Charles University in Prague
Faculty of Mathematics and Physics

BACHELOR THESIS

Marek Šabata

Nonlinear ARMA model

Department of Probability and Mathematical Statistics

Supervisor of the bachelor thesis:	Doc., RNDr. Petr Lachout, CSc.
Study programme:	Mathematics
Specialization:	General Mathematics

Prague 2015

I wish to thank Doc. RNDr. Petr Lachout, Ph.D., who introduced me to the topic of nonlinear ARMA models and provided me with helpful comments and suggestions. I would also like to thank my parents, because without their support it would be impossible for me to study and have the opportunity to write this thesis.

I declare that I carried out this bachelor thesis independently, and only with the cited sources, literature and other professional sources.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Coll., the Copyright Act, as amended, in particular the fact that the Charles University in Prague has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 paragraph 1 of the Copyright Act.

In .... on .... Author signature

Název práce: Nelineární ARMA model

Autor: Marek Šabata

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: Doc., RNDr. Petr Lachout, CSc.,
Katedra pravděpodobnosti a matematické statistiky

Abstrakt: Práce se zabývá teorií lineárních a nelineárních ARMA modelů a jejich aplikací na data finančních trhů.
Nejprve je uveden obecný rámec teorie časových řad. Následně je vyložena teorie lineárních ARMA modelů, která je základním kamenem i pro nelineární modely. Z nelineárních modelů je představen prahový autoregresivní model (TAR), autoregresivní podmíněně heteroskedastický model (ARCH) a zobecněný autoregresivní podmíněně heteroskedastický model (GARCH). U všech modelů je odvozena metoda pro odhad parametrů, jsou odvozeny asymptotické vlastnosti estimátorů a následně spolehlivostní oblasti a intervaly pro testy parametrů.
Teorie je aplikována na finanční data, konkrétně na index Standard and Poor's 500 (S&P500). Všechny modely jsou implementovány ve statistickém softwaru R.
Klíčová slova: Časové řady, ARMA model, Nelineární ARMA model

Title: Nonlinear ARMA model

Author: Marek Šabata

Department: Department of Probability and Mathematical Statistics

Supervisor of the bachelor thesis: Doc., RNDr. Petr Lachout, CSc.,
Department of Probability and Mathematical Statistics

Abstract: The thesis regards theory of nonlinear ARMA models and its application on financial markets data.
First of all, we present general framework of time series modeling. Afterwards the theory of linear ARMA models is layed out, since it plays a key role in the theory of nonlinear models as well. The nonlinear models presented are threshold autoregressive model (TAR), autoregressive conditional heteroscedastic model (ARCH) and generalized autoregressive conditional heteroscedastic model (GARCH). For each model, we derive a method for estimating the model's parameters, asymptotic properties of the estimators and consequently confidence regions and intervals for testing hypotheses about the parameters.
The theory is then applied on financial data, speficically on the data from Standard and Poor's 500 index (S&P500). All models are implemented in statistical software R.
Keywords: Time series, ARMA model, Nonlinear ARMA models

Contents

Introduction

Financial data can be viewed as a realization of a random process. It is therefore natural to employ methods of time series analysis to find a suitable model explaining the data, which could then help us predict the next move. In financial jargon, this area of stock market analysis falls behind technical analysis of stocks.

Among the most popular time series models in general is the linear autoregressive moving average model, or in short linear ARMA model. It is composed of two parts - autoregressive model, which models current value of a stochastic process as a linear combination its past values, and moving average model, which models the current value of the process as a linear combination of past values of a white noise process.

Linear models can be a good first step when analysing any kind of time series data. However, in reality linear models are hardly ever sufficient to capture the nonlinearities clearly present in financial time series. Financial data incorporate lot of nonlinear features such as nonlinear relationship between lagged variables, different growth in periods of recession and expansion, etc. It is therefore suitable to introduce nonlinearities in time series models. At the same time we would like to keep the simple structure of the linear model, which has usually very straightforward interpretation. Combining these two desires, we arrive at nonlinear ARMA models.

The thesis is divided as follows. In the first chapter, we will present general theory concerning time series modeling. In the second chapter, theory of linear ARMA models will be presented, since methods used in linear ARMA models analysis can be considered as building blocks for the nonlinear theory of ARMA models. In the third chapter, we present three nonlinear ARMA models. The first one is so called threshold autoregressive model (TAR), the second is autoregressive conditional heteroscedastic model (ARCH) and the last one is generalized conditional heteroscedastic model (GARCH). In the fourth chapter, we will fit the models on real financial data, namely on daily returns of S&P500 index and analyse soundness of the models. In the conclusion, we summarize the thesis and our findings.

All our analyses will be done in statistical software R. The script is implemented in such way that any stock can be analyzed in the same way as we've analysed the S&P500 index just by downloading the right data set.

Chapter 1

Time series characteristics

In this chapter, we will present general framework of time series analysis and basic tools commonly used in analysing time series data, that will enable us to proceed with statistical inference later on. Most of the definitions and theoreems we present in the thesis are loosely adapted from the books Fan and Yao (2003) and Hamilton (1994).

In the thesis, we suppose that Ω is a nonempty set - sample space, $\mathscr{A}$ is a σ-algebra such that $\mathscr{A} \subset 2^{\Omega}$ and P is a probability measure on $\mathscr{A}$, so $\mathrm{P}(\Omega) = 1$. Therefore $(\Omega, \mathscr{A}, \mathrm{P})$ is a probability space.

For the purpose of our analyses, we will consider only real random variables, that is measurable functions $X : (\Omega, \mathscr{A}) \to (\mathbb{R}, \mathscr{B})$, where $\mathscr{B}$ denotes the Borel σ-algebra of the real line $\mathbb{R}$. If we refer to a stochastic (or random) process, we mean a discrete sequence of random variables $\{X_t, t \in \mathbb{Z}\}$, where $X_t : (\Omega, \mathscr{A}) \to (\mathbb{R}, \mathscr{B})$ for each $t \in \mathbb{Z}$. Since our main concern will be daily financial data, discrete time processes will suffice for our analyses.

Throughout the thesis we will consider only parametric time series models. Parametric models make assumptions about distribution of the data, that is the data are from some family of distributions $\{f(\mathbf{x}, \theta), \theta \in \Omega_0\}$, where $\Omega_0 \subset \mathbb{R}^m$. The family of distributions is usually known. In contrary to nonparametric models, parametric models have fixed number of parameters, independent of the amount of data.

1.1 Stationarity

If a time series is stationary, we can say that it retains some time invariant properties. This enables us to proceed with statistical inference. Without stationarity, we can hardly make any statistical inference about the time series data, since, simply said, the distribution of the data throughout the time varies. In time series analysis, two types of stationarity are usually used. First one is the weak stationarity and second is the strict stationarity.

Definition 1. *A time series $\{X_t, t \in \mathbb{Z}\}$ is stationary if $\forall t$, $EX_t^2 < +\infty$ and:*
i) $\forall t$, $EX_t = c$, $c \in \mathbb{R}$
ii) $\mathrm{Cov}(X_t, X_{t+k})$ is independent of t for each k.

Definition 2. *A time series $\{X_t, t \in \mathbb{Z}\}$ is strictly stationary, if for all $n \in \mathbb{N}$ and any integer $k \in \mathbb{N}$ vectors $(X_1, \dots, X_n)$ and $(X_{1+k}, \dots, X_{n+k})$ have the same* *joint distribution.*

Note that in the definition of weak stationarity, we only impose some conditions on the first two moments and make no further assumption about distributions of the random variables. Hence other moments do not even have to exist. In order for a time series to be strictly stationary, it has to have equal not only the first two moments, but whole distribution function has to be the same throughout time. It can easily be seen that if a strictly stationarity series has finite second moments, then it is weakly stationary but not vice-versa.

In reality, weak stationarity is usually sufficient for linear time series models, where we are interested in linear relationship among the variables in time and focus on modeling the first moment. However, in context of nonlinear time series, strict stationarity is often required. If we want to make some statistical inference about nonlinear models, we usually have to look beyond the first two moments and here the strict stationarity comes in play.

1.2 White noise and causality

We now briefly introduce two basic types of time series processes, that are essential in time series analysis in general, since most of the more sophisticated models build on their theory. Namely these two are the white noise process and the Gaussian process.

Definition 3. *Stochastic process* $\{X_t\}$ *is called a white noise, denoted as* $\{X_t\} \sim \mathrm{WN}(0,\sigma^2)$*, if* $EX_t = 0$*,* $\mathrm{Var}(X_t) = \sigma^2$ *and* $Cov(X_t, X_s) = 0$ *for all* $s \neq t$*.*

White noise process interprets not directly observable information. It is used heavily in linear time series processes since we make assumptions only about the first two models, so it goes in hand with weak stationarity. We can easily see, that if we have a sequence of independent and indentically distributed (i.i.d.) random variables $\{X_t\}$ with $EX_t = 0$ and $\mathrm{Var}(X_t) = \sigma^2$, denoted by IID(0, σ^2), it is a special case of a white noise process.

Definition 4. *We say that a stochastic process* $\{X_t\}$ *is Gaussian, if all of its finite-dimensional distributions are normal.*

Since normal random variables are uncorellated if and only if they are independent, a sequence of i.i.d. normal random variables is automatically a Gaussian white noise process.

It can also be easily checked, that WN(0,σ^2) is stationary, but does not have to be strictly stationary. If, however, we have Gaussian white noise process that is weakly stationary, it is automatically strictly stationary.

Another important concept in time series analysis is causality. A causal time series is such that it is caused entirely by a white noise process. We may premise - using definition of moving average process from the following chapter - that causal time series can be expressed as MA($+\infty$).

Definition 5. *We say that time series* $\{X_t\}$ *is causal, if for all t*

$$X_t = \sum_{j=0}^{+\infty} d_j \epsilon_{t-j}, \quad \sum_{j=0}^{+\infty} |d_j| < +\infty,$$

where $\{\epsilon_t\} \sim WN(0, \sigma^2)$.

Causality will help us detemine whether a time series following a specific model is stationary later on.

We now proceed to explain how to measure linear relationship between the variables in a stochastic process $\{X_t\}$.

1.3 Autocorrelation

Autocorrelation tells us about linear relationships between random variables in the process $\{X_t\}$. The autocorrelation coefficient, defined as $\text{Cov}(X_{t+k}, X_t)$ measures linear dependence of the variables X_{t+k} and X_t.
When the process $\{X_t\}$ is stationary, we may then write

$$\text{Cov}(X_{t+k}, X_t) = \text{Cov}(X_k, X_0), \quad \text{for } k \in \mathbb{N},$$

so the correlation between X_{t+k} and X_t depends only on the time difference k. That leads us to the following definition.

Definition 6. *Let* $\{X_t\}$ *be a stationary time series. The autocovariance function (ACVF) of the process* $\{X_t\}$ *is defined as*

$$\gamma(k) = \text{Cov}(X_{t+k}, X_t), \quad k \in \mathbb{Z}.$$

The autocovariance function (ACF) of the process $\{X_t\}$ *is defined as*

$$\rho(k) = \frac{\gamma(k)}{\gamma(0)} = \text{Corr}(X_{t+k}, X_t), \quad k \in \mathbb{Z},$$

We would now like to know under what assumptions is a real valued function $\gamma(\cdot) : \mathbb{N} \to \mathbb{R}$ an ACVF function of a stationary time series.

Theorem 1. *A real valued function* $\gamma(.) : \mathbb{N} \to \mathbb{R}$ *is the ACVF of a stationary time series if and only if it is even and nonnegative definite in the sense that*

$$\sum_{i,j=1}^{n} a_i a_j \gamma(i-j) \geq 0$$

for any integer $n \in \mathbb{N}$ *and arbitrary numbers* $a_1, \dots a_n \in \mathbb{R}$.

Autocorrelation function will help us to identify parameters in linear models further on. For this reason, we would like to know, given observations $\{X_1, \dots, X_T\}$ generated by a stationary time series, how to estimate the ACF function.

We first show how to estimate the autocovariance function. Knowing how to estimate the autocovariance function, the autocorrelation function can then be estimated easily.

Since the autocovariance is defined in the means of covariance, it can be naturally estimated as follows.

$$\hat{\gamma}(k) = \frac{1}{T} \sum_{t=1}^{T-k} (X_t - \bar{X}_T)(X_{t+k} - \bar{X}_T), \quad k = 0, 1, \dots, T-1,$$

where $\bar{X}_T = 1/T \sum_{t=1}^{T} X_t$. The ACF function can be then estimated as

$$\hat{\rho}(k) = \frac{\hat{\gamma}(k)}{\hat{\gamma}(0)}, \quad k = 0, 1, \ldots, T-1.$$

However in this manner it is impossible to estimate $\gamma(k)$ and therefore $\rho(k)$ for $k \geq T$ and even for k slightly smaller than T the estimates are not really good, given only a few observations (X_t, X_{t+k}) available. Box and Jenkins (1970) proposed that $T \geq 50$ and $k \leq T/4$ for the estimation to be sound.

Another important measure that will give us some insights into the structure of a time series $\{X_t\}$ is so called partial autocorrelation function (PACF).

Definition 7. *Let $\{X_t\}$ be a stationary time series with $EX_t = 0$. Then the partial autocorrelation function is defined as $\pi(1) = \mathrm{Corr}(X_1, X_2) = \rho(1)$ and*

$$\pi(k) = \mathrm{Corr}(R_{1|2,\ldots,k}, R_{k+1|2,\ldots,k}), \quad \text{for } k \geq 2,$$

where

$$R_{j|2,\ldots,k} = X_j - (\alpha_{j2}X_2 + \ldots + \alpha_{jk}X_k), \text{ and}$$
$$(\alpha_{j2}, \ldots, \alpha_{jk}) = \arg\min_{\beta_2,\ldots,\beta_k} E[X_j - (\alpha_{j2}X_2 + \ldots + \alpha_{jk}X_k)]^2.$$

In other words, $R_{j|2,\ldots,k}$ denotes the residual from the linear regression of X_j on $(X_2, \ldots, X_k)$. It can be intuitively interpreted as additional information contained in X_j that is not already explained by $(X_2, \ldots, X_k)$.

We can see that for a Gaussian process, the PACF may be rewritten as

$$\pi(k) = E[\mathrm{Corr}(X_1, X_{k+1}|X_2, \ldots, X_k)].$$

The definition of PACF may seem quite opaque and it may seem hard to calculate the PACF function for a given time series $\{X_t\}$. The following theorem tells us that PACF is in fact entirely determined by the ACVF.

Theorem 2. *For any stationary time series $\{X_t\}$, it holds*

$$\pi(k) = \frac{\gamma(k) - \mathrm{Cov}(X_{k+1}, \mathbf{X}_{2,k}^T)\mathbf{\Sigma}_{2,k}^{-1}\mathrm{Cov}(\mathbf{X}_{2,k}, X_1)}{\gamma(0) - \mathrm{Cov}(X_1, \mathbf{X}_{2,k}^T)\mathbf{\Sigma}_{2,k}^{-1}\mathrm{Cov}(\mathbf{X}_{2,k}, X_1)},$$

where $\gamma(\cdot)$ is the ACVF of $\{X_t\}$, $\mathbf{X}_{2,k} = (X_k, \ldots, X_2)^T$ and $\mathbf{\Sigma}_{2,k} = \mathrm{Var}(\mathbf{X}_{2,k})$.

What follows is a theorem linking the PACF function with time series modeling. Althoug its meaning and importance might be unclear now, we will use it later on to determine one of the parameters in the linear ARMA modeling case.

Theorem 3. *Let $\{X_t\}$ be a stationary time series with $EX_t = 0$. Then $\pi(k) = b_{kk}$ for $k \geq 1$, where*

$$(b_{1k}, \ldots, b_{kk}) = \arg\min_{b_1,\ldots,b_k} E[X_t - b_1X_{t-1} - \ldots - b_kX_{t-k}]^2.$$

The theorem tells us, that $\pi(k)$ is in fact the last autoregressive coefficient in the autoregressive approximation for X_t by the variables $X_{t-1}, \ldots, X_{t-k}$.

Both PACF and ACF provide important information about the correlation structure of the series $\{X_t\}$ and are crucial for identification and estimation of various time series models. We will show more details in the fourth chapter in light of fitting concrete time series models.

In reality, both ACF and PACF are estimated by some standard algorithms implied in the modeling software. In our case, we will use a special library for time series modeling in R, namely tseries library, where the methods for estimating ACF and PACF are implemented.

We now proceed to theory of linear ARMA modeling.

Chapter 2

Linear ARMA models

In this chapter, we will introduce theory of linear autoregressive moving average (ARMA) processes. Methods used in estimation of linear ARMA models are heavily utilized in nonlinear ARMA models, so we can consider the linear processes as a building stone for nonlinear ARMA processes. We will also fit a linear model on the financial data in the fifth chapter, so we can see whether nonlinear models truly gives us better representation of a real financial data compared to the linear ones.

2.1 Introduction to linear ARMA models

We are going to present basic definitions concerning linear ARMA processes. We start with basic definitions of linear time series autoregressive process, moving average process and subsequently to combination of the two former - autoregressive moving average process.

Definition 8. *An autoregressive process of order $p \geq 1$ is defined as*

$$X_t = b_1 X_{t-1} + \ldots + b_p X_{t-p} + \epsilon_t$$

where $\{\epsilon_t\} \sim$ WN(0, σ^2). The time series $\{X_t\}$ generated from this model is called AR(p) process.

Autoregressive (AR) model represents the current value of the process X_t as a linear combination of past values of the process together with some white noise. That is, current state is completely determined by the past values $X_{t-1}, \ldots, X_{t-p}$ and some random error ϵ_t. The model can be viewed as a classical linear regression model without intercept. In practice modeling the time series without intercept is no restriction, since it is common in time series analysis to subtract the mean from the data before proceeding with further analyses.

Definition 9. *A moving average process with order $q \geq 1$ is defined as*

$$X_t = \epsilon_t + a_1 \epsilon_{t-1} + \ldots + a_q \epsilon_{t-q}$$

where $\{\epsilon_t\} \sim$ WN(0, σ^2). The time series $\{X_t\}$ generated from this model is called MA(q) process.

MA models represents the current value as a linear combination of a white noise process realizations, so the value of X_t can be considered completely random.

By combining AR and MA processes, we arrive at an autoregressive moving average process.

Definition 10. *The autoregressive moving average (ARMA) process of orders p and q is defined as:*

$$X_t = b_1 X_{t-1} + \ldots + b_p X_{t-p} + \epsilon_t + a_1 \epsilon_{t-1} + \ldots + a_q \epsilon_{t-q},$$

where $\{\epsilon_t\} \sim$ *WN(0,* σ^2*),* $p, q \geq 0$ *are integers. We write* $\{X_t\} \sim$ *ARMA(p,q). The time series* $\{X_t\}$ *generated from this model is called ARMA(p,q) process.*

It is often useful to represent linear ARMA processes using backshift operators. Namely, denote

$$b(z) = 1 - b_1 z - \ldots - b_p z^p \quad \text{and} \quad a(z) = 1 + a_1 z + \ldots a_q z^q,$$

for $z \in \mathbb{C}$ and further define the backshift opetaror B as

$$BX_t = X_{t-1}, \quad B^k X_t = (B^{k-1}) B X_t = X_{t-k}, \quad k \in \mathbb{N}.$$

We can then rewrite the ARMA process in a simple form as

$$b(B) X_t = a(B) \epsilon_t.$$

One advantage of using the polynomial representation of ARMA model is that lot of properties of ARMA models can be determined by exploring the polynomials $b(z)$ and $a(z)$. For example the following theorem, which helps us to identify whether a linear ARMA process is stationary, holds.

Theorem 4. *Stochastic process* $\{X_t\}$ *given by a linear ARMA(p,q) process is stationary if* $b(z) \neq 0$ *for all complex numbers z such that* $|z| \leq 1$.

Before proceeding to the estimation of parameters in the linear ARMA model, we provide one more definition that is useful for their analysis. In the previous chapter, we defined a causal time series. As we said, causality means that the process $\{X_t\}$ may be expressed as MA($+\infty$) process. Similar term to causality is invertibility of time series.

Definition 11. *We say that ARMA(p,q) is invertible, if* $a(z) \neq 0$ *for all complex number z such that* $|z| \leq 1$.

As opposed to causality, invertibility means that the white noise process $\{\epsilon_t\}$ can be expressed as an AR($+\infty$) process. That is, if ARMA(p,q) process is invertible, the white noise can be represented as

$$\epsilon_t = \sum_{j=0}^{+\infty} d_j X_{t-j} \quad \text{with} \quad \sum_{j=1}^{+\infty} |d_j| < +\infty,$$

so the current state of the white noise is completely determined by infinite past of the series $\{X_t\}$.

2.2 Linear ARMA modeling

When we want to fit ARMA(p,q) model, two main questions arise. First is how to determine the orders p and q, that is how to specify the model. The second question is how to estimate the parameters $b_1, \ldots, b_p$ and $a_1, \ldots a_q$. What is also very important in time series analysis is postfitting diagnostic checking on the validity of the fitted model. We will present methods showing how to solve all of these issues. Our main focus will be on the Gaussian maximum likelihood estimation method, which is applicable to any stationary time series, according to Fan and Yao (2003).

2.2.1 Models and background

Suppose we have $X_1, \ldots, X_T$ obervations from a causal ARMA(p, q) process, so

$$X_t = b_1 X_{t-1} + \ldots + b_p X_{t-p} + \epsilon_t + a_1 \epsilon_{t-1} + \ldots + a_q \epsilon_{t-q},$$

where $\{\epsilon_t\} \sim \mathrm{WN}(0, \sigma^2)$. We want to determine the orders p, q and coefficients of the AR part and MA part of the model and the variance of white noise σ^2. Without loss of generality, we asume that $EX_t = 0$, which can be obtained by subtracting the sample mean from the data before we fit the model.

As we already stated, we will proceed with maximum likelihood estimation when fitting the model. However, given the dependence in data, calculating the Gaussian likelihood function for ARMA model requires calculating inverse of a $T \times T$ covariance matrix. This may be computationally very challenging. Hence various approaches were developed to deal with this problem. We will explore one such method in the next few paragrahps.

2.2.2 Prewhitening - the best linear prediction

As we said, when calculating the maximum likelihood estimatior, we would have to calculate inverse of large matrices, which could be computationaly infeasible. One method how to deal with this problem is to prewhiten the data. That is to find the best linear predictor for X_t, based on $X_{t-1}, \ldots, X_1$ for each $t > 1$.

Definition 12. *Let $\{X_t\}$ be a stationary process with zero mean. We say that*

$$\hat{X}_{k+1} = \phi_{k1} X_k + \ldots + \phi_{kk} X_1$$

is the best linear predictor for X_{k+1} based on $X_k, \ldots, X_1$ if

$$E[X_{k+1} - \hat{X}_{k+1}]^2 = \min_{\{\psi\}} E\Big[X_{k+1} - \sum_{j=1}^{k} \psi_j X_{k-j+1}\Big]^2$$

Given a set of coefficients $\{\phi_{kj}\}$, we would like to know under what assumptions are the coefficients the best linear predictor for X_{k+1}. The following theorem gives us the answer.

Theorem 5. *A set of coefficients $\{\phi_{kj}\}$ is the best linear predictor for $\{X_{k+1}\}$ if and only if*

$$\sum_{j=1}^{k} \phi_{kj} \gamma(i-j) = \gamma(i), \quad i = 1, \ldots, k$$

where $\gamma(\cdot)$ is the ACVF of $\{X_t\}$.

Proof. See Fan and Yao (2003), page 92. ❑

In the proof, it can be seen that

$$\text{Cov}(\hat{X}_{k+1} - X_{k+1}, X_i) = 0, \quad i = 1, \dots, k$$

From the definition of prewhitening, $X_i - \hat{X}_i$ is linear combination of only $X_i, \dots, X_1$. If we define $\hat{X}_1 = 0$, then $\{X_t - \hat{X}_t, t = 1, \dots, T\}$ is a sequence of uncorrelated random variables. Subtracting the best linear predictor from each variable X_k from the original observations, we obtain uncorrelated sequence $\{X_t - \hat{X}_t, t = 1, \dots, T\}$ - this sequence is called prewhitening. We can easily see that $E(X_t - \hat{X}_t) = 0$ and

$$\nu_{t+1} = \text{Var}(X_{t+1} - \hat{X}_{t+1}) = E[(X_{t+1} - \hat{X}_{t+1})X_{t+1}] = \gamma(0) - \sum_{j=1}^{t} \phi_{tj}\gamma(j).$$

We will now show so called innovation algorithm, according to which we can easily calculate the predictive errors $\{X_t - \hat{X}_t\}$ and their variances $\{\nu_t\}$. For details about the algorithm, we refer the reader to Brockwell and Davis (1991).
Innovation algorithm:
i) Set $\nu_0 = \gamma(0)$.
ii) Recursively calculate:

$$\theta_{k,k-j} = \nu_j^{-1}\Big\{\gamma(k-j) - \sum_{i=0}^{j-1} \theta_{j,j-i}\theta_{k,k-i}\nu_i\Big\},$$

and

$$\nu_k = \gamma(0) - \sum_{j=0}^{k-1} \theta_{k,k-j}^2 \nu_j.$$

Then calculate the values of sequences $\{\theta_{ij}\}$ and $\{\nu_j\}$ in the order

$$\theta_{11}, \nu_1; \theta_{22}, \theta_{21}, \nu_2; \dots; \theta_{T-1,T-1}, \dots, \theta_{T-1,1}, \nu_{T-1}.$$

The best linear predictors are then given by $\hat{X}_1 = 0$ and

$$\hat{X}_{k+1} = \sum_{j=1}^{k} \theta_{kj}(X_{k+1-j} - \hat{X}_{k+1-j}), \quad \text{for } k = 1, \dots, T-1.$$

We can now proceed to the maximum likelihood estimation.

2.2.3 Maximum likelihood estimation

Before we present the maximum likelihood estimation method, we need to intrudoce a few notations.

Firstly, denote $\mathbf{X}_T = (X_1, \dots, X_T)^T$ and $\hat{\mathbf{X}}_T = (\hat{X}_1, \dots, \hat{X}_T)^T$. We denote

$$\mathbf{\Theta} = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 \\ \theta_{11} & 0 & 0 & \cdots & 0 \\ \theta_{22} & \theta_{21} & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \theta_{T-1,T-1} & \theta_{T-1,T-2} & \cdots & \theta_{T-1,1} & 0 \end{pmatrix}$$

It is then possible to write

$$\hat{\mathbf{X}}_T = \mathbf{\Theta}(\mathbf{X}_T - \hat{\mathbf{X}}_T)$$

and express $\mathbf{X}_T = (\mathbf{\Theta} + \mathbf{I}_T)(\mathbf{X}_T - \hat{\mathbf{X}}_T)$, where $\mathbf{I}_T$ is $T \times T$ identity matrix. We further denote $\mathbf{C} = (\mathbf{\Theta} + \mathbf{I}_T)$, $\mathbf{D} = \text{diag}(\nu_1, \dots, \nu_{T-1})$ and $\mathbf{\Sigma} = \text{Var}(\mathbf{X}_T)$. According to the equation for $\nu_t = \gamma(0) - \sum_{j=1}^{t} \phi_{tj}(j)$ we can then express

$$\mathbf{\Sigma} = \mathbf{C}\mathbf{D}\mathbf{C}^T \text{ with } \quad |\mathbf{\Sigma}| = |\mathbf{D}| = \prod_{j=1}^{T-1} \nu_j.$$

Hence if the process $\{X_t\}$ is a Gaussian causal ARMA process, then the likelihood function is the density of multivariate normal distribution

$$L(\mathbf{b}, \mathbf{a}, \sigma^2) \sim |\mathbf{\Sigma}|^{-\frac{1}{2}} \exp\Big(-\frac{1}{2} \mathbf{X}_T^T \mathbf{\Sigma}^{-1} \mathbf{X}_T \Big) =$$

$$= (\nu_0 \dots \nu_{T-1})^{-\frac{1}{2}} \exp\Big(-\frac{1}{2} \sum_{j=1}^{T} \frac{(X_j - \hat{X}_j)^2}{\nu_{j-1}} \Big) =$$

$$= \sigma^{-T} (r_0 \dots r_{T-1})^{-\frac{1}{2}} \exp\Big(-\frac{1}{2\sigma^2} \sum_{j=1}^{T} \frac{(X_j - \hat{X}_j)^2}{r_{j-1}} \Big)$$

where $\mathbf{b} = (b_1, \dots, b_p)^T$, $\mathbf{a} = (a_1, \dots, a_q)^T$ and $r_j = \nu_j/\sigma^2$. We just note here, that the values $X_t = X_t(\mathbf{b}, \mathbf{a})$ in fact, since we suppose the process is generated by linear ARMA(p,q) process. We call the distribution of $L(\mathbf{b}, \mathbf{a}, \sigma^2)$ the Gaussian likelihood function. The maximum likelihood estimator can be then obtained as a maximizer to

$$(\hat{\mathbf{b}}, \hat{\mathbf{a}}, \hat{\sigma}^2) = \arg \max_{(\mathbf{b},\mathbf{a}) \in \mathscr{B}, \sigma > 0} L(\mathbf{b}, \mathbf{a}, \sigma^2)$$

where

$$\mathscr{B} = \{(\mathbf{b}, \mathbf{a}) : b(z) \cdot a(z) \neq 0 \ \forall z : \ |z| \leq 1\}$$

By requiring $(\mathbf{b}, \mathbf{a})$ to be in the set $\mathscr{B}$, it guarantees that the model is invertible and causal.

We can also notice, that $\{r_i\}$ and $\{\phi_{ji}\}$ do not depend on σ^2. Therefore we can proceed with the maximization in two steps, namely maximazing over σ first and then search for $(\hat{\mathbf{b}}, \hat{\mathbf{a}})$. For simpler notation, we write

$$S(\mathbf{b}, \mathbf{a}) = \sum_{j=1}^{T} \frac{(X_j - \hat{X}_j)^2}{r_{j-1}}.$$

Then maximizing over σ, the maximum likelihood estimator can be written as

$$(\hat{\mathbf{b}}, \hat{\mathbf{a}}) = \arg \max_{(\mathbf{b},\mathbf{a})\in\mathscr{B}} \left(\ln(S(\mathbf{b}, \mathbf{a})) + \frac{1}{T}\sum_{j=1}^{T} \ln(r_{j-1}) \right), \quad \hat{\sigma}^2 = \frac{S(\hat{\mathbf{b}}, \hat{\mathbf{a}})}{T}.$$

Now we can see why prewhitening is such powerful tool. We don't have to calculate the inverse of covariance matrix $\mathbf{\Sigma}$. By prewhitening the time series, we substantially reduce the computational time of searching for $(\hat{\mathbf{b}}, \hat{\mathbf{a}})$.
Acording to Fan and Yao (2003), when $\{X_t\}$ is not Gaussian, the Gaussian likelihood function or distribution of $L(\mathbf{b}, \mathbf{a}, \sigma^2)$ written above, may still be regarded as a measure of goodness of fit to the data. In the first theorem of the following section, we will show that even if the data does not follow a normal distribution, but only $\{\epsilon_t\} \sim \text{IID}(0, \sigma^2)$ holds, then the maximum likelihood estimator derived above is asymptotically distribution free. However, when ϵ_t is not Gaussian, the maximum likelihood estimators are usually inefficient. When ϵ_t has heavy tails in the sense that $\text{Var}(\epsilon_t) = +\infty$ then the Gaussian likelihood estimation may lead to even inconsistent estimators. For more robust methods, see David et al. (1992).
We now proceed to asymptotic properties of the likelihood estimators.

2.2.4 Asymptotic properties

After deriving the maximum likelihood estimators, we are interested in their asymptotic properties. Let $\{W_t\} \sim \text{WN}(0,1)$ and define

$$b(B)U_t = W_t \quad \text{and} \quad a(B)V_t = W_t,$$

where $b(\cdot)$ and $a(\cdot)$ are the polynomials from linear ARMA process representation. Then $\{U_t\}$ is an AR(p) process with coefficients $b_1, \dots, b_p$ defined in terms of AR-coefficients and $\{V_t\}$ is an AR(q) process with coefficients $a_1, \dots, a_q$ defined in terms of MA-coefficients. We can see that the processes are correlated with each other, since they are both generated from the same white noise process. Denote the initial values of processes $\{U_t\}$ and $\{V_t\}$ as $\mathbf{Z} = (U_{-1}, \dots, U_{-p}, V_{-1}, \dots, V_{-q})$, and

$$\mathbf{W}(\mathbf{b}, \mathbf{a}) = \text{Var}(\mathbf{Z})^{-1}$$

Then the following theorem tells us about asymptotic distribution of the maximum likelihood estimator with identically distributed errors.

Theorem 6. *Let $\{X_t\}$ be ARMA(p,q) process with $\{\epsilon_t\} \sim$ IID(0, σ^2), $\sigma^2 > 0$ and let $(\boldsymbol{b}_0, \boldsymbol{a}_0) \in \mathscr{B}$ be the true values of parameters from the set $\mathscr{B}$ defined above. Denote $(\hat{\mathbf{b}}, \hat{\mathbf{a}})$ the maximum likelihood estimators and $\hat{\sigma}^2$ the variance from MLE. Then for $T \to +\infty$ it holds:*

$$T^{1/2} \begin{pmatrix} \hat{\mathbf{b}} - \mathbf{b}_0 \\ \hat{\mathbf{a}} - \mathbf{a}_0 \end{pmatrix} \xrightarrow{D} N\big(\boldsymbol{0}, \, \boldsymbol{W}(\boldsymbol{b}_0, \boldsymbol{a}_0)\big)$$

We can see from the theoreem, that the asymptotic distribution of the MLE is independent of σ^2. Hence we can say that the estimator of causal and invertible ARMA processes does not deteriorate with the magnitude of the white noise. Hannah (1973) showed that it holds even for general ARMA processes.

2.2.5 Confidence Regions

With the theorem above in hand, we can proceed to construction of confidence regions and intervals for testing hypotheses about parameters we obtained using the MLE for the ARMA process or AR and MA processes respectively. For causal and invertible ARMA process, we may use the asymptotic variance matrix $\mathbf{W}(\mathbf{b}, \mathbf{a})$ to calculate the standard errors of the MLE for the parameters. The asymptotic limit distribution for the estimated parameters in ARMA process can be used to construct confidence regions in the following way. An approximate $(1 - \alpha)$ confidence region for the AR coefficient $\mathbf{b}$ is constructed as:

$$\{\mathbf{b} = (b_1, \ldots, b_p)^T : (\hat{\mathbf{b}} - \mathbf{b})^T \hat{\mathbf{W}}_1^{-1} (\hat{\mathbf{b}} - \mathbf{b}) \leq \chi_p^2(1 - \alpha)/T\},$$

where $\hat{\mathbf{W}}_1$ is the upper-left $p \times p$ submatrix of $\mathbf{W}(\hat{\mathbf{b}}, \hat{\mathbf{a}})$, and $\chi_p^2(1 - \alpha)$ is the 100α-th percentile of χ^2-distribution with p degrees of freedom. If we want to test a single parameter, the approximate $(1 - \alpha)$ confidence interval for b_j can be calculated as:

$$\{b_j : |\hat{b}_j - b_j| \leq T^{-1/2} w_{jj}^{1/2} z(1 - \alpha/2)\},$$

where $z(\alpha)$ is the 100α-th percentile fo the standard normal distribution.
Similarly the approximate $(1 - \alpha)$ confidence region for the MA coefficient vector $\mathbf{a}$ is constructed as:

$$\{\mathbf{a} = (a_1, \ldots, a_q)^T : (\hat{\mathbf{a}} - \mathbf{a})^T \hat{\mathbf{W}}_2^{-1} (\hat{\mathbf{a}} - \mathbf{a}) \leq \chi_q^2(1 - \alpha)/T\},$$

where $\hat{\mathbf{W}}_2$ denotes the $q \times q$ lower-right submatrix of $\mathbf{W}(\hat{\mathbf{b}}, \hat{\mathbf{a}})$ and the rest is the same as with the parameter $\mathbf{b}$. For the single parameter a_i, we can construct the confidence interval as

$$\{a_i : |\hat{a}_i - a_i| \leq T^{-1/2} w_{p+i,p+i}^{1/2} z(1 - \alpha/2)\},$$

where $z(\alpha)$ is the 100α-th percentile fo the standard normal distribution.

As we now know to estimate the parameters $(\mathbf{b}, \mathbf{a})$ in the linear ARMA process given set of observations $\{X_1, \ldots, X_T\}$, the question is how to determine the orders p and q. In the next part of this chapter, we present general method of fitting a model to a set of data, so called Akaike information criterion. It can be used to find the orders p and q.

2.3 Model identification

We will present two methods showing how to determine the linear ARMA model orders p and q, specifically the Akaike Information Criterion (AIC) and Bayesian Information Criterion. As we said, these methods can be used to help us fit any unknown density function. We will present it in a compact manner here. Let's start with the AIC, which can be considered as a basic measure of a goodness of fit. Details can be found in Akaike (1973).

2.3.1 Akaike information criterion

Let's first describe the idea of how AIC works. Say that we want to approximate unknown density function g by a probability density function f. We define the Kullback-Leibler information as follows

$$I(g,f) = E[\ln\Big(\frac{g(X)}{f(X)}\Big)] = \int g(x)\ln\Big(\frac{g(x)}{f(x)}\Big)\mathrm{d}x, \text{ where } X \sim g.$$

We can see, using Jensen's inequality, that $I(g,f)$ is always positive, since

$$I(g,f) = -E[\ln\Big(\frac{f(X)}{g(X)}\Big)] \geq -\ln(E\Big[\frac{f(X)}{g(X)}\Big]) =$$

$$= -\ln\Big(\int \frac{f(x)}{g(x)}g(x)\mathrm{d}x\Big) = -\ln(1) = 0,$$

and the equality holds if and only if $f = g$.
Since we can rewrite the integral as

$$\int g(x)\ln\Big(\frac{g(x)}{f(x)}\Big)\mathrm{d}x = \int g(x)\ln(g(x))\mathrm{d}x - \int g(x)\ln(f(x))\mathrm{d}x,$$

we see that the right hand side does not depend on f. It is therefore suitable to choose f that minimizes the right hand side of the equation

$$-\int g(x)\ln(f(x))\mathrm{d}x = -E_g[\ln(f(X))].$$

However, in reality, we don't know the function g. Therefore, given set of observations $\{X_1, \ldots, X_T\}$, we will replace the expectation by the statistic

$$-\frac{1}{T}\sum_{j=1}^{T}\ln(f(X_j)),$$

which is an unbiased estimator of the expectation.

In reality, we usually choose f from a set of parametric family $\{f_m(\cdot|\theta_m)\}$ and typically the form of the function f_m is given for each m. For example in linear time series modeling, functions f_m may stand for ARMA family with order $(p,q) = m$ and $\theta_m = (b_1, \ldots, b_p, a_1, \ldots, a_q)$. To find the best approximation, we need to minimze the term

$$-\frac{1}{T}\sum_{j=1}^{T}\ln(f_m(X_j|\theta_m)).$$

Similarly as with maximizing the Gaussian likelihood function, we may view this process as a two step optimization, in which we first find the minimizer of θ_m for each fixed m and then to find m that minimizes the whole sum. It is not hard to see that the minimizer in the first step is the maximum likelihood estimator, which maximizes the following log-likelihood density function

$$\hat{\theta}_m = \arg\max_{\theta}\sum_{j=1}^{T}\ln(f_m(X_j|\theta_m)).$$

The second step of the optimization is to find $\hat{m}$ that minimizes the sum, that is

$$\hat{m} = \arg\min_{m \in \mathbb{N}} -\frac{1}{T}\sum_{j=1}^{T} \ln(f_m(X_j|\hat{\theta}_m)).$$

However, as Akaike pointed out, this approach has one drawback - the expression

$$-\frac{1}{T}\sum_{j=1}^{T} \ln(f_{\hat{m}}(X_j|\hat{\theta}_{\hat{m}}))$$

is no longer an unbiased estimator of $-E_g[\ln(f_m(X|\theta_m)]$, since the expression is overfitted, which is caused by using the same sample $\{X_1, \ldots, X_T\}$ twice. Once for the estimation of the log-likelihood and second for estimating the parameter $\hat{\theta}_m$.

Akaike (1973) proposed one solution to the problem. He added the bias to the sample likelihood function and then showed that it can be asymptotically approximated as p_m/T, where p_m is the number of estimated parameters. Formally written, he proved that

$$-E_g[\ln(f_m(X|\theta_m)\} + \frac{1}{T}\sum_{j=1}^{T} E_g[\ln(f_{\hat{m}}(X_j|\hat{\theta}_{\hat{m}})] \approx \frac{p_m}{T}.$$

In order to correct the bias, we should add the term $\frac{p_m}{T}$ to the minimized value, that is

$$-\frac{1}{T}\sum_{j=1}^{T} \ln(f_{\hat{m}}(X_j|\hat{\theta}_{\hat{m}})) + \frac{p_m}{T}.$$

Finally, by multiplying the expression by $2T$, we obtain the Akaike information criterion (AIC):

$$AIC(m) = -2\sum_{j=1}^{T} \ln(f_m(X_j|\hat{\theta}_m) + 2p_m.$$

The first part of AIC can be intuitivelly explained as a lack of fit, so more sophisticated models should decrease this term and vice versa. The second part penalizes us for increasing the number of parameters. So the optimum model minimizing AIC ratio is a well balanced trade-off between complexity and sophistication.

2.3.2 Bayesian information criterion

AIC is surely good first step when fitting model orders, but in reality it has several drawbacks. Due to the form of AIC, it often overestimates the number of parameters in the model, because it doesn't penalize the model enough for increasing the number of parameters. Moreover, according to Akaike (1970), selecting orders based on AIC does not lead to consistent order selection. Therefore we would like to find a procedure which treats the drawbacks of AIC, yet retains its advantages. Therefore we will present the Bayesian information criterion, which is more sensitive to increasing the number of parameters, does not lead to oveftiffing and the parameters estimated by BIC are strongly consistent. Basing the criterion on AIC, Bayesian information criterion (BIC) is defined as

$$BIC(m) = -2\sum_{j=1}^{T} \ln f_m(X_j|\hat{\theta}_m) + \ln(T)p_m.$$

From the equation above, we can see that BIC penalizes the model for overfitting much more than the AIC. As we've already proclaimed, Hannah (1980) showed that BIC is consistent estimator for the order of the model.

We now dispose of all necessary theory for identification of the best linear ARMA model. However, in reality no thorough modeling should be complete without post fitting diagnostic. After specifying a model, we need to check whether our model truly represents the reality well. This is the topic of next few paragraphs. Again as with the procedures for order determination, diagnostic checking does not apply only to linear models, but generally to all time series models.

2.4 Diagnostic checking

Since fitting a time series model is only approximation of reality, we should always conduct postfitting diagnostic to see whether the model explains the trends in the data well. We will present introduction to one method for diagnostic checking, namely the residual based method, which tests whether residuals from a fitted model behave like a white noise process. We will show the basic procedure for linear model, however the ideas hold even for nonlinear models.

Suppose we have a fitted ARMA(p,q) model. Then we can define the standardized residuals as follows:

$$R_j = \frac{(X_j - \hat{X}_j)}{(\sigma^2 r_{j-1})^{1/2}}, \quad j = 1, \ldots, T,$$

where $\hat{X}_j$ and r_{j-1} are the same as in the Gaussian likelihood function, so both depend on the parameters $\mathbf{b}$ and $\mathbf{a}$. If we replace them with the maximum likelihood estimators, we arrive at the standardized residuals:

$$\hat{R}_j = \frac{X_j - \hat{X}_j(\hat{\mathbf{b}}, \hat{\mathbf{a}})}{(\hat{\sigma}^2 r_{j-1}(\hat{\mathbf{b}}, \hat{\mathbf{a}}))^{1/2}}, \quad j = 1, \ldots, T.$$

We would like the sequence $\{\hat{R}_j\}$ to be similar to the sequence $\{R_j\} \sim$ WN(0,1). Moreover if the sequence $\{\epsilon_t\}$ is Gaussian, then $\{\hat{R}_j\} \sim$ N(0,1).

Since statistical tests for whiteness require theory of spectral densities, which is beyond the scope of this thesis, we will proceed with just visual diagnostics. Looking at the time series plot of $\{\hat{R}_j\}$ it should look "randomly" as a white noise process. In case there is a clear deviation from the zero mean or changing variance over time, the sequence is most probably not a white noise process.

By this paragraph, we can close the theory of linear ARMA models and move on to the theory of nonlinear ones.

Chapter 3

Nonlinear ARMA models

As we've already discussed in the introduction, linear ARMA models can be a good stepping stone to model a time series data, however in real life and financial time series particularly, hardly any phenomenon exhibits strictly linear features.

Time series in reality can have many nonlinear futures, such as nonnormally distributed errors, the series may exhibit cycles, nonlinear relationship between lagged variables, and so on. This in turns limit the power of linear models and hence we need to turn to nonlinear methods to take care of such nonlinearities.

In this chapter, we will present three nonlinear models - the first one is threshold autoregressive model, second one is conditional heteroscedastic model and third one is generalized conditional heteroscedastic model.

Let's begin with the threshold models.

3.1 Threshold Autoregressive model

As we stated, linear approximation is a good stepping stone for time series modeling, yet global linear law is often bound to have a few insufficiencies and is therefore inappropriate. For example in the financial time series context, it would be naive to assume that during the phase of recession, markets behave the same as in the phase of expansion. Therefore, we may try to model the nonlinear dynamics through dividing the state space into several subspaces and then on each of those subspaces proceed with a linear approximation. This leads us to the following definition.

Definition 13. *A threshold autoregressive process (TAR) with k ($k \geq 2$) regimes is defined as*

$$X_t = \sum_{i=1}^{k} \{b_{i0} + b_{i1}X_{t-1} + \ldots + b_{i,p_i}X_{t-p_i} + \sigma_i \epsilon_t\} I(X_{t-d} \in A_i)$$

where $\{e_t\} \sim IID(0,1)$, $d, p_1, \ldots, p_k$ are unknown positive integers, $\sigma_i > 0$, b_{ij} are unknown parameters, and $\{A_i\}$ forms a partition of $\mathbb{R}$ such that $\bigcup_{i=1}^{k} A_i = \mathbb{R}$ and $A_i \cap A_j = \emptyset$ for all $i \neq j$.

We can easily see, that TAR is an AR model on each A_i. Therefore it retains the simplicity of linear models, yet can capture the nonlinearities present in time series data. The random variable X_{t-d} which dictates the partition of $\mathbb{R}$ is called threshold variable and d is called a delay parameter.

A few questions arise, when we look at the definition. One is how to determine the partition of the subspace and then how to determine the threshold variable. Although we will use a special case of the TAR model based on economic interpretation when modeling the financial data in the fifth chapter, we will proceed with general approach of fitting the TAR processes.

3.1.1 Estimation and model identification

Suppose that we have observed values $X_1, \dots, X_T$ from the TAR process, with $k \in \mathbb{N}$ given. We now show the method how to estimate parameters b_{ij}'s, σ_i's and d and determine the orders p_i's together with partitions of $\mathbb{R}$, A_i's.

Firstly, we will derive the method for estimation assuming the partition $\{A_i\}$ and orders p_i's are known. Afterwards, we will show how to determine the partition of the state space and orders p_i.

For simplicity assume that $d \leq p = \max_{1 \leq i \leq k} p_i$. We define the square error function $L(\mathbf{b_i}, d; A_i)$ as follows:

$$L(\mathbf{b}_i, d; A_i) = \sum_{\substack{p<t\leq T \\ X_{t-d} \in A_i}} \{X_t - (b_{i0} + b_{i1}X_{t-1} + \ldots + b_{i,p_i}X_{t-p_i})\}^2.$$

We can then estimate the least squares estimators for autoregressive coefficients $\mathbf{b_i} = (b_{i0}, \dots, b_{i,p_i})^T$, $i = 1, \dots k$ and d as the minimizer $(\hat{\mathbf{b_1}}, \dots, \hat{\mathbf{b_k}})$ and $\hat{d}$ that minimize the overall square error function, defined as a sum of square error functions:

$$\sum_{i=1}^{k} L(\mathbf{b_i}, d; A_i).$$

Similarly as we've already done in previous MLE maximizations, we can minimize the overall square error function in two steps. First for each d, we minimize the value $L(\mathbf{b_i}, d; A_i)$. That can be easily done using the least squares method, so it can be obtained explicitly by minimizing the square errors $\epsilon_{i,t}^2$, where $\epsilon_{i,t} = \sigma_i \epsilon_t$ for $i = 1, \dots, k$.

Afterwards it remains to choose $\hat{d}$ to minimize the value of the overall square error function. If we have two minimizers, we choose the one with smallest d as an estimator for the delay parameter.

Next we would like to find an estimator for the variances σ_i^2. First denote $T_i = |\{t : p < t \leq T \text{ and } X_{t-\hat{d}} \in A_i\}|$ for $i = 1, ..., k$. We can then defince the variance estimator as in ordinary least squares manner, that is

$$\hat{\sigma}_i^2 = \frac{1}{T_i} L(\hat{\mathbf{b_i}}, \hat{d}; A_i).$$

We can now proceed the problem of choosing the partition $\{A_i\}$. When modeling the financial data in the fifth chapter, we deal with this problem by partitioning the state space based on economic interpretation. However, generally, we would proceed as follows.

For each partition $\{A_i\}$ of $\mathbb{R}$, denote $L(\{A_i\}) = \sum_{i=1}^{k} L(\hat{\mathbf{b_i}}, \hat{d}; A_i)$, the minimal value of the sum $\sum_{i=1}^{k} L(\mathbf{b}_i, d; A_i)$. We are then looking for a partition $\{\hat{A}_i\}$, that minimizes $L(\{A_i\})$. In practice, we assume that the partition is in form $A_i = (r_i, r_{i+1}]$ such that $-\infty = r_0 < r_1 < \ldots < r_k = \infty$ and k is usually small, such as $k = 2, 3, 4$ and the thresholds r_k are searched within certain inner sample range to ease the computational burden. The inner sample may be selected for example based on some qualitative criteria.

We can proceed and define the least squares estimator, minimizing the overall square error term with $A_i = \hat{A}_i$, $\hat{\mathbf{b_i}}$ and $\hat{d}$ and define

$$\hat{\sigma}_i^2 = \frac{1}{T_i} L(\hat{\mathbf{b}}_i, \hat{d}, \hat{A}_i), \quad i = 1, \ldots, k$$

The only thing remaining to determine are the autoregressive orders p_i's. To do so, we will use a generalized Aikaike ratio, which is defined as

$$AIC(\{p_i\}) = \sum_{i=1}^{k} [T_i \ln(\hat{\sigma}_i^2) + 2(p_i + 1)],$$

where $\hat{\sigma}_i^2 \equiv \hat{\sigma}_i^2(p_i)$. We then choose orders p_i, that minimize the generalized AIC ratio. The intuition behind the generalized AIC is basically the same as with the ordinary AIC defined in the previous chapter.

Before proceeding to the asymptotic properties of the estimator, it is good to note under what assumptions is the series generated by TAR stationary. One simple criterion, that can be easily seen, is that the TAR process admits a strictly stationary solution, if $\sigma_1 = \ldots = \sigma_p$ and $\sum_{j=1}^{p} \max_{1 \le i \le k} |b_{ij}| \le 1$ where $p = \max_{1 \le i \le k} p_i$. If this condition holds, $\{X_t\}$ is then causal and therefore stationary. Just note that this is sufficient condition for the process to be stationary. Now we can proceed to the asymptotic properties of the estimator.

3.1.2 Asymptotic properties of TAR estimator

From now on, we will suppose that the series $\{X_t\}$ generated by TAR process is strictly stationary with finite second moments. Then if the partition $\{A_i\}$ and the parameter d are given (which in our case will true), the least squares estimator for $\mathbf{b_i}$ is asymptotically normal in the sense that

$$\sqrt{T_i}(\hat{\mathbf{b}}_i(d) - \mathbf{b}_i) \xrightarrow{D} N(0, \sigma_i^2 \mathbf{W}_i^{-1})$$

where

$$\mathbf{W}_i = \begin{pmatrix} 1 & \mu \mathbf{1}^T \\ \mu \mathbf{1} & E(\xi_i \xi_i^T) \end{pmatrix}, \quad \xi_i = (\xi_1, \ldots, \xi_{p_i})^T$$

and $\mathbf{1}$ denotes $p_i \times 1$ vector $(1, \ldots, 1)$, $\mu = E\xi_t$ and finally

$$\xi_t = b_{i0} + b_{i1}\xi_{t-1} + \ldots + b_{i,p_i}\xi_{t-p_i} + e_t, \quad \{e_t\} \sim \mathrm{WN}(0, 1)$$

With this in mind, we can easily derive the asymptotic confidence region and intervals and test hypotheses about parameters in the model. The derivation of confidence regions and intervals is similar to the one described in the previous chapter about linear ARMA model.

Similarly as with the linear models, we would like to have a method for post fitting diagnostics of the model. Since TAR is linear on each subset of $\mathbb{R}$, the residual based method described in the previous chapter can be very well employed even for TAR models, so there is no need to present a new method.

Let's move on to another nonlinear ARMA models, that are closely related, namely to the ARCH and GARCH models.

3.2 GARCH and ARCH models

In traditional time series analysis - what have we been doing up to now, we usually try to somehow model the first moment. Autoregressive conditional heteroscedastic (ARCH) and generalized autoregressive conditional heteroscedastic (GARCH) focus insted on the conditional second moments of the time series. This makes ARCH and GARCH extremely popular in the financial time series modeling, since they may be used to explain and model risk and uncertinity by taking into account the dependency of conditional second moments into modeling consideration. In this chapter, we will outline the basic probability properties of ARCH and GARCH models and most frequently used statistical inference models will be presented.

3.2.1 ARCH process

Let's begin with definition of the ARCH model.

Definition 14. *An autoregressive conditional heteroscedastic (ARCH) process with order p, $p \geq 1$, is defined as*

$$X_t = \sigma_t \epsilon_t \quad \text{where} \quad \sigma_t^2 = c_0 + b_1 X_{t-1}^2 + \ldots + b_p X_{t-p}^2,$$

$c_0 \geq 0$, $b_j \geq 0$ *are constants,* $\{\epsilon_t\} \sim \mathrm{IID}(0,1)$, *and* ϵ_t *is independent of* $\{X_{t-k}, k \geq 1\}$ *for all* t. *A stochastic process* $\{X_t\}$ *defined by the equation above is called an ARCH(p) process.*

The intuition behind ARCH model is that the predictive distribution of X_t based on its past values is a scale transform of the distribution of ϵ_t, where the scaling constant is represented by σ_t that depends on the past values of the process. Therefore it enables us to model nonconstant variance throughout time.

Under such construction, if past observations are large, the variance will be large. This nicely reflects the phenomenon present in financial markets, that with higher prices comes greater volatility. This is in contrast with linear models, where conditional mean squared predictive errors are constants.
We proceed with a theorem that links stationarity with ARCH models.

Theorem 7. *i) The necessary and sufficient condition for ARCH model defining a unique strictly stationary process* $\{X_t, t \in \mathbb{Z}\}$ *with* $EX_t^2 < +\infty$ *is that* $\sum_{j=1}^{p} b_j < 1$.

Furthemore,

$$EX_t = 0 \quad \text{and} \quad EX_t^2 = \frac{c_0}{1 - \sum_{j=1}^{p} b_j}$$

and $X_t \equiv 0$ *for all* t *if* $c_0 = 0$.
ii) If $EX_t^4 < +\infty$ *and*

$$\max\{1, (E\epsilon_t^4)^{\frac{1}{2}}\} \sum_{j=1}^{p} b_j < 1,$$

then for the strictly stationary solution of the ARCH model holds that $EX_t^4 < +\infty$.

According to the previous theorem, stationary ARCH process $\{X_t\} \sim \mathrm{WN}(0, c_0/(1 - \sum_{j=1}^{p} b_j))$. We may rewrite the square of X_t as follows:

$$X_t^2 = c_0 + b_1 X_{t-1}^2 + \ldots + b_p X_{t-p}^2 + e_t,$$

where $e_t = (\epsilon_t^2 - 1)(c_0 + \sum_{j=1}^{p} b_j X_{t-j}^2)$. From this, we can conclude that

$$E(e_t | X_{t-k}, X_{t-k-1}, \ldots) = 0, \quad \text{for } k \geq 1.$$

Then calculation of the expectation of the second moment of X_{t+k} for $k > p$ can be obtained as:

$$E(X_{t+k}^2 | X_{t-m}, m \geq 0) = c_0 + \sum_{j=1}^{p} b_j E(X_{t+k-j}^2 | X_{t-m}, m \geq 0)$$

or

$$\mathrm{Var}(X_{t+k} | X_{t-m}, m \geq 0) = c_0 + \sum_{j=1}^{p} b_j \mathrm{Var}(X_{t+k-j} | X_{t-m}, m \geq 0).$$

More generally, for any $k \in \mathbb{N}$ we have

$$\mathrm{Var}(X_{t+k} | X_{t-m}, m \geq 0) = c_0 + \sum_{j=1}^{k-1} b_j \mathrm{Var}(X_{t+k-j} | X_{t-m}, m \geq 0) +$$

$$+ \sum_{j=k}^{p} b_j X_{t+k-j}^2.$$

What we've now derived is in fact called volatility clustering in the scope of financial time series analysis. From the above equation, we can see that if the series gets volatile, that is the variance $\mathrm{Var}(X_{t+k} | X_{t-m}, m \geq 0)$ is high, it won't fade out until after $k - 1$ periods.
We now show a theorem with some further properties of the ARCH model.

Theorem 8. *Let* $\{X_t\}$ *be a strictly stationary ARCH(p) process with* $c_0 > 0$ *and* $\sum_{j=1}^{p} b_j < 1$. *Then:*

i) $\{X_t\} \sim \mathrm{WN}(0, c_0/(1 - \sum_{j=1}^{p} b_j))$*, and the conditional variance function fulfills the equation*

$$Var(X_{t+k}|X_{t-m}, m \geq 0) = c_0 + \sum_{j=1}^{p} b_j Var(X_{t+k-j}|X_{t-m}, m \geq 0).$$

Under additional condition

$$\max\{1, (E\epsilon_t^4)^{\frac{1}{2}}\} \sum_{j=1}^{p} b_j < 1,$$

it holds that
ii) $\{X_t^2\}$ *is a linear causal AR(p) process, and its ACF is always positive if*

$$\sum_{j=1}^{p} b_j > 0$$

and
iii) X_t *exhibits heavier tails than those of* ϵ_t *in the sense that* $\kappa_x \geq \kappa_\epsilon$*, where* κ_x *and* κ_ϵ *are the kurtoses of random variables* X_t *and* ϵ_t *respectively.*

3.2.2 GARCH process

We now proceed to the properties of the GARCH model. The GARCH model is generalized ARCH model, which apart from second moments of the series takes into consideration second moments of moving averages. According to Fan and Yao (2003), ARCH(p) model defined in the previous section provides a reasonable fit to a financial time series only if the order p is large. This is caused by the fact, that conditional variance is dependent only on previous values of X_t^2. It is then quite natural to take into account not only past values X_j^2's but also past values of σ_j^2's. This leads us to the definition of the GARCH processes.

Definition 15. *A generalized autoregressive conditional heteroscedastic (GARCH) process with order p,* $p \geq 1$*, and q,* $q \geq 0$*, is defined as:*

$$X_t = \sigma_t \epsilon_t \quad \text{where} \quad \sigma_t^2 = c_0 + \sum_{i=1}^{p} b_i X_{t-i}^2 + \sum_{j=1}^{q} a_j \sigma_{t-j}^2,$$

where $c_0 \geq 0$*,* $b_i \geq 0$*,* $a_j \geq 0$ *are constants,* $\{\epsilon_t\} \sim \mathrm{IID}(0,1)$*, and* ϵ_t *is independent of* $\{X_{t-k}, k \geq 1\}$ *for all t. A stochastic proces* $\{X_t\}$ *defined by the equation above is called a GARCH(p,q) process.*

We will now show the relationship between linear ARMA model and GARCH model. It is possible to express X_t^2 in the following terms:

$$X_t^2 = c_0 + \sum_{i=1}^{p} b_i X_{t-i}^2 + \sum_{j=1}^{q} a_j \sigma_{t-j}^2 + \epsilon_t =$$

$$= c_0 + \sum_{i=1}^{\max(p,q)} (b_i + a_i) X_{t-i}^2 + e_t - \sum_{j=1}^{q} a_j e_{t-j},$$

where $b_{p+j} = a_{q+j} = 0$ for all $j \geq 1$ and

$$e_t = X_t^2 - \sigma_t^2 = (\epsilon_t^2 - 1)\left(c_0 + \sum_{i=1}^{p} b_i X_{t-i}^2 + \sum_{j=1}^{q} a_j \sigma_{t-j}^2 \right).$$

We can see, that the process $\{X_t^2\}$ is in fact ARMA($\max(p,q), q$) model. We know from the section about ARMA processes, that invertible ARMA(p,q) process with finite p and q can be transformed into AR($+\infty$) process. In the case of GARCH processes, it means that invertible GARCH process may be expressed as ARCH($+\infty$) process. This explains why even low order GARCH model, such as GARCH(1,1) may provide an easily comprehensible representation of even complexly autodependent structure of $\{X_t^2\}$, which would otherwise require an ARCH process with high order p to explain the data similarly well.
We will now provide a theorem linking stationarity with GARCH processes.

Theorem 9. *The necessary and sufficient condition for the series $\{X_t, t \in \mathbb{Z}\}$ with $EX_t^2 < +\infty$ generated by a GARCH(p,q) process to be unique and strictly stationary is*

$$\sum_{i=1}^{p} b_i + \sum_{j=1}^{q} a_j < 1$$

Furthemore, $EX_t = 0$ and

$$Var(X_t) = \frac{c_0}{1 - \sum\limits_{i=1}^{p} b_i - \sum\limits_{j=1}^{q} a_j}, \quad Cov(X_t, X_{t-k}) = 0, \quad \forall k \in \mathbb{Z} \setminus \{0\}$$

Additionaly, $EX_t^4 < +\infty$ holds, if

$$\max\{1, (E\epsilon_t^4)^{\frac{1}{2}}\} \frac{\sum\limits_{i=1}^{p} b_i}{1 - \sum\limits_{j=1}^{q} a_j} < 1$$

The previous theorem gives us sufficient and necessary condition for the GARCH process to be strictly stationary with finite second moments.

Under the condition $\sum\limits_{i=1}^{p} b_i + \sum\limits_{j=1}^{q} a_j < 1$, $\{X_t\} \sim \mathrm{WN}(0, \mathrm{c_0}/(1 - \sum\limits_{\mathrm{i=1}}^{\mathrm{p}} \mathrm{b_i} - \sum\limits_{j=1}^{q} a_j))$ and the ARMA representation is casual and invertible. Hence $EX_t^2 = E\epsilon_t^2 E\sigma_t^2 = E\sigma_t^2$ and

$$E(X_t | X_{t-1}, \ldots) = 0.$$

It also holds from the equation for e_t that

$$Ee_t = E(e_t | X_{t-1}, X_{t-2}, ...) = 0$$

As a consequence, we can write

$$\mathrm{Var}(X_t | X_{t-1}, \ldots) = E(X_t^2 | X_{t-1}, X_{t-2}, \ldots) =$$

$$c_0 + \sum_{i=1}^{p} b_i X_{t-i}^2 + \sum_{j=1}^{q} a_j \sigma_{t-j}^2 = \sigma_t^2.$$

Hence σ_t^2 is the conditional variance of X_t given its infinite past.

If $\{X_t\}$ is a strictly stationary GARCH(p,q) process and condition

$$\max\{1, (E\epsilon_t^4)^{1/2}\} \frac{\sum\limits_{i=1}^{p} b_i}{1 - \sum_{j=1}^{q} a_j} < 1 \quad (*)$$

holds, then $E\sigma_t^4 = EX_t^4 / E\epsilon_t^4 < +\infty$ and hence even $E\epsilon_t^4 < +\infty$. Therefore $\{X_t^2\}$ is causal and invertible ARMA(max(p,q),q) process. Note that in contrast to ARCH processes, the ACF of $\{X_t^2\}$ does not have to be always positive.

We can summarize our knowledge about GARCH in the following proposition.

Theorem 10. *i) If $\{X_t\}$ is a stationary GARCH(p,q) process, then it is also a white noise process, namely $\{X_t\} \sim$ WN(0, $c_0/(1 - \sum\limits_{i=1}^{p} b_i - \sum\limits_{j=1}^{q} a_j)$) and σ_t^2 is the conditional variance of $\{X_t\}$ given its infinite past, that is*

$$\sigma_t^2 = \mathrm{Var}(X_t | X_{t-1}, X_{t-2}, \ldots).$$

*ii) If $\{X_t\}$ is a strictly stationary GARCH(p,q) process, for which condition $(*)$ holds, then $\{X_t^2\}$ is a causal and invertible ARMA(*max*(p,q),q) process. Furthemore X_t exhibits heavier tails than those of ϵ_t in the sense that $\kappa_x \geq \kappa_\epsilon$, where again κ_x and κ_ϵ denotes the kurtosis of X_t and ϵ_t respectively.*

We now proceed to the estimation of ARCH and GARCH processes.

3.2.3 Estimation

Suppose we have observations $X_1, \ldots, X_T$. We assume, that $\{X_t\}$ is a strictly stationary solution of the GARCH process, that is

$$X_t = \sigma_t \epsilon_t \quad \text{and} \quad \sigma_t^2 = c_0 + \sum_{i=1}^{p} b_i X_{t-i}^2 + \sum_{j=1}^{q} a_j \sigma_{t-j}^2$$

where $p \geq 1$, $q \geq 0$, c_0, b_i, $a_j > 0$ and $\sum\limits_{j=1}^{p} b_i + \sum\limits_{j=1}^{q} a_j < 1$, $\{\epsilon_t\} \sim \mathrm{IID}(0,1)$. We'd like to estimate the conditional second moments, which are generally more difficult to estimate than conditional means. Although there are various methods for estimating the parameters c_0, b_i and a_j, we will present just one, namely the conditional maximum likelihood estimator.

Conditional maximum likelihood estimators

Similiar as for the ARMA processes, the most frequently used estimator for ARCH/GARCH models are derived from a Gaussian likelihood function. For a general GARCH(p,q) model with $p > 0$, $q \geq 0$, conditional variance σ_t^2 cannot be expressed as a finite number of past observations $X_{t-1}, X_{t-2}, \ldots$. We have to therefore truncate some observations. We can write:

$$\sigma_t^2 = \frac{c_0}{1 - \sum\limits_{j=1}^{q} a_j} + \sum_{i=1}^{p} b_i X_{t-i}^2 + \sum_{i=1}^{p} b_i \sum_{k=1}^{\infty} \sum_{j_1=1}^{q} \cdots \sum_{j_k=1}^{q} a_{j_1} \ldots a_{j_k} X_{t-i-j_1-\ldots-j_k}^2$$

It can be seen, that if $q = 0$, the multiple sum is equal to zero, hence for ARCH model, we can express the conditional variance as a finite sum of X_t's. Since the expected value of the series is finite and b_i's and a_j's are nonnegative, the sum converges almost surely.

In order to calculate the maximum likelihood estimator in real modeling, we take a truncated version of the expression above, that is:

$$\tilde{\sigma_t}^2 = \frac{c_0}{1 - \sum\limits_{j=1}^{q} a_j} + \sum_{i=1}^{p} b_i X_{t-i}^2 +$$

$$+ \sum_{i=1}^{p} b_i \sum_{k=1}^{\infty} \sum_{j_1=1}^{q} \cdots \sum_{j_k=1}^{q} a_{j_1} \ldots a_{j_k} X_{t-i-j_1-\ldots-j_k}^2 I(t - i - j_1 - \ldots - j_k \geq 1), \quad t > p$$

As we stated above, if $q = 0$, then $\tilde{\sigma_t}^2 = \hat{\sigma_t}^2$ holds.

Denote $\mathbf{b} = (b_1, \ldots, b_p)^T$ and $\mathbf{a} = (a_1, \ldots, a_q)^T$. Then we can write the conditional maximum likelihood estimator $(\hat{\mathbf{b}}, \hat{\mathbf{a}}, \hat{c_0})$ as a maximizer to

$$l_k(c_0, \mathbf{b}, \mathbf{a}) = \sum_{t=k}^{T} [\ln(\tilde{\sigma_t}^2) + X_t^2 / \tilde{\sigma_t}^2],$$

where $k > p$ is an integer. Formally, the estimator can be written as

$$(\hat{\mathbf{b}}, \hat{\mathbf{a}}, \hat{c_0}) = \arg\max l_k(c_0, \mathbf{b}, \mathbf{a}).$$

When we think that normal distribution is not sufficient for the modeling, as may be frequent in financial time series modeling, two other distributions are commonly used. Speficically the t-distribution and the Generalized Gaussian distribution. Generally we may write the maximum likelihood function as

$$l_k(c_0, \mathbf{b}, \mathbf{a}) = \sum_{t=k}^{T} [\ln(\tilde{\sigma}_t) - \ln(f(X_t / \tilde{\sigma}_t))].$$

The function f may then be the density of normal distribution, or we may use other forms, such as:
i) t-distribution with v degrees of freedom and density function:

$$f_v(x) = \frac{\Gamma((v+1)/2)}{(\pi v)^{1/2} \Gamma(v/2)} \left(\frac{v}{v-2} \right)^{1/2} \left(1 + \frac{x^2}{v-2} \right)^{-\frac{v+1}{2}},$$

where $v > 2$ may be viewed as a continuous parameter.

ii) Generalized Gaussian distribution and density function:

$$f_{v,\lambda}(x) = v \{ \lambda 2^{1+\frac{1}{v}} \Gamma(1/v) \}^{-1} \exp\Big(-\frac{1}{2} \frac{|x|^v}{\lambda^v} \Big),$$

where $\lambda = (2^{-\frac{2}{v}}\Gamma(1/v)/\Gamma(3/v))^{1/2}$ and $v \in (0,2)$.

Both distributions are normalized, so they have mean 0 and variance 1 and all have heavier tails than normal distribution, so they may be more suitable for modeling financial time series.

When searching for the most feasible density function, we may again use the Akaike information criterion or Bayesian information criterion to determine which one is the best.

Let's proceed to the asymptotic properties of the conditional MLE.

3.2.4 Asymptotic properties of conditional MLE

We now present in a very compact manner results about asymptotic distributions of the Gaussian likelihood estimation method as developed by Hall and Yao (2003).

Suppose again that $\{X_t\}$ is a strictly stationary solution of GARCH(p,q) process with $p \geq 1$ and $q \geq 0$, $c_0 > 0$, $b_j > 0$ for $j = 1, \dots, p$ and $a_i > 0$ for $i = 1, \dots, q$. Denote $(\hat{c}_0, \hat{\mathbf{b}}, \hat{\mathbf{a}})$ the maximum likelihood estimator derived from maximizing

$$l_k(c_0, \mathbf{b}, \mathbf{a}) = \sum_{t=k}^{T} [\ln(\tilde{\sigma}_t^2) + X_t^2/\tilde{\sigma}_t^2].$$

Assume, that $k = k(T) \to +\infty$ for $T \to +\infty$ and at the same time $k(T)/T \to 0$. For simplicity, denote further $\theta = (c_0, \mathbf{b}^T, \mathbf{a}^T)^T$, $\hat{\theta} = (\hat{c}_0, \hat{\mathbf{b}}^T, \hat{\mathbf{a}}^T)^T$ and $\mathbf{U}_t = \frac{\partial \sigma_t^2}{\partial \theta}$. Hall and Yao (2003) show, that $\mathbf{U}_t/\sigma_t^2$ has all of its moments finite. Write

$$\mathbf{M} = E(\mathbf{U}_t \mathbf{U}_t^T/\sigma_t^4) > 0$$

that is the matrix $\mathbf{M}$ is positive definite.

In order to present the theorem concerning asymptotic distribution of the MLE, we need to introduce one more definition.

Definition 16. *We say that distribution G is in the domain of attraction of a distribution F if*

$$\frac{1}{a_n}(S_n - b_n) \xrightarrow{D} F, \quad \text{as } n \to +\infty$$

where $S_n = \sum_{i=1}^{n} \xi_i$, $\{\xi_i\} \sim$ i.i.d. G, and $a_n > 0$ and b_n are some constants.

With this in mind, we can proceed to the theorems about asymptotic properties of the maximum likelihood estimator.

Theorem 11. *i) If $E(\epsilon_t^4) < +\infty$, then*

$$\frac{T^{1/2}}{(E[\epsilon_t^4] - 1)^{1/2}}(\hat{\theta} - \theta) \xrightarrow{D} N(0, \boldsymbol{M}^{-1})$$

.

ii) Denote $\lambda_T = inf\{\lambda > 0 : \; E[\epsilon_1^4 I(\epsilon_1^2 \leq \lambda)] \leq \lambda^2/T\}$. If $E(\epsilon_t^4) = +\infty$ and the distribuiton of ϵ_t^2 is in the domain of attraction of normal distribution, then

$$\frac{T}{\lambda_T}(\hat{\theta} - \theta) \xrightarrow{D} N(0, \boldsymbol{M}^{-1}).$$

3.2.5 Bootstrap regions and intervals

The last thing we need to construct are confidence regions and intervals for testing hypotheses about parameters. However the situation about confidence regions and intervals is quite complicated in connection with the ARCH and GARCH processes. Looking at the previous theorem, we see that the range of possible limit distributions for the Gaussian likelihood estimator can be pretty extensive. The limit distributions in case of MLE for ARCH and GARCH processes are not restricted to a family of distributions that can be described with a finite number of parameters, as in the case of linear ARMA process or TAR process. For this reason, it is impossible to perform statistical tests based on asymptotic distributions. Fan and Yao (2003) propose use of bootstrap methods. Since the theory of bootstrap methods is beyond the scope of this thesis, we refer the reader to Fan and Yao (2003), pages 163-166 for their treatement.

Chapter 4

Financial data modeling

In this chapter, we will aplly the theory laid in previous three chapters to model real financial data. Namely we will fit the linear ARMA model, TAR model, ARCH model and GARCH model on the historic data of the Standard & Poor's 500 (S&P500) index, that is an American stock market index based on the market capitalization of the 500 largest companies with stock listed on the New York Stock Exchange and National Association of Securities Dealers Automated Quotations. All the models, graphs and analyses were carried out in statistical software R. Before proceeding to our own analyses and models estimation, we will first present a brief introduction to financial data modeling.

Daily financial data usually consists of some unpleasant characterisic. We touched the first two According to Rydberg (2000), these can be:

i) **Heavy tails**

In the financial comunity, it is generally accepted, that the returns X_t have heavier tails than are the tails of a normal distribution. Usually we suppose that X_t only has a few number of finite moments. General agreement is that the daily returns have finite second moment, that is $EX_t^2 < +\infty$. That is also the prerequisite to GARCH and ARCH modeling.

ii) **Volatility clustering**

In the financial world, large price changes usually occur in clusters. This means that large volatility changes are usually followed by large changes and small volatility changes are usually followed by small changes. This observation in fact led to the developement of ARCH and GARCH models.

iii) **Asymmetry**

Assymetry means that the distribution of stock returns is skewed negatively. Rydberg lists one possible explanation that traders react more strongly to negative information than to positive one.

iv) **Long range dependence**

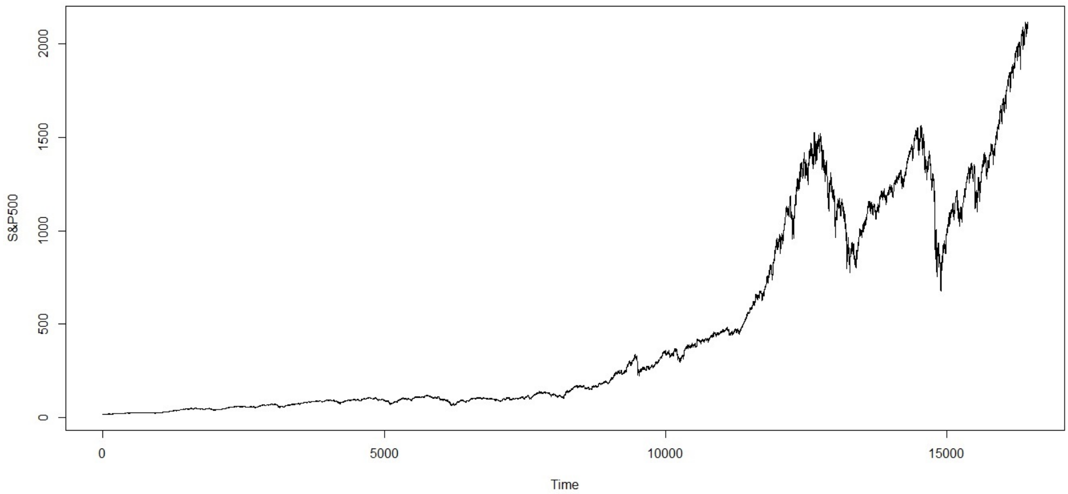
Returns usually don't show any serial correlation, which however doesn't mean they are unrelated. It shows that squared returns and even absolute returns exhibit persistent autocorrelations, which indicates long-memory dependence of these functions of returns.

4.1 Preliminary stock analysis

We obtain the S&P500 data directly from the yahoo's financial webpage for the S&P500 index, namely
http://finance.yahoo.com/q?s=ĜSPC. We downloaded daily data about closing values of the index from 3.1.1950 to 8.5.2015. The sample size is $T = 16443$. The data are downloaded directly using our script via the quantmod library in R.

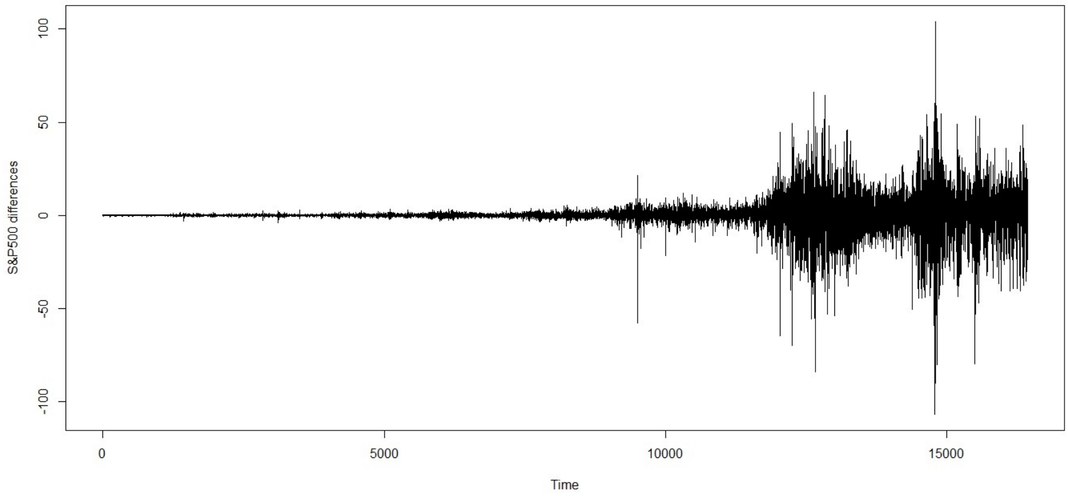
First we will do some graphical investigation of the data, that is we will take a look at plots of various indicators we described in the previous chapters. From the plots, we can identify seasonal trends, components and possible outliers. These informations should tell us whether, for example, it will be neccessary to difference the data to arive at a stationary time series.

The figure below shows evolution of the S&P500 index from 3.1.1950 till 1.5.2015. Despite a few eras of economic downturns, the index is continually rising, beggining at value 16.6 in 1950 and reaching 2108.3 on 1st May 2015. The two major declines depict the 2001 and 2008 global financial crises.
Time series plot of S&P500 index:

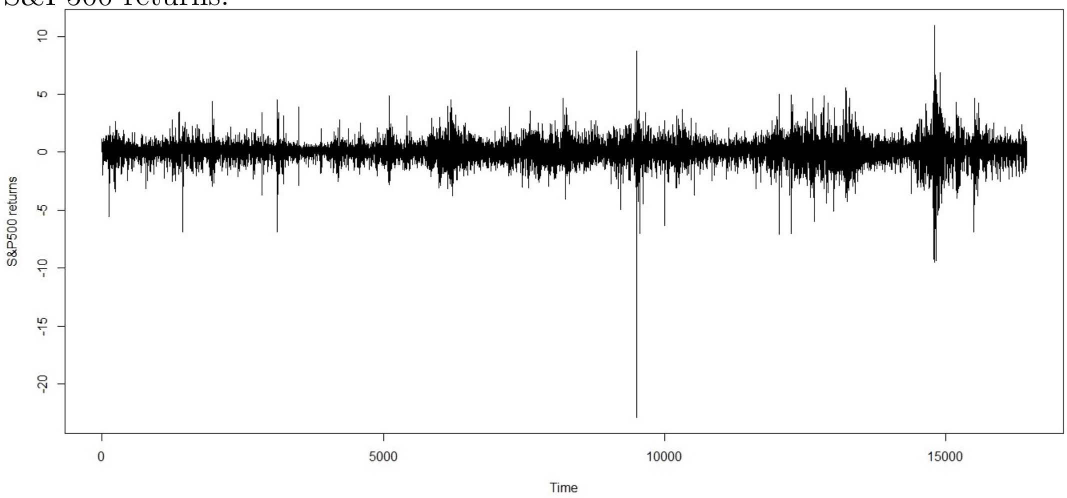


It is clearly visible, that the time series can't be stationary. Therefore we will take a look at daily differences, that is the values $X_t - X_{t-1}$, which are in fact daily gains or losses of the index. Looking at the figure below, we see that we got rid of the increasing trend, however the volatility is much larger the further we go in time. This can be well explained by the fact that the higher the value of the index is, the higher the volatility is.

S&P500 gains and losses:



We may try to take a look at just percentual changes, which should be independent of the magnitude of the data. That is, we define a new random varible $Y_t = 100(\ln(X_t) - \ln(X_{t-1}))$ and plot it against time. In fact, the process $\{Y_t\}$ captures daily returns of the S&P500. On the figure below, we see the returns of the index.
S&P500 returns:



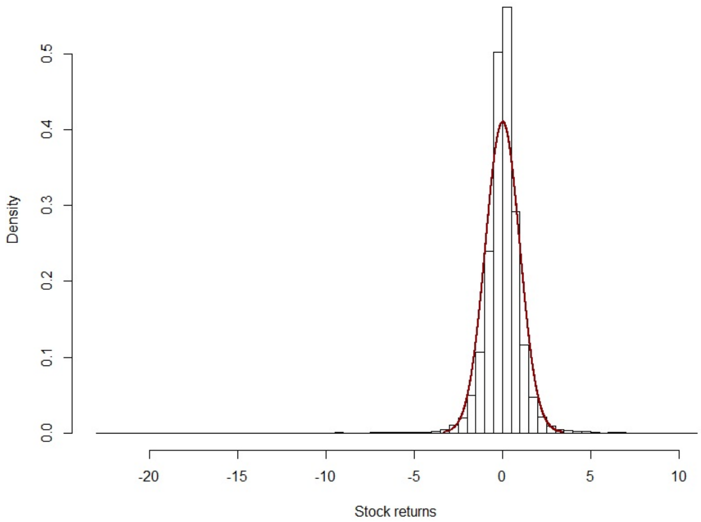
We see that this series looks pretty stationary. We check the stationarity by using the augumented Dickey-Fuller test. The null hypothesis states that the series is not stationary versus the alternative that it is. For a detailed treatment of the augumented Dickey-Fuller test, see Xiao and Phillips (1998). The Dickey-Fuller statistic for stock returns is -26,151 and the p-value is less than 0.01, hence we reject the hypothesis that the data are not stationary at even 0,99 level of confidence. With this knowledge, we can proceed with analysis of the daily returns of the S&P500 index.

Below we present basic statistical summaries concerning the returns of the index. The measures of location and variability we obtained are:

Minimum	1^{st} quartile	Median	Mean	3^{rd} quartile	Maximum
-22,9	-0,4119	0,0470	0,02951	0,4967	10,960

Variance	IQR
0,9439	0,9806

We also present the histogram for returns compared to the normal distribution N(0,02951; 0,9439). We see that there is a long stretch on the left hand side of the distribution, corresponding to the 1987 stock market crash. However, if we get rid of this outlier, the distribution is quite symmetric and does not deviate from the normal distribution a lot.
S&P500 returns histogram, compared with normal distribution:

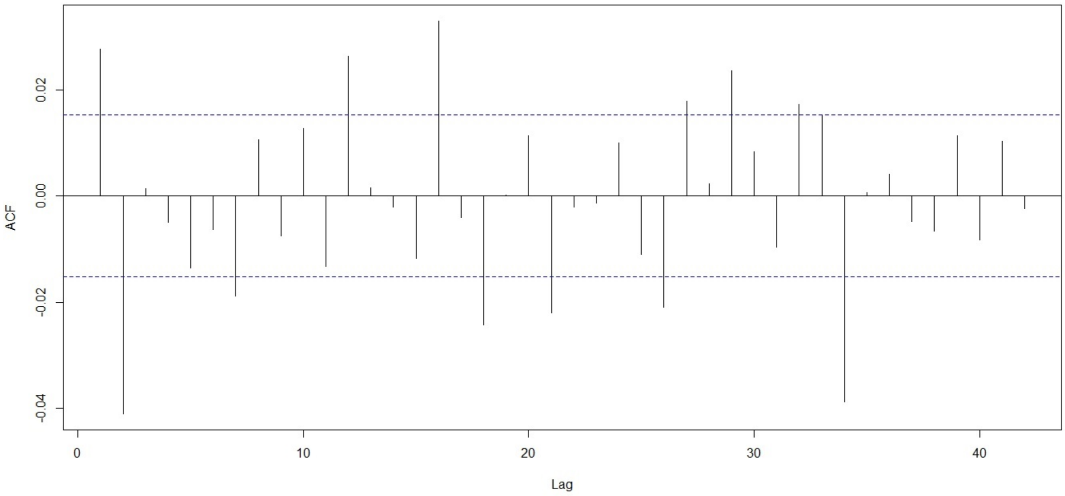


With this initial analysis in mind, we proceed to the linear ARMA modeling of the financial returns.

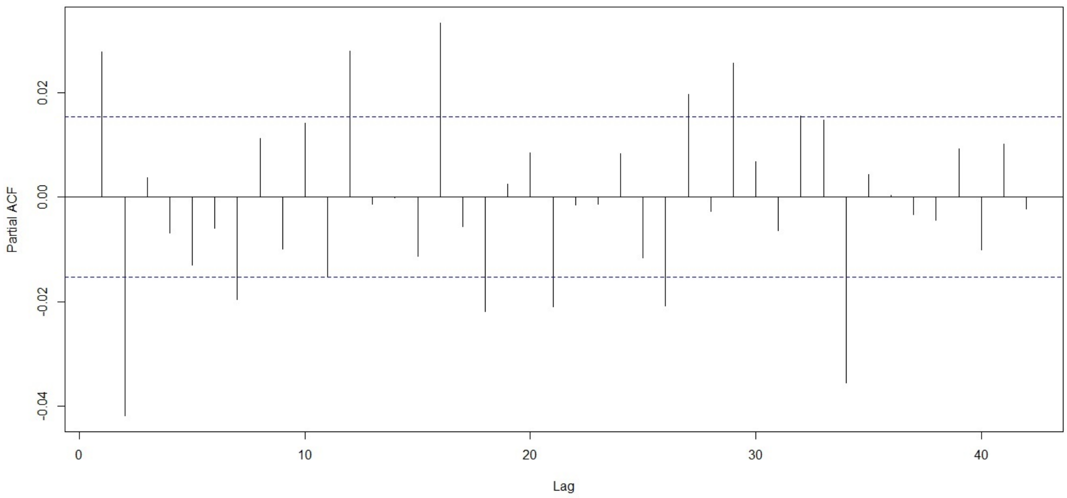
4.2 Linear ARMA modeling

We will first try to fit the daily returns of the index using linear ARMA model. To determine the orders p and q, we will plot the ACF and PACF of the daily returns. This will help us to determine the subsets $\{1, \ldots, \tilde{p}\}$ and $\{0, \ldots, \tilde{q}\}$ in which we will then search for the optimal parameters $\hat{p}$ and $\hat{q}$ respectively, that minimize BIC/AIC.
Let's take a look at the ACF plot for the time series.

Autocorrelation function of S&P500 returns:



We see from the figure above that the series displays relatively strong correlation even between large lags. It cuts off after 34 lags, which is quite a lot. Below we can see the partial autocorrelation function for the series.
Partial autocorrelation function of S&P500 returns:



From the figures above, we see that they both display very similar structure. The dashed lines on both plots are set at $\pm 1,96/\sqrt{T}$ (since $u(1-\alpha/2) = 1,96$ for $\alpha = 0,05$, where $u(\cdot)$ is the quantile function of standard normal distribution). We can then visualy test the hypothesis $\mathrm{H}_0 : \rho(k) = 0$ at the 5% significance level. It shows that $|\hat{\rho}(k)|$ and $|\hat{\pi}(k)|$ are beyond the $\pm 1,96/\sqrt{T}$ line for even quite large k's, namely for $k = 1, 2, 7, 12, 16, 18, 21, 26, 27, 29, 32, 34$. It seems that the value X_t depends even on very large lags. One possible explanation could be, given the nature of financial data and taking into consideration that we have daily returns, that present value of the index might be influenced by what happened one month ago.

Based on the analysis of ACF and PACF we will set $\tilde{p} = 2$ and $\tilde{q} = 2$ and will select the best model based on the BIC and AIC. Even though they are some significant correlations beyond the second lags, we won't fit a model with high orders p and q to prevent overfitting.

Using the both BIC and AIC ratio, the optimal orders of p and q of the

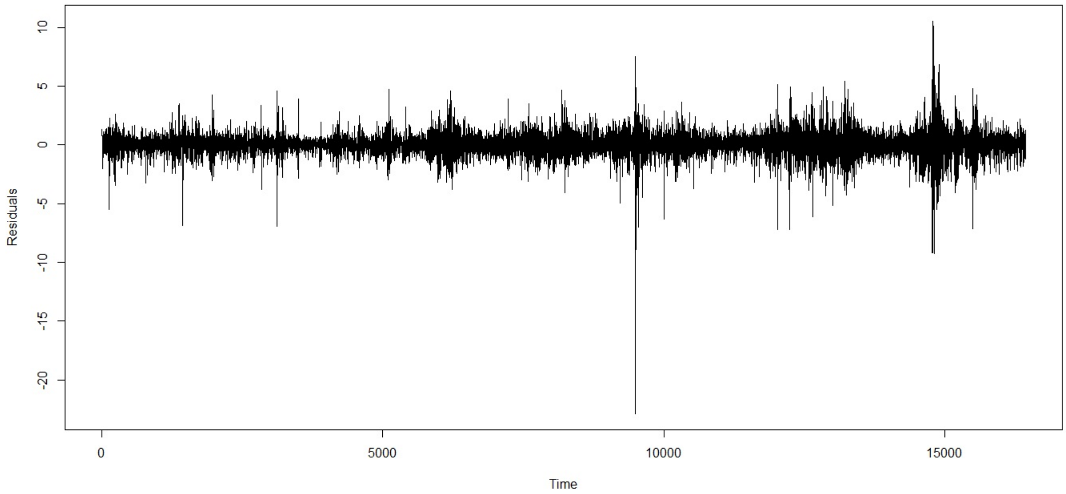
ARMA model was selected as $p = 2$ and $q = 2$. The fitted model looks as follows:

$$\hat{X}_t = b_1 X_{t-1} + b_2 X_{t-2} + \epsilon_t + a_1 \epsilon_{t-1} + a_2 \epsilon_{t-2} =$$

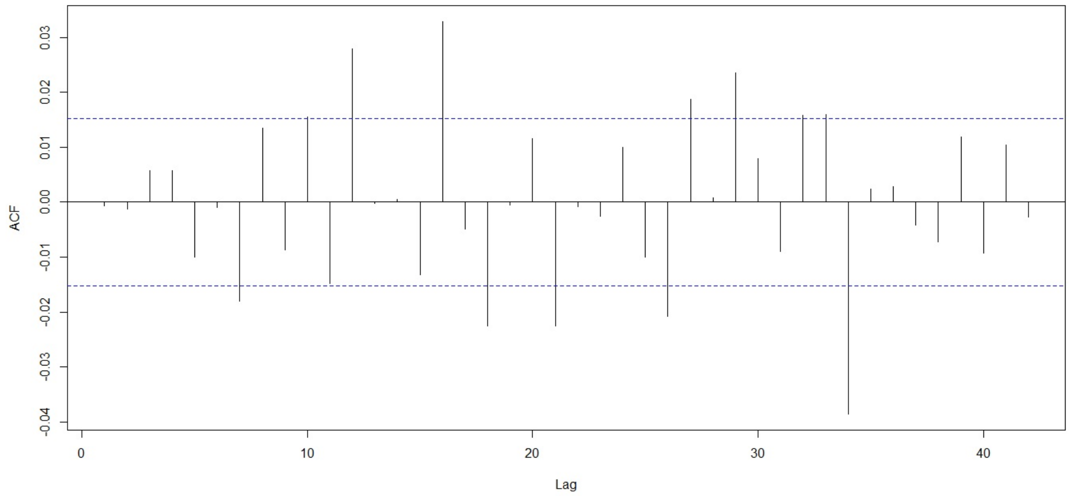
$$= 0,2881 X_{t-1} + 0,2267 X_{t-2} + \epsilon_t - 0,2585 \epsilon_{t-1} - 0,2756 \epsilon_{t-2},$$

where all coefficients are significant at more than 95% level of confidence. The standard errors are $0,15$ for b_1, $0,14$ for b_2, $0,14$ for a_1 and $0,11$ for a_2. The estimated $\hat{\sigma}^2 = 0,9413$, so $\{\epsilon_t\} \sim \text{WN}(0, 0{,}9413)$.

We now proceed to diagnostic checking to determine whether the model is sound. We will first check the structure of residuals defined as $\hat{\epsilon}_t = X_t - \hat{X}_t$ and see if they behave like a white noise process. The mean of the residuals series is 0,0005, so it is basically zero. Below we show time series graph of the residuals.
Residuals plot:



We see that apart from a few major jumps, which happen during volatile periods of the initial returns series, the process appears as a white noise process. We will further check the ACF function of the series $\{\hat{\epsilon}_t\}$ to see whether the variables are correlated.
ACF of residuals:



Looking at the graphs of residuals, we see that there are some significant correlations between the lags. This may be caused by the model we selected - by selecting the ARMA(2,2) model, we got nicely rid of the correlations in the first two terms, however the correlation structure beyond the second lag persisted.

Based on this information, we have to conclude that the series $\{\epsilon_t\}$ is not a white noise process, hence our model may have some insufficiencies.

We now proceed to the nonlinear ARMA models fitting. First we start with the TAR model.

4.3 TAR modeling

As we already stated in the theoretical part of the thesis, we are going to use a special case of the TAR model, which is based on economic representation. Specifically, we inspire ourselves with model by Tiao and Tsay (1994), who fitted TAR model on quarterly US gross national product in a following way. They divided the state space into four regimes that represent the economic phases of growth, contraction, recession and expansion. Mathematically written, if we have a time series data $\{X_1, \ldots, X_T\}$, representing the percentage growth of economy.

The phase of recession can be expressed as:
$X_{t-1} \leq X_{t-2} \leq 0$, meaning that the economy was in contraction and proceeded into even worse one. The phase of growth can be expressed as:
$X_{t-1} > X_{t-2}$ and $X_{t-2} \leq 0$, that is the economy declined, but improved in the following period. The phase of contraction can be expressed as: $X_{t-1} \leq X_{t-2}$ and $X_{t-2} > 0$, so the economy grew, but did not manage to sustain the growth in the next period. The last phase, phase of expansion, can be expressed as:
$X_{t-1} > X_{t-2} > 0$, so the economy grew in the previous term and managed to grow even more in the following one.

The same interpretation as for economic phases holds for stock market returns. Just note that the names of phases may be quite misleading in our case. Since we have daily financial data, calling a two consecutive days with decline a recession may seem exaggerated, as it is not such a big deal as in the case of economic gross national product data.

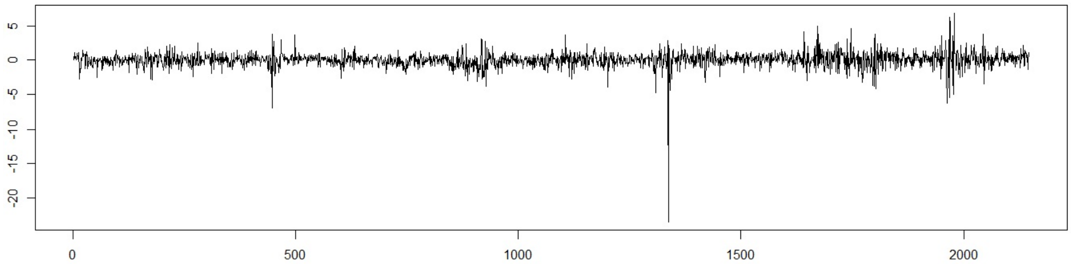
After analysing our data, we got that there were 2147 recessions, 6351 contraction periods, 5588 growth periods and 2355 periods of expansion. Below, we can see the estimated TAR model, where (**) means that the coefficient is significant at less than 0.001 level of significance, (*) at 0.01 significance level and others are insignificant at even 0.1 level of significance. The model fitted by TAR is as follows:

$$\hat{X}_t = -0.2^{**} - 0.14113X_{t-1}^{**} - 0.01303X_{t-2} + \epsilon_{1,t} \qquad X_{t-1} \leq X_{t-2} \leq 0$$
$$\hat{X}_t = 0.01968 + 0.03835X_{t-1}^{*} - 0.05964X_{t-2}^{**} + \epsilon_{2,t} \qquad X_{t-1} > X_{t-2}, X_{t-2} \leq 0$$
$$\hat{X}_t = -0.020224 + 0.019830X_{t-1} + 0.004788X_{t-2} + \epsilon_{3,t} \qquad X_{t-1} \leq X_{t-2}, X_{t-2} > 0$$
$$\hat{X}_t = 0.15573^{**} + 0.01044X_{t-1} - 0.11422X_{t-2}^{*} + \epsilon_{4,t} \qquad X_{t-1} > X_{t-2} > 0,$$

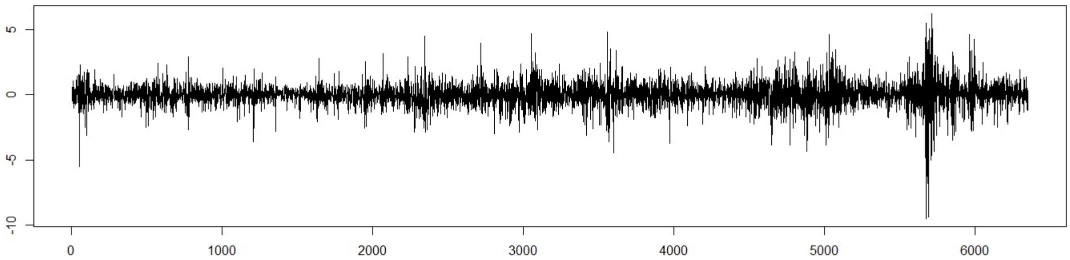
where $\epsilon_{1,t} \sim \text{WN}(0,\ 1.283)$, $\epsilon_{2,t} \sim \text{WN}(0,\ 1.0567)$, $\epsilon_{3,t} \sim \text{WN}(0,\ 0.82425)$ and $\epsilon_{4,t} \sim \text{WN}(0,\ 0.6367)$.

We may see, that in the phase of contraction, there is no significant coefficient. The lowest p-value for any coefficient for the contraction period is 0,2. Based on this finding, we can conclude that if the stocks are in contraction period, then their developement for the next day is completely random. However, we should take the fitted model with slight skepticism, since the R-squared for the model at any of the four regimes is below 0,1.

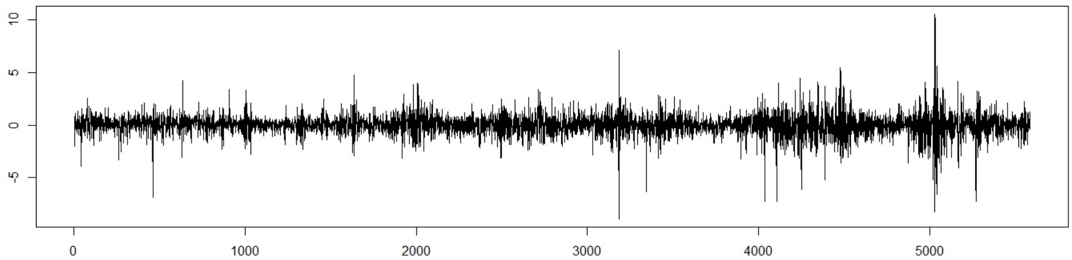
Last thing we need to do is check whether residuals from the model at least visually correspond to the white noise process. Again our post-diagnostic checking will be just very basic. Means of the residuals's series $\{\epsilon_{i,t}\}$ for each of the four models ($i = 1, 2, 3, 4$) are basically zero, with the largest one being $1,142636 \times e^{-16}$. Hence it is fair to assume that the errors follow a white noise processes with zero first moment.
Below we can see four graphs, all residuals from model on each subspace.
Recession:



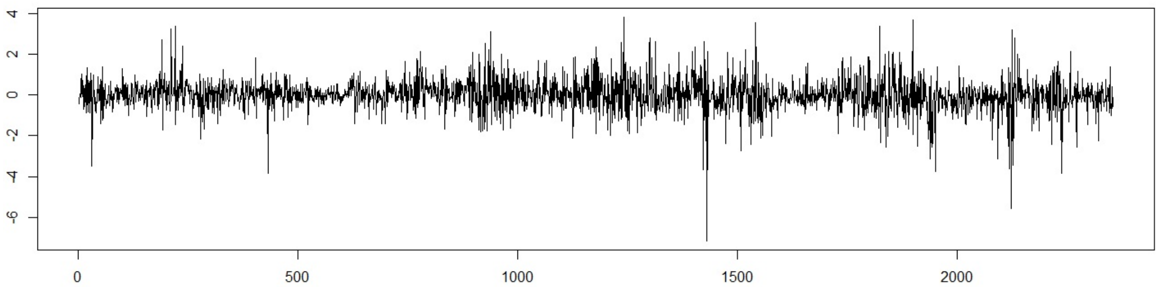
Contraction:



Growth:



Expansion:

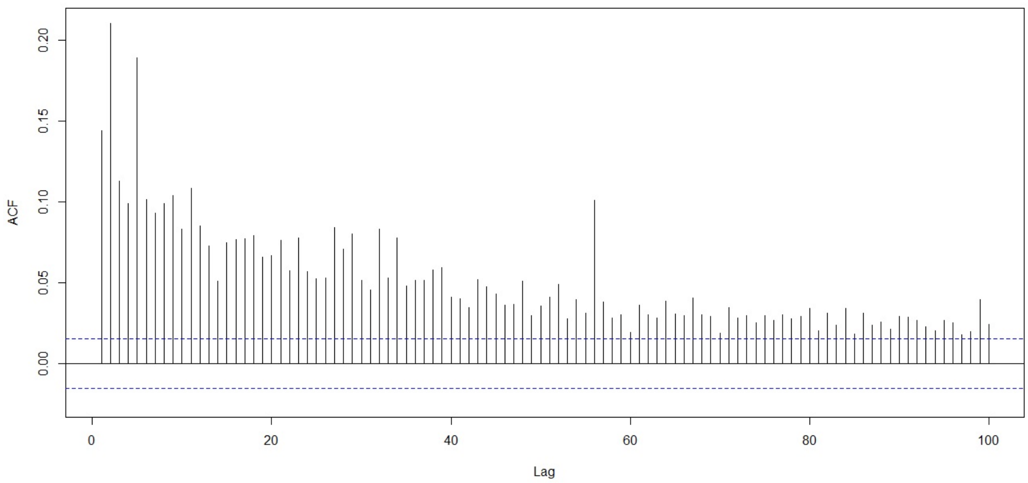


Although all four of them are more volatile in the periods of large volatility of the initial data, overall we may say that they appear fairly random. Hence it wasn't a big mistake to think of the series $\{\epsilon_{i,t}\}$, $i = 1, 2, 3, 4$ as a white noise processes.

4.4 ARCH and GARCH modeling

Looking at the possible problems arising in financial data, which were presented at the beggining of this chapter, ARCH and GARCH models are constructed as to deal with the first two problems. According to theorems 8 and 10, ARCH and GARCH proceses with normal errors are naturally heavy-tailed. Also, the models are specifically created as to deal with the volatility clustering.

Some variations of GARCH process have been introduced to deal with all the problems present in financial data, however in this thesis we will proceed with just the models presented in previous chapters. We refer interested reader to see Fan and Yao (2003), page 170, where some advanced models are presented. Before proceeding to fitting the models, we will take a look at ACF function of squared returns, which might be helpful in the scope of ARCH modeling. Autocorrelation function of squared returns of S&P500:



Comparing the result with the autocorrelation in the series $\{X_t\}$ of returns, wee see that the autocorrelation is much stronger in the squared series $\{X_t^2\}$.

We will first fit the GARCH(p,q) model with Gaussian error, that is $\{\epsilon_t\} \sim$ i.i.d N(0,1), based on the maximum likelihood ratio method we presented in the fourth chapter. The orders p and q will be again determined by the AIC/BIC. If the optimal model select $q = 0$, the GARCH model will in fact transofm into ARCH model.

Based on both AIC and BIC, we selected GARCH(1,3) model with estimated conditional standard deviation

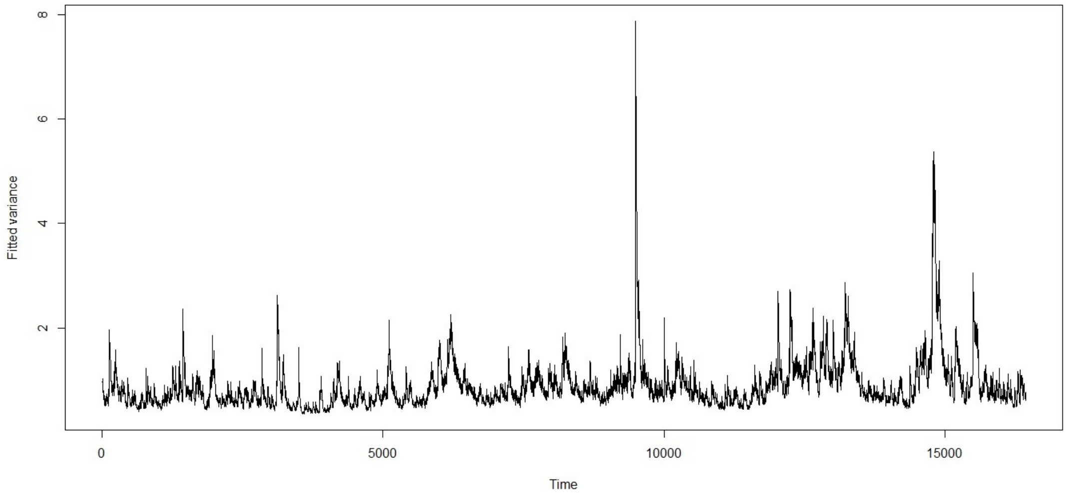
$$\hat{\sigma}_t^2 = c_0 + b_1 X_{t-1}^2 + a_1\sigma_{t-1}^2 + a_2\sigma_{t-2}^2 + a_3\sigma_{t-3}^2 =$$

$$= 0.048 + 0.1103X_{t-1}^2 + 0.6602\sigma_{t-1}^2 + 10^{-9}\sigma_{t-2}^2 + 0.2209\sigma_{t-3}^2.$$

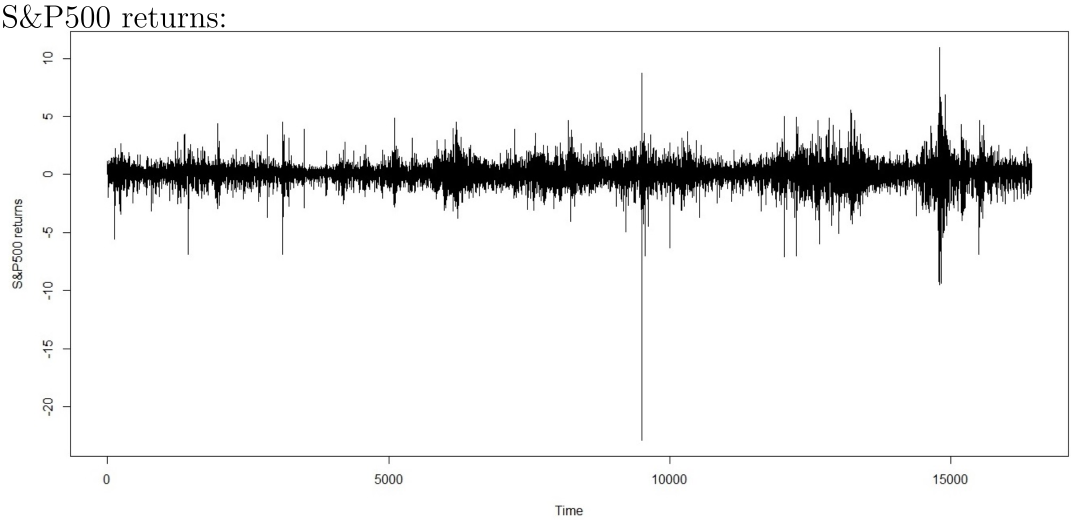
The standard errors of the estimated coefficients of the equation above are, 0.005 for c_0, 0.08 for b_1, 0.08 for a_1, 0.12 for a_2 and 0.07 for a_3 respectively and were calculated based on the asymptotic normal distribution of the estimator. The coefficient β_2 is not significantly different from zero, therefore we may remove it. All other coefficients are significant at less than 0.01 significance level.

Below we can see the graph of fitted variance $\hat{\sigma}_t^2$ according to our GARCH model. We added the graph of S&P500 returns to check whether the estimated variance behaves according the trends in the data. We may conclude that it models the volatility in the returns series very well. We can also see the plot of the residuals defined as $\hat{\epsilon}_t = X_t/\hat{\sigma}_t$. The series $\{\hat{\epsilon}_t\}$ is definitely less volatile than the initial series and at first look appear as a white noise.

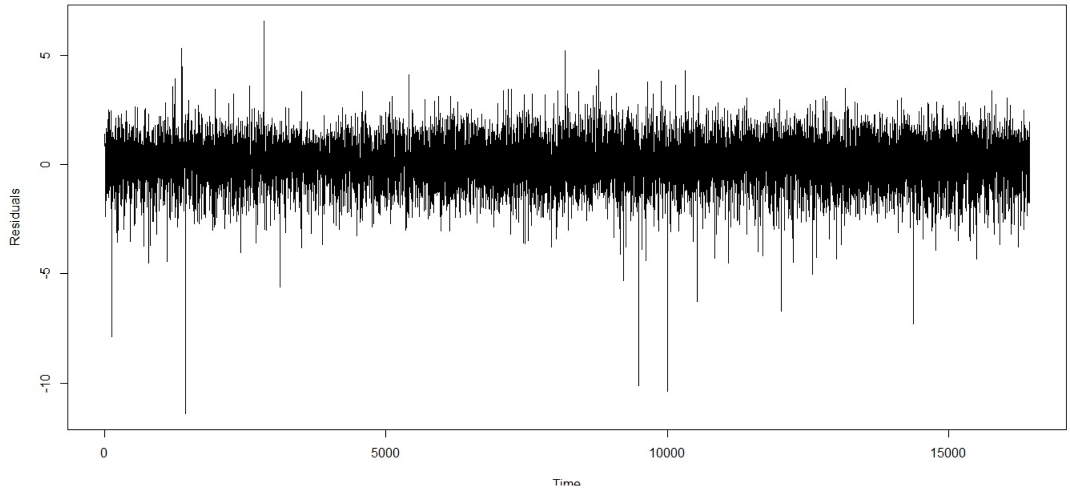
Fitted variance:



S&P500 returns:

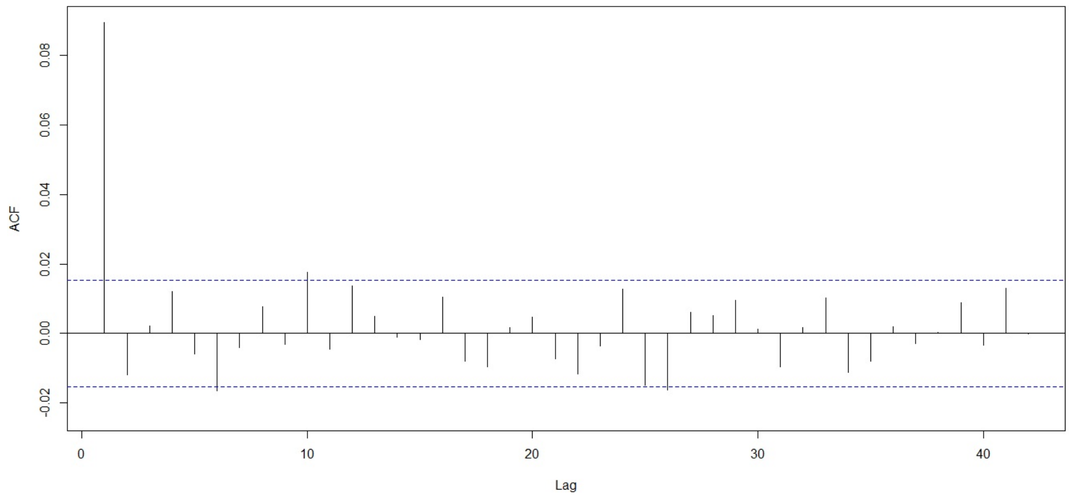


Residuals:



Since we don't have any proper test for whiteness of the residuals $\hat{\epsilon}_t$, we will at least check whether the series has zero mean and whether the terms are mutually uncorrelated. The mean of the series $\{\hat{\epsilon}_t\}$ is 0,035, so it seems sufficiently close to zero. To check whether the values are independent, we take a look at the ACF function of the series $\{\hat{\epsilon}_t\}$.

ACF of residuals:



Looking at the ACF plot of the residuals series, we see that there is relatively weak correlation between the lags, indicating that they may be dependent, which would violate the assumption of their independence. Note that if our model is perfect, the residuals should behave like a Gaussian white noise.

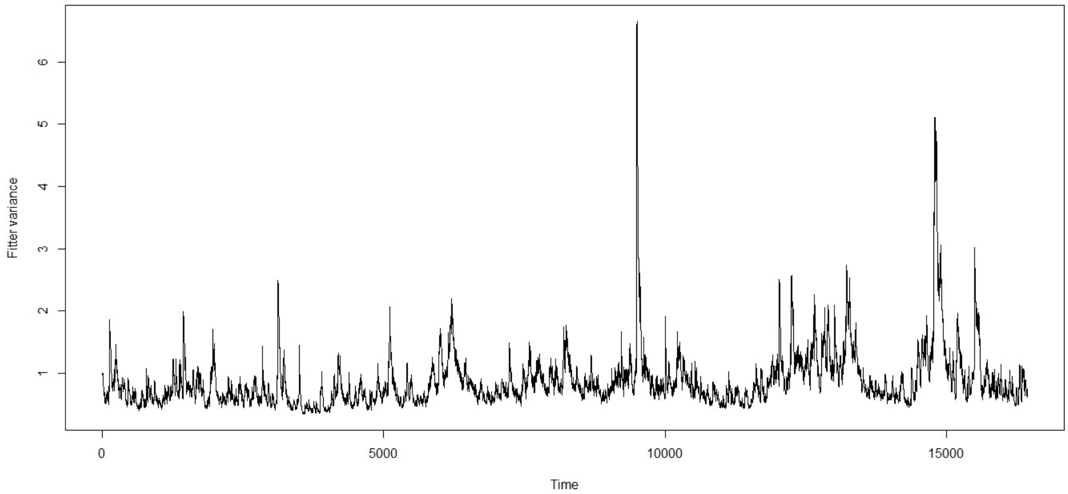
After all our analyses, we may conclude, that the model fitted with normal conditional distribution is quite sound. We will try to fit the model with t-distribution described in the previous chapter instead of normal distribution. We will again select the model based on AIC and BIC. The method for estimating the model will again be the maximum likelihood estimation we presented in the fourth chapter, however now we will use the t-distribution as a density function.

The optimal model was selected as GARCH(1,1) by both AIC and BIC. The GARCH(1,1) model for the S&P500 returns, which was fitted using the maximum likelihood method looks as follows:

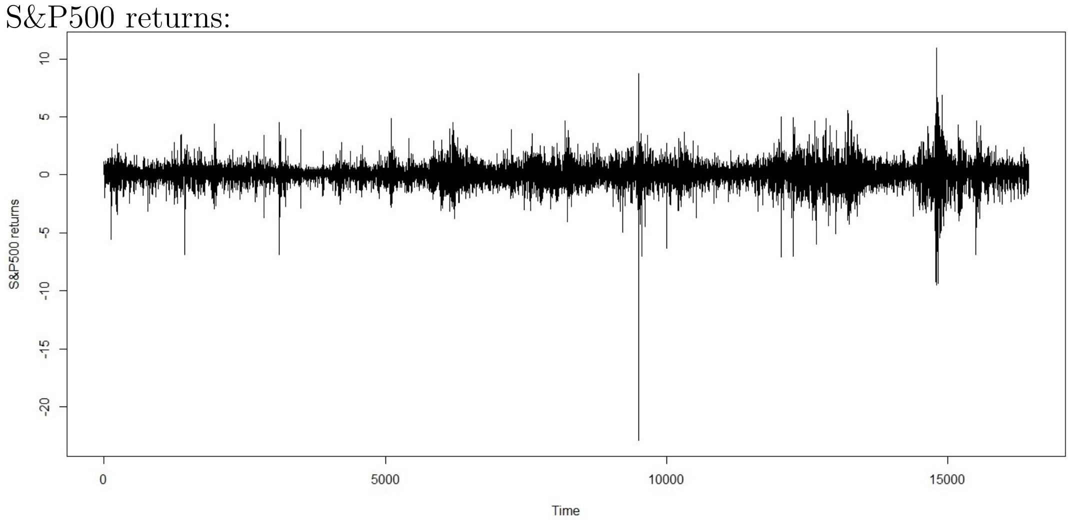
$$\hat{\sigma}_t^2 = c_0 + b_1 X_{t-1}^2 + a_1 \sigma_{t-1}^2 =$$
$$= 0,0569 + 0,0751 X_{t-1}^2 + 0,9194 \sigma_{t-1}^2,$$

where all coefficients are statistically significant at less than 0,001 level of significance based on the t-tests. The standard errors of the estimated coefficients of the equation above are 0,005 for c_0, $0,004$ for b_1 and $0,004$ for a_1 respectively.

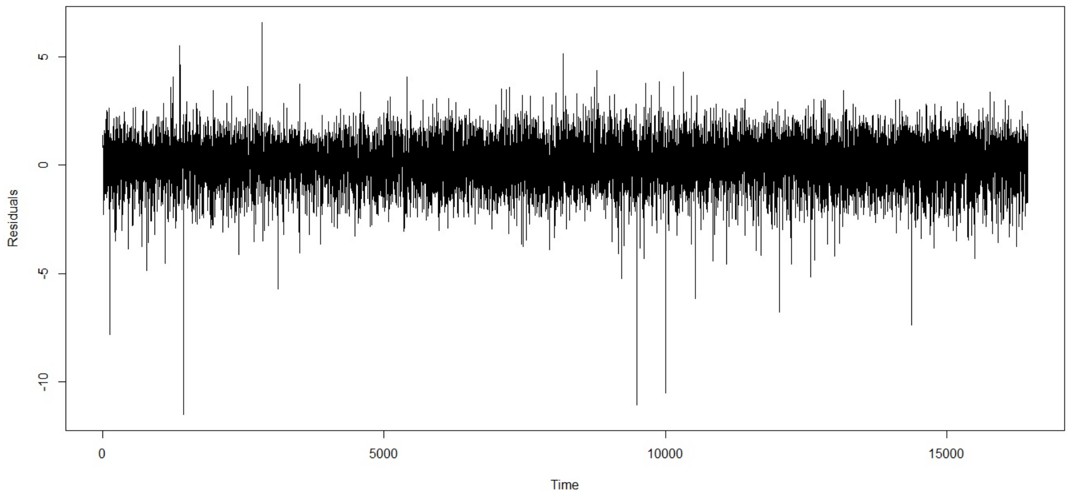
We again take a look at the plot of the conditional standard deviation $\hat{\sigma}_t$, now fitted based on the GARCH(1,1) model with t-distribution. Visually comparing the volatility with S&P500 returns, we see that the model catches the volatility in the data well. We may see that the volatility fitted with t-distribution is somehow higher in the more volatile periods and lower in less volatile periods than when we fitted the model with normal distribution. Looking at the residuals $\hat{\epsilon}_t = X_t/\hat{\sigma}_t$, we see that the series is much less volatile than the initial values of returns.
Fitted variance:



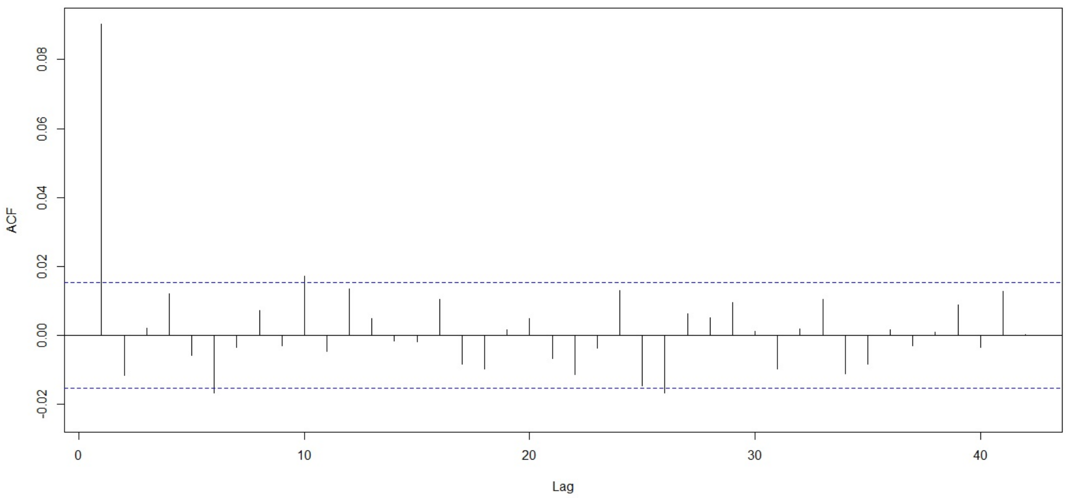
S&P500 returns:



Residuals:



We will again check the ACF function of the residuals series $\{\hat{\epsilon}_t\}$ based on the model with t-distribution. The mean of the series is 0,036, again not far from zero. Below we can see the plot of the ACF function.
ACF of residuals:



We see that the residuals correlation structure is very similiar to the previous one. Hence, all our conclusions from the model with normal distribution apply even for the model with t-distribution. On the other hand, we can't say that by using the t-distribution as a conditional distribution in the MLE, we improved the model. The model is simpler in the sense that it has lower order q and therefore may be easier to interpret.

Chapter 5

Conclusion

In this bachelor thesis, we dealt with theory of nonlinear ARMA models and its application to financial time series. The main objective of this thesis was to acquaint with the theory of nonlinear ARMA models and then apply it on real financial time series data, speficically on Standard and Poor's 500 index returns.

After introduction to general time series modeling in the first chapter, we presented the theory of linear ARMA models in the second chapter, since many concepts are used in the nonlinear theory as well. In the third chapter, the theory of three nonlinear ARMA models was presented - namely the threshold autoregressive model (TAR), autoregressive conditional heteroscedastic (ARCH) and generalized autoregressive conditional heteroscedastic (GARCH). Main concern was put on the estimation of the models using the maximum likelihood method, asymptotics of the estimators and then confidence regions and interval for testing hypothesis about the parameters in the model. Finally, in the fourth chapter, we fitted linear ARMA model, TAR model and two GARCH models, one using standard normal distribution in the maximum likelihood method and the second using t-distribution.

The thesis can have a few possible extensions. First of all we may explore methods, which provide mathematically reasonable diagnostic checking and thereafter select better models to fit the reality. Secondly, we may try to forecast the values of the index or of the volatility of the index using our models, giving them very usefull utilization. Thirdly, we may try to employ another nonlinear ARMA models, such as bilinar models, which are not so standard in the scope of financial time series modeling and see how they behave. We also didn't touch the theory of nonparametric models, which could give us completely different look at the financial time series modeling.

We may conclude that ARMA modeling is nevertheless very interesting topic of financial time series analysis.

Bibliography

H. Akaike. Statistical predictor identification. *Annals of the Institute of Statistical Mathematics*, 22(1):203–217, 1970.

H. Akaike. Information theory and an extension of the maximum likelihood principle. *Second International Symposium in Information Theory*, 1(1):276–281, 1973.

G.E.P Box and G.M Jenkins. *Time series analysis, Forecasting and control.* Fourth edition. Holden-Day, San Francisco, 1970. ISBN 978-0-470-27284-8.

P. Brockwell and R. Davis. *Time Series: Theory and Methods.* 2nd edition. Springer-Verlag, New York, 1991. ISBN 978-1441903198.

R.A. David, K. Knight, and J. Liu. M-estimation for the autoregression with infinite variances. *Stochastic Processes and Their Application*, 40(1):145–180, 1992.

J. Fan and Q. Yao. *Nonlinear Time Series.* III. series. Springer, New York, 2003. ISBN 0-387-95170-9.

P. Hall and Q. Yao. Inference in arch and garch models with heavy-tailed errors. *Econometrica*, 71(1):285–317, 2003.

J.D. Hamilton. *Time series analysis.* 1st edition. Princetion University Press, Princeton, 1994. ISBN 860-1300372280.

E.J. Hannah. The asymptotic theory of linear time-series models. *Journal of Applied Probability*, 10(1):130–145, 1973.

E.J. Hannah. The estimation of the order of an arma process. *The Annals of Statistics*, 8(1):1071–1081, 1980.

T. Rydberg. Realistic statistical modelling of financial data. *International Statistical Review*, 68(1):233–258, 2000.

G.C. Tiao and R.S. Tsay. Some advances in nonlinear and adaptive modeling in time series. *Journal of Forecasting*, 13(1):109–131, 1994.

Z. Xiao and P. Phillips. An adf coefficient test for a unit root in arma models of unknown order with empirical applications to the us economy. *Econometrics Journal*, 1(1):27–43, 1998.