pm \sqrt{\phi_1^2 + 4\phi_2}}{2}. \tag{1.25}$$

According to stationarity condition, the poles $|\lambda_1| \leq 1$ and $|\lambda_2| \leq 1$. This yields

$$|\lambda_1\lambda_2| = \left| \frac{\phi_1 + \sqrt{\phi_1^2 + 4\phi_2}}{2} \cdot \frac{\phi_1 - \sqrt{\phi_1^2 + 4\phi_2}}{2} \right| \tag{1.26}$$

$$= \left| \frac{\phi_1^2}{4} - \frac{\phi_1^2 + 4\phi_2}{4} \right| = |\phi_2| \leq 1, \tag{1.27}$$

and

$$|\lambda_1 + \lambda_2| = |\phi_1| \leq 2. \tag{1.28}$$

Now let us divide proof into two cases corresponding to real and complex roots. The roots will be real if and only if $\phi_1 + 4\phi_2 \geq 0$. Consider that roots are real. This implies

$$-1 \leq \frac{\phi_1 - \sqrt{\phi_1^2 + 4\phi_2}}{2} \leq \frac{\phi_1 + \sqrt{\phi_1^2 + 4\phi_2}}{2} \leq 1 \tag{1.29}$$

$$-2 \leq \phi_1 - \sqrt{\phi_1^2 + 4\phi_2} \leq \phi_1 + \sqrt{\phi_1^2 + 4\phi_2} \leq 2. \tag{1.30}$$

Consider the first inequality. Now $-2 \leq \phi_1 - \sqrt{\phi_1^2 + 4\phi_2}$ if and only if $\sqrt{\phi_1^2 + 4\phi_2} \leq \phi_1 + 2$ if and only if $\phi_1^2 + 4\phi_2 \leq \phi_1^2 + 4\phi_1 + 4$. This yields

$$\phi_2 - \phi_1 \leq 1. \tag{1.31}$$

The second inequality is treated similarly and leads to

$$\phi_2 + \phi_1 \leq 1. \tag{1.32}$$

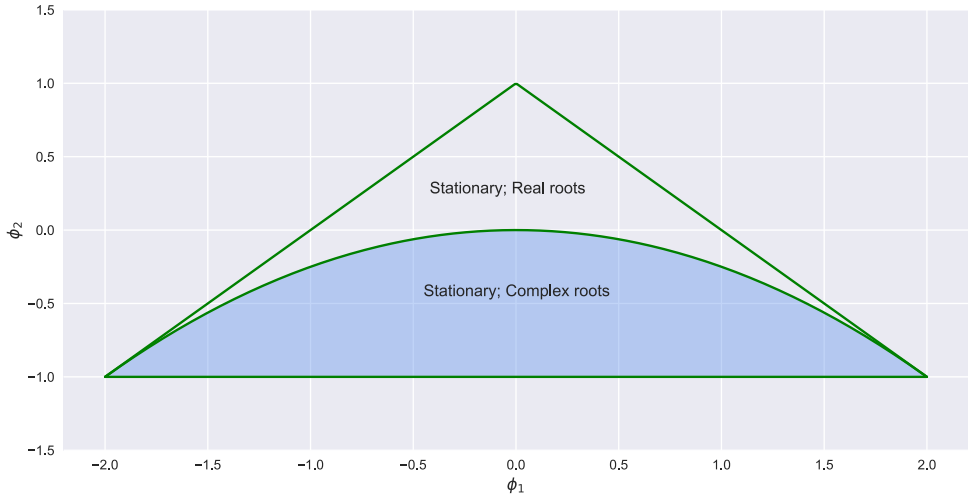


Figure 1.6: The stationarity regions for AR(2) process. For the values of $\phi_{1,2}$ outside this triangle AR(2) process will be explosive.

These inequalities together with $\phi_1 + 4\phi_2 \geq 0$ define the region of stationarity for the real root case.

Now consider complex roots, i.e. $\phi_1^2 + 4\phi_2 < 0$. Here λ_1 and λ_2 will be complex conjugates and $|\lambda_1| = |\lambda_2| \leq 1$ if and only if $|\lambda_2|^2 \leq 1$. However,

$$|\lambda_2|^2 = \frac{\phi_1^2 - \phi_1^2 - 4\phi_2}{4} = -\phi_2, \tag{1.33}$$

so that $\phi_2 \geq -1$. This together with the inequality $\phi_1^2 + 4\phi_2 < 0$ defines the part of the stationary region for complex roots. The stationarity region for AR(2) process is displayed in Figure 1.6.

Investigating the properties of higher-order processes is more complex and requires deeper analysis.

1.3 AR processes in frequency domain

We can consider representing the variability in a time series in terms of harmonic components at various frequencies. This is a natural way to look at a time series and it is called **frequency domain analysis**. The aim of the analysis is to estimate the strength of different frequency components (power

spectrum) of time-domain process. For instance, a simple model for a time series Y_t exhibiting cyclic fluctuations with a known period p is [6]

$$Y_t = \alpha \cos(\omega t) + \beta \sin(\omega t) + \varepsilon_t,$$

where ε_t is a white noise, $\omega = \frac{2\pi}{p}$ is the known frequency of cyclic oscillations, and parameters α and β.

When we examine the properties of the stochastic process using autocovariances, it is an analysis in the *time domain*, whereas examining properties by considering the frequency components of a process is analysis in the *frequency domain*. This kind of analysis could help us describe the behaviour of the stochastic process in terms of oscillations and seasonality.

1.3.1 Spectrum

Let Y_t be a stationary stochastic process with autocovariances (γ_k). For any such sequence generated by a stationary process, there exists a function F such that

$$\gamma_k = \int_{-\pi}^{\pi} e^{ik\lambda}\, dF(\lambda), \tag{1.34}$$

where F is the unique function called **spectral distribution function** [6] or **spectrum** on $[-\pi, \pi]$ such that

1. $F(-\pi) = 0$
2. F is non-decreasing and right-continuous
3. For any $0 \leq a < b \leq \pi$,

$$F(b) - F(a) = F(-a) - F(-b).$$

The function F has many of the properties of a probability distribution function, however $F(\pi) = 1$ is not required.

If the function F is everywhere continuous and differentiable, then

$$f(\lambda) = \frac{dF(\lambda)}{d\lambda} \tag{1.35}$$

is called the **spectral density function** and we can rewrite Equation 1.34 as

$$\gamma_k = \int_{-\pi}^{\pi} e^{ik\lambda} f(\lambda)\, d\lambda \tag{1.36}$$

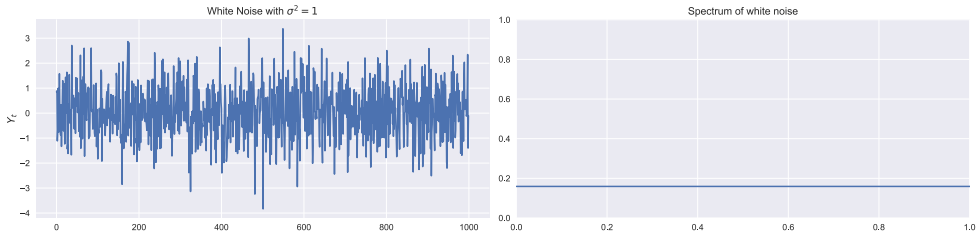


Figure 1.7: Generated white noise with its theoretical spectrum.

If the sum $\sum |\gamma_k|$ is absolutely convergent, then it can be shown that f always exists [6] and is given by

$$f(\lambda) = \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} \gamma_k e^{i\lambda k} = \frac{\gamma_0}{2\pi} + \frac{1}{\pi} \sum_{k=1}^{\infty} \gamma_k \cos(\lambda k). \tag{1.37}$$

Example: White Noise Consider the zero-order autoregressive process $Y_t = \varepsilon_t$, where $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$. As shown in Section 1.2.1, for the process Y_t the autocovariances are $\gamma_0 = \sigma^2$, $\gamma_k = 0$ for any $k \neq 0$. From (1.37) we immediately get

$$f(\lambda) = \frac{\sigma^2}{2\pi}, \tag{1.38}$$

which is independent from λ.

All frequencies receive the equal weight in a spectral representation of white noise and this explains the origin of this name. In physics, the spectrum of white light contains components at all frequencies across the visible spectrum.

Example: AR(1) Spectral Density From (1.15) and (1.17) we know that $\gamma_0 = \frac{\sigma^2}{1-\phi_1^2}$ and $\gamma_k = \phi_1^k \gamma_0$ for $k \geq 0$. Knowing that $e^{-i\lambda} + e^{i\lambda} = 2\cos(\lambda)$, we get

$$\begin{aligned} f(\lambda) &= \frac{1}{2\pi}\gamma_0 \sum_{k=-\infty}^{\infty} \phi_1^{|k|} e^{i\lambda k} \\ &= \frac{\gamma_0}{2\pi} + \frac{\gamma_0}{2\pi} \sum_{k=1}^{\infty} \phi_1^k e^{i\lambda k} + \frac{\gamma_0}{2\pi} \sum_{k=1}^{\infty} \phi_1^k e^{-i\lambda k} \\ &= \frac{\gamma_0}{2\pi}\Big(1 + \frac{\phi_1 e^{i\lambda}}{1-\phi_1 e^{i\lambda}} + \frac{\phi_1 e^{-i\lambda}}{1-\phi_1 e^{-i\lambda}}\Big) \\ &= \frac{\gamma_0(1-\phi_1^2)}{2\pi(1-2\phi_1\cos(\lambda)+\phi_1^2)} \\ &= \frac{\sigma^2}{2\pi(1-2\phi_1\cos(\lambda)+\phi_1^2)}. \end{aligned} \tag{1.39}$$

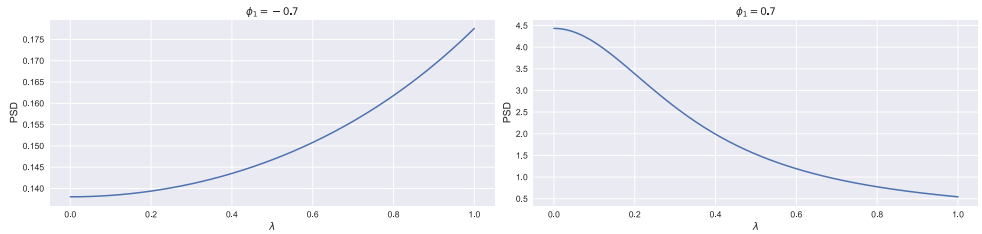


Figure 1.8: Spectra of AR(1) processes with the positive and negative regression coefficient ϕ_1.

In Figure 1.8 we can see, that in the case $\phi_1 > 0$ the spectral density is a decreasing function of λ. That is, the power is concentrated at low frequencies, corresponding to long-range fluctuations [6]. While in the case $\phi_1 < 0$ the spectral density increases as a function. That means, that the power is concentrated at higher frequencies, which explains the tendency to oscillate among AR(1) processes with negative ϕ.

Chapter 2

AR Modelling

In this chapter, we will go through all steps of time series modelling using the autoregressive model. The main goal of time series modelling is *forecasting*, that is predicting the future behaviour of some variable of interest based on past and present data. The most common approach to time series modelling is **Box-Jenkins method** [5], named after after the statisticians George Box and Gwilym Jenkins.

This approach consists of 3 main steps:

1. model specification
2. model fitting
3. model diagnostics

In the first step, different time series models are selected that may be appropriate for a given case. To choose one, we analyze the progress of the series, extract different statistics from the data, and apply any domain knowledge matter in which the data arise, such as economy, biology, or healthcare. Since this thesis is dedicated to autoregressive models, we will skip this phase and will not describe the techniques of choosing an appropriate model for time series, except for one stage of this phase – stationarity testing.

Every model involves parameters whose values must be estimated from the given time series. This step is called *model fitting*. In case of the autoregressive model, we need to estimate the model order p and regression coefficients $\phi = [\phi_0, \phi_1, \ldots, \phi_p]^\intercal$. We will outline the common approaches of parameter estimation and compare the results of individual methods in the next chapter.

The last step consists of assessing the quality of the chosen model with the estimated parameters. We make sure that the model fits the data well and that the assumptions of the model are reasonably well satisfied. If some problems were found, we return to the model specification and repeat the cycle. In this way, we loop through these three steps until we find an acceptable model.

In this chapter we will use the following notation:

- **Set of time indices** $T = \mathbb{N}$
- **Observable series** $X = \{X_t \,|\, t \in T\}$
- **Sample statistics** use hat-notation (i.e., sample mean $\hat{\mu}$ or sample autocorrelation function $\hat{\rho}$)
- **Matrices** will be denoted by bold variables (i.e., $\mathbf{X}$ or $\mathbf{I}$)

2.1 Sample statistics

Now, we consider an autoregressive model that will be initially fitted to some time series X. Thus, instead of deriving the theoretical moments of the process, we can estimate the sample statistics, e.g, the **simple mean** $\hat{\mu}$ of the time series X of length n is given by

$$\hat{\mu} = \frac{1}{n} \sum_{t=1}^{n} X_t, \tag{2.1}$$

and sometimes referred as average.

The **sample variance** is given by

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{t=1}^{n} (X_t - \hat{\mu})^2, \tag{2.2}$$

where the term $\frac{1}{n-1}$ is called *Bessels's correction* [7].

Essential diagnostic tool for examining dependence is the **sample autocorrelation function**. The natural way to compute the sample ACF is to take the sample correlation between the pairs h units apart in time, e.g, $(X_1, X_{1+h}), (X_2, X_{2+h})), \ldots, (X_{k-n}, X_n)$, for each k. However, if we consider the series X to be stationary, this implies a common mean and variance for the series and we define the sample ACF at lag h as [1]

$$\hat{\rho}(h) = \frac{\sum_{t=k+1}^{n} (X_t - \hat{\mu})(Y_{t-k} - \hat{\mu})}{\sum_{t=1}^{n} (Y_t - \hat{\mu})^2}. \tag{2.3}$$

Note that the denominator is a sum of n squared terms while the numerator contains only $n - k$ cross products. For a variety of reasons, this has become the standard definition for the sample autocorrelation function [1].

2.2 Stationarity Tests

When we model some time-dependent process, we need to determine whether the observed time series X is (weakly) stationary or not. It affects the choice of the model. Now, let us show the most common approaches to determining the stationarity of a given time series.

2.2.1 ACF method

Statistically speaking, this method is not a hypothesis test. Usually, the simple ACF computed from a given non-stationary time series could indicate its non-stationarity [1].

In Figure 2.1 you can see that the values of ACF of the non-stationary series do not decrease rapidly as the lag increases. This behaviour is typical for ACF of a non-stationary series [1]. Furthermore, the definition of the sample autocorrelation (2.3) implicitly assumes stationarity. For instance, we use the lagged products of deviations from the overall mean, and the denominator assumes a constant variance over time. Thus, it is not at all clear what the sample ACF is estimating for a non-stationary process. Thus, the general rule is that when the values of ACF are slowly drifting either up or down, there is a high probability of the series being non-stationary.

This method is not as accurate as the following ones, but it might give an idea of the series' stationarity.

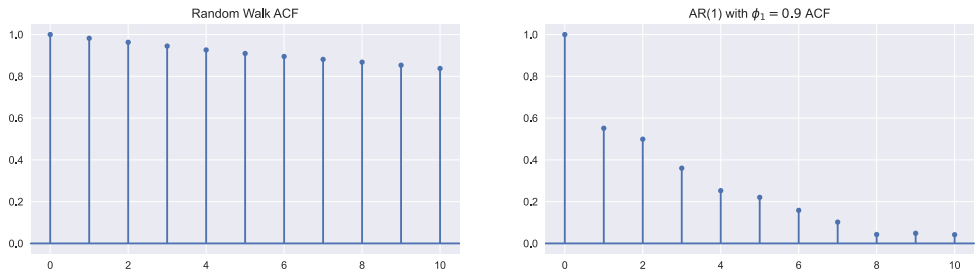


Figure 2.1: Example of ACF for non-stationary (Random Walk) and stationary AR(1) processes.

2.2.2 Augmented Dickey-Fuller Test

Augmented Dickey-Fuller test (ADF) is used to determine the presence of a unit root[2] in the series. Recall the stationarity condition from the Theorem 1. It is clear that the presence of a unit root breaks the condition and implies a non-stationarity of the underlying process.

The intuition behind test hypotheses is the following

H_0 : The process Y has a unit root.

H_1 : The roots of the characteristic equation lie outside the unit circle.

The testing procedure is applied to the model

$$\Delta X_t = \alpha + \beta_t + \gamma X_{t-1} + \delta_1 \Delta X_{t-1} + \ldots + \delta_{p-1} \Delta X_{t-p+1} + \varepsilon_t, \tag{2.4}$$

where α is the constant, β the coefficient on a time trend, and $\Delta X_t = X_t - X_{t-1} = (1 - B)X_t$. Operator Δ is called the *first-difference operator*. The order p could be determined by the techniques described in the Section 2.3, usually the *Akaike information criterion* (AIC) is used (see Subsection 2.3.2).

The hypothesis of the presence of a unit root is equivalent to $\gamma = 0$, that is, when the lagged value X_{t-1} has no direct affect on increment ΔX_t. Only an increment ΔX_{t-1} has an effect on ΔX_t. Thus, the hypotheses are

$$\begin{aligned} H_0 &: \quad \gamma = 0 \\ H_1 &: \quad \gamma < 0. \end{aligned}$$

The test statistic for ADF is

$$DF = \frac{\hat{\gamma}}{SE(\hat{\gamma})}, \tag{2.5}$$

where $\hat{\gamma}$ is the estimation of γ and $SE(\hat{\gamma}) = \frac{\sigma_{\hat{\gamma}}}{\sqrt{n}}$ is the standard error of this estimation. If the calculated statistic DF is less than the critical value for Dickey-Fuller t-distribution, then the null hypothesis is rejected and no unit root is present.

The result of this test is not usually used alone. Instead, we combine the result of ADF with the Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test and then decide whether the observed series X is stationary or not.

Before describing the KPSS test, let us introduce the **trend-stationary processes**.

[2] A stochastic process is said to have a unit root if 1 is a root of the process's characteristic equation.

Definition 7 (Trend-stationary process) *The process Y is said to be trend-stationary if*

$$Y_t = f(t) + Z_t,$$

where $f(t)$ is a trend value of the process at time t and $Z = \{Z_t \,|\, t \in T\}$ is a stationary process.

Note that the process with a trend cannot be stationary since its first moment changes over time. Removing the trend from trend-stationary series gives us a stationary series. There are other techniques which transform the observed series into a stationary one. They will be described later.

2.2.3 Kwiatkowski–Phillips–Schmidt–Shin test

KPPS test is used for testing that an observable time series is stationary around a deterministic trend. The intuition behind the test's hypotheses is

$$\begin{aligned} H_0: &\quad \text{The process Y is trend-stationary.} \\ H_1: &\quad \text{The process Y has a unit-root.} \end{aligned}$$

Note that in this test the hypotheses are inverted. Contrary to most unit root tests, the presence of a unit root is not a null hypothesis but an alternative.

The KPSS test is based on linear regression. It breaks up a series into three parts: a deterministic trend β_t, a random walk r_t, and a stationary error ε_t [8]

$$X_t = r_t + \beta_t + \varepsilon_t, \tag{2.6}$$

where $r_t = r_{t-1} + u_t$ and u_t is white noise with variance σ_u^2.

To test if X_t is a trend stationary process, we use the following hypotheses [9]

$$\begin{aligned} H_0: &\quad \sigma_u^2 = 0. \\ H_1: &\quad \sigma_u^2 > 0. \end{aligned}$$

The null hypothesis means that the intercept is a fixed element. The residuals e_t for $t \in \{1, 2, \ldots, n\}$ are from the regression of X on an intercept and time trend, so $e_t = \varepsilon_t$.

Let the partial sum process of e_t be $S_t = \sum_{j=1}^{t} e_j$, and σ_e^2 be the long-run variance of e_t [9]. Then the KPSS statistics is given by

$$KPSS = \frac{1}{n^2 \sigma_e^2} \sum_{t=1}^{n} S_t^2. \tag{2.7}$$

The p-value is computed then and compared with the significance level. If p-value is less then the chosen significance level, then the null hypothesis is rejected, thus the series is not stationary around the deterministic trend.

Table 2.1: Interpretation of the stationarity tests results.

ADF	KPSS	Result
Insignificant	Insignificant	The series is trend-stationary
Insignificant	Significant	The series is non-stationary
Significant	Insignificant	The series is stationary
Significant	Significant	The series is difference stationary

As we mentioned, ADF and KPSS tests are used together and then we decide whether the observed series is stationary or not. Interpretation of the results is provided in the Table 2.1.

We have mentioned the trend and difference stationary series, that is, the series is stationary after removing the trend, respectively, after differencing. Differencing is a common way to make non-stationary time series stationary by computing the differences between consecutive observations [10] and can be written as

$$X'_t = X_t - X_{t-1}. \tag{2.8}$$

Note that the series X' will have only $n-1$ values, since it is not possible to calculate a difference X'_1 for the first observation. Differencing helps to stabilize the mean of a time series by removing changes in the level of a time series, thus, eliminating trend and seasonality [10].

If the differenced data do not appear to be stationary, it may be necessary to difference the data a second time to obtain a stationary series [10].

$$\begin{aligned} X''_t &= X'_t - X'_{t-1} \\ &= (X_t - X_{t-1}) - (X_{t-1} - X_{t-2}) \\ &= X_t - 2X_{t-1} - X_{t-2}. \end{aligned}$$

Recall that the process should have a constant mean and variance to be stationary. Differencing is a technique that helps to stabilize the mean of the series. To stabilize the variance of the observed series, the logarithm transformation is usually applied [10].

2.3 Model Order

One of the parameters we want to estimate is the model order p. To do that, we need to know how many past values affect the current one. Fortunately, this is what ACF measures. However, we cannot use the ACF plot to detect the model order. The autocorrelations of AR(p) model do not become zero after a certain number of lags – they die off rather than cut off [1]. The value X_t is affected by the X_{t-1}, which is affected by X_{t-2}. This effect propagates to ACF values at higher lags and makes determining the model order from the ACF plot impossible. To get rid of this influence, we introduce the partial autocorrelation function.

2.3.1 Partial autocorrelation function

The partial autocorrelation function (PACF) at lag h measures the correlation between X_t and X_{t-h} after removing the effect of the intervening variables $X_{t-1}, X_{t-2}, \ldots, X_{t-h+1}$.

Definition 8 (Partial autocorrelation function) *Let X be a time series of length n. The partial autocorrelation of lag h r_h is given by*

$$r_h = Corr(X_t, X_{t-h} | X_{t-1}, X_{t-2}, \ldots, X_{t-h+1}).$$

That is, r_h is the correlation in the bivariate distribution of X_t and X_{t-k} conditional on $X_{t-1}, X_{t-2}, \ldots, X_{t-h+1}$.

In Figure 2.2 examples of PACF of simulated processes of 500 samples are provided. Note that for stationary processes of order p, the values of PACF at lags $k \leq p$ are close to the values of regression coefficients and cuts off after lag p, contrary to the ACF. It is remarkable that PACF of non-stationary random walk process is almost equal to 1 at lag 1. In fact, random walk is nothing but AR(1) process with $\phi_1 = 1$ and PACF reflects it.

2.3.2 Akaike information criterion

Another method to determine the parameter p uses Akaike's information criterion. In this method, we fit some models of different orders, e.g., from 1 to 15, and then choose the model that minimizes

$$AIC = 2p - 2\ln(\textit{maximized likelihood}) \approx 2p - 2\ln(\sigma_p^2). \tag{2.9}$$

Originally, the criterion uses the maximized likelihood of the model, however it can be estimated by σ_p^2 [5], where $\sigma_p^2 = \frac{1}{n-1}\sum_{t=1}^{n}(\hat{X} - X)^2$ is the prediction error variance. The term $2p$ serves as a "penalty function" to avoid high-order

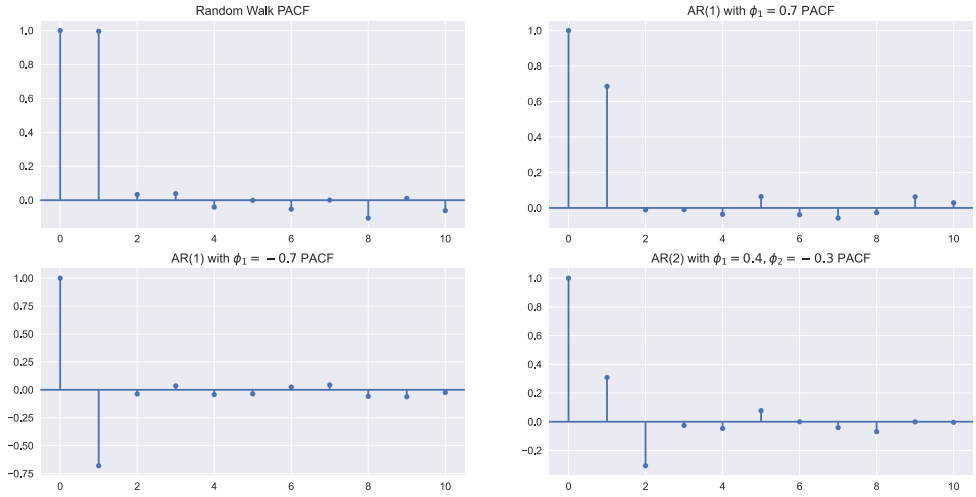


Figure 2.2: Examples of Partial Autocorrelation function for simulated processes of 500 samples.

models. In other words, AIC deals with the trade-off between the goodness of fit of the model and the simplicity of the model.

It has been shown that for a relatively small sample size, AIC underestimates the true order. Nonetheless, the probability of under estimation reduces as the sample size grows [11].

2.4 Fitting AR Model

After the model order $\hat{p}$ is determined, we want to estimate the regression coefficients $\phi = [\phi_0, \phi_1, \ldots, \phi_{\hat{p}}]^\intercal$. The most common approaches to the estimation of p are the ordinary least squares method and the method of moments, which uses Yule-Walker equations. We will describe both of them and in the following chapter will compare the results of these methods.

2.4.1 Least-Squares method

This method of parameter estimation uses least-squares minimization technique. Our goal is to find the vector of regression coefficients $\hat{\phi}$, such that the values of the error sequence $\varepsilon = [\varepsilon_1, \ldots, \varepsilon_n]^\intercal$ are as minimal as possible. Recall, that AR(p) process is given by

$$X_t = \phi_0 + \sum_{k=1}^{p} \phi_k X_{t-k} + \varepsilon_t.$$

If we express the noise term we get

$$\varepsilon_t = X_t - \phi_0 - \sum_{k=1}^{p} \phi_k X_{t-k}. \tag{2.10}$$

Let us convert this expression to matrix notation

$$\begin{pmatrix} \varepsilon_{p+1} \\ \varepsilon_{p+2} \\ \vdots \\ \varepsilon_n \end{pmatrix} = \begin{pmatrix} X_{p+1} \\ X_{p+2} \\ \vdots \\ X_n \end{pmatrix} - \begin{pmatrix} 1 & X_p & X_{p-1} & \dots & X_1 \\ 1 & X_{p+1} & X_p & \dots & X_2 \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & X_{n-1} & X_{n-2} & \dots & X_{n-p} \end{pmatrix} \begin{pmatrix} \phi_0 \\ \phi_1 \\ \vdots \\ \phi_p \end{pmatrix} \tag{2.11}$$

$$\varepsilon = \mathbf{Y} - \mathbf{X}\phi. \tag{2.12}$$

Our aim is to find the vector of regression coefficients $\hat{\phi}$ which minimizes the squared error. The function we want to minimize is given by

$$RSS(\hat{\phi}) = \sum_{t=p+1}^{n} \varepsilon_t^2 = \varepsilon^\intercal \varepsilon \tag{2.13}$$

$$= \|\varepsilon\|^2, \tag{2.14}$$

$$= \|\mathbf{Y} - \mathbf{X}\hat{\phi}\|^2. \tag{2.15}$$

The optimization problem can be then expressed as

$$\hat{\phi} = \arg\min_{\phi \in \Phi} RSS(\phi). \tag{2.16}$$

The optimum regression coefficients can be obtained by applying an *orthogonality principle* [12]. Let X be an unknown random vector which is to be estimated based on the observation vector Y. One wishes to construct a linear estimator $\hat{X} = \mathbf{M}Y + \varepsilon$ for some matrix $\mathbf{M}$ and vector ε. Then, the orthogonality principle states that an estimator $\hat{X}$ achieves the minimum mean square error if and only if

- $\mathbb{E}[(\hat{X} - X)Y^\intercal] = 0$
- $\mathbb{E}[\hat{X} - X] = 0$.

Thus, to minimize MSE, we want the vector ε to be orthogonal to each explanatory vector $X_{\bullet,k}$ for $k \in \{1, \dots p+1\}$, i.e., to each column vector in matrix $\mathbf{X}$. This yields

$$\mathbf{X}^\intercal \varepsilon = \mathbf{X}^\intercal(\mathbf{Y} - \mathbf{X}\phi) = 0 \tag{2.17}$$

$$\mathbf{X}^\intercal \mathbf{Y} - \mathbf{X}^\intercal \mathbf{X}\phi = 0 \tag{2.18}$$

$$\mathbf{X}^\intercal \mathbf{X}\phi = \mathbf{X}^\intercal \mathbf{Y} \tag{2.19}$$

$$\hat{\phi} = (\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{X}^\intercal \mathbf{Y}. \tag{2.20}$$

Equation (2.20) is usually referred to as the *normal equation*.

Note that matrix $\mathbf{X}^\intercal\mathbf{X}$ is not necessarily regular, thus might not be invertable. In such cases, the *Moore-Penrose inverse* $(\mathbf{X}^\intercal\mathbf{X})^+$ is used, which is the most common generalization of the inverse matrix [13].

Recursive least squares estimation In time series analysis, it is natural that as time passes we get more data and need to tweak the model parameters. We can completely recalculate the estimate, but if the number of measurements becomes large, then the computational effort could become prohibitive [14]. So, instead of recomputing, we rather adjust our model parameters to new data. This approach is called *online learning*. The online learning approach applicable for this particular problem is called *recursive least squares* estimation. Let us describe the algorithm of recursive estimation.

Recursive least squares estimation algorithm [14]

1. Initialize the estimator as follows:

$$\begin{aligned}\hat{\phi} &= \mathbb{E}[\phi] \\ P_0 &= \mathbb{E}[(\phi - \hat{\phi}_0)(\phi - \hat{\phi})^\intercal],\end{aligned}$$

   where P_0 is estimation-error covariance.

2. For $k = 1, 2, \ldots$ perform the following.

   a) Obtain the measurements y_k, assuming that y_k is given by the equation

$$y_k = \mathbf{X}_k\phi + \varepsilon_k,$$

   where ε_k is zero-mean random vector with variance σ_k^2. Further, assume that the measurement noise at each time step k is independent, that is, $\mathbb{E}[\varepsilon_i\varepsilon_k] = \sigma_k^2\delta_{i,k}$, where $\delta_{i,k}$ is the Kronecker delta function. This implies that the measurement noise is white.

   b) Update the estimate of ϕ and the estimation-error covariance as follows:

$$\begin{aligned}K_k &= P_{k-1}\mathbf{X}_k^\intercal(\mathbf{X}_k P_{k-1}\mathbf{X}_k^\intercal + \sigma_k^2)^{-1}, \\ \hat{\phi}_k &= \hat{\phi}_{k-1} + K_k(y_k - \mathbf{X}_k\hat{\phi}_{k-1}), \\ P_k &= (\mathbb{I} - K_k\mathbf{X}_k)P_{k-1}(\mathbb{I} - K_k\mathbf{X}_k)^\intercal + K_k\sigma_k^2K_k^\intercal,\end{aligned}$$

   where K_k is estimator gain matrix, term $(y_k - \mathbf{X}_k\hat{\phi}_{k-1})$ is called correction term, and $\mathbb{I}$ is identity matrix.

This method allows us to adjust the previous estimation of $\hat{\phi}$ to the new data as we get it. This makes this model applicable to problems with real-time processing of streaming data.

2.4.2 Method of moments

This method uses Yule-Walker equations. These equations relate the autocorrelation sequence for a time series to the model parameters ϕ [3].

Definition 9 (Yule-Walker equation) *Let Y_t be a stochastic process with autocovariances $(\gamma)_\tau$ and noise variance σ_ε^2. The Yule-Walker equations are given by*

$$\gamma_t = \sum_{k=1}^{p} \phi_k \gamma_{m-k} + \sigma_\varepsilon^2 \delta_{t,0}, \tag{2.21}$$

where $\delta_{m,0}$ is the Kronecker delta function.

These equations express the relationship between the autocovariance function and model parameters. Note that these equations use the theoretical values of ACVF, which should be replaced by estimates $\hat{\gamma}_t$ during the modelling process. The term $\sigma_\varepsilon^2 \delta_{t,0}$ is only present in equation for γ_0, thus the equations for γ_t, where $t > 0$ can be solved in matrix form

$$\begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_p \end{pmatrix} = \begin{pmatrix} \gamma_0 & \gamma_{-1} & \cdots & \gamma_{1-p} \\ \gamma_1 & \gamma_0 & \cdots & \gamma_{2-p} \\ \vdots & \vdots & \ldots & \vdots \\ \gamma_{p-1} & \gamma_{p-2} & \cdots & \gamma_0 \end{pmatrix} \begin{pmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_p \end{pmatrix} \tag{2.22}$$

$$\mathbf{r} = \mathbf{R}\phi. \tag{2.23}$$

Recall, that $\gamma_{-t} = \gamma_t$, which implies the symmetry of matrix $\mathbf{R}$. Similarly, multiplying expression (2.23) by $\mathbf{R}^{-1}$ gives us the estimation of the optimal regression coefficients

$$\hat{\phi} = \mathbf{R^{-1}r}. \tag{2.24}$$

The recursive method generally used to solve the Yule-Walker equations is called Levinson-Durbin algorithm [15]. The advantage of the recursive method is that it produces solutions for all model orders lower than chosen $\hat{p}$, thus making the the selection of $\hat{p}$ easier.

Either of the methods may be used to estimate the model parameters. There are also other methods to extract reasonable estimates of the model parameters from a sequence of data, but all other methods are derived from these two basic methods [16].

If the regression coefficients are constant, i.e., time-invariant, then the classical estimation methods presented in this thesis are used. If they vary slowly with time and their evolution model is unknown, then either the windowing or forgetting methods are used [17]. An example of windowing method is provided in Section 3.5. If time evolution model for parameters is known, then the state-space methods based on Kalman filtering or H_∞ filtering should be used [15].

2.5 Forecasting

One of the main reasons of time series modelling is forecasting. We want to predict the future development of the observed process. Forecasting using an autoregressive model is straightforward, since the current value of the series is a linear combination of its lagged values. After we estimate the model order $\hat{p}$ and the vector of regression coefficients $\hat{\phi}$, we can recursively predict the future values applying the Definition 5 as follows

$$X_{t+1} = \sum_{k=1}^{\hat{p}} \hat{\phi} X_{t-k+1}. \tag{2.25}$$

Note that after $\hat{p}$ predictions, further forecasting will be based on predicted values, not the real ones. An example of forecasting with differnced time series is shown in Figure 3.5. However, an autoregressive models are barely used for modelling real-world processes, which usually have trend and seasonality terms. These terms are not considered in AR models. Instead, more general models (which contain an AR model) should be used for this kind of modelling, like ARIMA (Autoregressive integrated moving average) or SARIMA (Seasonal autoregressive integrated moving average) models.

2.6 Diagnostics

Estimation of the order p and regression coefficients is followed by model model diagnostics. The tests are usually based on the statistical properties of the prediction errors. The most commonly tested properties are whiteness and normality [16]. The correlation of residuals is also computed and checked.

Let us briefly introduce the most common methods of model diagnostics

2.6.1 Q-Q plot

Normality of the residuals can be checked by plotting the quantile-quantile (Q-Q) plot. The plot displays the quantiles of data against theoretical quantiles of standardized normal distribution. If the data are normally distributed, the

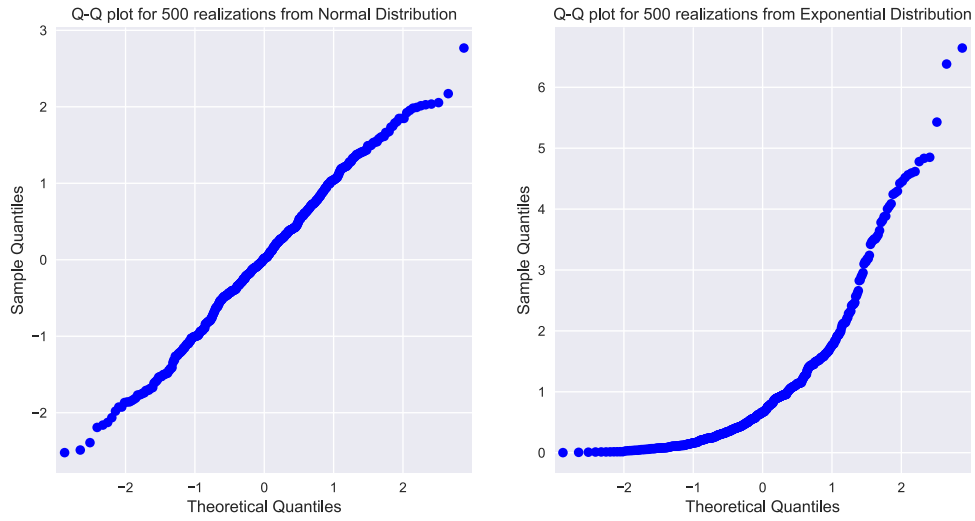


Figure 2.3: Examples of Q-Q plots.

Q-Q plot should look like a straight line. Figure 2.3 shows the examples of Q-Q plot for realizations of normal and exponential distribution.

2.6.2 Jarque-Bera test

Another method for testing the normality of residuals is Jarque-Bera test. The test is based on the fact that a normal distribution has zero skewness and zero kurtosis [1]. Test considers the following hypotheses

$$\begin{aligned} H_0: &\quad \text{Data are normally distributed.} \\ H_1: &\quad \text{Data follow some other distribution.} \end{aligned}$$

Test statistic is based on the two mentioned moments and is approximately distributed as χ^2_2 distribution [1]. Test statistic is given by

$$JB = n\left(\frac{S^2}{6} + \frac{(K-3)^2}{24}\right), \tag{2.26}$$

where n is the number of samples, S is the sample skewness, and K is a sample kurtosis.

2.6.3 Ljung-Box test

We want to find such a model, so the residuals produced are uncorrelated. If the residuals turn out to be correlated, that means that the data has some dependencies which the model does not consider, thus the predictions are not as accurate as they could be.

The Ljung-Box test is commonly used to check whether any of autocorrelations of a time series is different from zero, or

H_0 : Data are independently distributed.
H_1 : Data exhibit serial correlation.

The test statistic is

$$LB = n(n+2)\sum_{k=1}^{h}\frac{\hat{\rho}_k}{n-k}, \tag{2.27}$$

where n is the number of samples, $\hat{\rho}_k$ is sample autocorrelation at lag k, and h is the number of lags being tested. The statistic LB is approximately distributed as χ^2_h.

The test statistic uses sample autocorrelations. Note, that this is autocorrelation of residuals, not the origianl data. We can plot ACF function for residuals. In ideal case, values of ACF for lags $k > 0$ should be insignificant.

Chapter 3

Application example

This chapter is dedicated to the practical application of the autoregressive model and the comparison between the methods introduced in Chapter 2. We are going to apply those methods to the real-world time series related to the COVID-19 pandemic course in Czech Republic.

It should be noted that the AR model is a very simple model and we do not expect the forecasting accuracy on the real data to be high. We conduct these experiments not to get as accurate results as possible, but to show the usage of the methods and compare them.

For analysis and modelling purposes, we will use Python programming language and packages pandas, matplotlib, and statsmodels.

3.1 Data

For the application example we will use data provided by the Czech Ministry of Health. Data are publicly available and can be obtained from [18].

In this file, there are three columns:

- **datum** - date when measurements were taken
- **prirustkovy_pocet_nakazenych** - new cases for this day
- **kumulativni_pocet_nakazenych** - cumulative sum of the previous column

datum	prirustkovy_pocet_nakazenych	kumulativni_pocet_nakazenych
2020-01-27	0	0
2020-01-28	0	0
...	...	...
2021-05-04	2419	1639265
2021-05-05	1869	1641134

Table 3.1: Structure of the data.

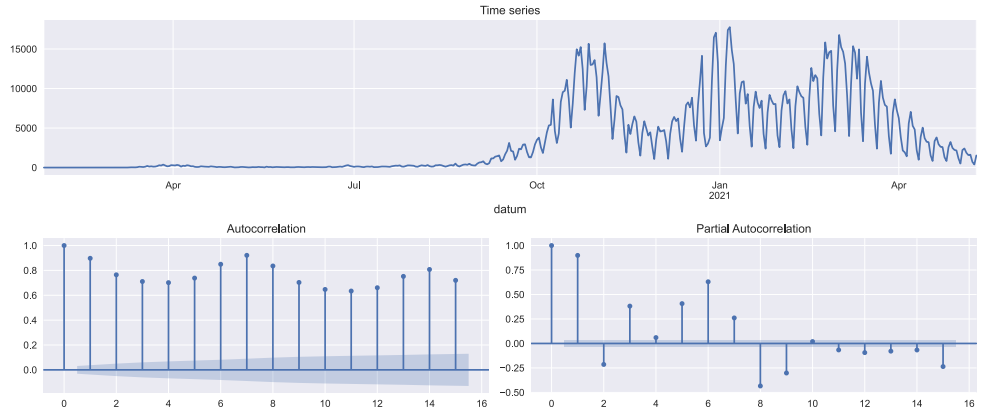


Figure 3.1: Time series X of daily new cases of COVID-19 decease in Czech Republic and corresponding ACF and PACF.

Some data samples are provided in Table 3.1. The column of our interest is prirustkovy_pocet_nakazenych and we will use it as our time series X (see Figure 3.1).

3.1.1 Data properties

Before the modelling process, let us explore the data and get some insight of the underlying structure. The time series with corresponding ACF and PACF are displayed in Figure 3.1. We can see that the observed series has evolving mean and variance. Furthermore, the ACF function gives us an evidence of non-stationarity, since its values do not decrease rapidly as the lag decreases (see Section 2.2.1).

Main statistics extracted from the data are provided by pandas.DataFrame class method *summary()* and are displayed in Table 3.2.

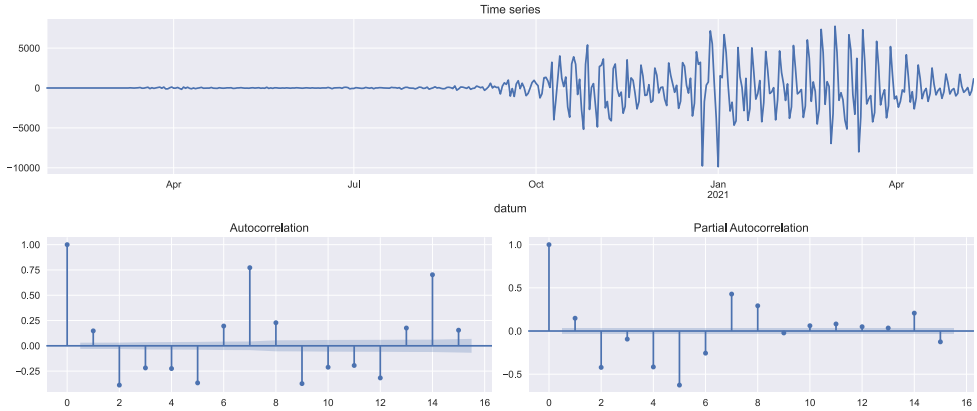


Figure 3.2: Differenced series X' and corresponding ACF and PACF.

3.2 Stationarity testing

We have already mentioned that there is evidence of the series non-stationarity. Let us apply the ADF and KPSS tests described in the Section 2.2. We will use the tests implemented in *statsmodels* package. Since the series X is most likely non-stationary, we will not only test the original series, but the differenced series X' (see Figure 3.2) as well. The results of the tests are provided in Table 3.3.

Let α be the significance level for the tests. Recall that the smaller the p-value of the test is, the stronger the evidence that we should reject the null hypothesis. In our case, let $\alpha = 0.05$, thus in order to reject null hypothesis p-value should be lower than 0.05.

For the original series X, the result of the ADF test is insignificant ($0.12 > \alpha$) and result of the KPSS test is significant ($0.01 < \alpha$). According to Table 2.1, we can assume that the original series X is non-stationary. As for the differenced series X', the null hypothesis of the ADF test is rejected, whereas

count	469
mean	3508.4
std	4523.1
min	0
25%	91
50%	1074
75%	5861
max	17771

Table 3.2: Statistics of the series from *summary()* method.

the result of the KPSS test is insignificant. This implies the stationarity of the differenced series X'.

In this example, we have shown how to make non-stationary series stationary by differencing. Differencing stabilizes the mean, e.g., in our case the mean has changed drastically from 3508.04 (see Table 3.2) to 0.81. However, the variance of X' does not seem to be constant over time, the differenced series is considered stationary and may be used for AR modelling.

Time series	ADF p-value	KPSS p-value
X	0.120	0.010
X'	0.001	0.100

Table 3.3: Results of the stationarity tests.

3.3 Determining order

We are going to determine the optimal model order p for the differenced series. For this purpose, we will use the Akaike information criterion that was introduced in Section 2.3. In *statsmodels* package, there is a function *ar_select_order(data, maxlag, ic)*, which fits the models from order 1 to *maxlag* and uses information criterion *ic*, which is AIC in our case.

Let us not use the whole series, but only the measurements taken between 01.12.2020 and 30.04.2021. Let us denote this shorter series X_{short}. In this interval the behaviour is different from the beginning of the series, and considering that AR models alone do not generalize well, we have decided to limit to this interval.

For the series X_{short} we computed AIC value for orders 1..15. The results are shown in Figure 3.3. The optimal order estimated by AIC is $\hat{p} = 8$ with $AIC(8) = 15.07$.

3.4 Fitting AR model

In this section, we will try two different methods of regression coefficients estimation introduced in Section 2.4. The first method uses OLS estimation and was implemented by the author of this thesis. The method of moments uses the Levinson-Durbin algorithm. Statsmodels implementation was used.

The parameters are going to be estimated on the train set, which is the X_{short} series without the last 50 values. Then we will predict values for the last 50 days and compare the prediction errors. In addition, the estimated coefficients

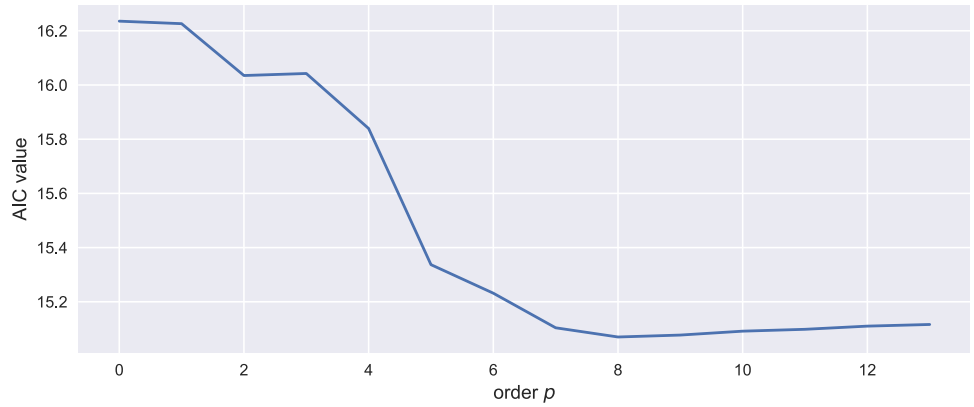


Figure 3.3: Values of AIC for different p.

will be compared. For measuring the prediction error, the RMSE criterion was chosen. RMSE is given by

$$RMSE(X, \hat{X}) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{X}_i - X_i)^2}.$$

The results for both methods are following

OLS Estimation

- $\hat{\phi}_{OLS} = [-0.242, -0.393, -0.264, -0.271, -0.349, -0.166, 0.450, 0.145]^\intercal$
- $RMSE = 3979.84$

Method of moments

- $\hat{\phi}_{LD} = [-0.218, -0.486, -0.355, -0.346, -0.418, -0.232, 0.310, 0.035]^\intercal$
- $RMSE = 3553.74$

In Figure 3.4 the coefficients are plotted next to each other. We can see that the coefficients estimated by OLS tend to be greater than the estimates of the method of moments. Moreover, the coefficients estimated by the Levinson-Durbin algorithm produced lower RMSE on testing. In Figure 3.5 predictions of individual models are shown.

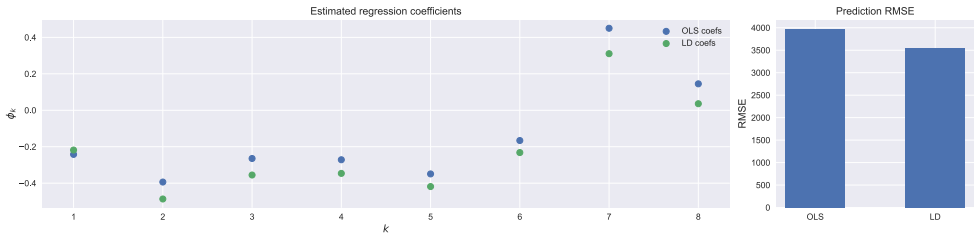


Figure 3.4: Comparison between OLS and Covariance estimation methods. Estimations of ϕ are plotted on the right plot. The left plot shows the prediction RMSE for two methods.

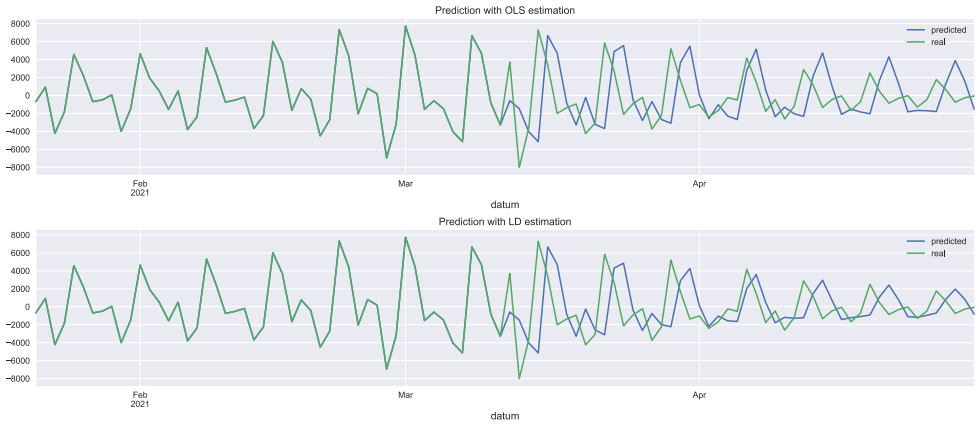


Figure 3.5: Predictions of last 50 points with different estimations of ϕ

3.5 Windowing method forecasting

As we have seen before, autoregressive model is not well suited for forecasting of some real-world processes. To increase the quality of the prediction, we suggest to use a sliding-window based approach.

Let $w_{i,l}(X)$ be a window of length l defined as

$$w_{i,l}(X) = \{X_i, X_{i+1}, \ldots, X_{i+l}\}.$$

Then the forecasting procedure is following

1. Let X be observed time series of length n, l be the width of the window. For $k = 1, \ldots, n - l$,
   a) Fit an AR model on window data $w_{k,l}(X)$ and get coefficients $\hat{\phi}_k$.
   b) Calculate the value $\hat{X}_{k+l}$ by applying the AR model definition.

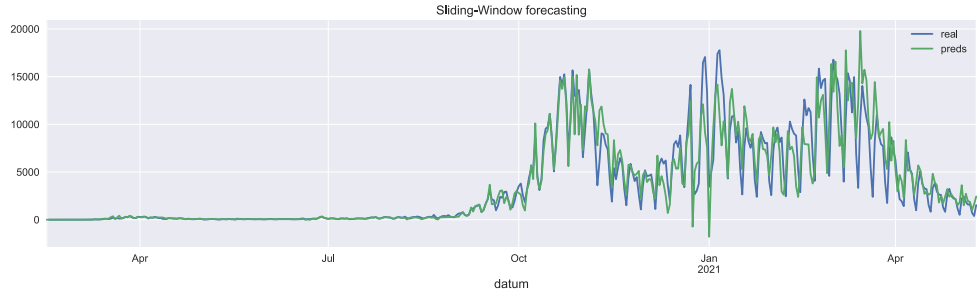


Figure 3.6: Predictions of 450 points produced by sliding-window forecasting technique.

Window Width	RMSE
10	4620.52
14	2146.77
18	2063.63
21	1823.79

Table 3.4: RMSE values for different window widths.

Note that the prediction sequence $\hat{X}$ will be shorter by l values.

We have implemented this approach and applied it to our original COVID-19 time series X. For each window, the order p_k was selected automatically by statsmodels function *ar_select_order*. In Table 3.4 RMSE for different window widths are provided. The negative correlation between the window width and RMSE is obvious. Note that this approach for wider windows outperforms the models trained on differenced data. Predictions are plotted along with the original time series in Figure 3.6.

In Figure 3.7 standardized residuals and their diagnostics are provided. Estimated density remarkably differs from the standard normal distribution. Correlogram, which is an ACF of standardized residuals, shows us some positive correlations between the residuals. Ideally, the residuals should be independent identically distributed. The Ljung-Box test was performed for lags 1-10 (see Table 3.5). For every lag, the p-value is lover than α, so the null hypothesis about independent distribution is rejected. P-value for the Jarque-Bera test is $2.3 \cdot 10^{-153}$, therefore the null hypothesis that the data is normally distributed is also rejected.

These results indicate that our model is missing some dependencies in the data. However, the RMSE notably decreased with sliding-window approach. This indicates that these models are best suited for fast online modelling, whether

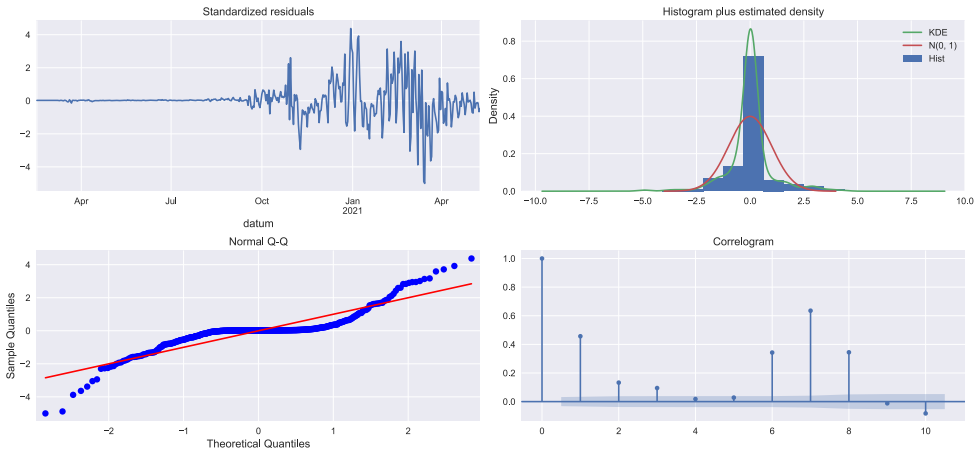


Figure 3.7: Diagnostics of residuals.

lag	LB statistic	LB p-value
1	94.52	$2.42 \cdot 10^{-22}$
2	102.52	$5.46 \cdot 10^{-23}$
3	106.61	$5.85 \cdot 10^{-23}$
4	106.76	$3.55 \cdot 10^{-22}$
5	107.13	$1.65 \cdot 10^{-21}$
6	160.94	$3.74 \cdot 10^{-32}$
7	346.27	$7.73 \cdot 10^{-71}$
8	400.89	$1.20 \cdot 10^{-81}$
9	400.95	$8.55 \cdot 10^{-81}$
10	404.10	$1.25 \cdot 10^{-80}$

Table 3.5: Ljung-Box test results for lags 1-10.

with RLS or sliding-window approaches. Nevertheless, AR model ability to fit the data and produce reliable prediction sequences is low compared to more general (S)AR(I)MA or GARCH (Generalized autoregressive conditional heteroskedasticity) models. The AR model simplicity is its advantage, since it can be used in real-time processing of streaming data, whereas more general models are more computationally and memory expensive, thus, unsuitable for this purpose.

Conclusion

In this thesis, we have studied autoregressive processes and models, which is a fundamental concept in the domain of time series analysis.

In Chapter 1, we have presented the autoregressive process and studied its characteristics with regard to the probability theory. For the processes of low order, some key characteristics were derived. The condition of stationarity was introduced and stationarity regions were defined for the first and second-order autoregressive processes. Furthermore, the theoretical basis of spectral estimation of the stationary stochastic process was outlined. The tendency to oscillate among the first-order autoregressive processes with negative regression coefficient was explained by constructing the corresponding spectra.

Chapter 2 introduced the main techniques of time series modelling with emphasis on AR models. The topics such as determining model order via AIC, regression coefficients estimation with OLS, and the method of moments, and model diagnostics were covered. Statistical tests for determining stationarity were described as well and interpretation of the test results was also provided. Furthermore, we have described an approach for online learning of the AR model (RLS). The model simplicity allows us to use and apply this approach in real-world use cases.

In Chapter 3 an application example was provided. We have used the relevant COVID-19 time series of daily new cases in Czech Republic. To compare the methods of parameter estimation, we trained the model on differenced data and compared the predictions produced by different estimations. Coefficients estimated by the method of moments had lower RMSE. However, this is not the real use case of AR modelling, since the data are too complex for such a simple model. This was performed to show the difference between the two methods, not to get as accurate predictions as possible. Finally, forecast-

ing with sliding-window technique and one-step prediction was implemented. The error sequence produced by this forecasting technique has a remarkably lower RMSE than two previous methods. However, the residuals diagnostics indicated the correlation between residuals.

Future work

The field of time series analysis is very large and in high demand. There is a variety of models to be studied, such as moving-average models, which together with AR models construct (S)AR(I)MA model family. These models are more general and "powerful" in terms of generalization ability and can be used in the modelling of real-world processes. These models take into consideration not only the most past values, but also the development of error, seasonality, and trend.

According to Takalo, Hytti, and Ihalainen, the AR models are used for for high-resolution spectral estimation of a short time series and is preferred to discrete Fourier transform. In biomedical engineering, AR modelling is used especially in the spectral analysis of heart rate variability and electroencephalogram tracings [16]. Spectral estimation requires analysis in the frequency domain and might be considered as an extension to this thesis.

Bibliography

[1] Cryer, J. D.; Chan, K.-S. *Time Series Analysis*. Springer-Verlag New York, second edition, 2008, ISBN 978-0-387-75958-6.

[2] Florescu, I. *Probability and stochastic processes*. John Wiley & Sons, third edition, 2014, ISBN 978-1-118-59313-4.

[3] Yule, U. On a method of investigating periodicities disturbed series, with special reference to Wolfer's sunspot numbers. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 1927: pp. 267–298, ISSN 0264-3952.

[4] Chatfield, C. *The analysis of time series: an introduction*. CRC Press, 6th edition, 2004, ISBN 978-0-203-49168-3.

[5] Box, G.; Jenkins, G. M. *Time Series Analysis: Forecasting and Control*. Holden-Day, fifth edition, 1976, ISBN 978-1-118-67502-1.

[6] Reinert, G. Time series. *Hilary Term, USA*, 2010.

[7] Reichmann, W. J. *Use and Abuse of Statistics*. Penguin, 1964, ISBN 978-0-140-20707-1.

[8] Kwiatkowski, D.; Phillips, P. C.; et al. Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of Econometrics*, 1992: pp. 159–178, ISSN 0304-4076.

[9] Wang, W. *Stochasticity, nonlinearity and forecasting of streamflow processes*. Ios Press, 2006, ISBN 978-1-586-03621-8, 32-33 pp.

[10] Hyndman, R. J.; Athanasopoulos, G. *Forecasting: principles and practice*. OTexts, 2018, ISBN 978-0-987-50711-2.

[11] Liew, V. On autoregressive order selection criteria. *Universiti Putra Malysia*, 2004.

[12] Kay, S. M. *Fundamentals of statistical signal processing*. Prentice Hall PTR, 1993, ISBN 978-8-131-72900-7.

[13] Campbell, S. L.; Meyer, C. D. *Generalized inverses of linear transformations*. SIAM, 2009, ISBN 978-0-898-71904-8.

[14] Simon, D. *Optimal state estimation: Kalman, H infinity, and nonlinear approaches*. John Wiley & Sons, 2006, ISBN 978-0-470-36053-8, 84 - 86 pp.

[15] Durbin, J. The fitting of time-series models. *Revue de l'Institut International de Statistique*, 1960: pp. 233–244.

[16] Takalo, R.; Hytti, H.; et al. Tutorial on Univariate Autoregressive Spectral Analysis. *Journal of Clinical Monitoring and Computing*, 2005: pp. 401–410, ISSN 1573-2614.

[17] Dedecius, K.; Nagy, I.; et al. Parameter tracking with partial forgetting method. *International Journal of Adaptive Control and Signal Processing*, volume 26, no. 1, 2012: pp. 1–12.

[18] Komenda, M.; Karolyi, M.; et al. COVID-19: Přehled aktuální situace v ČR. Onemocnění aktuálně [online]. 2020, [cit. 04 May 2021]. Available from: `https://onemocneni-aktualne.mzcr.cz/api/v2/covid-19`

Appendix A

Acronyms

AR Autoregressive

ACVF Autocovariance function

ACF Autocorrelation function

WSS Wide-sense stationarity

ADF Augmented Dickey-Fuller test

KPSS Kwiatkowski–Phillips–Schmidt–Shin test

OLS Ordinary Least Squares

RLS Recursive Least Squares

Appendix B

Contents of enclosed SD

- `env` ..... the directory with environment configs
  - `requiremnts.txt` ..... required packages
- `notebooks` ..... the directory with Jupyter notebooks
- `source` ..... the directory of source codes
- `thesis` ..... the thesis directory
  - `tex_src` ..... the directory of LaTeX source codes of the thesis
  - `thesis.pdf` ..... the thesis text in PDF format
- `README.txt` ..... the file with SD contents description