UNIVERSITY OF TARTU
FACULTY OF SCIENCE AND TECHNOLOGY
INSTITUTE OF MATHEMATICS AND STATISITCS

Kseniia Guskova

Fractional ARIMA processes and applications in modeling financial time series

Financial Mathematics

Master Thesis (15 ECTS)

Supervisor: Raul Kangro, University of Tartu

Tartu 2017

Fractional ARIMA processes and applications in modeling financial time series

Master's thesis

Kseniia Guskova

Abstract: Time-series analysis is widely used in forecasting future trends on financial markets. There is a family of models which represent the property of long memory. In this thesis we aim at introducing fractionally differentiated ARIMA model in forecasting future returns of market index. In theoretical part the description of long-memory processes and statistical testing of given data are provided. In practical part we fit the models without differencing, with differencing and with fractional differencing to the market data and compare its forecast accuracy with observed values.

CERCS research specialization: P160 Statistics, operations research, programming, actuarial mathematics.

Key words: Financial mathematics, time-series analysis, long memory processes, ARFIMA processes

Autoregressiivsed murruliselt integreeritud liikuva keskmisega protsessid ja nende rakendamine finantsaegridade modelleerimiseks

Magistritöö

Kseniia Guskova

Lühikokkuvõte: Aegridade analüüs leiab laialdas kasutamist selleks, et kirjeldada ja prognoosida finantsturgude tulevikukäitumist. ARFIMA mudelid võimaldavad kirjeldada protsesse, millel on pika mälu omadus, st mille puhul avalduvad minevikusündmused mõju küllalt pika ajavahemiku jooksul. Käesolevast töös tutvustatakse ARFIMA mudeleid turuindeksi tulevikutulususte prognoosimise näitel. Töö teoreetilises osas kirjeldatakse vaadeldavat mudelit ja ning kirjeldatakse statistilisi teste, mida mudeli sobitamisega soetud otsuste tegemisel on võimalik kasutada. Praktilises osas sobitatakse vaadeldavale aegreale ilma diferentsita mudel, tavalise diferentsiga mudel ja murrulise diferentsiga mudel ning võrreldakse saadud mudelite abil saadavate prognooside täpsuseid.

CERCS teaduseriala: P160 Statistika, operatsioonianalüüs, programmeerimine, finants- ja kindlustusmatemaatika.

Märksõnad: finantsmatemaatika, aegridade analüüs, pika mäluga protsessid, ARFIMA protsessid

Table of content

Introduction

The presence of long memory is an important aspect in modeling financial time-series. It influences on the behavior of investors, which can make their decisions based on different investment horizons. First studies of possibility of statistical dependence in asset returns was Mandelbrot(1971). After that, for example, Greene & Fielitz (1977) discovered long-range dependence in daily returns of assets listed on NYSE.

The property of long memory is usually related to the persistence that is shown by the sample autocorrelations of certain stationary time series, which decrease at a very slow rate, but finally converge towards zero, indicating that the innovations of these series have transient effects but last for a long time. This behavior is not compatible neither with the stationary models, which have exponential decrease in autocorrelations and therefore in effects of the innovations[1], nor with the integrated models, where innovations have permanent effects.

The model, which shows this evidence, was introduced by Granger and Joyeux in 1980 as a generalized version of ARIMA models. It was shown, that for suitable values of fractional differencing factor the processes can model long-term persistence.

In this work our aims will be:

- to introduce the fractional ARIMA model as one of the models which has the property of the long memory dependence in autocorrelations;
- to test the model on the data;
- to see how this model perform in forecasting in comparison with the standard ARIMA models.

The thesis is divided into two main parts. First part is theoretical, which consists of review of long memory models and description of statistical tests which are used for the data examination. In the practical part, using rescaled range statistic (R/S) test we find evidence of long memory in absolute returns of daily values of S&P500 index. Using the Geweke-Porter-Hudak estimator we compute the fractional differencing parameter of ARFIMA model. After that we fit long-memory models to market data and compare the accuracy of of forecasting with long and short memory models based on RMSE statistic.

[1] Innovation - the difference between the observed value of a variable at time t and the optimal forecast of that value based on information available prior to time t.

1. Long memory processes

1.1 Literature review

The empirical evidence of the long memory processes goes back to Hurst (1951) in field of hydrology. However, interest in long memory models for economic series arises from the works of Granger and Joyeux (1980), who noted that many such series are apparently not stationary in mean, and yet, the differentiated series usually present clear evidence of overdifferencing.

For example, if the series exhibits any long-term trends, this will produce positive autocorrelations out to a high number of lags in the ACF plot. Therefore, the series should be differenced until the data is stationary. At the same time, if we take the difference over the stationary process, the result of overdifferencing can be more complicated model. Problem appears because the difference of a stationary series is not invertible. For example, if $Y_t = 0.5Y_{t-1} + \varepsilon_t$, so that Y_t is the stationary AR(1), then the first difference Z_t is the non-invertible ARMA(1,1) process[2] $Z_t = 0.5Z_{t-1} + \varepsilon_t - \varepsilon_{t-1}$, which has more parameters than original process. Because of the non-invertability of Z_t, its parameters will be hard to estimate, and it makes difficult to make the forecast of Z_{t+h}. (Hurvich, Differencing and unit root test, NYU)

Coming back to visible evidence of overdifferencing, the differencing in general introduces negative correlation to the series, driving the autocorrelation of the lag-1 term towards negative values. If the lag-1 autocorrelation becomes negative, the series does not need to be differenced further. If the lag-1 autocorrelation becomes less than -0.5, it is possible the series is over-differenced. For the given example this evidence is not observed.

Therefore, to model this type of series, the differentiation seems "excessive" but non-differentiation is not adequate either. To cover this gap between the extreme cases of ARIMA models with unit roots, typically used to model non-stationary series whose level evolves in time, and stationary ARMA models where the mean level is constant and the series returns relatively quickly to that level, Granger and Joyeux (1980) and Hosking (1981) proposed a class of intermediate processes in which the integration order is fractional. These are ARMA processes fractionally integrated, ARFIMA (p, d, q), where d is a real number. By allowing the order of integration, d, to be a non-integer number, these models act as a "bridge" between

[2] An ARMA (p, q) is invertible if the largest root θ of the equation $z^q + b_1 z^{q-1} + \cdots + b_q = 0$ satisfies $|\theta| < 1$, where $b_1, \ldots, b_q$ are the MA parameters.

the processes with ARIMA unit roots (d = 1) and stationary ARMA processes (d = 0). When 0 <d <1/2, the ARFIMA processes are stationary, that is, its mean level is constant, but deviations from the series over this level have a longer duration than when d = 0.

The presence of long memory in economic series may be justified by what Granger (1966) (Granger, 1966) called the "typical" form of the spectrum of the series, which is characterized by not being bounded in the low frequencies and decreasing hyperbolically to zero. This must be added to the results on aggregation of Robinson (1978) and Granger (1980), which shows that the sum of independent AR (1) processes, whose coefficients follow a Beta-type distribution, is a fractionally integrated process. Many economic variables are aggregates of other variables, this result could explain the presence of long memory in certain economic series. Other alternative explanation of the existence of long memory in the economic aggregates can be seen in Parke (1999).

The empirical evidence on the presence of long memory in economic series and financial services is extensive. To name a few examples, Greene and Fielitz (1977) use the statistic of rescaled rank to contrast the presence of long memory in 200 series of yields and find evidence in a large number of them. Subsequently, Lo (1991) detects long memory in financial returns using a modification of said statistic. Also, Cheung (1993) and Baillie and Bollerslev (1994) find evidence of long memory in the prices of assets. In macroeconomic series, e.g. Diebold and Rudebush (1989) and Sowell (1992) find long memory in quarterly series of the American Gross National Product, and Hassler and Wolters (1995) and Baillie, Chung and Tieslau (1996) - in different monthly series of inflation. On the other hand, the existence of seasonal long memory has been empirically observed, among others, by Porter-Hudak (1990), Ray (1993), or Franses and Ooms (1997). In recent years, there has also been a great deal of interest in the use of long memory processes for modeling the volatility of financial series.

The pioneering work of Ding, Granger and Engle (1993) revealed that sample autocorrelations of certain transformations of absolute yields of the S&P500 stock index decline very slowly towards zero, in line with the long memory property. Later works, such as Crato and de Lima (1994), Bollerslev and Mikkelsen (1996) and Lobato and Robinson (1998) have confirmed the evidence of long memory in the squares of different financial series. Motivated by these a number of models have recently been proposed which seek to represent the property of long memory in conditioned moments of second order.

1.2 Long memory processes definition

The long memory models have played a significant role already at least since 1950, and have been used by statisticians from different fields starting from physical science, hydrology or climatology, when the presence of long memory within data recorded over both time and space was recognized, to econometricians in early 1980-s. The presence of long memory can be defined from an empirical, data-oriented approach in terms of the persistence of observed autocorrelations. In the case of a long memory process the behavior of the autocorrelations is essentially consistent with a stationary process, but decay way slower than the exponential rate associated with the ARMA class of models.

This phenomenon was described in different data series by Hurst (1951, 1957), Mandelbrot (1972), and McLeod and Hipel (1978) among others. If considered as observations of the time series of a stochastic process, the autocorrelation function of those series exhibits persistence that is neither consistent with an I(1) process nor an I(0) process.

A significant success in econometrics has been obtained from using the ARMA class of models which impose an exponential, or geometric, rate of decay on the Wold decomposition[3] coefficients. At the same time, there is no conceptual reason for restricting attention to exponential rates of decay in the Wold decomposition, and there are indeed both theoretical and economic reasons for considering slower rates, such as hyperbolic decay[4]. While a most of recent works have emphasized the role of persistence of shocks, significant part of it has been directed towards testing for the presence of unit roots in autoregressive representations of univariate and vector processes. However, the sharp distinction between I(0) and I(1) processes may be too restrictive. The fractionally differenced process can be regarded as a compromise between the I(0) and I(1) paradigms. One attraction of long memory models is that they imply different long run predictions and effects of shocks to conventional macroeconomic approaches.

The origin of interest in processes with a long memory is connected with the examination of data in the physical sciences and preceded interest from economists. Perhaps

[3] Wold's decomposition, or the Wold representation theorem, says that every covariance-stationary time series can be written as the sum of two time series, one deterministic and one stochastic.
Formally, $Y_t = \sum_{j=0}^{\infty} b_j \, \varepsilon_{t-j} + \eta_t$, where Y_t - time series, ε_t - uncorrelated sequence, b - the possibly infinite vector of moving average weights (coefficients or parameters), η_t - deterministic time series.

[4] The sample autocorrelations $\hat{\rho}(k) = \frac{1}{n}\sum_{i=1}^{n-|k|}(x_i - \bar{x})(x_{i+|k|} - \bar{x})$, where $\bar{x} = \frac{1}{n}\sum x_i$, decays slowly with increasing lag k, and such decay of $\hat{\rho}(k)$ is called hyperbolicwith a rate k^{-a} for $0 < \alpha < 1$. (Beran, Long-memory processes, 2013)

the most well-known example has seen in hydrology. It included tidal flows and the inflows into reservoirs and was originally documented by Hurst (1951).

The property of long memory can be described, for example, using a discrete time series process Y_t with autocorrelation function ρ_j at lag j.

According to McLeod & Hipel (1978), the process possesses long memory if the quantity

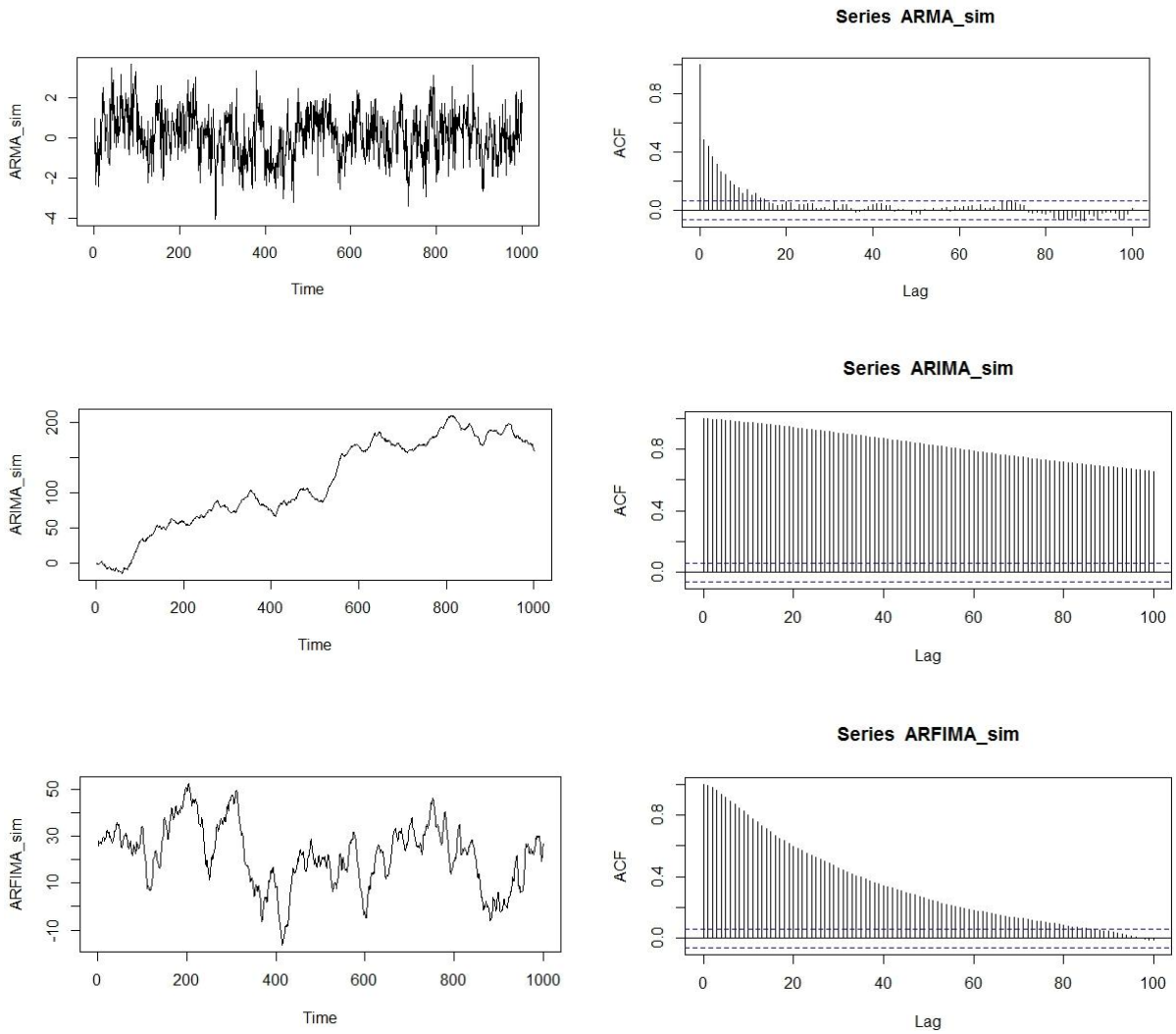
$$\lim_{n\to\infty} \sum_{j=-n}^{n} |\rho_j|$$

is nonfinite. A stationary and invertible ARMA process has autocorrelations which are geometrically bounded, i.e., $|\rho_k| \leq cm^{-k}$, for large k, where 0 < m < 1 and is hence a short memory process (Baillie, 1996).

To see how autocorrelation function behave in different models, we simulated ARMA(1,1), ARIMA(1,1,1) and ARFIMA(1,0.45,1) series with equal coefficients: ϕ=0.9 and θ=-0.6. From the Scheme 1.1 we can observe the decay in autocorrelations close to linear in autocorrelations in integrated model and the exponential decay of autocorrelations in fractionally differenced model. Also we can see that the model ARMA(1,1) has the distribution around the constant zero, ARIMA(1,1,1) does not have a constant mean value, and the fractionally integrates series ARFIMA(1,0.45,1) shows the transitional behaviour between those two models.

Scheme 1.1

Comparison of autocorrelation functions of ARMA(1,1), ARIMA(1,1,1) and ARFIMA(1,0.45,1)



1.3 ARFIMA processes

ARIMA models were introduced by Box and Jenkins (1976). They are one of the general class of models for forecasting a time series, which consists the integrated part to allow also the modelling of non-stationarity. A random process describing a time series is stationary if its statistical properties are all constant over time. A stationary series has no trend, the variance around its mean is constant in time, its autocorrelations remain constant

over time. A random variable of this form can be described as a combination of signal and noise, and ARIMA model can be viewed as a filter that tries to separate the signal from the noise, and the signal is then extrapolated into the future to obtain forecasts.

In case of financial forecasting, there is a need for a models which have the property of long-memory. There is a family of model which does meet these property, by generalizing the ARIMA model. The generalization consists of permitting the degree of differencing *d* to take any real value rather than being restricted to integral values; it turns out that for a suitable values of *d*, specifically $0 < d < 1/2$, these 'fractionally differenced' processes are capable of modeling long-term persistence.

The population characteristics of ARFIMA processes have been extensively studied by Granger(1980), Granger and Joyeux (1980), and Hosking (1981). For $0<d<1/2$ the process Y_t is covariance stationary and the moving average coefficients decay at a relatively slow hyperbolic rate compared with the stationary and invertible ARMA process where the moving average coefficients decline exponentially with increasing lag.

The general form of ARFIMA (p, d, q) model can be written as:

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

where $\varepsilon_t \sim iid(0,\sigma^2)$, B is the backward-shift operator,

$$\Phi(\mathrm{B})=1-\phi_1 B - \cdots - \phi_p B^p,$$

$$\Theta(\mathrm{B})= 1-c_1 B + \cdots + \theta_q B^q,$$

and $(1-\mathrm{B})^{d}$ is the fractional differencing operator defined by the Tailor expansion:

$$(1-\mathrm{B})^{d} = \sum_{\mathrm{k}=0}^{\infty}\frac{\mathrm{d}(\mathrm{d}-1)\ldots(\mathrm{d}-\mathrm{k}+1)}{\mathrm{k}!}(-\mathrm{B})^{\mathrm{k}} = 1-\mathrm{dB}-\frac{1}{2}\mathrm{d}(1-\mathrm{d})\mathrm{B}^2-\frac{1}{6}\mathrm{d}(1-\mathrm{d})(2-\mathrm{d})\mathrm{B}^3-\cdots.$$

X is both stationary and invertible if the roots of Φ(B) and $\Theta(B)$ are outside the unit circle and $\mathrm{d}<|1/2|$. The Wold decomposition and autocorrelation coefficients will both exhibit a very slow rate of hyperbolic decay. When d=0, an ARFIMA process reduces to an ARMA process.

Hosking (1981) showed that the autocorrelation, ρ(.), of an ARFIMA process satisfies $\rho(\mathrm{k})\sim k^{2d-1}$ for $0<\mathrm{d}<1/2$ as $\mathrm{k}\to\infty$. Thus the memory property of a process depends crucially on the value of d. When $\mathrm{d}\in(0,1/2)$, the autocorrelations do not have a finite sum. When $\mathrm{d}\leq 0$, the autocorrelations have a finite sum; that is, ARFIMA processes with $\mathrm{d}\in(0,1/2)$ display long memory.

For $-1/2 < d < 0$, the sum of absolute values of the processes autocorrelations tends to a constant, so that it has short memory according to definition of long memory. In this case the ARFIMA(0, d, 0) process is 'antipersistent' or has 'intermediate memory', and all its autocorrelations, except lag zero, are negative and decay hyperbolically to zero.

The effect of the *d* parameter on distant observation decays hyperbolically when the lag increases, while the effects of the ϕ_i and θ_j parameters decay exponentially. Thus *d* may be chosen to describe the high-lag correlation structure of a time series while the ϕ_i and θ_j parameters are chosen to describe the low-lag correlation structure.

There are several approaches of estimating the fractional differencing parameter. Graphical method based on R/S statistics and the variance-time plot was described by Leland et al.(1994). The common parametric method is based on assumptions that d-th fractional difference of series follows a standard ARMA model, the order (p,q) is already known and model parameters including d are estimated by likelihood procedure (Fox & Taqqu, 1986). The algorithmic aspects computing the likelihood estimates are discussed by Hosking (1984), Haslett and Raftery (1989), Sowell (1992). In this thesis will be used the most widely studied non-parametric method, described by Geweke and Porter-Hudak (1983).

ARFIMA models have also a lot of extensions comparisons of their relative forecasting performance. For example, Franses & Ooms (1997) proposed the periodic ARFIMA(0, d, 0) model where d can change seasonally. Ravishanker & Ray (2002) described the estimation and forecasting of multivariate ARFIMA models. Baillie & Chung (2002) considered the linear trend-stationary ARFIMA models, and Beran et al. (2002) extended this model to allow for nonlinear trends. Souza & Smith (2002) investigated the effect of different timeframes on estimates of the long-memory parameter d, such as monthly and quarterly ones. Similarly, Souza & Smith (2004) looked at the effects of temporal aggregation on estimates and forecasts of ARFIMA processes. Within the context of statistical quality control, Ramjee et al. (2002) introduced a hyperbolically weighted moving average forecast-based control chart, designed specifically for non-stationary ARFIMA models. (De Gooijer & Hyndman, 2006) in their research paper summarized the works and the results gained since the year 1982.

1.4 Testing for long memory

There are various methods that are frequently used for the recognition of a long memory. Mandelbrot has suggested to use the range over standard deviation or R/S statistics, also called "rescaled range". It uses the Hurst exponent, which was produced by a British hydrologist Harold Hurst in 1951 during his studies of river discharges.

The main idea behind the R/S analysis is that one looks at the scaling behavior of the rescaled cumulative deviations from the mean. The R/S analysis first estimates the range R for a given n (Lo, 1991):

$$R_n = \max_{m=1,\dots,n} \sum_{j=1}^{m} (Y_j - \bar{Y}) - \min_{m=1,\dots,n} \sum_{j=1}^{m} (Y_j - \bar{Y}),$$

where R_n is the range of accumulated deviation of Y_t over the period of n and $\bar{Y}$ is the overall mean of the time series. Let $S_n = \left[\frac{1}{n}\sum_j (Y_j - \bar{Y})^2\right]^{1/2}$ - the usual standard deviation estimator.

As n increases, the following holds:

$$\text{Log } [R_n / S_n] = \log \alpha + H \log n$$

This implies that the estimate of the Hurst exponent H is the slope. Thus, H is a parameter that relates mean R/S values for subsamples of equal length of the series to the number of observations within each equal length subsample. H is always greater than 0. When 0<H<1, the long memory structure exists. If H ≥ 1, the process has infinite variance and is non-stationary. If 0 < H < 1/2, anti-persistence structure exists. If H = 1/2, the process is white noise. (Wei & Leuthold, 2000)

1.5 Statistical testing of data

Jarque-Bera is a test statistic for testing whether the series is normally distributed (Jarque & Bera, 1980). The test statistic measures the difference of the skewness and kurtosis of the series with those from the normal distribution. The statistic is computed as:

$$JB = \frac{n}{6}\left(S^2 + \frac{(K-3)^2}{4}\right),$$

where S is the skewness and K is the kurtosis. Under the null hypothesis of a normal distribution, the Jarque-Bera statistic is distributed as χ^2 with 2 degrees of freedom. The reported probability is the probability that a Jarque-Bera statistic exceeds (in absolute value)

the observed value under the null hypothesis—a small probability value leads to the rejection of the null hypothesis of a normal distribution.

For testing for presence of unit root (that is, for the need of differencing of the series for it to become stationary), we are going to use ADF test. It is an augmented version of the Dickey–Fuller test (Dickey & Fuller, 1979) for a larger and more complicated set of time series models.

The testing procedure for the ADF test is applied to the model:

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \delta_1 \Delta y_{t-1} + \cdots + \delta_{p-1} \Delta y_{t-p+1} + \varepsilon_t,$$

where α is a constant, β the coefficient on a time trend and p the lag order of autoregressive process. (Cheung & Lai, 1995) The null hypothesis is $\gamma = 0$, and the alternative hypothesis is $\gamma < 0$. Once a value for the test statistic

$$DF = (\hat{\gamma} - 1)/SE(\hat{\gamma})$$

is computed, it can be compared to the relevant critical value for the Dickey–Fuller Test. The output value is a negative number, and the more negative is, it means the stronger rejection of the hypothesis that there is a unit root.

Important aspect is the choice of p, the number of AR terms. There are several papers outlining the possible approaches. One possible approach is to test down from high orders and examine the t-values on coefficients. (Corbae & Ouliaris, 1988) An alternative approach is to examine information criteria such as for example, the Akaike information criterion (AIC), as it is advised by Brockwell and Davis (1991, 2002). In this thesis we use the default value of the parameter in R, which is calculated inside the function 'adf.test' as 'trunc((length(x)-1)^(1/3))', where x is a numeric vector or time series.

To compare the goodness of fit of chosen models, we use the Akaike information criterion (AIC). The measuring the goodness of fit for some particular model can be done by balancing the error of the fit against the number of parameters in the model. It provides the measure of information lost when a given model is used to describe reality. AIC values provide a means for model selection and cannot say anything about how well a model fits the data in an absolute sense. If the entire candidate models fit poorly, AIC will not give any warning of that. The AIC is defined as

$$AIC = 2k - 2\ln(\hat{L}),$$

where *k* is the number of parameters in the statistical model, and *L* is the maximized value of the likelihood function for the estimated model. The AIC is applied in model selection in which the model with the least AIC is selected as the best candidate model.

To distinguish the best model after procedure of prediction, the RMSE as the most frequently used measure to draw conclusions about forecasting methods. To calculate the RMS (root mean squared) error, the individual errors are squared, added together, divided by the number of individual errors, and then square rooted. This gives a single number that summarizes the overall error.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}}$$

where $(\hat{y}_i - y_i)$ - residuals, and y_i are observed values, $\hat{y}_i$ - predicted values.

2. Fitting the models

2.1 Data downloading and statistical testing

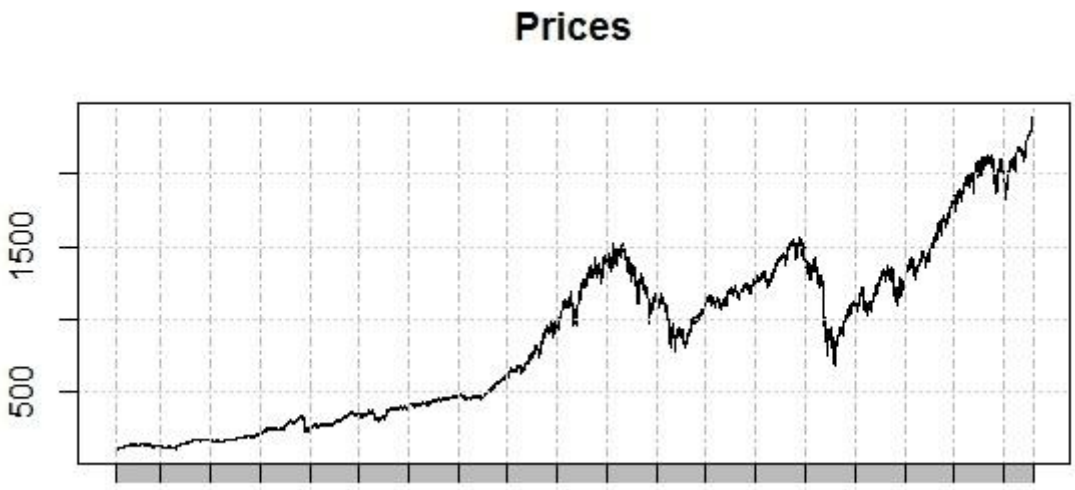
The data set that will be analyzed in this chapter is S&P 500 stock market daily closing price index. The software for computation is R. R programming language as a software environment for statistical computing and graphics has many capabilities which are extended by various packages. The series was obtained from Yahoo Finance with R package 'quantmod'. There are altogether 9340 observations from 30.04.1980 to 03.15.2017. Last 10 observations will be subtracted from the main part for the purpose of comparing of forecasting accuracy of applied models.

```
#getting the data (SP500 index)
getSymbols('^GSPC',src='yahoo',from='1980-03-01',to='2017-03-15')
## [1] "GSPC"
head(GSPC)
##            GSPC.Open GSPC.High GSPC.Low GSPC.Close GSPC.Volume
## 1980-03-03    113.66    114.34   112.01     112.50    38690000
## 1980-03-04    112.50    113.41   110.83     112.78    44310000
## 1980-03-05    112.78    113.94   110.58     111.13    49240000
## 1980-03-06    111.13    111.29   107.85     108.65    49610000
## 1980-03-07    108.65    108.96   105.99     106.90    50950000
## 1980-03-10    106.90    107.86   104.92     106.51    43750000
##            GSPC.Adjusted
## 1980-03-03        112.50
## 1980-03-04        112.78
## 1980-03-05        111.13
## 1980-03-06        108.65
## 1980-03-07        106.90
## 1980-03-10        106.51
Prices <- GSPC$GSPC.Close
Returns <- diff(log(Prices))
Returns_vect <- as.vector(Returns)
Returns_vect <- Returns_vect[-1]
```

On the Scheme 2.1 there is the plot of original time series - the daily prices index. On the Scheme 2.2 simple log-returns of prices were plotted.

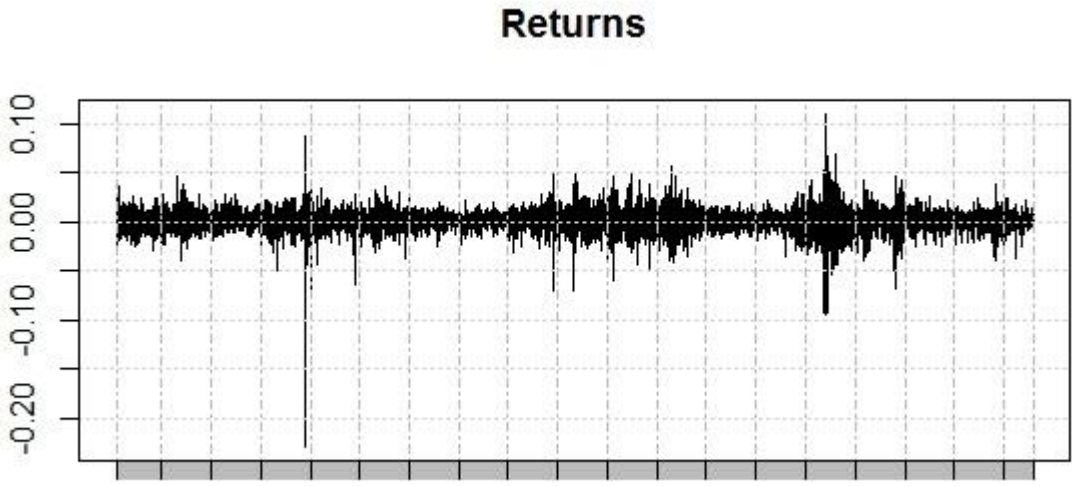
Scheme 2.1

S&P 500 daily price index 01.03.1980 - 15.03.2017



Scheme 2.2

S&P 500 daily returns 01.03.1980 - 15.03.2017



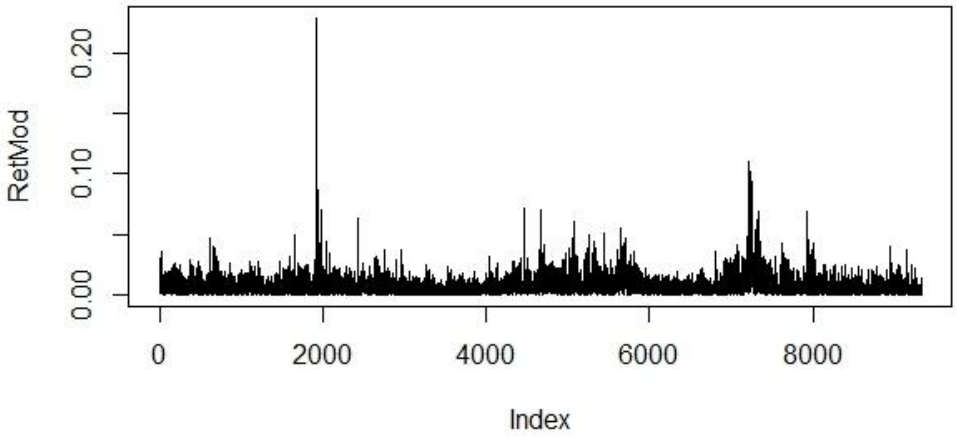
For analysis there were picked absolute returns as a typical transform of the return series (Scheme 2.3). It is one of stylized statistical properties of asset returns that the autocorrelation function of absolute returns decays slowly (Cont, 2001). To see this property and also the behavior of partial autocorrelations, we plot the ACF and PACF graphs.

```
RetMod_all <- abs(Returns_vect)#absolute log-returns
RetMod <- RetMod_all[-(9331:9340)]#substract the values we are going
```

```
to forecast
#now we have returns up to 2017-03-01
plot(RetMod, type = 'l')
acf(RetMod,lag.max = 200)
pacf(RetMod,lag.max = 200)
```

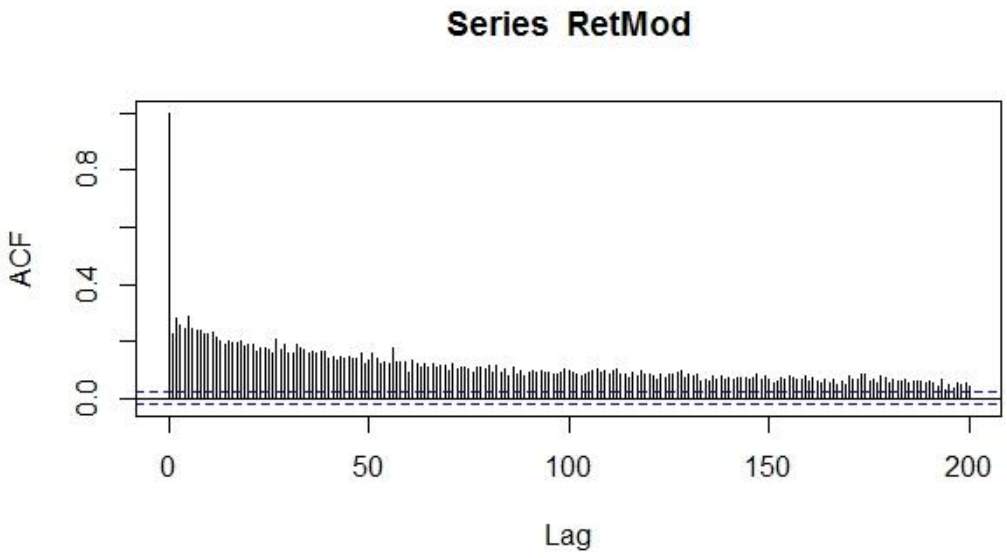
Scheme 2.3

S&P 500 daily absolute returns 30.04.1980 - 28.04.2017



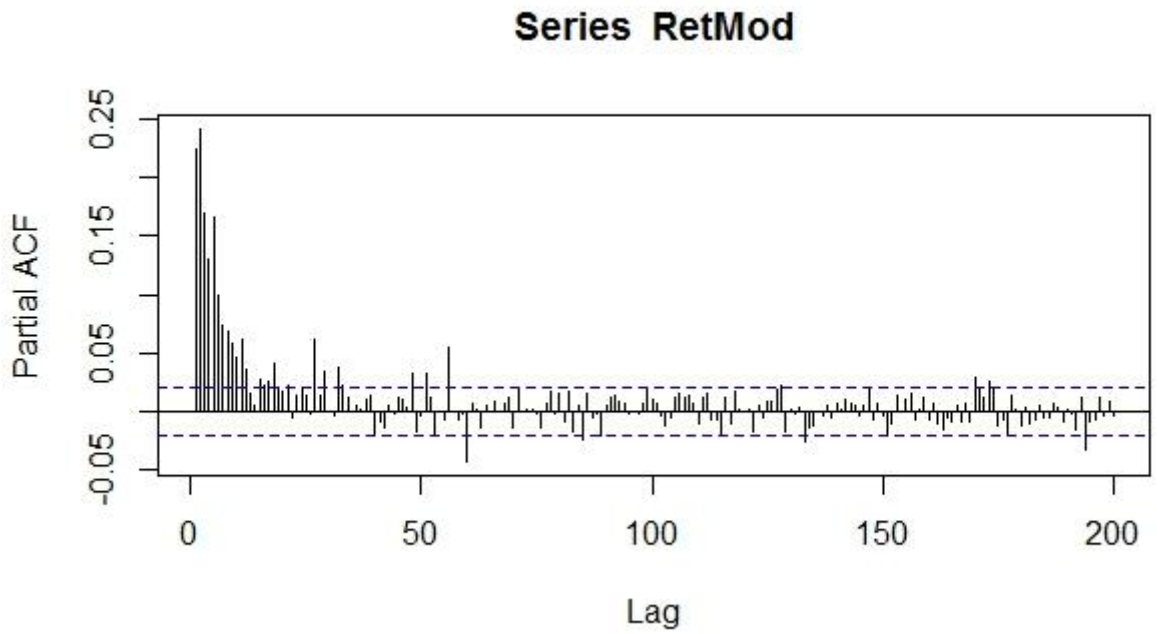
Scheme 2.4

Autocorrelations of absolute returns up to lag 200



Scheme 2.5

Partial autocorrelations of absolute returns up to lag 200



Schemes 2.4 and 2.5 graph the first 200 autocorrelation and partial autocorrelation coefficients for the absolute returns respectively. The autocorrelations exhibit a clear pattern of persistence and slow decay which is typical of a long-memory process.

Before starting with model estimation, we can take a look at some statistics of this time-series.

```
summary(Returns_vect)
##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
## -0.2290000 -0.0046270  0.0005319  0.0003270  0.0056620  0.1096000
skewness(Returns_vect)
## [1] -1.156088
kurtosis(Returns_vect)
## [1] 29.67046
jarque.bera.test(Returns_vect)
##
##  Jarque Bera Test
##
## data:  Returns_vect
## X-squared = 278900, df = 2, p-value < 2.2e-16
```

Of course, we have to be careful about the results of the three last values since the data does not correspond to independent samples from a distribution but are serially correlated, but

the output values still give some feeling about the nature of the data in the series. We can see from output that kurtosis of 29.67 is higher than that of a normal distribution which is 3. It shows the characteristic 'fat-tailed' behavior compared with a normal distribution. The Jarque-Bera normality test statistic is far beyond the critical value which suggests that absolute returns series is far from a normal distribution.

To check the stationarity, the Augmented Dickey Fuller (ADF) test is mostly used. The ADF test examines the null hypothesis that a time series is stationary against the alternative that it is non-stationary.

```
adf.test(RetMod, alternative = "stationary")
## Warning in adf.test(RetMod, alternative = "stationary"): p-value
smaller
## than printed p-value
##
##  Augmented Dickey-Fuller Test
##
## data:  RetMod
## Dickey-Fuller = -10.309, Lag order = 21, p-value = 0.01
## alternative hypothesis: stationary
```

The result shows, that since the p-value is smaller than 0.05, we can reject the null hypothesis that the series has a unit root. If there are no unit roots, then we conclude the series is stationary.

2.2 Choosing the model

The presence of long memory is tested using Hurst exponent produced by the Rescaled range analysis. The value of H indicates that the absolute returns have long memory structure since 0.5<H<1.

```
hurstexp(RetMod)
## Simple R/S Hurst estimation:         0.7323002
## Corrected R over S Hurst exponent:   0.8441684
## Empirical Hurst exponent:            0.9168726
## Corrected empirical Hurst exponent:  0.8889158
## Theoretical Hurst exponent:          0.5270019
```

The long memory parameter d is estimated with Geweke-Porter-Hudak model. The estimated value of the parameter, its asymptotic deviation value and regression standard deviation values are reported in the following output:

```
fdGPH(RetMod, bandw.exp = 0.5)
## $d
## [1] 0.4610339
##
## $sd.as
## [1] 0.07104335
##
## $sd.reg
## [1] 0.06560436
```

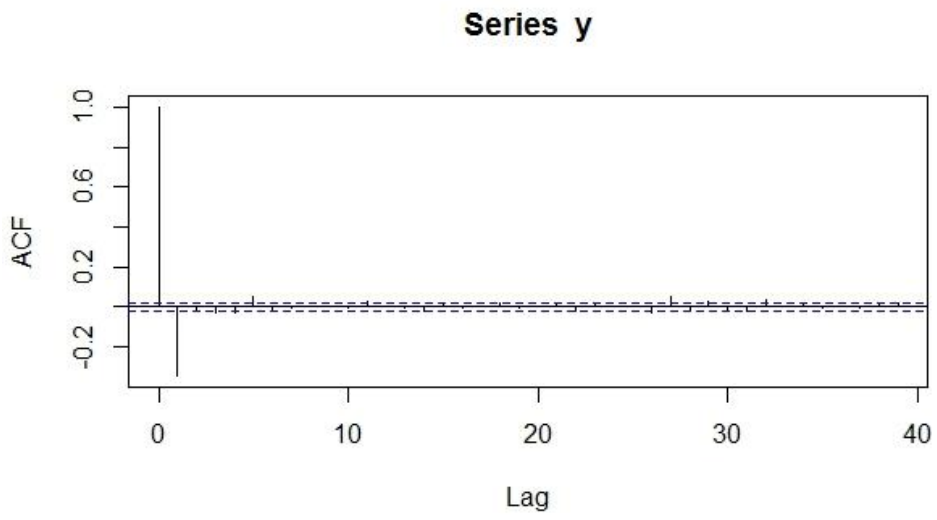
The value of d, 0<0.46<0.5 shows, that we can use this coefficient for estimation of ARFIMA model.

To choose the appropriate candidate model, we fit ARFIMA with fixed fractional parameter, and then look at ACF and PACF plots of residuals.

```
library(arfima)
x <- arfima(RetMod, fixed=list(frac=0.46),dmean=FALSE)
## Note: only one starting point.  Only one mode can be found.
## Beginning the fits with 1 starting values.
y <- resid(x)
y <- as.numeric(unlist(y))
acf(y)
pacf(y)
```

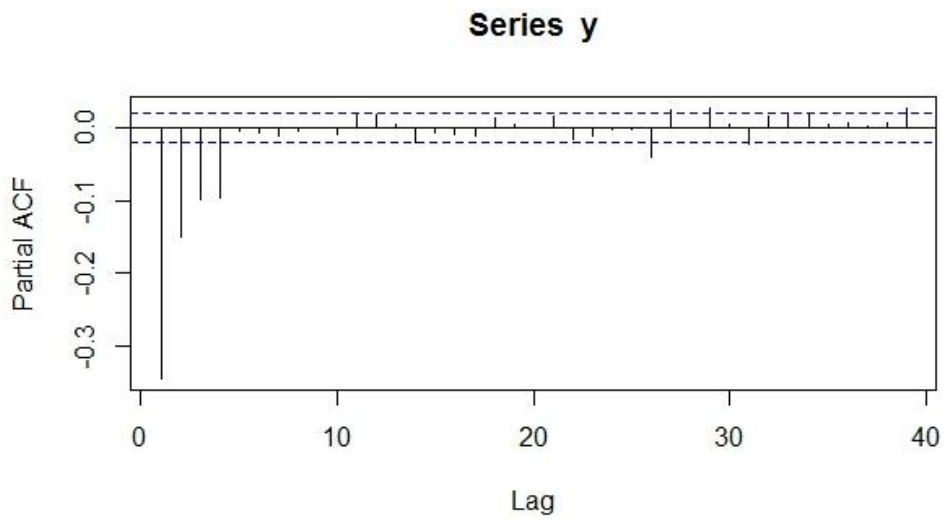
Scheme 2.6

Autocorrelation function of residuals



Scheme 2.7

Partial autocorrelation function of residuals



Looking at ACF and PACF plots (Scheme 2.6 and 2.7), we can see the evidence of MA(1) term on the first plot, and AR(4) on the second one. Therefore, two candidate models for fitting are ARFIMA(4,0.46,0) and ARFIMA(1,0.46,0).

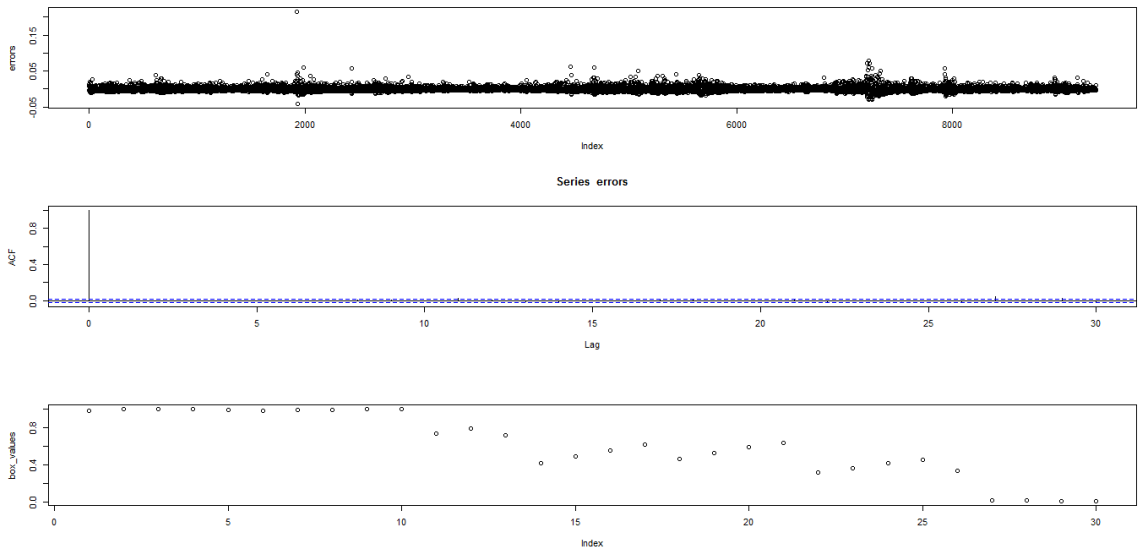
```
fitArfima1 <-
arfima(RetMod,order=c(4,0,0),fixed=list(frac=0.46),dmean=FALSE)
fitArfima2 <-
arfima(RetMod,order=c(0,0,1),fixed=list(frac=0.46),dmean=FALSE)
fitArfima1
## Number of modes: 1
## Warning in rbind(coeff, ses): number of columns of result is not
a multiple
## of vector length (arg 2)
##
## Call:
## arfima(z = RetMod, order = c(4, 0, 0), dmean = FALSE, fixed =
list(frac = 0.46))
##
## Coefficients for fits:
##           Coef.1:       SE.1:
## phi(1)    -0.422087      0.0103064
## phi(2)    -0.207168      0.0111031
## phi(3)    -0.137832      0.0111043
## phi(4)    -0.0959633     0.0103073
## d.f        0.46
## zbar       0.00749157
## logl       45745.6
## sigma^2    5.51229e-05
```

```
## phi_p(1)  -0.346046
## phi_p(2)  -0.149939
## phi_p(3)  -0.0982315
## phi_p(4)  -0.0959633     0.0103064
## Starred fits are close to invertibility/stationarity boundaries
fitArfima2
## Number of modes: 1
## Warning in rbind(coeff, ses): number of columns of result is not
a multiple
## of vector length (arg 2)
##
## Call:
## arfima(z = RetMod, order = c(0, 0, 1), dmean = FALSE, fixed =
list(frac = 0.46))
##
## Coefficients for fits:
##           Coef.1:       SE.1:
## theta(1)   0.443533      0.00991675
## d.f        0.46
## zbar       0.00749157
## logl       45717.8
## sigma^2    5.54353e-05   0.00991675
## Starred fits are close to invertibility/stationarity boundaries
AIC(fitArfima1)
## [1] -91477.28
AIC(fitArfima2)
## [1] -91427.54
```

The output shows, that the model ARFIMA(4,0.46,0) has the smaller value of Akaike criterion and bigger log-likelihood value. So, it is the model which will be used for predictions. The statistics for the model in on the Scheme 2.8. P-values are significant for more than 20 days, which shows the evidence of long memory of the model.

Scheme 2.8

Statistics for ARFIMA(4,0.46,0)



Next step is choosing model without differencing and with ordinary differencing to compare the accuracy of prediction. Looking at the plot of ACF and PACF (Scheme 2.4 and 2.5), it is hard to identify the right type of model, since it does not correspond to simple AR or MA, so to find the best suitable model the function 'auto.arima' was used. Also, since we have daily market data, it is reasonable to allow a seasonal part with period 5 (weekly dependence).

```
library(forecast)
seas <- ts(RetMod, frequency = 5)#returns with weekly seasoning
auto1 <- auto.arima(seas,max.p = 10,max.q =
10,stationary=TRUE,stepwise=FALSE,approx=TRUE,seasonal = TRUE)
auto1
## Series: seas
## ARIMA(1,0,3)(1,0,0)[5] with non-zero mean
##
## Coefficients:
##          ar1      ma1     ma2      ma3    sar1    mean
##       0.9853  -0.9478  0.0755  -0.0356  0.0620  0.0075
## s.e.  0.0026   0.0107  0.0143   0.0109  0.0112  0.0005
##
## sigma^2 estimated as 5.516e-05:  log likelihood=32505.83
## AIC=-64997.65   AICc=-64997.64   BIC=-64947.67
auto2 <- auto.arima(seas,max.p = 10,max.q =
10,stepwise=FALSE,approx=TRUE,seasonal = TRUE)
auto2
```

```
## Series: seas
## ARIMA(2,1,2)(1,0,0)[5]
##
## Coefficients:
##          ar1     ar2      ma1     ma2    sar1
##       0.8614  0.0707  -1.8234  0.8256  0.0497
## s.e.  0.0144  0.0115   0.0104  0.0104  0.0109
##
## sigma^2 estimated as 5.525e-05:  log likelihood=32493.53
## AIC=-64975.06   AICc=-64975.06   BIC=-64932.22
```

The result of autofitting is two models - ARIMA(1,0,3) and ARIMA(2,1,2) with seasonal order of differencing 1 and period 5. These models will be fitted to the data and be compared with output of ARFIMA(4,0,0).

```
fit_arima1 <- arima(RetMod,order=c(1,0,3),seasonal = list(order =
c(1, 0, 0), period = 5))
fit_arima2 <- arima(RetMod,order=c(2,1,2),seasonal = list(order =
c(1, 0, 0), period = 5))
```

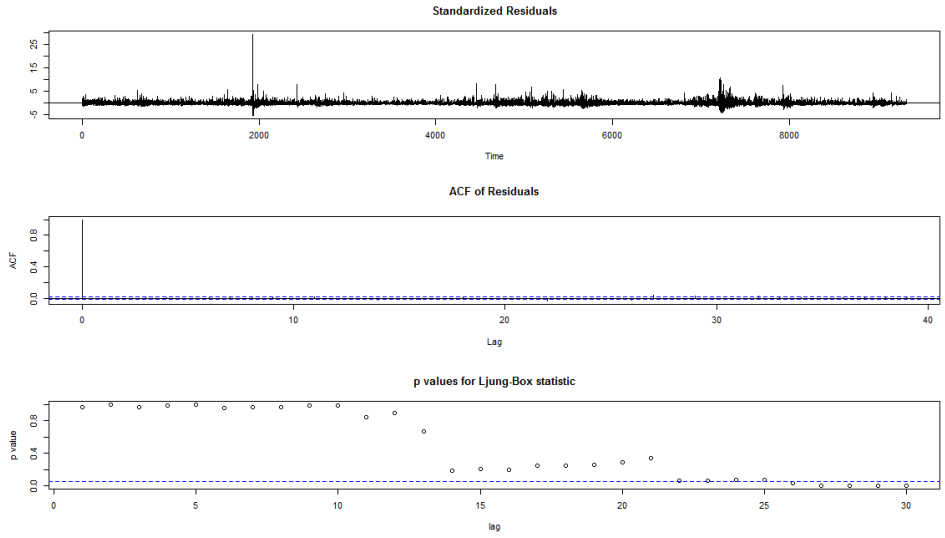
The evidence of long memory can be seen with plotting time-series diagnosis:

```
tsdiag(fit_arima1,30)
tsdiag(fit_arima2,30)
```

The output is on the Scheme 2.9 and 2.10.

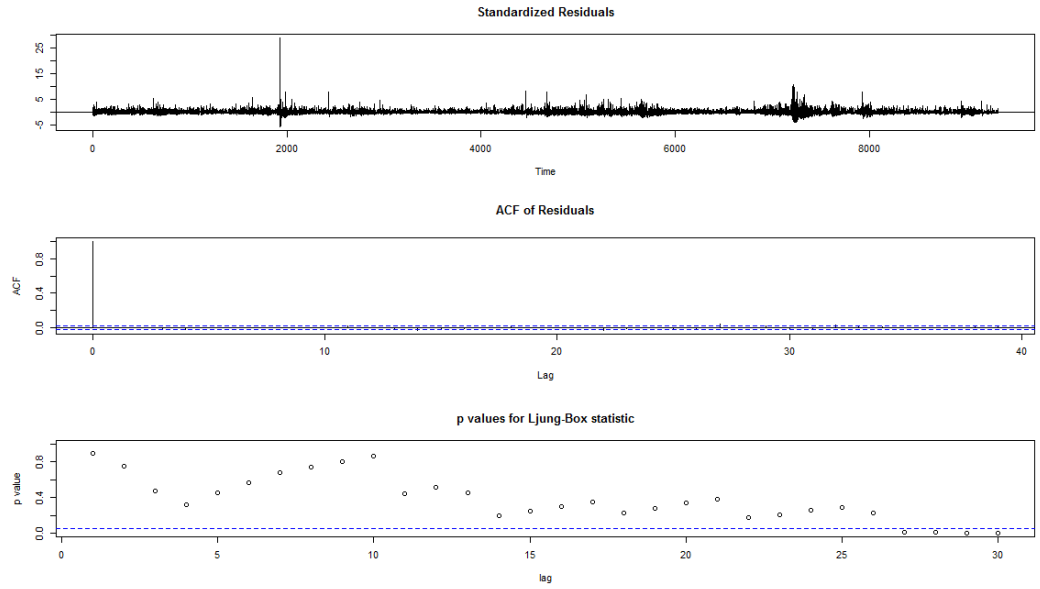
Scheme 2.9

Statistics for ARIMA(1,0,3)



Scheme 2.10

Statistics for ARIMA(2,1,2)



For both models p-values are significant more that for 20 days.

2.3 Computing predictions

For each model were made 10 step-by-step computations for one-step prediction and for 5 step prediction. The result of computations is in the Table 2.1.

Table 2.1

The observed and predicted values of absolute log-returns of S&P500

№	Observed	Predicted_1_ARIMA103	Predicted_1_ARIMA212	Predicted_1_ARFIMA400	Predicted_5_ARIMA103	Predicted_5_ARIMA212	Predicted_5_ARFIMA400
1	0.00587712	0.0035503	0.0031414	0.00338823	0.0035805	0.0031242	0.00340476
2	0.00050375	0.0045788	0.0040841	0.00437971	0.0035552	0.0030635	0.00333052
3	0.00328262	0.0042352	0.0039931	0.00411947	0.00341	0.0028731	0.00318093
4	0.00291763	0.0039482	0.0037454	0.00384276	0.0034915	0.0029362	0.00327164
5	0.00228683	0.0047123	0.0042442	0.00463897	0.0049926	0.0045304	0.00488985
6	0.00079958	0.0041453	0.003743	0.00423936	0.0047274	0.0043876	0.00480784
7	0.00326334	0.0035878	0.0032605	0.00361828	0.0041353	0.0037314	0.00415593
8	0.00036657	0.0036324	0.0031731	0.0035112	0.0042806	0.0037457	0.00417836
9	0.00338475	0.0036339	0.0031851	0.00343625	0.0042258	0.0036423	0.00409057
10	0.00833988	0.0034029	0.0029496	0.00322844	0.0040528	0.0034847	0.0038865

After that all models were tested for best fitting. For this purpose the root mean squared error (RMSE) test is used. It is a frequently used statistical measure of difference between predicted and actually observed values. The lower values of RMSE indicate better fit. RMSE is a good measure of how accurately the model predicts the response, and is an important criterion for fit if the main purpose of the model is prediction.

```
library(forecast)
acc_ARIMA_1_103 <- accuracy(Predicted_1_ARIMA103,Observed,d=0,D=1)
acc_ARIMA_1_103
##                       ME        RMSE         MAE      MPE     MAPE
## Test set -0.0008404959 0.002761757 0.002293254 -220.722 240.4797
acc_ARIMA_1_212 <- accuracy(Predicted_1_ARIMA212,Observed,d=0,D=1)
acc_ARIMA_1_212
##                       ME        RMSE         MAE       MPE    MAPE
## Test set -0.0004497313 0.002663837 0.002115445 -186.2934 209.727
acc_ARFIMA_1_400 <- accuracy(Predicted_1_ARFIMA400,Observed,d=1,D=0)
acc_ARFIMA_1_400
##                      ME        RMSE         MAE       MPE    MAPE
## Test set -0.000738062 0.002761932 0.002258127 -212.6304 233.358
acc_ARIMA_5_103 <- accuracy(Predicted_5_ARIMA103,Observed,d=0,D=1)
acc_ARIMA_5_103
##                       ME       RMSE       MAE       MPE     MAPE
## Test set -0.0009429649 0.00269902 0.0022597 -226.7709 244.8672
acc_ARIMA_5_212 <- accuracy(Predicted_5_ARIMA212,Observed,d=0,D=1)
acc_ARIMA_5_212
##                       ME        RMSE         MAE       MPE    MAPE
## Test set -0.0004497071 0.002597764 0.002053237 -188.1892 211.696
acc_ARFIMA_5_400 <- accuracy(Predicted_5_ARFIMA400,Observed,d=1,D=0)
acc_ARFIMA_5_400
##                       ME        RMSE        MAE       MPE    MAPE
## Test set -0.0008174846 0.002696999 0.00222297 -217.7912 237.504
```

Although all values are very small and the prediction horizon is only 10 days, looking at the root mean square error values and also at mean average error values we can conclude that the best model for prediction in this case for both 1-step and 5-step predictions is ARIMA(2,1,2).

Conclusion

Application of fractionally integrated models in forecasting future values of time-series is widely used. Many studies are focused on measuring forecast performance of ARIMA and ARFIMA models for stationary type series that exhibit long memory properties.

In the first part of the work provided brief background on important concepts used in thesis. Overview of literature in the field provides knowledge about long memory processes and fractionally integrated models. In that part, there was defined the long memory processes, fractionally integrated autoregressive moving average model was introduced. Description of statistical indicators explains the methods used in measuring forecasting accuracy.

The second part consists of empirical study of implementing ARFIMA model on the real market data. The result gained during numerous computations is not very obvious. Due to the small values of daily absolute returns and short horizon of prediction it is hard to distinguish the best model for future predictions. One of the difficulties is that programming tool for ARFIMA modeling (package 'arfima' for R by Justin Q. Veenstra) is still under developing, the functions are not optimally defined and computations take much more time comparing with popular ARIMA models. Theoretically it is also not clear if fractional differenced type of models captures the long-memory tendencies better than the models, where the differencing parameter is an integer. For example, (Ray, 1993) made such a comparison between ARFIMA models and standard ARIMA models. The results show that higher order AR models are capable of forecasting the longer term well when compared with ARFIMA models.

In final conclusion it can be stated that the evidence of long memory in fractionally integrated time-series was found. The ARFIMA model was applied on the market data and the forecasting using this model performed better that applying non-differenced model. The ARFIMA model was not found to be better than ARIMA model as indicated by model diagnostic tools. The estimated forecast values from ARFIMA model is as closely reflect the changing in absolute returns as indicated by the forecast evaluation tools applied on both non-integrated and integrated ARIMA models. Empirical studies show that further analysis is necessary for finding the advantages of using this model instead of ordinary ARIMA models.

References

Baillie, R. (1996). Long memory processes and fractional integration. *Journal of Econometrics 73* .

Baillie, R., & Bollerslev, T. (1994). The long memory of the forward premium. *Journal of International Money and Finance 13* , 565-571.

Baillie, R., & Chung, S.-K. (2002). Modeling and forecasting from trend-stationary long memory models with applications to climatology. *International Journal of Forecasting 18* , 215-226.

Baillie, R., Chung, C., & Tieslau, M. (1996). Analysing inflation by the fractionally integrated ARFIMA-GARCH model. *Journal of Applied Econometrics 11* , 23-40.

Beran, J. (2013). *Long-memory processes.* Springer-Verlag Berlin Heidelberg.

Beran, J., Feng, Y., Ghosh, S., & Sibbertsen, P. (2002). On robust local polynomial estimation with long-memory errors. *International Journal of Forecasting 18* , 227–241.

Bollerslev, T., & Mikkelsen, H. (1996). Modeling and pricing long memory in stock market volatility. *Journal of Econometrics 73* , 151-184.

Box, G., & Jenkins, G. (1976). *Time-series analysis: forecasting and control.* San Fransisco: Holden-Day.

Brockwell, P. J., & Davis, R. A. (1991). *Time Series: Theory and Methods (2nd ed.).* Springer.

Brockwell, P., & Davis, R. (2002). *Introduction to time-series and forecasting.* Springer.

Cheung, Y. (1993). Long memory in foreign-exchange rates. *Journal of business and economic statistics 11* , 93-101.

Cheung, Y. W., & Lai, K. S. (1995). Lag order and critical values of the augmented Dickey–Fuller test. *Journal of Business & Economic Statistics 13* , 277-280.

Cont, R. (2001). Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative Finance vol.1* , 223-236.

Corbae, D., & Ouliaris, S. (1988). Cointegration and Tests of Purchasing Power Parity. *The Review of Economics and Statistics 70* , 508-511.

Crato, N., & de Lima, P. (1994). Long range dependence in the conditional variance of stock returns. *Economics letters 45* , 281-285.

De Gooijer, J. G., & Hyndman, R. J. (2006). 25 Years of Time Series Forecasting. *International Journal of Forecasting 22* , 443 – 473.

Dickey, D., & Fuller, W. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association 74* , 427–431.

Diebold, F., & Rudebusc, G. (1989). Long-memory and persistence in aggregate output. *Journal of Monetary Economics 24* , 189-209.

Ding, Z., Granger, C. W., & Engle, R. F. (1993). A long memory property of stock market returns and a new model. *Journal of Empirical Finance 1* , 83-106.

Fox, R., & Taqqu, M. (1986). Large-sample properties of parameter estimates for strongly dependent stationary Gaussian time series. *The Annals of Statistics 14* , 517-532.

Frances, P., & Ooms, M. (1997). A periodic long memory model for quarterly UK inflation. *International Journal of Forecasting 13* , 117-126.

Geweke, J., & Porter-Hudak, S. (1983). The estimation and application of long memory time-series models. *Journal of time-series analysis vol.4, No.4* , 221-238.

Granger, C. (1980). Long memory relationships and the aggregation of dynamic models. *Journal of Econometrics 14* , 227-238.

Granger, C. (1966). The typical spectral shape of an economic variable. *Econometrica 34* , 150-161.

Granger, C., & Joyeux, R. (1980). An introduction to long-memory time series models and fractional differencing. *Journal of Time Series Analysis 1* , 15–29.

Greene, M., & Fielitz, B. (1977). Long-term dependence in common stock returns. *Journal of Financial Economics 4* , 339-349.

Haslett, J., & Raftery, A. (1989). Space-time modelling with long-memory dependence: assesing Ireland's wind power resource. *Appl. Statist. 38* , 1-50.

Hassler, U., & Wolters, J. (1995). Long memory in inflation rates: international evidence. *Journal of Business and Economic Statistics 13* , 37-45.

Hosking, J. (1984). Modelling persistence in hydrological time-series using fractional differencing. *Water Resourses Res. 20* , 1898-1908.

Hosking, J. R. (1981). Fractional differencing. *Biometrika 68* , 165–176.

Hurst, H. (1957). A suggested statistical model of some time series which occur in nature. *Nature 180* , 494.

Hurst, H. (1951). Long-term storage capacity of reservoirs. *Transactions of the American Society of Civil Engineers* , 770–799, 800–808.

Hurvich, C. (n.d.). *Differencing and unit root test*. Retrieved May 02, 2017, from NYU Stern: http://people.stern.nyu.edu/churvich/Forecasting/Handouts/UnitRoot.pdf

Jarque, C. M., & Bera, A. K. (1980). Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics Letters 6* , 255–259.

Leland, W. E., Taqqu, M., Willinger, W., & Wilson, D. (1994). On the self-similar nature of Ethernet traffic (extended version). *IEEE/ACM Transactions on Networking 2* , 1-15.

Lo, A. W. (1991). Long-term memory in stock market prices. *Econometrica 59* , 1279–1313.

Lobato, I., & Robinson, P. (1998). A non-parametric test for I(0). *Review of Economic Studies 65* , 475-495.

Mandelbrot, B. B. (1972). Statistical methodology for non-periodic cycles: From the covariance to R/S analysis. *Annals of Economic and Social Measurement* , 259–290.

Mandelbrot, B. (1971). When Can Price Be Arbitraged Efficiently? A Limit to the Validity of the Random Walk and Martingale Models. *Review of Economics and Statistics 53* , 225-236.

McLeod, A., & Hipel, K. (1978). Preservation of the rescaled adjusted range, Part 1, A reassessment of the Hurst phenomenon. *Water Resources Research 14* , 491-508.

Parke, W. (1999). What is fractional integration? *Review of Economics and Statistics 81* , 632-638.

Porter-Hudak, S. (1990). An application of the seasonal fractionally differenced model to the monetary agregates. *Journal of the American Statistical Assosiation, Applic. Case Studies 85* , 338-344.

Ramjee, R., Crato, N., & Ray, B. (2002). A note on moving average forecasts of long memory processes with an application to quality control. *International Journal of Forecasting 18* , 291–297.

Ravishanker, N., & Ray, B. (2002). Bayesian prediction for vector ARFIMA processes. *International Journal of Forecasting 18* , 207–214.

Ray, B. (1993). Modeling long-memory processes for optimal long-range prediction. *Journal of Time Series Analysis 14* , 511–525.

Robinson, P. (1978). Alternative models for stationary stochastic processes. *Stochastic Processes and their Applications 8* , 141-152.

Souza, L., & Smith, J. (2002). Bias in the memory for different sampling rates. *International Journal of Forecasting 18* , 299-313.

Souza, L., & Smith, J. (2004). Effects of temporal aggregation on estimates and forecasts of fractionally integrated processes: A Monte-Carlo study. *International Journal of Forecasting 20* , 487–502.

Sowell, F. (1992). Maximum likelihood estimation of stationary univariate fractionally integrated time series models. *Journal of econometrics 53* , 165-188.

Sowell, F. (1992). Modeling long-run behavior with the fractional ARIMA model. *Journal of Monetary Economics 29, ,* 277–302.

Wei, A., & Leuthold, R. M. (2000). Agricultural future prices and long memory processes. (pp. 1-53). University of Illinois at Urbana-Champaign.

License

Non-exclusive licence to reproduce thesis and make thesis public

I, **Kseniia Guskova** (date of birth: 1991.05.03),

(*author's name*)

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to:

   1.1. reproduce, for the purpose of preservation and making available to the public, including for addition to the DSpace digital archives until expiry of the term of validity of the copyright, and

   1.2. make available to the public via the web environment of the University of Tartu, including via the DSpace digital archives until expiry of the term of validity of the copyright,of my thesis

Fractional ARIMA processes and applications in modeling financial time series,
(title of thesis)

supervised by Raul Kangro,

(supervisor's name)

2. I am aware of the fact that the author retains these rights.

3. I certify that granting the non-exclusive licence does not infringe the intellectual property rights or rights arising from the Personal Data Protection Act.

Tartu, **16.05.2017**