DEGREE PROJECT IN MATHEMATICS,
SECOND CYCLE, 30 CREDITS
STOCKHOLM, SWEDEN 2018

Online Outlier Detection in Financial Time Series

ROBIN SEDMAN

KTH ROYAL INSTITUTE OF TECHNOLOGY
SCHOOL OF ENGINEERING SCIENCES

Online Outlier Detection in Financial Time Series

ROBIN SEDMAN

Degree Projects in Financial Mathematics (30 ECTS credits)
Degree Programme in Applied and Computational Mathematics
KTH Royal Institute of Technology year 2018
Supervisor at Fjärde AP-fonden: Victor Tingström, Oscar Blomquist
Supervisor at KTH: Anja Janssen
Examiner at KTH: Anja Janssen

TRITA-SCI-GRU 2018:071
MAT-E 2018:21

Royal Institute of Technology
School of Engineering Sciences
KTH SCI
SE-100 44 Stockholm, Sweden
URL: www.kth.se/sci

Online Outlier Detection in Financial Time Series

Robin Sedman

Abstract

In this Master's thesis, different models for outlier detection in financial time series are examined. The financial time series are price series such as index prices or asset prices. Outliers are, in this thesis, defined as extreme and false points, but this definition is also investigated and revised. Two different time series models are examined: an autoregressive (AR) and a generalized autoregressive conditional heteroskedastic (GARCH) time series model, as well as one test statistic method based on the GARCH model. Additionally, a nonparametric model is examined, which utilizes kernel density estimation in order to detect outliers. The models are evaluated by how well they detect outliers and how often they misclassify inliers as well as the run time of the models.

It is found that all the models performs approximately equally good, on the data sets used in thesis and the simulations done, in terms of how well the methods find outliers, apart from the test static method which performs worse than the others. Furthermore it is found that definition of an outlier is very crucial to how well a model detects the outliers. For the application of this thesis, the run time is an important aspect, and with this in mind an autoregressive model with a Student's t-noise distribution is found to be the best one, both with respect to how well it detects outliers, misclassify inliers and run time of the model.

Online Outlier Detektering i Finansiella Tidsserier

Robin Sedman

Sammanfattning

I detta examensarbete undersöks olika modeller för outlierdetektering i finansiella tidsserier. De finansiella tidsserierna är prisserier som indexpriser eller tillgångspriser. Outliers är i detta examensarbete definierade som extrema och falska punkter, men denna definition undersöks och revideras också. Två olika tidsseriemodeller undersöks: en autoregressiv (AR) och en generel autoregressiv betingad heteroskedasticitet[1] (GARCH) tidsseriemodell, samt en hypotesprövning[2] baserad på GARCH-modellen. Dessutom undersöks en icke-parametrisk modell, vilken använder sig utav uppskattning av täthetsfunktionen med hjälp av kärnfunktioner[3] för att detektera outliers. Modellerna utvärderas utifrån hur väl de upptäcker outliers, hur ofta de kategoriserar icke-outliers som outliers samt modellens körtid.

Det är konstaterat att alla modeller ungefär presterar lika bra, baserat på den data som används och de simuleringar som gjorts, i form av hur väl outliers är detekterade, förutom metoden baserad på hypotesprövning som fungerar sämre än de andra. Vidare är det uppenbart att definitionen av en outlier är väldigt avgörande för hur bra en modell detekterar outliers. För tillämpningen av detta examensarbete, så är körtid en viktig faktor, och med detta i åtanke är en autoregressiv modell med Students t-brusfördelning funnen att vara den bästa modellen, både med avseende på hur väl den detekterar outliers, felaktigt detekterar inliers som outliers och modellens körtid.

[1] Generalized Autoregressive Conditional Heteroskedastic

[2] Test Statistic

[3] Kernel Density Estimation

Acknowledgements

First of all I would like to thank Anja Janssen, my supervisor and examiner at KTH Royal Institute of Technology, for the support and feedback throughout the thesis work.

Moreover, I would like to thank Victor Tingström and Oscar Blomquist at Fjärde AP-fonden (AP4) for their feedback, comments, and interest in the project. I would also like to express my sincere appreciation to AP4 for making it possible to carry out this thesis and for providing me with the data necessary.

Last but not least I would like to thank my family and especially my girlfriend Eveliina for all the support and encouragement during my five years at KTH.

Stockholm, May 2018

Robin Sedman

Contents

List of Figures

List of Tables

Nomenclature

$\mathbb{1}_A$	Indicator function for the event A.
$\mathbb{C}$	All complex numbers.
$\mathbb{N}$	All natural numbers.
$\mathbb{N}^+$	The natural numbers excluding zero, i.e. $\mathbb{N}^+ \overset{\text{def}}{=} \mathbb{N}\backslash\{0\}$.
$\mathbb{R}$	All real numbers.
$\mathbb{Z}$	All integers.
$\mathcal{F}_t$	σ-algebra at time t.
$\mathcal{N}(\mu, \sigma^2)$	Gaussian distribution with expectation μ and variance σ^2.
$\mathcal{U}(a, b)$	Uniform distribution on the interval $[a, b]$.
$\overset{\text{def}}{=}$	Is defined as.
$\overset{d}{=}$	Equal in distribution.
$\Phi(\cdot)$	CDF for a $\mathcal{N}(0, 1)$ r.v.
P_t	The price of an asset at time t.
t_ν	Standard Student's t-distribution, with ν degrees of freedom.
$t_\nu(\mu, \sigma)$	Student's t-distribution with degrees of freedom ν, location μ and scale σ.

Abbreviations

acf Autocorrelation Function

CDF Cumulative Distribution Function

IID Independent Identically Distributed

pacf Partial Autocorrelation Function

PDF Probability Density Function

r.v. Random Variable

Chapter 1

Introduction

Time series show up in all kinds of applications in the real world. In engineering, economics, business, environmental and other applications of science, data can often be collected as time series. By time series it is meant that some data is collected over an interval of time, with periodic intervals. It could for instance be temperature collections, the daily price of a stock index, financial asset or consumption of some specific product in a country, [33].

There is no single definition of an outlier, but an intuitive definition can be done in several ways. Aggarwal, [1], defines an outlier in the following way: *"An outlier is a data point that is significantly different from the remaining data."* Hawkins, [22], also tries to define an outlier: *"an observation which deviates so much from other observations to arouse suspicions that it was generated by a different mechanism."* In plain words an outlier could be defined in many ways and when it comes down to a mathematical definition there exist no unique definition of an outlier, [29].

Outlier, or anomaly, detection is a very broad field within statistics. In any scientific field, both natural and social, an outlier may have a significant impact on a conclusion of an analysis. Hence, an important step in the process of analyzing data is taking care of possible outliers in a suitable way. The way a detected outlier should be handled is of course up to the process owner. Within financial applications, which will be the main focus in this thesis, outliers may have an impact on the conclusions when computing e.g. the risk of a financial position or when computing the performance of a financial portfolio, [35].

An online (or on-line) algorithm, in contrast to an offline (or off-line) algorithm, does not have any information about future data. For instance an offline algorithm is given a whole time series while an online algorithm only is given parts of it or even just the latest observation, [28]. In finance an online algorithm could be any algorithm that handles streams of for instance stock prices, index data or interest rate data.

There are several methods available for finding outliers in data, some approaches more naïve than others. Many methods use some unsupervised "machine learning" approach, such as distance- or density-based methods for mining outliers, [31], or more sophisticated ways such as Voronoi diagrams, [35]. Other methods for outlier detection, in financial data, are based on time series models such as the GARCH model, [15]. Simpler time series models, such as the AR model, can also be used for outlier detection in financial data, [36]. In some cases one method might not detect all outliers, and it might "detect" outliers which are not present, then a combination of methods could be good to use, so-called ensemble methods, [31].

1.1 Problem Statement

The problem in this thesis is to find, or develop, a model, or an algorithm, which detects outliers in financial data, mainly in asset prices and index data. Fjärde AP-fonden (AP4) are each business day given new financial data from an external provider which has to be checked for outliers before the data is used for e.g. analysis. The data points which are considered as outliers by AP4 are outliers in the sense that the values of them are false and they are very extreme. Extreme in the sense that the deviate a lot from the other data points. AP4 has detected these outliers manually before and now wants to do it in a more systematized way.

The algorithm developed in this thesis will be online, in the sense that not all past observations will be available and the new observations will be processed daily once it is given to the algorithm, i.e. no future observations will be available to the algorithm. The task is to use some of the available past observations, say the last n observations $p_1, \ldots, p_n$, to build a model and then use the old data plus the model to decide whether the new observation, p_{n+1}, is an outlier or not. If one finds that the new observation indeed is an outlier then AP4 will require new data from the external provider. If, however, the new observation, p_{n+1}, is not considered an outlier then the model will be updated based on the observations $p_2, \ldots, p_{n+1}$, and so on. This technique is sometimes called "sliding window", [37]. The reason that observation p_1 is not kept in the model is because of memory restrictions. AP4 has several financial assets which has to be checked for outliers, say N assets, and one model will be built for each asset, i.e. there will be N models.

1.1.1 The Data

AP4 will provide data to the project which will be used for both training, testing and performance evaluation. Furthermore it is desired by AP4 that the model will be implemented in `Python`. The algorithm will firstly be developed with a generated training data set. This data set is based on the American stock market index Standard & Poor 500, often denoted as S&P 500. The data spans from 2000-01-03 to 2018-04-09, a set of 4595 data points. Outliers are then inserted to the price series by AP4 in a stochastic manner. Consider that an outlier is inserted at a random time point τ, at this point the "unaffected" price, P_τ, is replaced by the price with an outlier, $\hat{P}_\tau$, according to the following equation

$$\hat{P}_\tau = P_\tau \cdot (1 + 0.15 \cdot t_3)\,, \tag{1.1}$$

where t_3 is a Student's t-distributed r.v. with three degrees of freedom.

The S&P 500 index itself with outliers inserted is shown in Figure 1.1. The logarithmic returns, described by equation (2.3), of the training data can be seen in Figure 1.2a and 1.2b. In the former no outliers are present, although one can see the higher volatility of the index in late 2008, where the last financial crisis occurred[1]. In the latter some outliers have been added by AP4. Please note that the scale on the vertical axis of the two mentioned figures is not the same.

[1] Financial Crisis, Investopedia, `https://www.investopedia.com/terms/f/financial-crisis.asp`.

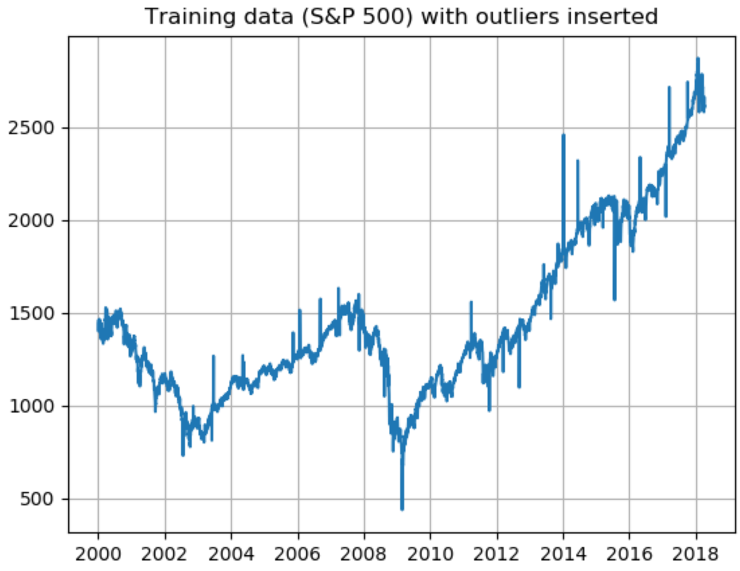
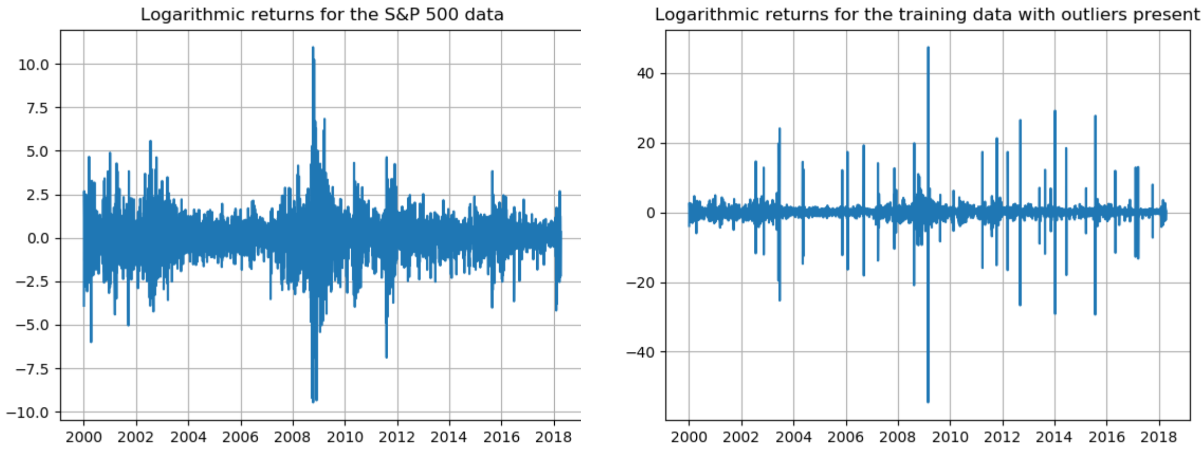


Figure 1.1: Training data (S&P 500) with outliers inserted.



(a) No outliers present.

(b) Some outliers present.

Figure 1.2: Logarithmic returns (in percent) of the training data (S&P 500).

As additional validation one of the oldest[2] stock indices, the Dow Jones Industrial Average (DJIA) will be used and outliers will be added in a similar manner as with the S&P 500 index. This data set will also be provided by AP4.

1.2 Assumptions and Limitations

Any possible dependence between different financial assets will not be taken into account. Furthermore it is assumed that the data is equally spaced, with a daily frequency. Further restrictions include not using all of the historic data possible, because of memory restrictions, as mentioned earlier.

[2] Dow Jones Industrial Average - DJIA, Investopedia, `https://www.investopedia.com/terms/d/djia.asp`.

1.3 Outline of the Thesis

The outline of the thesis will be as follows: in Chapter 2 some basic properties of time series will be presented, as well as an introduction to outliers and some properties of financial data.

In Chapter 3 two particular time series models will be introduced, the ARMA and the GARCH model, as well as some techniques for outlier detection based on these models.

In the next chapter, Chapter 4, one nonparametric method will be introduced, namely kernel density estimation, along with one outlier detection technique for the estimated probability density.

In Chapter 5 the result for the different outlier detection techniques will be presented. A discussion related to the presented data will also be held throughout the chapter.

Chapter 6 is the last chapter of the thesis and some final conclusions as well as some critique and possible extensions of the project will be presented. Additional information can be found in Appendix A, B & C.

Chapter 2

Background

2.1 Basic Properties of Time Series

Here a few definitions of a generic time series, $\{X_t\}$, will be presented. These are all basic properties and can be found in most books related to time series, e.g. [10, 33].

Definition. A *time series*, $\{X_t\}$, is a sequence of random variables, of which $\{x_t\}$ is a realization of the sequence of random variables.

Definition. The *mean function* of a time series, $\{X_t\}$, is defined by $\mu_X(t) = \mathbb{E}[X_t]$.

Definition. Let $\{X_t\}$ be a time series with $\mathbb{E}[X_t^2] < \infty$. The *covariance function* of $\{X_t\}$ is then $\gamma_X(r,s) = \mathrm{Cov}(X_r, X_s)$, for all integers r and s.

Definition. $\{X_t\}$ is a *(weakly) stationary* time series if

1. $\mu_X(t) = \mu, \quad \forall t \in \mathbb{Z}$ and
2. $\gamma_X(t+h,t) = \gamma_X(h), \quad \forall h, t \in \mathbb{Z}$.

When analyzing time series there are two tools which are very important, namely the acf (autocorrelation function) and the pacf (partial acf). Both are defined below.

Definition. Let $\{X_t\}$ be a stationary time series. The *autocovariance function* at lag h is defined as

$$\gamma_X(h) = \mathrm{Cov}[X_{t+h}, X_t].$$

The *autocorrelation function* (acf) of $\{X_t\}$ at lag h is defined by

$$\rho_X(h) \equiv \frac{\gamma_X(h)}{\gamma_0(h)}.$$

Definition. Let $\{X_t\}$ be a stationary time series. Then the *partial autocorrelation function* (pacf) is defined by

$$\alpha(0) = 1, \quad \alpha(h) = \phi_{hh}, \;\; h \geq 1,$$

where ϕ_{hh} is the last component of $\phi_h = \Gamma_h^{-1}\gamma_h$. Here $\Gamma_h = [\gamma_X(i-j)]_{i,j=1}^h$, is the covariance matrix of $(x_1, \ldots, x_n)$, and $\gamma_h = [\gamma_X(1), \ldots, \gamma_X(h)]^T$. The interpretation of the pacf is the correlation between x_t and x_{t-h} given the observations $(x_{t-1}, \ldots, x_{t-h+1})$, i.e. $\alpha(h) = \mathrm{Corr}(x_t, x_{t-h} | x_{t-1}, \ldots, x_{t-h+1})$.

Both the acf and pacf can be approximated from a set of observations, $\{x_i\}_{i=1}^n$. First the sample mean is introduced, then the sample acf and sample pacf, see definitions below.

Definition. The *sample mean* is defined as

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{2.1}$$

Definition. The *sample autocovariance function* is defined as

$$\hat{\gamma}_X(h) = \frac{1}{n}\sum_{i=1}^{n-|h|}(x_{i+|h|} - \bar{x})(x_i - \bar{x}), \quad -n < h < n,$$

and the *sample autocorrelation function* is defined as

$$\hat{\rho}_X(h) = \frac{\hat{\gamma}_X(h)}{\hat{\gamma}_X(0)}, \quad -n < h < n.$$

Definition. The *sample partial autocorrelation function* (sample pacf) is defined as

$$\hat{\alpha}(0) = 1, \quad \hat{\alpha}(h) = \hat{\phi}_{hh}, \;\; h \geq 1,$$

where $\hat{\phi}_{hh}$ is the last component of $\hat{\phi}_h = \hat{\Gamma}_h^{-1}\hat{\gamma}_h$. $\hat{\Gamma}_h$ is defined as $[\hat{\gamma}_X(i-j)]_{i,j=1}^{h}$, the sample covariance matrix, and $\hat{\gamma}_h$ is defined as $[\hat{\gamma}_X(1), \ldots, \hat{\gamma}_X(h)]^T$.

Brockwell and Davis, [10], show that the pacf for an AR(p) process is zero for lags greater than p, i.e. $\alpha(h) = 0$, for $h > p$. Hence a good way of selecting the order of an AR(p) model would be to choose p to be the largest h where $\hat{\alpha}(h)$ is non-zero. Numerically this of course will not be completely true since there is always some noise in the data. One way to deal with this is to check that a fraction of $\hat{\alpha}(h)$ falls within some bounds for $h > p$. For instance if one chooses a confidence of 95% for $\hat{\alpha}(h)$, then 95% of $\hat{\alpha}(h)$ for $h > p$ should fall within $\pm 1.96/\sqrt{n}$, where n is the number of samples, then order p would be a good choice for our AR model, [10].

2.2 Different Type of Outliers

As mentioned in Chapter 1, there is no formal definition of an outlier, but one can still divide the outliers into several categories. Below a few type of outliers are presented for a generic time series, $\{X_t\}$, for an outlier of magnitude $\gamma \in \mathbb{R}$.

The output of an outlier detection algorithm is either in a probabilistic form, if it assigns a probability to a point being an outlier, or as a binary detection where the algorithm says that either the point is an outlier or not, [1].

2.2.1 Additive Outliers

An additive outlier is when there is only one point in the time series which is affected. Consider the generic time series, $\{X_t\}$, one then observes a new time series, $\{Z_t\}$, defined by, [33],

$$Z_t = \begin{cases} X_t, & t \neq s, \\ X_t + \gamma \mathbb{1}_{\{t=s\}}, & t = s. \end{cases}$$

An outlier of this type could for instance be a measurement error and an example is provided in Figure 2.1. The three consequent graphs (Figure 2.1, 2.2 & 2.3) are based on the Swedish stock index OMXS30[1] spanning from 2000-01-03 to 2000-10-17.

[1] The data can be fetched from Nasdaq Nordic, `http://www.nasdaqomxnordic.com/`.

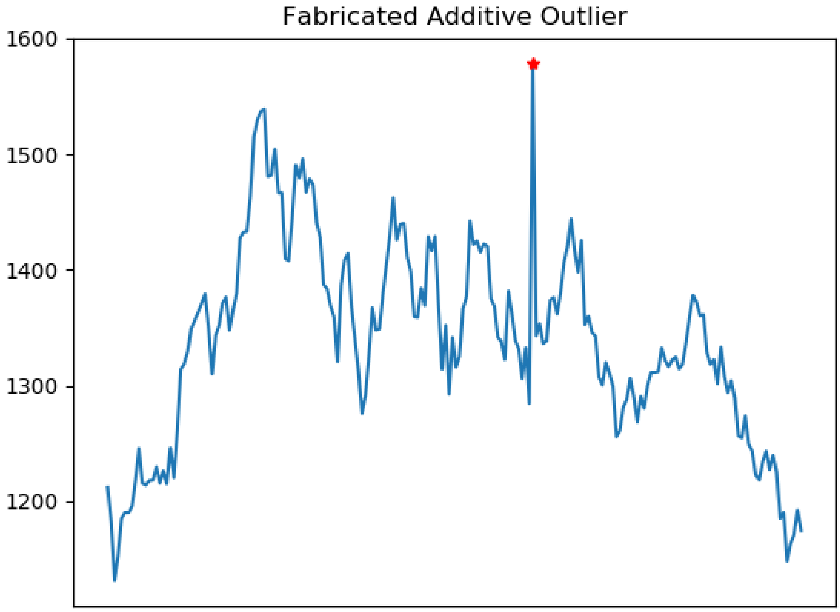


Figure 2.1: Example of an additive outlier.

2.2.2 Innovative Outliers

An innovative outlier (also innovational outlier) is produced by some change in the noise of the process. The representative impact of an innovational outlier is an initial impact of a single observation and then a few consequent observations are also affected. The specific impact on a time series is decided by its coefficients and the length of the impact is dependent on the memory of the process, [11, 33]. A fabricated innovative outlier is provided in Figure 2.2.

2.2.3 Level Shifts

An outlier of the type level shift is an outlier where the mean level of the time series suddenly changes and then the time series keeps evolving in the same way as previously. Again, consider the generic time series $\{X_t\}$, then one observes $\{Z_t\}$ defined in [33],

$$Z_t = \begin{cases} X_t, & t < s, \\ X_t + \gamma, & t \geq s. \end{cases} \tag{2.2}$$

In Figure 2.3 an example of an outlier of the type level shift is given. An outlier of this type could for instance be generated by new information provided by a company about their performance which in turn could impact the price of their shares. The outlier described by equation (2.2) could also be called a “change point”, which is a point where the distribution of the time series changes.

From the problem statement, section 1.1 of this thesis, one could see that the only outlier that is possible to detect when only looking at the next point is an additive outlier. Hence this type of outlier will be the focus of this thesis.

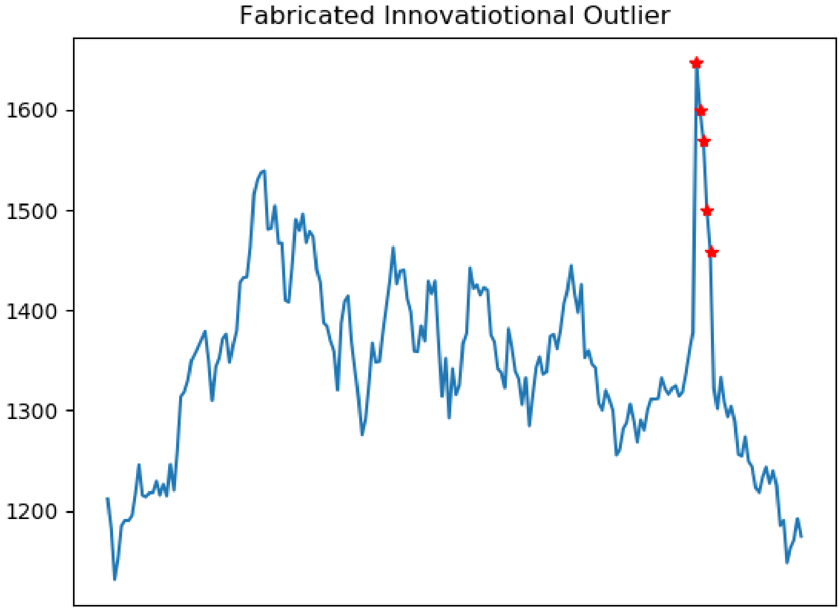


Figure 2.2: Example of an innovative outlier.

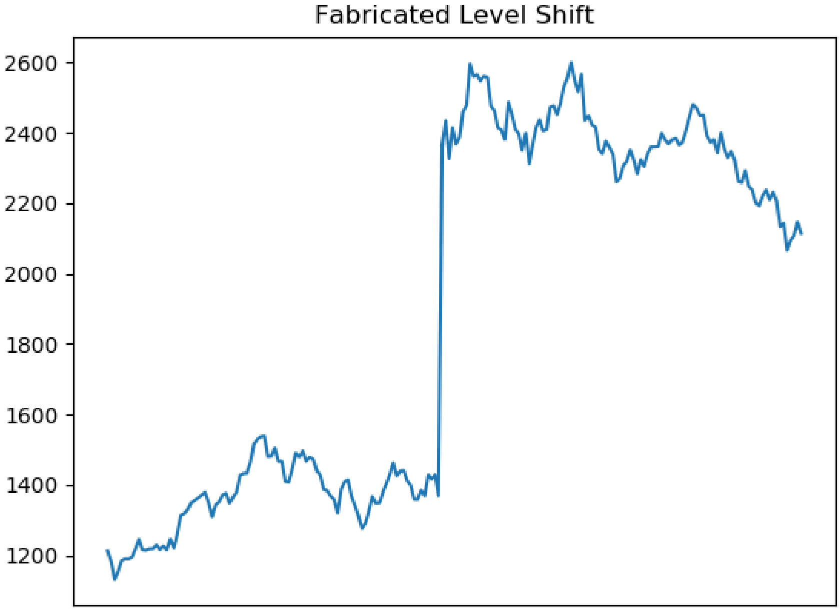


Figure 2.3: Example of a level shift.

2.3 Financial Background

As a first assumption one has no particular reason to believe that financial data[2] has some specific statistical properties, but Jondeau, Poon and Rockinger, [26], state six different properties, specific to financial data, which have been found by empirical studies. All properties are defined for the log returns, that is if P_t is the price at time t, and P_{t-1} is the price at time $t-1$ then the log return R_t can be defined as $R_t = \log \frac{P_t}{P_{t-1}}$, [36]. The six properties, holding for the returns, are the following

1. *Heavy (or fat) tails:* The unconditional distribution has heavier tails than the expected from a normal distribution.
2. *Asymmetry:* The conditional distribution is negatively skewed, suggesting that large negative returns occurs more often than large positive returns.
3. *Aggregated normality:* As the frequency of the returns decrease, the return distribution get closer to a normal distribution.
4. *Absence of serial correlation:* Returns generally do not show any significant serial correlation.
5. *Volatility clustering:* The volatility of returns are serially correlated, suggesting that a large positive (negative) return tends to be followed by another positive (negative) return. In other words, the absolute values of returns are serially correlated.
6. *Time-varying cross-correlation:* Meaning that correlation between assets changes over time. The cross-correlation tends to increase during high volatile periods, especially during market crashes.

These properties suggest that returns of financial assets may be stationary, this even though the volatility clustering is present. The volatility clustering does *not* suggest a lack of stationarity, just that the conditional variance in the process might have some dependence, [36].

One other important tools used when dealing with financial data is normalization, which is very common. This is done in order to work with returns instead of the nominal price of an asset. Logarithmic returns, or log returns, will be used in this thesis is in percent, defined as, [42],

$$R_t = 100 \log \frac{P_t}{P_{t-1}}. \tag{2.3}$$

The models presented in Chapter 3 assume that the expectation of the returns, $\mathbb{E}[R_t]$, is zero. This can be a problem since the market often has a positive or negative direction over a longer time horizon, say a few years. These market conditions are often referred to as a bull[3] or bear[4] market, respectively. In a bull market the mean of the returns, over a longer time horizon, is positive, and similarly in a bear market there it is negative. Here one often assumes a linear trend for the market which can be removed by a mean correction in order to have $\mathbb{E}[R_t] = 0$. Consider n price samples of an asset, $\{p_1, \ldots, p_n\}$, and consider the log return series of these samples, $\{r_2, \ldots, r_n\}$, $n-1$ return samples. The mean, $\bar{r}$, of the return series is computed with equation (2.1). Then $\bar{r}$ is simply subtracted from the return series,

$$r_t^* = r_t - \bar{r}.$$

Another important property of a financial asset for investors is the 'volatility' of an asset. The volatility can be compared to the statistical term standard deviation but it can also be

[2] By financial data the author refers to index, commodity, stock prices or exchange rates.

[3] Bull, Investopedia, `https://www.investopedia.com/terms/b/bull.asp`.

[4] Bear, Investopedia, `https://www.investopedia.com/terms/b/bear.asp`.

explained as how much a price of an asset changes over a specific amount of time. In general assets with high volatility are seen as 'riskier' than assets with low volatility.[5]

[5] Volatility, Investopedia, https://www.investopedia.com/terms/v/volatility.asp.

Chapter 3

Parametric Models

For any statistical procedure, some model assumptions are made, but for parametric statistics (models) there are a finite number of parameters that can be chosen. These parameters could for instance be mean, variance and degrees of freedom for a particular distribution. Another example of an assumption is that the data belongs to a family of distributions, such as a the exponential family of distributions, or more particular a Gaussian distribution or a Student's t-distribution. This is generally what characterizes parametric models, [13]. Aggarwal, [1], states that the key in parametric statistics is the assumption that is being made about the underlying probability distribution. All statistical inference that is being made will be based on the chosen distribution, hence why it is such an important decision.

Furthermore, time series models such as AR (autoregressive), MA (moving-average), ARMA (autoregressive-moving-average), ARCH (Autoregressive Conditional Heteroskedasticity) and GARCH (Generalized ARCH) are models that can be considered to be parametric. The reason for this is that it is assumed that the time series follows some specific model and there is also an assumption that the noise of the process follows some specific probability distribution, [18].

In section 3.1 a short introduction to ARMA (autoregressive-moving-average) processes will be made, in section 3.2 an AR model will be presented along with an outlier detection technique for AR models. An AR model can be considered a somewhat naïve way of modelling financial returns, but can be used as a benchmark for the more sophisticated GARCH time series model, which will be presented in section 3.3 together with two different outlier detection techniques based on the GARCH process.

3.1 ARMA Models

An ARMA model is one of the most common time series models. It is used to model linear time series processes. In financial applications however, especially return series, the ARMA model is not that common, but the ARMA model is sometimes used when modelling volatility, [42]. An introduction to an ARMA model can be found in most time series literature, e.g. [10, 33].

Definition. $\{X_t\}$ is an *ARMA(p,q) process* if $\{X_t\}$ is stationary and if for every t,

$$X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q}, \text{ or equivalently}$$
$$\phi(B)X_t = \theta(B)Z_t,$$

with $\{Z_t\} \sim \text{IID}(0, \sigma^2)$ and the polynomials $(1 - \phi_1 z - \cdots - \phi_p z^p)$ and $(1 - \theta_1 z - \cdots - \theta_q z^q)$ have no common factors.

Here σ^2 is the variance of the noise process and B is the backward shift operator, $B^j X_t = X_{t-j}$. A Gaussian distribution is a common choice for the noise distribution but other distributions, such as the Student's t-distribution, are also possible, [10].

3.2 AR Models

An AR model is essentially an ARMA model with $q = 0$. It is one of the most intuitive time series models that one can think of. The next data point, X_t, simply depends on a linear combination of the previous ones, $X_{t-1}, \dots, X_{t-p}$, and some additional noise term ϵ_t. One can consider the time series $\{X_t\}$ to be logarithmic returns of an asset, then ϵ_t can be interpreted as "new information" and this information can be considered independent of yesterdays information, hence ϵ_t is modelled as an IID r.v., [36]. The AR(p) process is a linear process by definition, see below, [10].

Definition. $\{X_t\}$ is an *AR(p) process* if $\{X_t\}$ is stationary and if

$$\begin{aligned} &X_t = \sum_{i=1}^{p} \phi_i X_{t-i} + \epsilon_t, \quad \epsilon_t \sim \text{IID}(0, \sigma^2), \text{ or equivalently} \\ &\phi(B)X_t = \epsilon_t, \quad \phi(z) = 1 - \phi_1 z - \dots - \phi_p z^p. \end{aligned}$$

Both [7] and [10] show that for an AR(p) process to be stationary it is required that

$$\phi(z) = 1 - \phi_1 z - \dots - \phi_p z^p \neq 0, \quad \forall z \in \mathbb{C} \text{ with } |z| = 1.$$

This is equivalent to saying that the polynomial $\phi(z)$ should have no roots on the unit circle. Hence the AR(p) defined above exists if and only if this condition is fulfilled.

A common choice for the noise process ϵ_t is of course $\mathcal{N}(0, \sigma^2)$, but one could also consider a Student's t distribution, t_ν. For a Student's t-distribution the degrees of freedom, ν, has to be chosen and then from the Yule-Walker equations, presented in section 3.2.2, the variance for the noise process can be estimated. The point of using the Student's t-distribution is that it has heavier tails than a normal distribution, which could be beneficial when analyzing financial data. It is also possible to show that for $\nu \to \infty$ a t_ν distribution converges to a $\mathcal{N}(0, 1)$ distribution, [12]. For this reason it is reasonable to choose a rather low ν, but the variance still has to be finite for this analysis, which requires $\nu > 2$. With this in mind ν will be chosen to be five and the reason for this choice is discussed briefly by Ruppert and Matteson, [36], the authors of the article mention that a Student's t-distribution with $\nu = 4, 5, 6$ is a much better choice than a normal distribution for the returns of financial assets.

Given σ^2 as variance for the noise process, it is possible to compute the scale parameter, $c > 0$, for the noise process with Student's t-distribution. Consider the noise process

$$\epsilon_t \sim c \cdot t_\nu.$$

Then consider the equation $\text{Var}[\epsilon_t] = \text{Var}[c \cdot t_\nu]$. Since $\text{Var}[\epsilon_t] = \sigma^2$ it is possible to compute the scale parameter c as follows

$$c = \sigma\sqrt{\frac{\nu - 2}{\nu}},$$

where σ is estimated by the Yule-Walker equations and $\frac{\nu}{\nu-2}$ is the variance of a t_ν r.v.

In the financial literature it is very common to assume that returns[1] are weakly stationary, [42]. The one thing that an AR(p) will miss out on is the volatility clustering, which the more complex GARCH(p,q) model will handle better. Hence an AR(p) model could be considered a somewhat naïve model for returns of financial assets.

[1] Both log returns and absolute returns.

3.2.1 Order Selection

The choice of the order of the model, i.e. selecting p in this case, is often referred to as "order selection". Tsay, [42], gives numerous examples that daily log returns show minor serial correlation[2], whilst monthly log returns often do not show any serial correlation. The author even suggests that *"for some daily return series, a simple AR model might be needed"*[3]. In this specific project the given data has a daily frequency, hence a low order AR(p) would be a suitable choice. An assumption will be made here, which is that an AR(1) will give a sufficient model for the naïve approach.

Brockwell and Davis, [10], show that an AR(1) process,

$$X_t = \phi_1 X_{t-1} + \epsilon_t,$$

is stationary for $|\phi_1| < 1$, and in this case it can be rewritten as a so-called MA(∞), see equation (3.1).

$$X_t = \sum_{i=0}^{\infty} \phi_1^i Z_{t-i}. \tag{3.1}$$

From this is follows directly that $\mathbb{E}[X_t] = 0$. Hence the mean correction of the logarithmic returns has to be done, as described in section 2.3.

3.2.2 Parameter Estimation

Parameter estimation refers to estimation of the coefficients for the AR(p) model, i.e. estimation of $\phi_1, \ldots, \phi_p$ and σ^2. These can be estimated from observations of a time series, $x_1, \ldots, x_n$. For an autoregressive time series there are mainly two algorithms that are being used for parameter estimation. The Yule-Walker equations and Burg's algorithm, [10]. An empirical investigation has been made by the author, comparing the Yule-Walker equations to Burg's algorithm. The result is that there are almost no differences between the estimated parameters for the analyzed data. Hence Yule-Walker is chosen due to the easier implementation. For more information see Appendix B.

The Yule-Walker equations are derived in many time series books, e.g. [8, 10]. The Yule-Walker equations are defined as follows for an AR model of order p

$$\begin{cases} \hat{\phi} & = \left(\hat{\phi}_1, \ldots, \hat{\phi}_p\right)^T = \hat{\Gamma}_p^{-1}\hat{\gamma}_p, \\ \hat{\sigma}^2 & = \hat{\gamma}_X(0) - \hat{\gamma}_p^T \hat{\Gamma}_p^{-1}\hat{\gamma}_p. \end{cases}$$

For order $p = 1$ this simplifies to

$$\hat{\phi}_1 = \frac{\hat{\gamma}_X(1)}{\hat{\gamma}_X(0)}, \quad \hat{\sigma}^2 = \hat{\gamma}_X(0)\left(1 - \left(\frac{\hat{\gamma}_X(1)}{\hat{\gamma}_X(0)}\right)^2\right).$$

3.2.3 Outlier Detection in AR(1) Models

One way of outlier detection in any time series model, would be to compute the conditional distribution, $X_t|\mathcal{F}_{t-1}$, then given the latest observation, x_t, one can compute how extreme the latest observation is by computing

$$p^* = \mathbb{P}(X_t > x_t|\mathcal{F}_{t-1}) \quad \text{and} \quad p_* = \mathbb{P}(X_t \leq x_t|\mathcal{F}_{t-1}). \tag{3.2}$$

After this check if

$$p^* < p_{\text{threshold}} \text{ or } p_* < p_{\text{threshold}}, \tag{3.3}$$

[2] Serial correlation is another term for autocorrelation.

[3] Needed referring to when one wants to model financial returns.

for some small value of $p_{\text{threshold}}$, e.g. 0.005. If either one of these conditions hold then the latest observation, x_t, can be labeled as an outlier.

In section 3.2 two different options for the distribution of ϵ_t were mentioned. Given these choices Grunwald et al., [17], states the conditional distribution, $X_t|\mathcal{F}_{t-1}$. For the AR(1) process with mean zero, $X_t = \phi_1 X_{t-1} + \epsilon_t$

$$X_t|\mathcal{F}_{t-1} \overset{d}{=} \mathcal{N}(\phi_1 X_{t-1}, \sigma^2), \qquad \epsilon_t \sim \text{IID } \mathcal{N}(0, \sigma^2), \tag{3.4}$$

$$X_t|\mathcal{F}_{t-1} \overset{d}{=} \phi_1 X_{t-1} + t_\nu, \qquad \epsilon_t \sim \text{IID } t_\nu. \tag{3.5}$$

Here X_{t-1} can be considered deterministic since one has conditioned on $\mathcal{F}_{t-1}$. In the case of a Gaussian distribution it is possible to scale it to a $\mathcal{N}(0,1)$ r.v. If $X_t|\mathcal{F}_{t-1}$, in equation (3.4), has the distribution mentioned above then

$$Z_t \overset{def}{=} \frac{X_t - \phi_1 X_{t-1}}{\sigma}\bigg|\mathcal{F}_{t-1} \sim \mathcal{N}(0,1).$$

With this in mind the equation (3.2) can be rewritten as follows

$$p^* = 1 - \Phi\left(\frac{x_t - \phi_1 x_{t-1}}{\sigma}\right), \quad p_* = \Phi\left(\frac{x_t - \phi_1 x_{t-1}}{\sigma}\right). \tag{3.6}$$

Here $\Phi(\cdot)$ is the CDF of a $\mathcal{N}(0,1)$ r.v., a similar equation can be set up when ϵ_t has a Student's t-distribution.

3.3 GARCH Models

Economic and financial time series have proven to be difficult to model, their behaviour is often characterized by non-stationarity and heteroskedasticity[4], which in plain words mean that the conditional variance of the time series changes with time. This typical behaviour gives the motivation to model our data with either ARCH or GARCH models. With these two models one does not only model the time series itself but also the variance of the time series, [4]. A GARCH model is a generalization of the ARCH model, hence the GARCH model will be investigated further. The GARCH(p,q) process was first introduced in 1986 by Bollerslev, [5], and it is a nonlinear process by definition, see below.

Definition. *GARCH(p,q) process.* Let

$$Z_t = \sqrt{h_t} e_t, \quad e_t \sim \text{IID}(0,1). \tag{3.7}$$

Furthermore let

$$h_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i Z_{t-i}^2 + \sum_{j=1}^{q} \beta_j h_{t-j}, \tag{3.8}$$

$$\alpha_0 > 0, \;\; \alpha_i \geq 0, \;\; i = 1, \ldots, p, \quad \beta_j \geq 0, \;\; j = 1, \ldots, q.$$

Then $\{Z_t\}$ is a GARCH(p,q) process. The $\{Z_t\}$ are also called innovations. Note that for $q = 0$, $\{Z_t\}$ becomes an ARCH(p) process, [5, 10].

The noise, e_t, is modelled by some distribution with expectation zero and variance one and needs not to be Gaussian even though it is a popular choice, [26]. As mentioned earlier, in Chapter 1, financial returns have heavy tails, which is also mentioned in [10], hence it could be interesting to test with a noise distribution which has heavier tails than the Gaussian distribution, such as

[4] Heteroskedasticity, Investopedia, `https://www.investopedia.com/terms/h/heteroskedasticity.asp`.

the Student's t-distribution. But even if the noise is modelled with a Gaussian r.v., the GARCH process exhibits heavy tails, and with Student's t-distribution the tails will be even heavier, [36]. One could for instance consider

$$e_t \sim \mathcal{N}(0,1), \quad \text{or} \quad e_t \sim \sqrt{\frac{\nu-2}{\nu}} t_\nu, \;\; \nu > 2,$$

where the factor $\sqrt{\frac{\nu-2}{\nu}}$ is a scale term such that $\mathrm{Var}[e_t] = 1$.

The distribution of $Z_t|\mathcal{F}_{t-1}$ is of interest since if one knows the distribution of the next point, then it is possible to compute how "extreme" the next point is. After this computation it is then possible to either label the next point as an outlier or an inlier[5]. For both mentioned distributions of e_t above the distribution of $Z_t|\mathcal{F}_{t-1}$ is as follows, [5, 6, 23, 27].

If $e_t \sim \mathcal{N}(0,1)$ then $Z_t|\mathcal{F}_{t-1} \sim \mathcal{N}(0,h_t)$, i.e. a Gaussian distribution with variance h_t.

If $e_t \sim \sqrt{\frac{\nu-2}{\nu}} t_\nu$ then $Z_t|\mathcal{F}_{t-1} \sim t_\nu(0,h_t) \sim \sqrt{\frac{h_t(\nu-2)}{\nu}} t_\nu$, i.e. t-distribution with variance h_t.

See Appendix A for more details.

It is possible to show that the conditional expectation of the GARCH(p,q) process is zero, i.e. that

$$\mathbb{E}[Z_t|\mathcal{F}_{t-1}] = \mathbb{E}[\sqrt{h_t} e_t|\mathcal{F}_{t-1}] = \{h_t \in \mathcal{F}_{t-1}, \;\; e_t \text{ independent of } \mathcal{F}_{t-1}\} = \sqrt{h_t}\mathbb{E}[e_t] = 0.$$

This is why the mean correction has to be done, described in section 2.3.

There is also an alternative to the mean correction, which would be to add a constant into the GARCH model, simply by defining some return variable, R_t as follows

$$R_t = \mu + Z_t.$$

For the implementation of the GARCH(p,q) model the mean correction is not necessary, since the `Python` library `ARCH`[6] has this built-in.

3.3.1 Order Selection

The order selection procedure is often done with the help of an information criteria, common ones are Akaike information criterion (AIC), AICC (bias corrected AIC) and Bayesian information criterion (BIC). All of these statistics have some "penalty factor"[7], which means that a more complex (complex as in more parameters) model gets penalized compared to a simpler (simpler as in fewer parameters) model. The selection procedure when using an information criterion is to fit several models, of different order, and then choose the model with lowest information criterion. In detail [10] proposes to use AICC for the GARCH model, defined as

$$\mathrm{AICC} \overset{\text{def}}{=} -2\frac{n}{n-p}\log L + 2n\frac{p+q+2}{n-p-q-3}.$$

Here L is the conditional likelihood, which depends on the distributional choice of e_t, n is the number of observations used to fit the model and (p,q) is the order of the GARCH model. One should choose the pair of (p,q) which yields the lowest AICC, i.e. the pair which yields the highest conditional likelihood, [10, 27].

A related method is described in [14] also uses an information criteria, according to the modellers preference, but this method requires manual inspection of the acf and pacf of Z_t^2. However, this is not feasible for a large number of assets.

[5] An inlier is equivalent to a non-outlier.

[6] arch 4.3.1.

[7] Similar to Occam's razor.

The method which will be used for order selection for the GARCH model is the first mentioned one, with the help of AICC. Several low order GARCH models will be fitted to the given training data and then the most common model will be chosen according to the lowest AICC. The computation of the AICC is built-in to the `Python` library `ARCH` and will be used for implementation.

3.3.2 Parameter Estimation

There is not one single way to estimate the parameters in a GARCH model. In [14], quasi-likelihood is described as a way to estimate the parameter, $\hat{\theta}_n = (\alpha_0, \ldots, \alpha_p, \beta_1, \ldots, \beta_q)$. This is also called Gaussian quasi-likelihood which can be used regardless of the noise distribution of e_t. No explicit assumption is being made about the distribution of the GARCH process but the PDF for a Gaussian r.v. is utilizied, hence the name, [14]. Given the order of the model, (p,q), the quasi-likelihood is maximized by adjusting the parameters $\hat{\theta}_n$. Francq, [14], also proves that if θ_0 denotes the true parameters of the GARCH model and n is the number of observations used to estimate $\hat{\theta}_n$, then

$$\hat{\theta}_n \to \theta_0, \quad \text{as } n \to \infty, \tag{3.9}$$

i.e. the estimated parameter converges to the true parameter as the number of observations, n, increases. Another, but similar, method is maximum likelihood estimation (MLE), which is stated in [10, 14], both for Gaussian and Student's t-distributed e_t. It is also shown that the convergene also holds for MLE, that is equation (3.9) holds. The parameter estimation is also built-in to the `Python` library `ARCH` and the built-in estimation will be used.

3.3.3 Stationarity of a GARCH Process

One assumption for estimating the parameters of a GARCH(p,q) process is that the process is stationary, so of course it could be interesting to check if this condition holds true when a GARCH(p,q) model is implemented as well. The condition for covariance-stationarity for a GARCH(p,q) process is well known in the literature and can be found in e.g. [27, 41] and is as follows

$$\sum_{i=1}^{p} \alpha_i + \sum_{j=1}^{q} \beta_j < 1.$$

3.3.4 Outlier Detection in GARCH Models

One approach that can be used for detecting outliers in a GARCH model is the approach already described in section 3.2.3 for AR models. This is straightforward to apply for the GARCH(p,q) model as well, since the conditional distribution of $Z_t | \mathcal{F}_{t-1}$ is known from above.

3.3.5 Outlier Detection with a Test Statistic

Another approach is presented by Franses & Ghijsels, [15], and is also further examined by Charles & Darné, [11]. The reader of the thesis should be aware that this method is in the two articles used for an offline application, i.e. that future values of the time series are available, but it is also possible to modify it to an online application. This outlier detection approach utilizes the fact that a squared GARCH(p,q) process can be rewritten as an ARMA(r,p) with non IID noise process (v_t), where $r = \max(p, q)$. The procedure is shown in e.g. [14, 27]. Define $v_t = Z_t^2 - h_t$, then equation (3.7) & (3.8) can be rewritten as

$$Z_t^2 = \alpha_0 + \sum_{i=1}^{r} (\alpha_i + \beta_i) Z_{t-i}^2 + v_t - \sum_{j=1}^{p} \beta_j v_{t-j}.$$

Which is an ARMA(r,p) process for Z_t^2 with non IID noise process v_t. For a GARCH(1,1) this is simplified to

$$Z_t^2 = \alpha_0 + (\alpha_1 + \beta_1) Z_{t-1}^2 + v_t - \beta_1 v_{t-1}.$$

Charles & Darné, [11], then propose to compute a test statistic $\hat{\tau}$, based on [15], and compare this statistic with some threshold for outlier detection. If the test statistic is sufficiently large, then the point is labeled as an outlier. Instead of observing the series $\{Z_t^2\}$ it is assumed that one observes

$$\hat{Z}_t^2 = Z_t^2 + \gamma \mathbb{1}_{\{t=s\}},$$

i.e. an additive outlier at time $t = s$. As mentioned earlier, the focus in this thesis is to determine if the latest point is an outlier or not, which simplifies the expressions given in [11], and this is also confirmed by [27]. The hypothesis here is that there is no outlier present, i.e. $\gamma = 0$. With this in mind the proposed test statistic is

$$\hat{\tau} = \frac{e_n}{\hat{\sigma}_v},$$

Here e_n is the last point in the noise process and can be computed from equation (3.7) since Z_t is directly observable and h_t is given from the `Python` library `ARCH` mentioned earlier. The parameter $\hat{\sigma}_v^2$ is the estimated variance for the process v_t defined above[8]. The variance (or standard deviation) can be estimated in several ways but the most common, which can be found in most elementary statistics books e.g. [12], is

$$\hat{\sigma}_v^2 = \frac{1}{m-1} \sum_{j=1}^{m} (v_j - \bar{v}_t)^2.$$

Here m is the number of data points used and $\bar{v}_t$ is the sample mean of v_t, defined by equation (2.1). The problem with estimating the variance with this equation is that it is very sensitive to outliers, [27, 32]. However it is possible to estimate the variance other ways, one is the so-called "omit-one" method. Here the point where an outlier is suspected to be present is neglected when estimating $\hat{\sigma}_v^2$ since an outlier has a significant impact on the estimated variance, [11, 15]. This could be difficult since its not necessarily known where the outlier is. For this reason a method called "Median Absolute Deviation" (MAD) will be used, which is presented in [27, 32], see equation (3.10)

$$\hat{\sigma}_v = b \cdot \text{median}\left(|v_j - \text{median}(v_t)|\right). \tag{3.10}$$

Here b is a parameter depending on the distribution of the underlying data, and it can be computed as follows, [32]

$$b = \frac{1}{F_V^{-1}(p)}, \quad p = 0.75.$$

Here $F_V^{-1}(p)$ is the inverse CDF, or the quantile function, for the r.v. V at level $p \in [0, 1]$. Since the true distribution of the data is not known is not possible to calculate the quantile analytically. The method presented above is called the Median Absolute Deviation (MAD) method.

The problem with the analytical quantiles being unknown is however rather easy to solve. One can replace the quantile function with empirical quantiles, presented by Hult, Lindskog, Hammarlind and Rehn, [24]. Consider n samples $\{x_1, \ldots, x_n\}$ generated by a r.v. X with CDF $F_X(x)$. Then consider these samples ordered such that $x_{1,n} \geq \cdots \geq x_{n,n}$. Then the empirical quantile, $\hat{F}_X^{-1}(p)$, can be expressed as one of the ordered samples

$$\hat{F}_X^{-1}(p) = x_{\lfloor n(1-p) \rfloor + 1, n}, \quad p \in [0, 1]. \tag{3.11}$$

[8] Based on [11, 14] this detail is not entirely clear but seems to be the most intuitive way based on notation in the articles.

Where $\lfloor \cdot \rfloor$ is the floor function. Hult et al., [24], also proves that the empirical quantile converges to the true quantile, i.e.

$$\lim_{n\to\infty} \mathbb{P}\left(\left| \hat{F}_X^{-1}(p) - F_X^{-1}(p) \right| \leq \epsilon \right) = 1, \quad \epsilon > 0, \;\; \forall p \in [0,1].$$

The next, and last, step is to choose the threshold for when a point is labeled an outlier. A point is labeled an outlier if

$$|\hat{\tau}| > C, \tag{3.12}$$

for some threshold C. Here Franses & Ghijsels, [15], suggests $C = 4$ while Charles & Darné, [11], suggests $C = 10$. Clearly this range is quite large and the chosen threshold is completely up to the modeller. The range of C, for this application, will be found by the means of an empirical investigation, simply by testing a range for the parameter C, the result is then presented in Chapter 5.

Chapter 4

Nonparametric Models

Nonparametric models, or more generally nonparametric statistics, conversely to parametric statistics is the part of statistics where no parametric assumptions are being made about, for instance, the distribution of the underlying data. An alternative way of viewing this is to say that the number of parameters in a nonparametric model is infinite. Although there is no precise limit between parametric and nonparametric statistics this is one way to describe the difference. Examples of methods that can be considered nonparametric are the empirical distributing function, histograms, kernel density estimation[1] and k-nearest neighbours (kNN), [2, 13].

Sadik & Gruenwald, [37], states that *"it is very difficult to select an appropriate auto-regression model for data streams[2]"*. Furthermore it is said that the cut-off point chosen for outlier detection also depends on the chosen model. This would give a motivation for selecting some nonparametric model and compare its performance to the presented parametric models.

4.1 Distance Based Outlier Detection

A distance based outlier detection method is a subfamily within the family of proximity-based outlier detection methods. A proximity-based outlier detection method is a method which defines a data point as an outlier if its neighbourhood is sparsely populated. Within this family there are also density based algorithms and cluster based algorithms. All of these algorithms are quite similar, hence the family name: proximity-based algorithms, [1].

One example of a density based model, which also can be interpreted as a distance based model, is the local outlier factor (LOF) model. This method measures the local deviation of a given data point to its kNN. By then comparing the density of a point to the densities of its neighbors it is possible to detect outliers, [1].

A well known method such as kNN is one example of a distance based model. The method, kNN, is most often used in supervised machine learning but can also be adapted for outlier detection. In an outlier detection manner the method computes the distance to the k:th neighbor, $k \in \mathbb{N}^+$, and then the computed distance in used as an outlier score, [1, 25]

Another distance based method is presented by Sadik & Gruenwald, [37], and is based on estimating the PDF of the data. One way, and probably the most well-known way, to estimate the PDF is by using a histogram. However a histogram is of best use when one wants to inspect the probability distribution visually, which for the scope of this thesis would be infeasible since the number of assets, N, might be rather large. An improvement of the histogram, which is a very common way to estimate the PDF, is kernel density estimation, [13, 40, 43]. Latecki, Lazarevic and Pokrajac, [30], also present kernel estimation for outlier detection. It is found that the method outperforms well-established methods such as LOF. The first two steps in

[1] Which essentially is estimating the PDF.

[2] Data stream refering to a never ending time series, [37].

using kernels to estimate the density is to select the kernel function and to select the so-called bandwidth. This will be presented in section 4.1.1 and section 4.1.2 respectively. Then in section 4.1.3 the outlier detection technique will be presented.

4.1.1 Kernel Selection

A kernel function is, as mentioned, used to estimate the PDF of a r.v., based on samples from this r.v. What the kernel function essentially does it to redistribute the samples from point masses to a spread out density. One could see it as a transformation from a discrete distribution (samples) to a continuous distribution. Just like with a PDF, the kernel function, K, should satisfy

$$\int_{-\infty}^{\infty} K(x)dx = 1.$$

In almost all cases, the kernel function is a non-negative symmetric unimodal probability density, such as the normal density. Symmetric means that

$$K(-u) = K(u), \quad u \in \mathbb{R},$$

and unimodal refers to that the kernel only has one single mode, [13, 40].

The article which this idea is based on, [37], does not explicitly mention which kernel function that has been used but only that there are several options available. Firstly, given n data points $\{x_i\}_{i=1}^{n}$ the kernel density estimate with bandwidth $h > 0$ is defined by

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x_i - x}{h}\right) \tag{4.1}$$

Some common kernel functions are presented by Fan and Yao in [13], one is the Gaussian kernel

$$K(u) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{u^2}{2}\right), \quad u \in \mathbb{R},$$

and another one is from the symmetric Beta family

$$K_\phi(u) = \frac{1}{\beta\left(\frac{1}{2}, \phi + 1\right)}(1 - u^2)^\phi \mathbb{1}_{\{|u| \leq 1\}}.$$

Where $\beta(x, y)$ is the beta function (also called the Euler integral of the first kind) defined in e.g. [3] by

$$\beta(x, y) = \int_0^1 t^{x-1}(1 - t)^{y-1}dt, \quad x, y > 0.$$

For different ϕ, the kernel function $K_\phi(u)$ has different names. The choices $\phi = 0, 1, 2$ and 3 correspond to uniform, Epanechnikov, biweight and triweight respectively. One can see that the kernels have different "concentration", by this the author refers to how the kernel functions distribute their mass over different parts of $\mathbb{R}$. For instance the uniform kernel has "concentration" over the interval $[-1, 1]$, while Epanechnikov, biweight and triweight has much shorter "concentration"[3]. The Gaussian meanwhile has "concentration" over $\mathbb{R}$, meaning that the Gaussian kernel function fulfills $K(u) > 0, u \in \mathbb{R}$, i.e. that it has more mass on the tails than the others, [13]. This would give a reason for choosing the Gaussian kernel since, as mentioned earlier, financial returns have heavy tails. However, both empirical and theoretical results from the literature show that the choice of kernel function does not have a large impact on the estimated PDF, [13, 37].

[3] Most of the Epanechnikov, biweight and triweight kernel functions mass is closer to zero compared to the uniform kernel.

In Figure 4.1 a visualization is made which shows the different presented kernel functions, the data is 500 samples from a $\mathcal{N}(0, 1)$ r.v., the bandwidth is chosen according to equation (4.3) below. Here one can see that the choice of kernel function is *not* that important, all estimations of the PDF looks rather similar, though the Gaussian kernel is the most “smooth” kernel due to its “concentration” over $\mathbb{R}$.

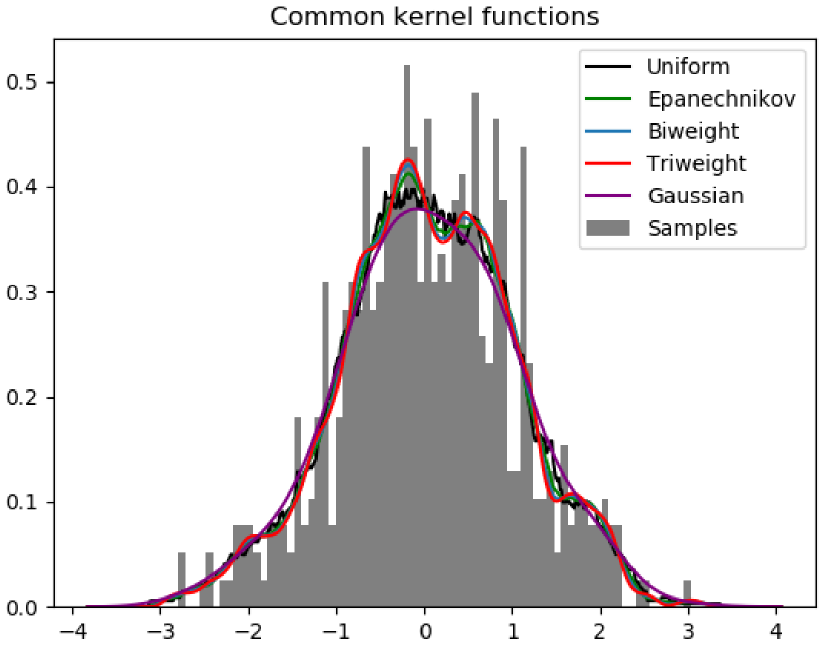


Figure 4.1: Visualization of the presented kernel functions[4].

4.1.2 Bandwidth Selection

The so-called bandwidth is the parameter which is more important when choosing kernel function and bandwidth. What the bandwidth, h, controls is how much the distribution gets smoothed out. If one selects a too small h then this will result in the estimated PDF having a lot of modes. For a too large h the shape of the estimated PDF will be oversmoothed and the properties of the data might be destroyed. Properties such as peaks in the densities and multimodalities will be underestimated and tail probabilities might be overestimated, i.e. a too large bandwidth might create large biases in the density estimation, [12, 13, 37]. In Figure 4.2 a visualization is made showing how important the bandwidth choice is, which is based on the same data as in Figure 4.1. With the bandwidth too large the probability mass in the tails is overestimated and for a too small bandwidth the estimated PDF becomes multimodal, just as the theory says.

[4] With the help of built-in function from the StatsModels library.

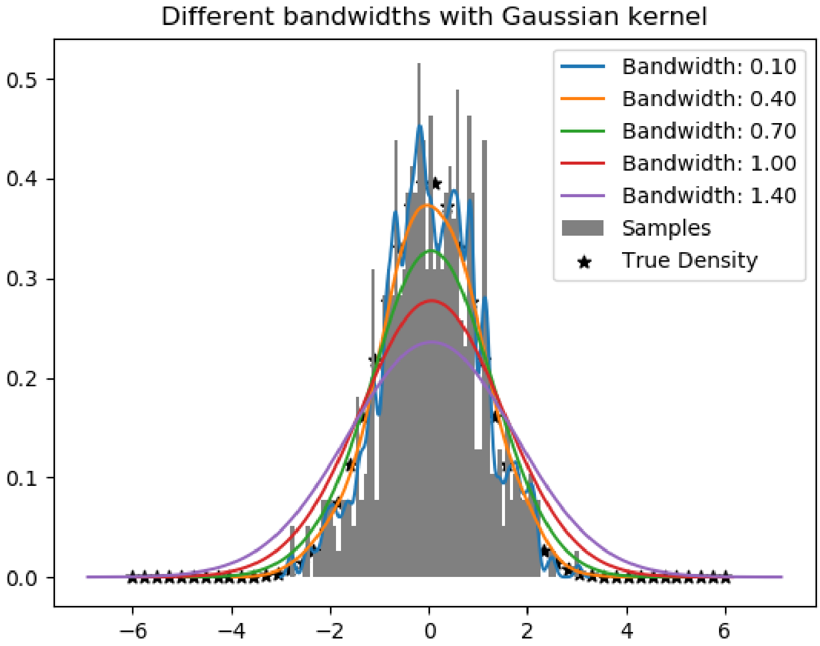


Figure 4.2: Gaussian kernel for different bandwidths[5].

The most common, or the theoretically optimal, way of choosing the bandwidth, h, is to minimize the so-called Mean Integrated Square Error (MISE). Let the estimated PDF be denoted by $\hat{f}_h(x)$ and the true PDF be denoted by $f(x)$. Then the problem is to minimize

$$\mathrm{MISE}(h) = \mathbb{E}\left[\int \left(\hat{f}_h(x) - f(x)\right)^2 dx\right].$$

However, this is not possible to solve analytically since the true PDF $f(x)$ is unknown, [13]. Scott, [38], instead proposes to choose the bandwidth according to

$$h = 3.49\hat{\sigma}n^{-1/3}, \tag{4.2}$$

where $\hat{\sigma}$ is the sample standard deviation and n is the number of sample points used to estimate the PDF. This choice of bandwidth is based on a Gaussian density, but the assumption is not so strong as using a parametric Gaussian distribution, i.e. if the bandwidth in equation (4.2) is used with non-Gaussian data then the resulting PDF will not look like a Gaussian density, [38].

Scott, [38], also states that the data-based approach in order to find the bandwidth, as above, tends to overestimate the bandwidth for a non-Gaussian distribution, which in turn gives a smoother PDF than the true PDF. Furthermore Scott also mentions that the rule, in equation (4.2), is not recommended to use as it is but rather to modify it slightly, one should also be aware that this equation is more or less a rule of thumb.

Fan and Yao, [13], presents an approach similar to equation (4.2), but also states that the rule presented is a rule of thumb as well and that the rule might lead to oversmoothing if the underlying distribution is asymmetric. As presented in section 2.3 financial returns indeed have the property of heavy tails and asymmetry, i.e. using rules such as these two mentioned above might lead to oversmoothing.

In [39], by Scott, the same author as above, another rule is presented, which is more robust compared to the above. This idea was first presented by Freedman and Diaconis, [16]. Here the

[5] With the help of built-in function from the StatsModels library.

sample standard deviation, $\hat{\sigma}$ is replace by the so-called "interquantile range" (IQR) and then optimal bandwidth is formulated as

$$h = 2 \cdot \text{IQR} \cdot n^{-1/3}. \tag{4.3}$$

Where IQR is defined as

$$\text{IQR} = F_X^{-1}(0.75) - F_X^{-1}(0.25).$$

Here the quantiles can be estimated by equation (3.11) presented earlier. The replacement of $\hat{\sigma}$ by IQR is a good choice in this specific application from a robustness point of view, since the IQR measure is less sensitive to outliers than the sample standard deviation, [43].
With all this in mind it is now possible to compute the bandwidth as of equation (4.3).

4.1.3 Outlier Detection with Kernels

The idea presented in this section is based on an article by the authors Sadik & Gruenwald, [37], and the idea is rather intuitive and easy to understand. Given n old data points, $\{x_i\}_{i=1}^n$ the PDF is estimated as $\hat{f}_h(x)$. Then consider that a new point is given, denoted $\tilde{x}$, and then the probability mass of its neighbourhood is computed as

$$p(\tilde{x}, r) = \int_{\tilde{x}-r}^{\tilde{x}+r} \hat{f}_h(x) dx,$$

where r is some fixed radius around the point $\tilde{x}$. The new point $\tilde{x}$ is then labeled according to the following rules

$$\begin{aligned} &\text{If } p(\tilde{x}, r) < q, && \tilde{x} \text{ is labeled an outlier,} \\ &\text{Else} && \tilde{x} \text{ is not labeled an outlier.} \end{aligned} \tag{4.4}$$

Here q is some chosen threshold. The interpretation here would be that for a large $p(\tilde{x}, r)$ it is a rather large probability that the point $\tilde{x}$ would occur and hence its not likely to be an outlier, and vice versa if $p(\tilde{x}, r)$ would be small.

Furthermore Sadik & Gruenwald, [37], also proposes to weigh the data points based on "freshness", i.e. when the PDF is estimated one values the recent points more than the historical points. This seems like a reasonable approach for financial data because of the volatility clustering of financial returns, mentioned in section 2.3. This approach should be able to quicker adapt to changes in the data compared to a approach where the points are not weighed based on "freshness".

Consider the historical data points $\{x_i\}_{i=1}^n$ where x_1 is the oldest point and x_n is the most recent one. Then consider to so-called "forgetting factor", γ, where $0 \leq \gamma \leq 1$. Then the data points $\{x_1, \dots, x_n\}$ are weighed by $\{\gamma^{n-1}, \gamma^{n-2}, \dots, 1\}$ respectively. This mean that $\gamma = 0$ would correspond to just using the most recent observations, which is not feasible since it is not possible to estimate the PDF with just one data point, and $\gamma = 1$ would correspond to equal weights on all n data points, [9, 37].

Choosing the forgetting factor, γ, can be rather technical, but Brailsford, Penm and Terrell, [9], presents an example where they investigate which γ would be suitable for the Australian stock market, the index All Ordinaries Index (AOI), and the initial guess $\gamma = 0.99$, seems to be a rather good choice. With $\gamma = 0.99$ the most recent observation will get almost full weight while the older ones still will have an impact, depending on the choice of n, on the estimation of the PDF. This value will be used in this thesis as well, since Brailsford et al., [9], also uses financial data.

The last step is now to incorporate the forgetting factors into the computation of the estimated PDF, see equation (4.5). Since each point does not have one as weight (as in equation

(4.1)) the factor $\frac{1}{n}$ is replaced by the sum of all "forgetting factors".

$$\hat{f}_h(x) = \frac{1}{h} \frac{\sum_{i=1}^{n} \gamma^{n-i} K\left(\frac{x_i - x}{h}\right)}{\sum_{i=1}^{n} \gamma^{n-i}}. \tag{4.5}$$

Here $\sum_{i=1}^{n} \gamma^{n-i}$ is the sum of all weights which is placed in the denominator as a normalization constant and h is the bandwidth chosen according to equation (4.3).

The problem with the presented method in [37] is the curse of having two parameters which can be varied, both q and r is completely up to the modeller to choose. The authors of the paper do not present any way of choosing these parameters, it is just mentioned that if one can assume continuity of the PDF $f(x)$, then the problem can be reduced to one parameter. But the procedure for this is not mentioned either. For this reason the author of this thesis has chosen to compute the parameter r in a data driven manner and then optimize with respect to the threshold parameter q. The parameter r is a radius around the latest observed data point, and given n historic data points it is possible to compute IQR, but the IQR is rather large and this would likely be a too large value for r. Hence r is chosen to be IQR/10. The reader of this thesis should have in mind that this choice is not definite and that there might be other options which give better result. Some details related to the implementation of this method are stated in Appendix B.

Lastly Sadik & Gruenwald, [37], mentions that the presented algorithm "outperforms the existing algorithms in terms of accuracy, but requires more time to execute." So it is possible that the approach is quite good but as mentioned earlier, although there are some pieces missing which might be critical for the performance of this outlier detection algorithm.

Chapter 5

Result and Discussion

In this chapter the results of the different outlier detection techniques will be presented. For each of the methods presented there is (at least) one parameter that can be optimized in order to achieve the most desirable result. Of course one could "optimize" the result of the algorithms in different ways, but in general a good outlier detection technique maximizes true positive (TP) and minimizes false positive (FP), for a visualization of these properties see the so-called "confusion matrix" below, [37]. In the confusion matrix a "+1" means that an outlier is present or predicted while a "-1" means that an outlier is not present or was not predicted.

		Prediction	
		+1	**-1**
Truth	**+1**	True Positive, TP	False Negative, FN
	-1	False Positive, FP	True Negative, TN

In plain words the different classes can be described as follows in this specific context:

- **TP:** An outlier was present and was predicted correctly.
- **FN:** An outlier was present but it was not detected.
- **FP:** No outlier was present but an outlier was predicted.
- **TN:** No outlier was present and no outlier detected.

With this in mind what a good outlier detection algorithm should do, as in what was mentioned above, seems almost obvious, but one should be aware that there is almost always a trade-off when optimizing an outlier detection method. With these properties in mind it is possible to define the so-called "true positive rate" (TP rate) and the "false positive rate" (FP rate), [30].

$$\text{True positive rate} = \frac{TP}{TP + FN}.$$
$$\text{False positive rate} = \frac{FP}{FP + TN}.$$

The interpretation of the TP rate is the frequency of outliers which were classified as outliers. The FP rate is the frequency of points which were classified as outliers even though no outlier was present. One way to present the result, for an outlier detection algorithm, is by the so-called receiver operating characteristic curve (ROC curve), which is achieved by plotting the TP rate versus the FP rate and then vary some parameter. For this thesis this parameter could be $p_{\text{threshold}}$, C or q (see equation (3.3), (3.12) & (4.4) respectively) of which all are some threshold for outlier detection. The ROC curve is known to be a robust and efficient performance measure for comparing different classification methods, or as in this thesis different outlier classifiers, [1, 34]. With an ROC curve it is sometimes possible to see by eye inspection which of some outlier detection algorithm is the best, simply by looking at which curve dominates the other ones. However for some algorithms this cannot be seen clearly by eye inspection of the graph. One can then look at the area under the ROC curve, (ROC AUC). If the area is larger for algorithm A than for algorithm B, then A should be preferred over B, [1]. See Figure 5.1 for an example of this. Aggarwal, [1], also presents an interpretation for ROC AUC which is that ROC AUC is equal to the *"probability that a randomly selected outlier-inlier pair is ranked correctly"*. That is, if one picks one outlier and one inlier at random, then the probability that both are classified correctly corresponds to ROC AUC of the algorithm used.

In Figure 5.1 the optimal ROC curve for outlier detection is shown in red, however this optimal ROC curve is hardly achieved in practice, [30]. James, Witten, Hastie and Tibshirani, [25], states that an ROC AUC of 0.95 "would be considered very good". So the problem in its entirety is essentially a trade-off problem between the probability of classifying outliers correctly and misclassifying inliers.

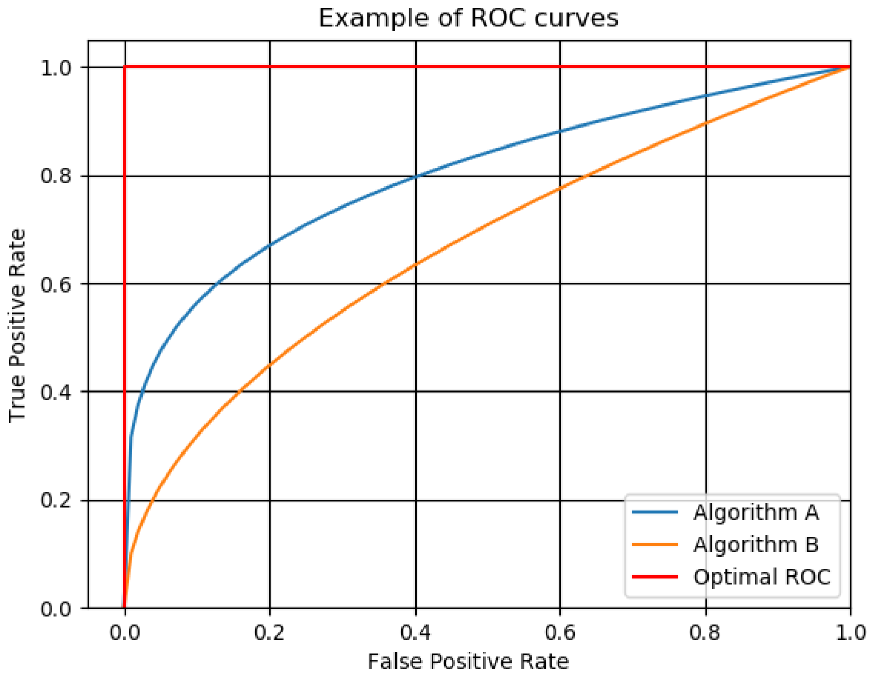


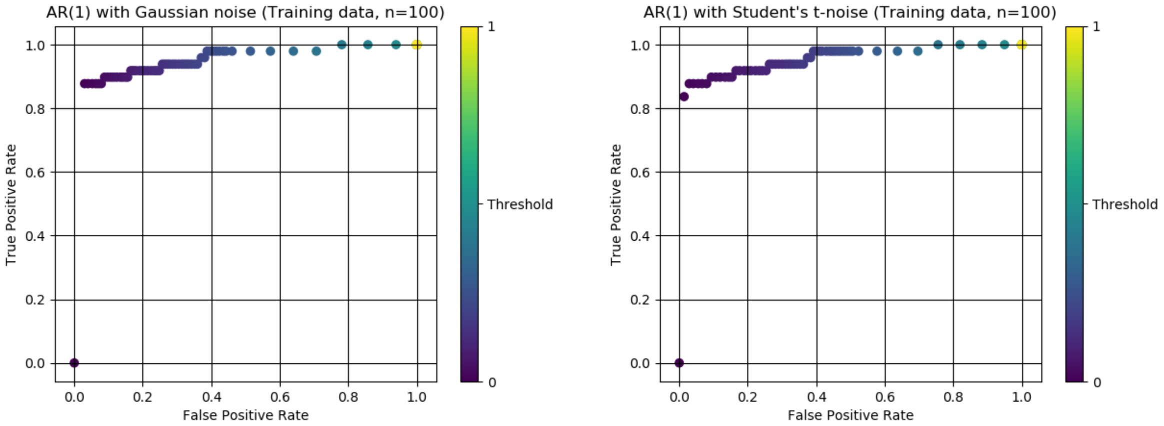
Figure 5.1: Example of ROC curves.

So the "best" algorithm, among the ones presented, can be chosen with the procedure above which would give the "best" algorithm over the range of a parameter. However it still remains to choose a specific threshold, i.e. a value for $p_{\text{threshold}}$, C or q. This can be done by considering an upper bound on the FP rate, say 5%, and then choosing the threshold such that the FP rate is at most 5%. One could also choose to set a lower bound on the TP rate and then choose a specific threshold. The "best" algorithm would then be the one which has the highest TP rate at this point. Another point that one should have in mind when choosing the "best" algorithm is the speed (run time) of the algorithms. If the run time for the algorithm is critical for the application then an algorithm that is faster might be preferable even though it does not perform as well as another algorithm.

The chapter is structured as follows, in section 5.1, 5.2 & 5.3 the results of the different models are presented and shortly discussed. In section 5.4 the results of the models are further examined and discussed. Two models, among the ones presented, are chosen for further validation and investigation, this is presented in the sections 5.4.1, 5.4.2 & 5.4.3. Furthermore in Appendix C some further investigation of the definition of an outlier is being done.

5.1 AR Models

Figure 5.2a & 5.2b are based on the training set, provided by AP4, with a total of 4595 data points, and $n = 100$ to fit each AR(1) model. A total of 50 outliers where put in into the data and the result where computed for $p_{\text{threshold}} \in [0, 1]$.



(a) Gaussian noise process.

(b) Student's t-noise process.

Figure 5.2: Performance for the AR(1) models.

The area under the curve for AR with Gaussian noise is 0.9461 and for AR with Student's t-noise 0.9526, indicating that the overall performance for the model with Student's t-noise performs slightly better.

As a comparison the AR(1) model with Gaussian noise has been evaluated for $n = 500$ as well, in order to see what impact n has on the result. The result can be seen in Figure 5.3. From this graph it is not entirely clear if an increase of n gives better performance. The ROC AUC for the AR(1) model with $n = 500$ is computed to be 0.9345, which is slightly lower than for $n = 100$. The most probable explanation for this is due to the non-stationarity of financial data, with a longer time horizon, $n = 500$, the oldest observation does not contribute with more information but rather just more noise which decreases the overall performance. This result would also suggest that $n = 100$ is "enough" observations in order to make "good" estimates of the AR(1) model parameters, since the ROC curves for $n = 100$ and $n = 500$ are so close.

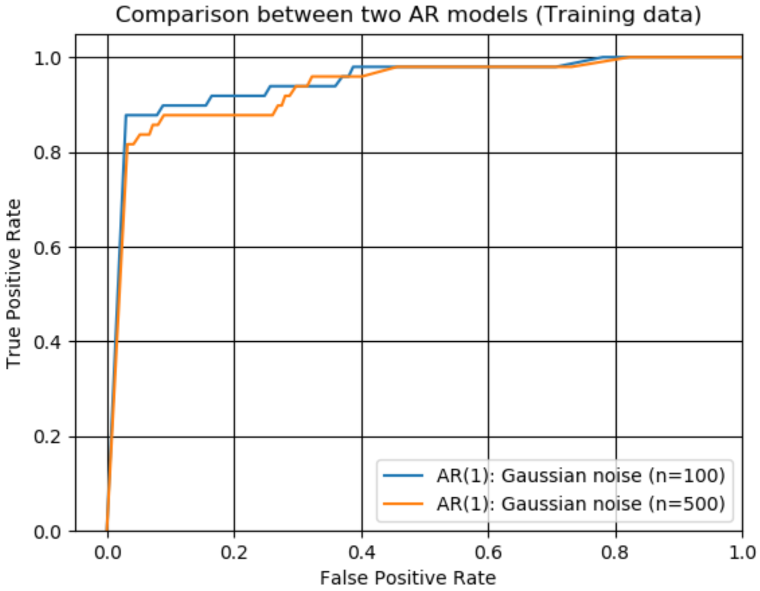


Figure 5.3: AR(1) models with Gaussian noise for two different n.

5.2 GARCH Models

The first part of using a GARCH(p,q) model was to select p and q for the process. An empirical investigation has been done based on the training data set, described in section 5.1. The different orders tested and then evalutated based on AICC was: GARCH(1,1), GARCH(2,1), GARCH(1,2) and GARCH(2,2). The result can be seen in Figure 5.4.

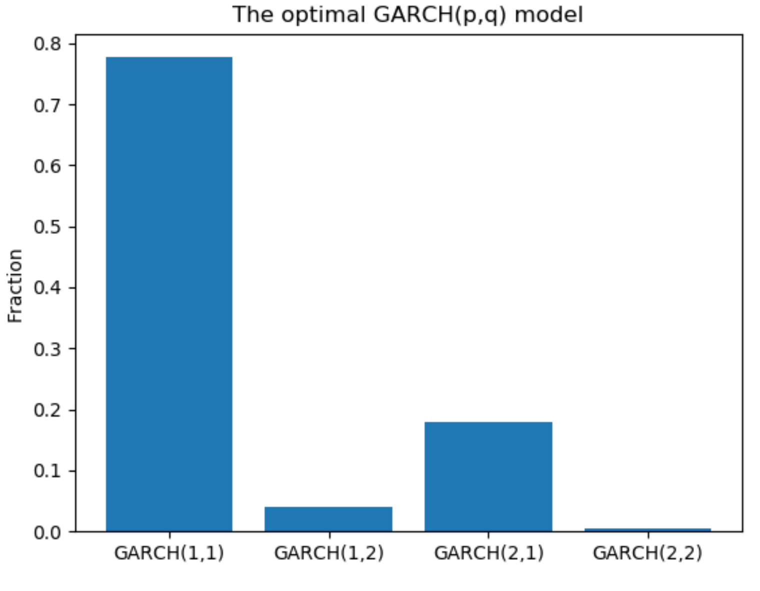
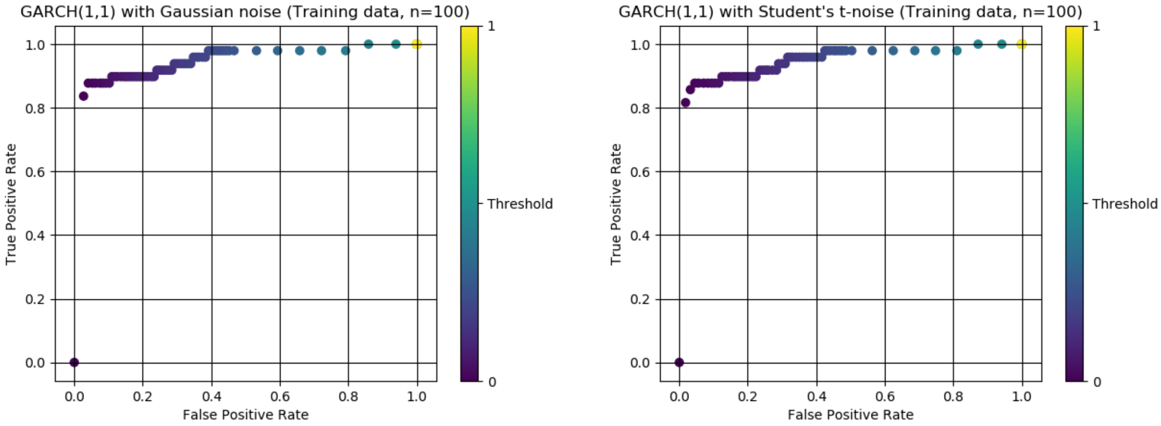


Figure 5.4: Optimal GARCH model based on AICC for the training data.

As one can see the most common model, by far, is the GARCH(1,1) model with it being the optimal model in over 78% of the cases according to AICC. To fit *each* GARCH model $n = 100$ samples were used and the noise process was set to be Gaussian. The models were fitted with a "sliding window" approach, consider a data set of M points: $\{x_i\}_{i=1}^{M}$. Then the first set of models were fitted to $\{x_i\}_{i=1}^{n}$ and the second set of models were fitted to $\{x_i\}_{i=2}^{n+1}$, and so on. Since the fraction of GARCH(1,1) is rather high the author of this thesis will argue to use this order for all the GARCH models in the thesis. One argument to choose one specific GARCH order is that the computational cost that one has to pay when fitting several models (and see which order is the optimal one) is too high and the algorithm would be too slow when testing several models. For this reason a GARCH(1,1) model will be used, throughout the thesis.

The result for the two GARCH(1,1) models, with different noise distributions, can be seen in Figure 5.5a & 5.5b. The figures are based on the given training data and the threshold was varied between 0 and 1, i.e. $p_{\text{threshold}} \in [0, 1]$.



(a) Gaussian noise process.

(b) Student's t-noise process.

Figure 5.5: Performance for the GARCH(1,1) models.

It is hard to see which models performs the best by visual inspection and the area under these curves also tells that there is more or less no difference. The area under the curve for GARCH with Gaussian noise is 0.9426 and for GARCH with Student's t-noise 0.9455, indicating that the overall performance for both models are almost identical if one would neglect the run time. The run time of the different models will be presented in section 5.4.

Another comparison, similar to the AR(1) models, has been made for the GARCH(1,1) models with Gaussian noise. Here the models use $n = 100$ and $n = 500$ historical observations respectively. In Figure 5.6 the result can been seen and the result indicated, just as for the AR(1) model, that there is no real difference between the $n = 100$ and $n = 500$. ROC AUC for the GARCH(1,1) with $n = 500$ observations is computed to be 0.9435, which confirms the visual inspection of a small or negligible difference.

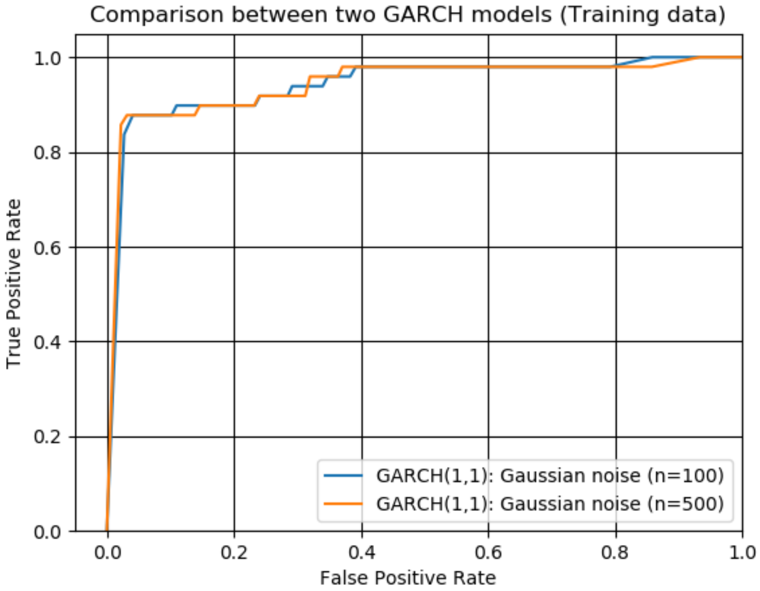


Figure 5.6: GARCH(1,1) models with Gaussian noise for two different n.

5.2.1 Test Statistic Method

Based on the empirical investigation of the training data above a GARCH(1,1) model will be used here as well. In Figure 5.7 the performance of the test statistic method can be seen, and it is rather clear that the methods presented previously performs far better than the test statistic method. The graph is based on the given training data. The parameter C was varied in the interval $[0, 10]$ to create the plot.

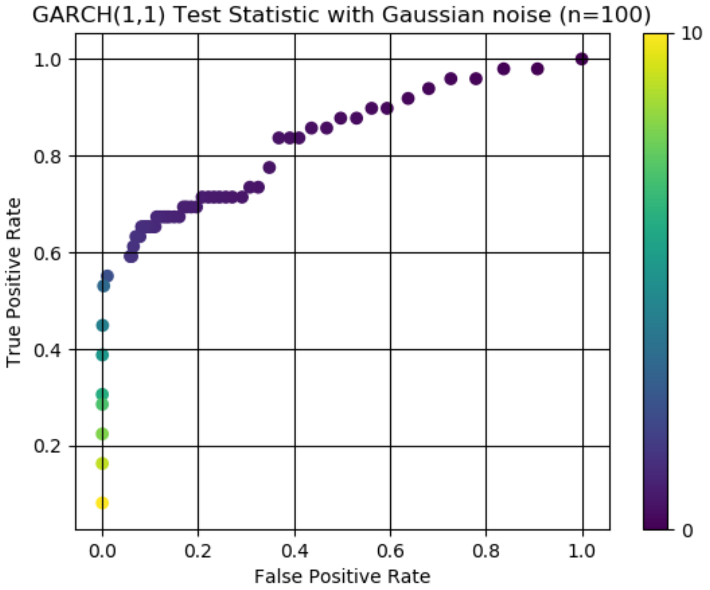
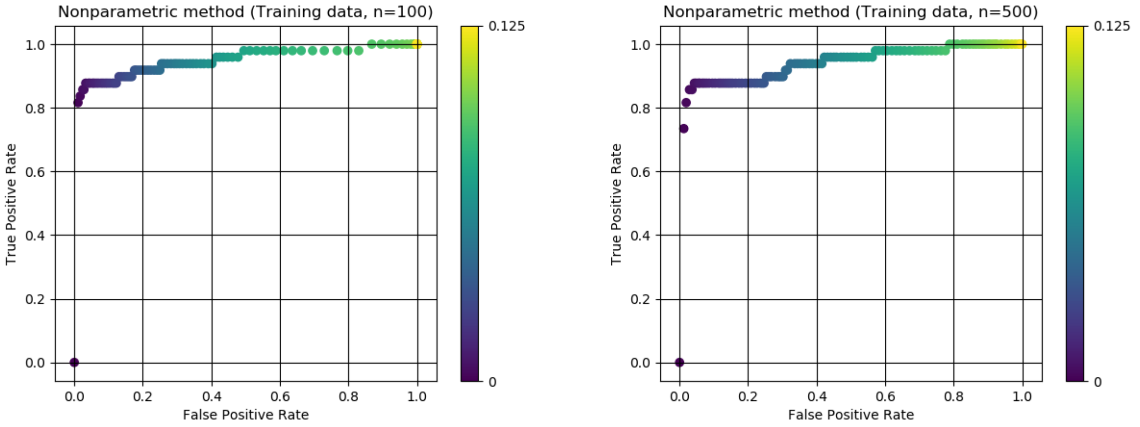


Figure 5.7: GARCH(1,1) test statistic method with Gaussian noise.

ROC AUC is computed to be 0.8364 which confirms the visual inspection of the graph. The overall performance of the test statistic method is worse than the methods presented previously. For this reason this method will be omitted from now on and will not be investigated any further.

5.3 Nonparametric Model

In Figure 5.8 one can see the result of the distance based outlier detection algorithm, based on kernel estimation. The threshold q was varied in the interval $[0, 0.125]$ in order to create the plots. The radius r was set to IQR/10 as mentioned in Chapter 4.



(a) $n = 100$.

(b) $n = 500$.

Figure 5.8: Performance of the nonparametric method.

The ROC AUC for Figure 5.8a was computed to 0.9474 and for Figure 5.8b to 0.9384. Similar to the parametric time series models the conclusion is the same, more observations used does not necessarily lead to a better overall performance. In Figure 5.9 it is also possible to see that the difference between $n = 100$ and $n = 500$ is almost negligible.

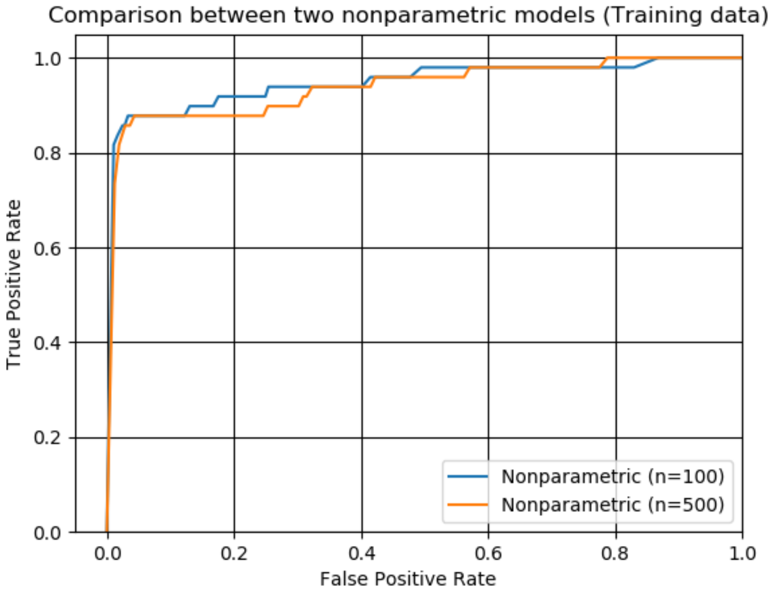


Figure 5.9: Comparison of the nonparametric methods for $n = 100$ and $n = 500$.

In order to see how the choice of the parameter r impacts the result of the nonparametric method a 3D plot has been made, see Figure 5.10. The scale on the x-axis is the factor, a, that IQR is multiplied by in order to get r, i.e. $r = a \cdot \text{IQR}$. One can from this plot draw the conclusion that the choice of r does not have a significant impact on the result since the surface looks rather similar in the direction of the IQR factor.

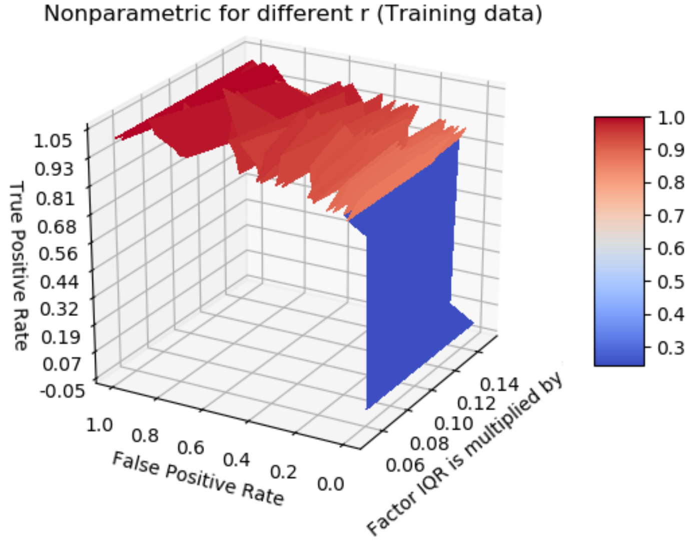


Figure 5.10: A 3D plot of the performance for the nonparametric model. The scale on the x-axis is the factor a: $r = a \cdot \text{IQR}$.

5.4 Algorithm Selection

In Figure 5.11 the results from the AR(1) and GARCH(1,1) with Gaussian and Student's t-noise is presented. As the plot and the ROC AUC indicate, see Table 5.1, all of the models perform very similar. The result here is probably not what one would expect, in particular the author of this thesis expected an advantage to the GARCH models. The reason for this expectation, is based on the fact GARCH was created in order to model prices in finance and the AR model is the simplest possible time series model, [5].

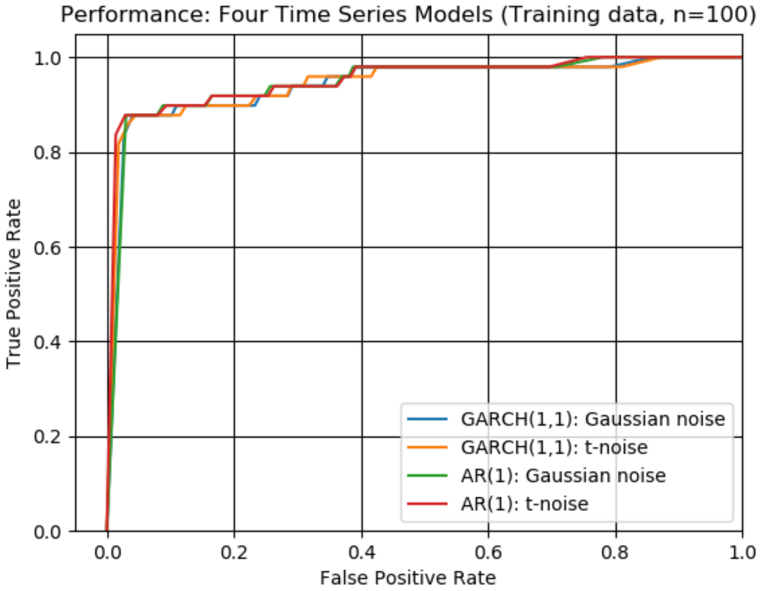


Figure 5.11: Comparison between AR(1) and GARCH(1,1) with two different noise distributions.

One conclusion that one can draw based on Figure 5.3 & 5.6 is that using $n = 100$ is probably enough with historical data in order to classify a point as an inlier or outlier. One would probably expect that an increase in the number of historical observations would increase the performance of the outlier detection algorithm but this is not the case based on this empirical investigation. This can probably be explained by the fact that in financial data, and especially asset and index prices, there is a lot of noise in the data and that more observations does not give the model more information. Another possible explanation would be that data is non-stationary and that the model parameters have changed when changing n. Moreover, the more recent data points are more relevant than the old ones, i.e. the older data points does not contribute with more information. This conclusion seems plausible since the training data, of course, does not come from a true AR or GARCH model, these time series models are just models and none of the models will fit perfectly to the data.

Based on the numbers provided in Table 5.1 one can compute the percentage change of ROC AUC when n is changed. The AR(1) with Gaussian noise has an decrease of ROC AUC of 1.24% whilst then GARCH(1,1) with Gaussian noise has an decrease of 0.10% when increasing n from 100 to 500. For the nonparametric method the decrease is 0.95%. For all of the models the change of ROC AUC is almost negligible and the same conclusion made earlier also holds here, an increase of n does not necessarily give better outlier detection. For this reason the author will recommend to use $n = 100$ historical observations when implementing the models, and $n = 100$ will also be used when the investigation of the models continue.

Model	ROC AUC
AR(1) Gaussian noise ($n = 100$)	0.9461
AR(1) Student's t-noise ($n = 100$)	0.9526
GARCH(1,1) Gaussian noise ($n = 100$)	0.9426
GARCH(1,1) Student's t-noise ($n = 100$)	0.9455
GARCH(1,1) Test statistic ($n = 100$)	0.8364
AR(1) Gaussian noise ($n = 500$)	0.9345
GARCH(1,1) Gaussian noise ($n = 500$)	0.9435
Nonparametric ($n = 100$)	0.9474
Nonparametric ($n = 500$)	0.9384

Table 5.1: ROC AUC for the different models based on training data.

Based on the data in Table 5.1 one can see no particular advantage, or reason why to choose any of the algorithms, since all of the algorithms have very similar ROC AUC apart from the GARCH(1,1) test statistic. This method will, as mentioned, be disregarded from further investigation.

Based on the training data and the graphs presented above one has to choose a specific threshold for outlier detection. By looking at the models presented so far, neglecting the GARCH test statistic model, and by setting an upper bound on the FP rate of 5% (0.05), the results are as follows, see Table 5.2. Since the interval for the "Threshold" is discretized the FP rate will not be exactly 0.05 in the table below, but rather as close as the discretization of the "Threshold" allows.

Model	FP Rate	TP Rate	Threshold
AR(1) Gaussian noise ($n = 100$)	0.0488	0.8776	0.0136
AR(1) Student's t-noise ($n = 100$)	0.0486	0.8776	0.0177
GARCH(1,1) Gaussian noise ($n = 100$)	0.0484	0.8776	0.0132
GARCH(1,1) Student's t-noise ($n = 100$)	0.0499	0.8776	0.0177
AR(1) Gaussian noise ($n = 500$)	0.0494	0.8367	0.0142
GARCH(1,1) Gaussian noise ($n = 500$)	0.0489	0.8766	0.0180
Nonparametric ($n = 100$)	0.0459	0.8776	0.0088
Nonparametric ($n = 500$)	0.0487	0.8776	0.0076

Table 5.2: Performance of the models for an upper bound of 5% on the FP rate, based on the training data.

The algorithms implemented by the author have different run time, and based on the simulations that have been done the performances are as follows. Consider the AR model with Gaussian noise, if this model takes T units of time to run then the run time of the other models are approximately as in Table 5.3.

Model	Run time [Time units]
AR(1) Gaussian noise	$1T$
AR(1) Student's t-noise	$\approx 1T$
GARCH(1,1) Gaussian noise	$\approx 6T$
GARCH(1,1) Student's t-noise	$\approx 12T$
Nonparametric	$\approx 100T$

Table 5.3: Approximate run time for the different models.

At first sight it might be strange that the GARCH model with Student's t-noise takes twice the time as the one with Gaussian noise, but this is probably due to the fact that the algorithm has to estimate the degrees of freedom for the noise process as well. The observant reader may here think that the AR model with Student's t-noise then should take twice as much time as the AR model with Gaussian noise. However, for the AR model with Student's t-noise the degrees of freedom, ν, is fixed to five and the Yule-Walker equations (used to estimate the parameters) estimates the variance of the noise process, regardless of its distribution. Another observation is of course that the nonparametric model, based on kernel density estimation, takes huge amount of times to run, but it performs rather well, just as the authors of the article, [37], wrote about their method.

Based on the data presented above two of the models seems particularly interesting to further investigate. The AR(1) with Student's t-noise has the best overall ROC AUC and also performs well at a 5% upper bound on the FP rate. Furthermore the run time of the model is very competitive compared to the others. Another reason for choosing the AR(1) model with Student's t-noise instead of the one with Gaussian noise is the fact that the former has heavier tails than the latter and this is an important property of financial data, mentioned in section 2.3.

The other one which seems interesting to investigate further is the GARCH(1,1) with Gaussian noise. The GARCH(1,1) with Student's t-noise is neglected since the performance are almost identical to the one with Gaussian noise and it has twice the run time. Furthermore the GARCH(1,1) with Gaussian noise performance well at the chosen upper bound on the FP rate and the GARCH models already exhibits heavy tails with a Gaussian noise distribution, [36]. One argument for not only investigating the AR(1) model with Student's t-noise is that, as mentioned the GARCH model should fit well to the returns of financial prices and therefore the author believes that this model requires some further investigation. In Figure 5.12 the two chosen models can be seen.

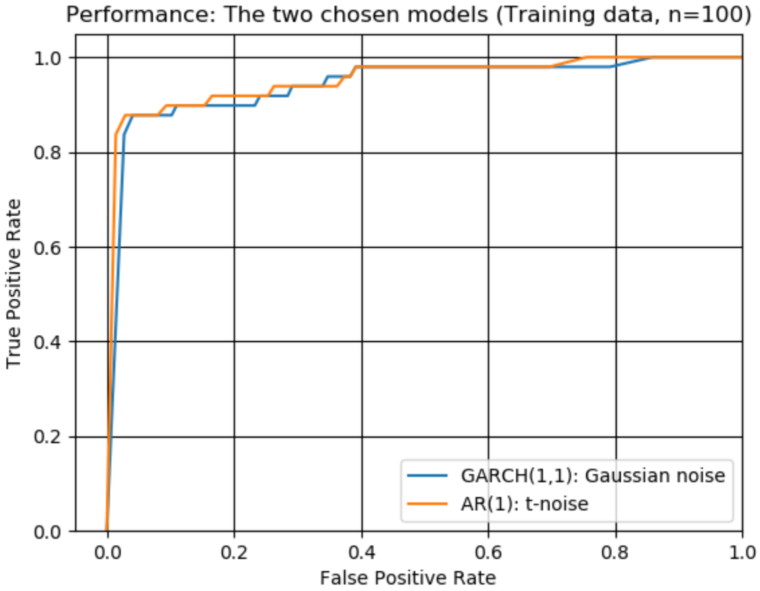


Figure 5.12: Comparison between AR(1) with Student's t-noise and GARCH(1,1) with Gaussian noise distributions.

An additional comment about the nonparametric method for outlier detection is required here. The overall performance of this method is very good and similar to the two chosen time series models, but as one can see in Table 5.3 the algorithm requires a *very* long run time and is therefore neglected since time is one important factor for the application of this thesis.

Lastly a threshold for the two models has to be chosen and one can find these thresholds in Table 5.2. Here one finds that the AR(1) with Student's t-noise will have the threshold $p_{\text{threshold}} = 0.0177$ and the GARCH(1,1) with Gaussian noise will have $p_{\text{threshold}} = 0.0132$, which should give an FP rate close to 5%.

5.4.1 Validation

The first step of the validation is to use, as mentioned in section 1.1.1, data from Dow Jones Industrial Average (DJIA) with outliers inserted. In Figure 5.14a the logarithmic returns of the DJIA data can be seen and in Figure 5.14b the same data is shown, with outliers inserted. Similar to Figure 1.2b the period with higher volatility during late 2008 can clearly be seen in Figure 5.14a. In addition the DJIA index itself is shown in Figure 5.13, with outliers inserted. The DJIA data set consists of 4595 data points in total, and 50 outliers are inserted, similar to the S&P 500 data set.

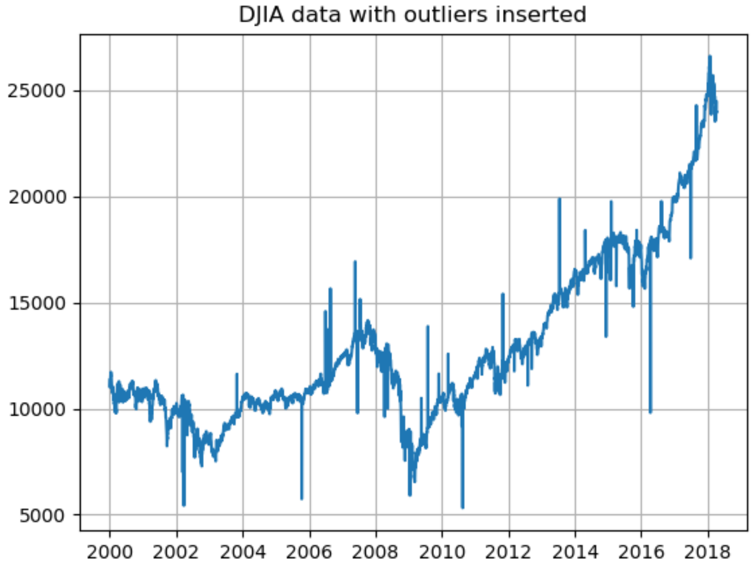
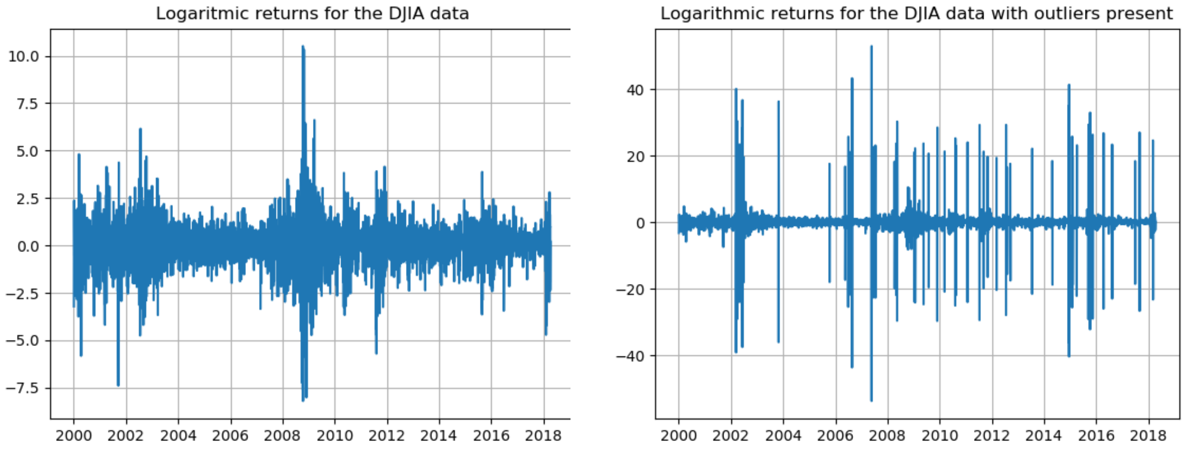


Figure 5.13: The price of the DJIA with outliers inserted.

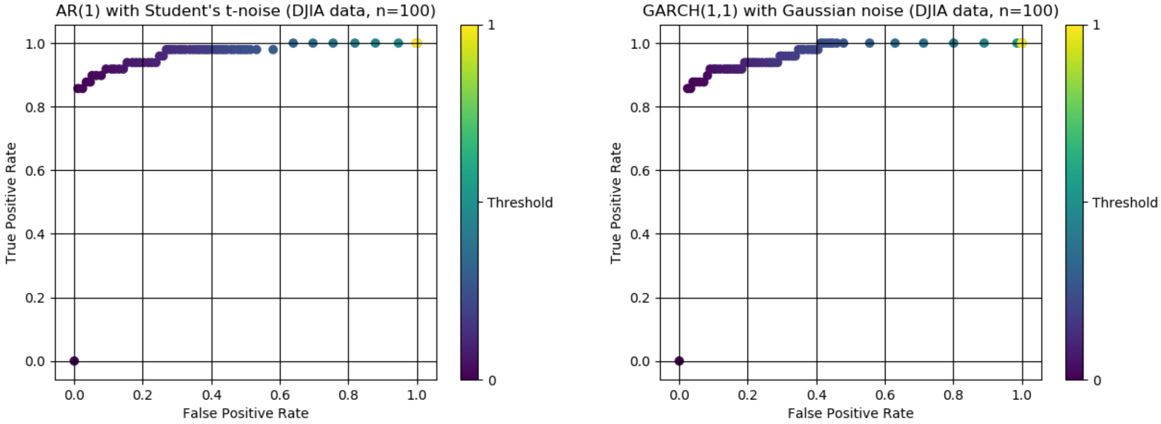


(a) No outliers present.

(b) Some outliers present.

Figure 5.14: Logarithmic returns (in percent) of the DJIA data.

In Figure 5.15a the performance of the AR(1) model with Student's t-noise is shown. Similar graphs are created here as for the training data set, but ultimately the model is evaluated at the chosen threshold. For this AR(1) model the ROC AUC is computed to be 0.9661, which has to be considered a very good performance. For the GARCH(1,1) model with Gaussian noise a similar figure has been created, see Figure 5.15b. Here the ROC AUC is computed to 0.9604, and the performance is almost identical to the AR(1) model.



(a) AR(1) with Student's t-noise.

(b) GARCH(1,1) with Gaussian noise.

Figure 5.15: Performance of the chosen models based on the DJIA data.

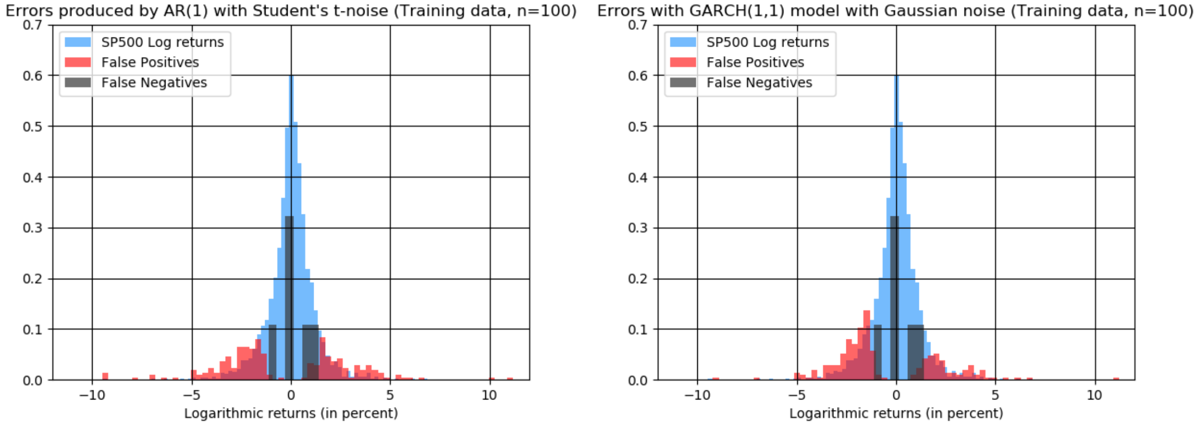
Below, in Table 5.4, the performance for the two chosen models is shown, for the threshold that were chosen based on the training data. Here one can see that the performance on the DJIA data set is very similar to the performance on the S&P 500 data set, in terms of the FP and TP rate at the specific thresholds. This is probably what one would expect since the two stock indices have a high (positive) correlation and the inserted outliers are generated in a similar way. Overall the result of the validation with the DJIA data is close to the result for the S&P 500 data set, which of course is a good sign, that the models do work on another, but similar, stock index.

Model	**FP Rate**	**TP Rate**	**Threshold**
AR(1) Student's t-noise ($n = 100$)	0.0468	0.8776	0.0177
GARCH(1,1) Gaussian noise ($n = 100$)	0.0481	0.8776	0.0132

Table 5.4: Performance of the two chosen models at the chosen threshold, based on the DJIA data set.

5.4.2 Further Investigation

In order to understand more about the chosen models and their performance some further investigation is necessary. The two things that one could want to investigate is which types of points do the models classify falsely, i.e. what are the characteristics of the points that falsely gets classified as outliers (false positives) and falsely classified as inliers (false negatives). This is done for the two chosen thresholds. The result is presented in Figure 5.16a & 5.16b and these plots are created based on the training data. From these two figures one can see the main characteristics of the two errors that the models make. In general, the false positives, i.e. the points that are labeled to be outliers but are not, are in general found in the tails of the distribution. On the other hand the false negatives, i.e. the points that truly are outliers and are not "detected" by the models are found in the center of the distribution.



(a) AR(1) with Student's t-noise.

(b) GARCH(1,1) with Gaussian noise.

Figure 5.16: Errors produced by the two chosen model (at the chosen thresholds) based on the training (S&P 500) data.

There is a rather intuitive explanations why the errors appear in this specific way, one type of error in the tail of the distribution and one in the center of the distribution. This all comes down to the how an outlier is defined, and outliers in this thesis are defined according to equation (1.1). From this equation it is possible to see that the value of an outlier defined at a random time point, τ, could have very similar vaue as without the outlier. This will happen *if* the r.v. with Student's t-distribution with three degrees of freedom, t_3, in equation (1.1) is simulated to be close to zero. The consequence of this specific event is that these points are set to be outliers but the points are not extreme in the sense that outliers are thought of in this thesis. Hence the models have a hard time "finding" these points which yields the error in the center of the distribution, the "false negative" errors.

The other errors are the points which are labeled as outliers but are not according to the definition of outliers in this thesis, see equation (1.1). As mentioned these points are found in the tails of the distribution and they are just "extreme" movements in the stock index on a daily basis. Ideally one would not want the models to "detect" these points since the points are not "wrong" in the way that AP4 think of outliers. But they are rather extreme and this is why the models label these points as outliers, yielding the false positive errors.

With this in mind it is necessary to revise the way that outliers are defined, since for AP4 the points that are thought of as outliers are "extreme" and the points are also false which means that it should be possible to eliminate most of the false negative errors, the errors that occur in the center of the distribution.

5.4.3 Revision of the Outlier Definition

Consider a revision of equation (1.1), where the t_3 r.v. is replaced by another r.v., X.

$$\hat{P}_\tau = P_\tau \cdot (1 + 0.15 \cdot X) \tag{5.1}$$

The goal with this new r.v. X is to only yield values $\hat{P}_\tau$ that substantially deviate from P_τ, such that the false negative errors are eliminated. One way of doing this is to only let X take values that are in the tails of the t_3 distribution. This can be done with the help of a symmetric Bernoulli r.v., denoted $\text{Be}(\frac{1}{2})$, and the so-called quantile transform. The former defined in most books related to probability theory, e.g. [19], while the latter is defined in [24]. See the following definition and proposition.

Definition. Let Z be a *symmetric Bernoulli random variable*, then Z has the probability mass function

$$\mathbb{P}(Z = -1) = \frac{1}{2}, \quad \mathbb{P}(Z = 1) = \frac{1}{2},$$

which is denoted $Z \sim \text{Be}(\frac{1}{2})$.

Proposition. Let $U \sim \mathcal{U}(0, 1)$, furthermore let a r.v. Y have CDF $F_Y(y)$ and quantile function $F_Y^{-1}(p)$, then

$$F_Y^{-1}(U) \sim Y.$$

Which means that if one lets the argument of the quantile function be the uniform r.v. $U \sim \mathcal{U}(0, 1)$ then the distribution is of this quantile function is the same as the distribution of Y.

With this knowledge it is now possible to simulate the r.v. X such that it only takes values from the tails of the Student's t-distribution with three degrees of freedom, t_3. Consider the following

1. Simulate a random variate Z from $\text{Be}(\frac{1}{2})$

   (a) If $Z = -1$, then simulate a random variate U from $\mathcal{U}(\alpha_1, \alpha_2)$

   (b) If $Z = 1$ then simulate a random variate U from $\mathcal{U}(\beta_1, \beta_2)$.

2. Finally use the quantile transform to compute one random variate X, $X = F_{t_3}^{-1}(U)$.

Here $0 \leq \alpha_1 < \alpha_2 \leq \beta_1 < \beta_2 \leq 1$ are bounds that define which parts of the t_3 distribution that X can attain. Furthermore $F_{t_3}^{-1}$ is the quantile function for a t_3 r.v. In the three following sections, section 5.4.3.1, 5.4.3.2 & 5.4.3.3 three different revisions of the definition of an outlier will be done, all utilizing the idea presented above.

5.4.3.1 First Revision of the Outlier Definition

For the first revision the author of the thesis has together with AP4 chosen the bounds to be as follow: $\alpha_1 = 0.05$, $\alpha_2 = 0.15$, $\beta_1 = 0.85$ and $\beta_2 = 0.95$, which will generate rather extreme values of the t_3 distribution. The outliers that AP4 has found manually are often wrong by a factor of two or four which would be very extreme outliers, so this definition here should be rather conservatitve. To illustrate the resulting distribution of X the procedure described above has been repeated $100,000$ times and the resulting distribution is shown in the histogram below, see Figure 5.17.

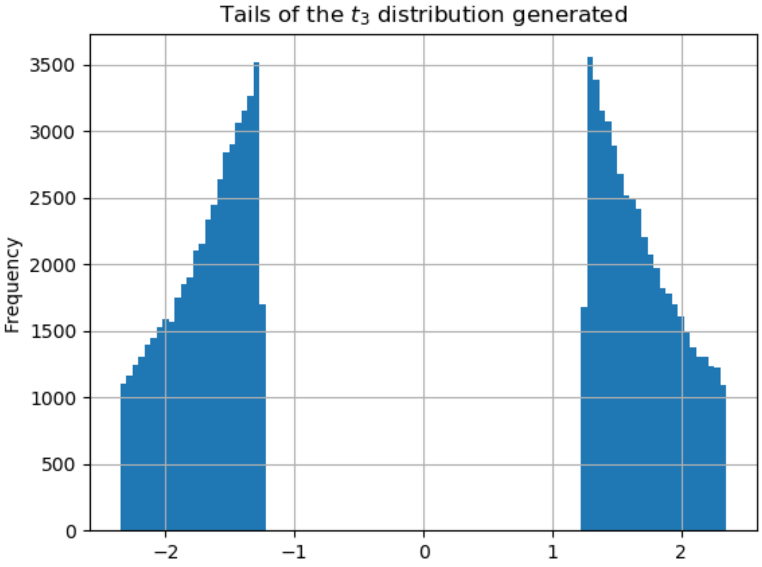
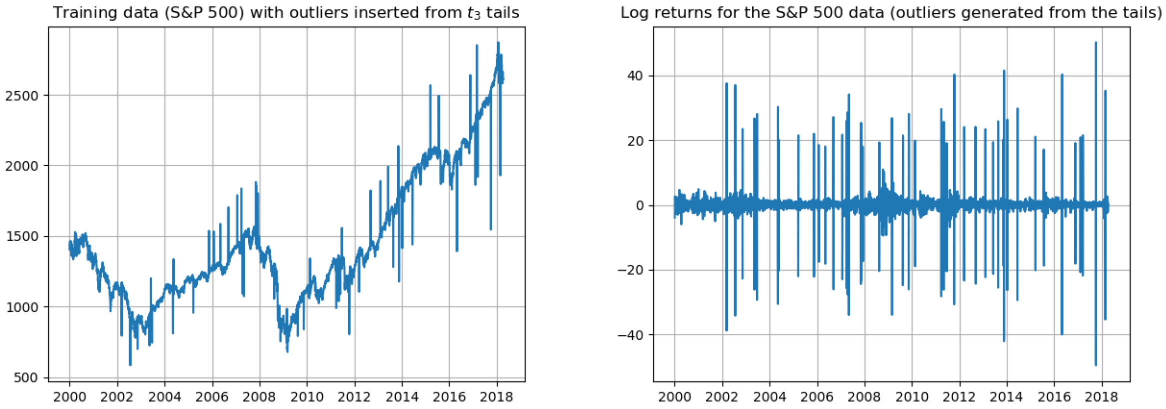


Figure 5.17: 100, 000 samples from tails of the Student's t_3-distribution.

In Figure 5.18a one can see the S&P 500 index together with 50 inserted outliers, the outliers were generated from the tails of the t_3 distribution, just as described in this section. In addition, in Figure 5.18b the logarithmic returns (in percent) of the data mentioned in the former figure can be seen. Here one can see that the outliers are more or less of the same magnitude as before (cf. Figure 1.2b), the one difference is that every outliers is "extreme".

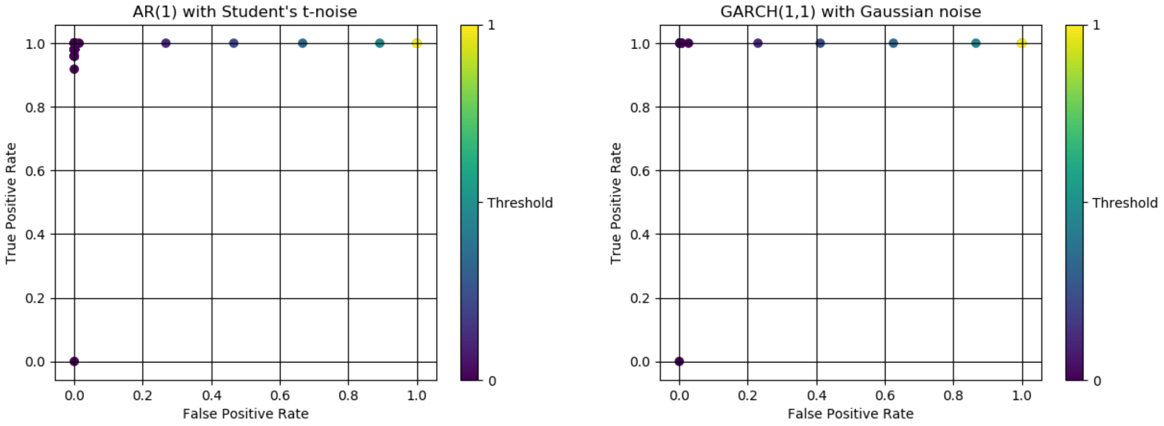


(a) The S&P 500 index with outliers.

(b) Logarithmic return (in percent).

Figure 5.18: Outliers inserted from the tails of the t_3 distribution into the S&P 500 data, according to equation (5.1).

Figure 5.19a & 5.19b are based on the S&P 500 data and the outliers are generated only from the tails of the t_3 distribution, as described in this section. As one can see the performance is almost "perfect" in the sense that one can see, by eye inspection, that the ROC AUC is very close to one. These two graphs tell one that the performance of a certain method is very much up to how outliers are defined, or how outliers are generated. An optimal threshold when outliers for these two method would, as the colorbar suggests, be a threshold very close to zero. In Table 5.5 the performance is shown for a threshold, for a lower bound on the TP rate of 100% (1.0). From this table it is also clear that the AR(1) model outperforms the GARCH(1,1) model when looking at the FP rate, by a factor of approximately five.



(a) AR(1) model with Student's t-noise. **(b)** GARCH(1,1) model with Gaussian noise.

Figure 5.19: Performance based on the S&P 500 data with outliers defined as in equation (5.1).

Model	FP Rate	TP Rate	Threshold
AR(1) Student's t-noise	$4.50 \cdot 10^{-4}$	1.0	$1.08 \cdot 10^{-4}$
GARCH(1,1) Gaussian noise	$2.25 \cdot 10^{-3}$	1.0	$1.68 \cdot 10^{-6}$

Table 5.5: Performance when outliers are generated by equation (5.1).

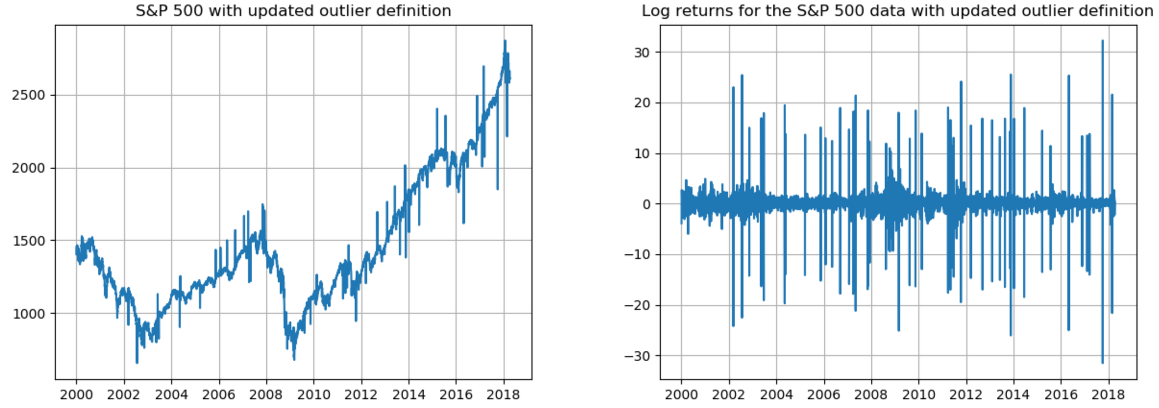
These, very good, results presented above make it further interesting to investigate how the definition of an outlier affects the performance of the AR(1) and the GARCH(1,1) model. The investigation will be made in two steps, firstly the way an outlier is defined will be revised, once again. Then, in the second step, the updated outlier definition will be kept but the distribution of X will be changed. The goal with these two steps in the investigation is to see how good the models can detect outliers with a smaller magnitude than previously.

5.4.3.2 Second Revision of the Outlier Definition

For the second revision, equation (5.1) will be modified as follows

$$\hat{P}_\tau = P_\tau \cdot (1 + 0.1 \cdot X)\,, \tag{5.2}$$

which simply means that the outliers inserted, will not be as extreme as before. Here the distribution of X will be kept as described in section 5.4.3, simulated from the tails of the t_3 distribution. In Figure 5.20 one can see the outliers that are inserted into the S&P 500 data and one can also see that the magnitude of the outliers are smaller compared to Figure 5.18.

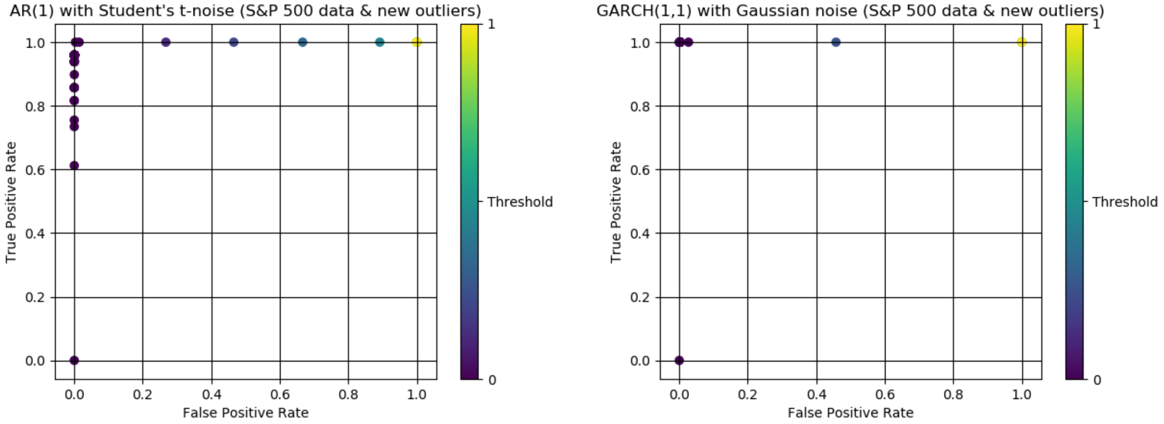


(a) The S&P 500 index with outliers.

(b) Logarithmic return (in percent).

Figure 5.20: Outliers inserted from the tails of the t_3 distribution into the S&P 500 data according to equation (5.2).

In Figure 5.21 one can see that the performance still is close to "perfect" when outliers are defined by equation (5.2), i.e. that the ROC AUC is very close to one, which can be seen by eye inspection.



(a) AR(1) model with Student's t-noise.

(b) GARCH(1,1) model with Gaussian noise.

Figure 5.21: Performance based on the S&P 500 data with outliers defined as in equation (5.2).

Furthermore in Table 5.6 the performance is presented for a lower bound on the TP rate of 100% (1.0). Here it seems that the GARCH(1,1) model outperforms the AR(1) model in terms of the FP rate, by a factor of approximately 5, which is the opposite result compared to the previous section.

Model	**FP Rate**	**TP Rate**	**Threshold**
AR(1) Student's t-noise	$3.37 \cdot 10^{-3}$	1.0	$1.04 \cdot 10^{-3}$
GARCH(1,1) Gaussian noise	$6.75 \cdot 10^{-4}$	1.0	$2.56 \cdot 10^{-8}$

Table 5.6: Performance when outliers are generated by equation (5.2).

5.4.3.3 Third Revision of the Outlier Definition

The third, and last, step in the investigation the outlier definition as in equation (5.2) will be kept, but the distribution of X will be changed such that it is closer to zero, i.e. the values that X can attain, will not be as extreme as before. The bounds are changed as follows:

$$\alpha_1 = 0.1, \quad \alpha_2 = 0.2, \quad \beta_1 = 0.8, \quad \beta_2 = 0.9.$$

The updated distribution of the r.v. X can be seen in Figure 5.22 where, compared to Figure 5.17, the probability mass is closer to zero on the horizontal axis.

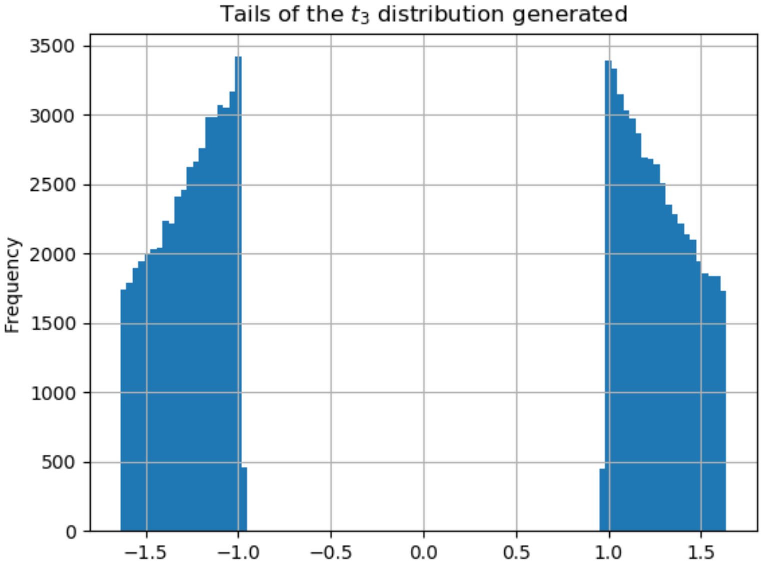
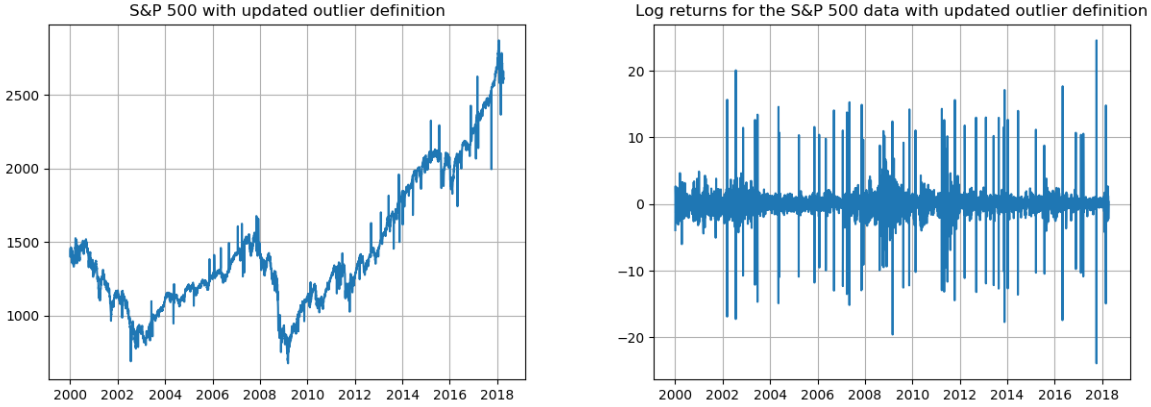


Figure 5.22: 100, 000 samples from tails of the Student's t_3-distribution.

With the distribution of X updated, the S&P 500 index with outliers can be seen in Figure 5.23a and the logarithmic returns of the same is shown in Figure 5.23b. Once again, the magnitude of the outlier has decreased compared to the previous definitions (cf. Figure 5.18 & 5.20).

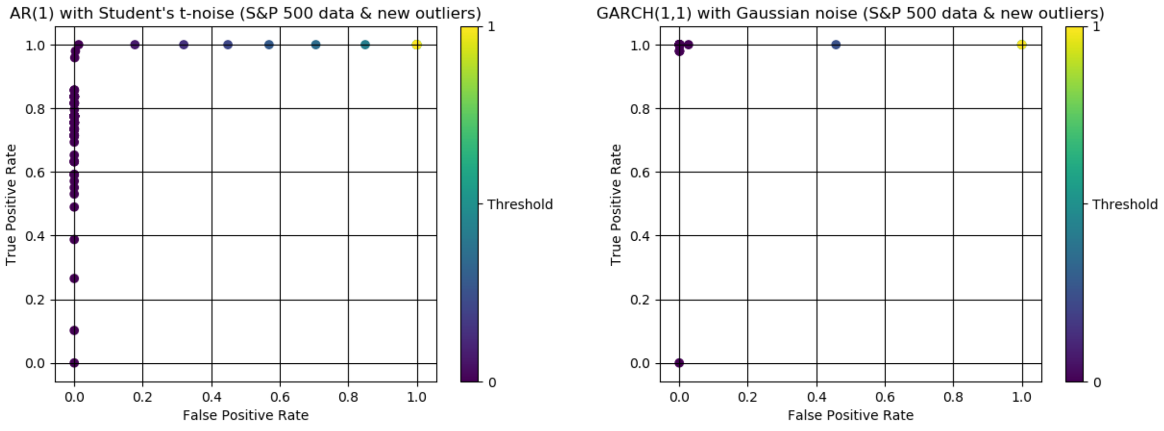


(a) The S&P 500 index with outliers.

(b) Logarithmic return (in percent).

Figure 5.23: Outliers inserted from the tails of the t_3 distribution into the stock index S&P 500 according to equation (5.2) & distribution of X closer to zero.

In the following two figures, Figure 5.24a & 5.24b, the result from the AR(1) and the GARCH(1,1) model is shown, when the distribution of X is set to be closer to zero. The figures do not say much, apart from the fact that the overall performance of the models are close to “perfect”, just as in section 5.4.3.1 & 5.4.3.2.



(a) AR(1) model with Student’s t-noise.

(b) GARCH(1,1) model with Gaussian noise.

Figure 5.24: Performance based on the S&P 500 data with outliers defined as in equation (5.2) & distribution of X closer to zero.

In Table 5.7 one can see the performance at one specific threshold, chosen such that there is a lower bound on the TP rate of 100% (1.0). Once again it appears that the GARCH(1,1) model outperforms the AR(1), by a factor of approximately 15.

Model	FP Rate	TP Rate	Threshold
AR(1) Student’s t-noise	$9.90 \cdot 10^{-3}$	1.0	$3.72 \cdot 10^{-3}$
GARCH(1,1) Gaussian noise	$6.75 \cdot 10^{-4}$	1.0	$1.03 \cdot 10^{-8}$

Table 5.7: Performance when outliers are generated by equation (5.2) & distribution of X closer to zero.

Based on the data presented in Table 5.5, 5.6 & 5.7 and how the magnitude of the inserted outliers have been changed between the presented data one could make a rather interesting conclusion. In the first of the mentioned tables, the magnitude of the outliers are the most extreme and here the AR(1) model outperforms the GARCH(1,1) model. As the magnitude of the outliers decrease, the GARCH(1,1) model performs better compared to the AR(1) model. It is possible that the GARCH(1,1) performs better, compared to the AR(1) model, when the magnitude of the outliers decrease and vice versa.

In Appendix C two additional investigations have been carried out. In particular in section C.1 an outlier is redefined to be an “extreme” movement in the price of a financial asset and not false point as previously. Moreover in section C.2 two time series models has been simulated and then “extreme” and false outliers are added and detected by the models.

Chapter 6

Conclusion

Several models were constructed and used for outlier detection. The different models used were, two different time series models (AR and GARCH model) with different submodels as well as one nonparametric model. The performance was evaluated by how well the models classified the points, comparing the "false positive rate" and "true positive rate" between the different models. Overall, the performance for all models was good, apart from one model that used a so-called "test statistic" that underperformed compared to the other models. The result was analyzed and discussed from several different standpoints. One important point was, of course, how well the models detected outliers and how well they classified points that were not outliers. Another important property for the models were their run time (or execution time) which has to be taken into account when the final implementation, or model selection, is carried out.

The first time series model, the AR model, was analyzed for both a Gaussian noise distribution as well as a Student's t-noise distribution. The results were that the performance was almost identical for the different noise distribution and since the Student's t-distribution has heavier (which in turn gives the AR model heavier tails), this one was chosen for further investigation. Furthermore both models had similar run time so this aspect did not impact the model choice here.

The second time series model, the GARCH model, was also analyzed for the two mentioned noise distributions. Here, as well, the results was almost identical but the run time was different between the two noise distributions which had an impact on the model choice. The GARCH model with Student's t-noise had twice the run time, compared to the GARCH with Gaussian noise, hence the model with Gaussian noise was chosen for further investigation. Lastly, as mentioned above, the "test-statistic" method performed poorly compared to the previous presented models so it was ruled out for further use.

Lastly, the nonparametric model was examined and it was equally good compared to the chosen AR and GARCH models. However, the run time was much longer than the AR and the GARCH models (see Table 5.3) so this model was also ruled out for further investigation.

To continue there were only two models left, the AR with Student's t-noise distribution and the GARCH with Gaussian noise distribution. Another interesting aspect that was investigated was to see what properties the points had that were not classified correctly. It was shown that the points which was classified as outliers but were not outliers (false positives) were of large magnitude, i.e. the false positive points deviate a lot the majority of the points. It was also shown that the points that were outliers but the models classified them as not being outliers (false negatives) was found close to the mean of all points. With this knowledge acquired the definition of an outlier was revised in several steps and it was shown that the way an outlier is defined is *very crucial* to how well a model is able to detect the outliers. Moreover the two chosen models were found to perform very well when the outlier definition was revised. From the short empirical investigation it was found that the GARCH model was found to detect outliers of smaller magnitude (but still substantially deviate from the true value) better than the AR model did. On the other hand the AR model was better at detecting outliers of larger magnitude compared to the GARCH. With this in mind the author believes the AR model with Student's t-noise distribution to be the best choice for this application considering how it performs and how the run time is compared to the GARCH model.

6.1 Critique

There are, of course, several aspects of the project that should be given some critique. One of these aspects is briefly discussed by Francq and Zakoïan, [14], and is related to the selection of the order of the GARCH(p,q) process. They state that the GARCH(1,1) model is overrepresented and that some financial time series (such as the S&P 500) may require a higher order GARCH model in order to capture the properties of these complex time series. They finally show that the GARCH(1,1) model is overrepresented in financial applications. However, as mentioned in section 5.2, the procedure of testing which GARCH order is the most optimal, according to some information criterion is not feasible, for this application, with respect to the run time.

Similarly for the autoregressive (AR) model one could have tried to fit models of different order to see which one would fit best to the data according to some information criterion. The approach with testing several models would be more feasible for the AR model compared to the GARCH due to the lower run time of the AR model.

One could also argue that there should have been more outliers inserted into the data, e.g. in the S&P 500 data set the fraction of outliers was close to 1%, and similarly for the Dow Jones Industrial Average data. One reason for having more outliers in the data would be that the ROC curve would be more smooth which probably would impact the ROC AUC for all the models. On the other hand the training data should be representative of the real data, and in the real data outliers are "very rare" according to AP4.

6.2 Possible Extensions

One extension of the project would be to try to find a method that can detect outlier for so-called "holding data". Holding data is the time series which states how many shares of company that is held by an investor. This data does, most often, not look like financial prices and do not share the typical properties that one can see in financial prices. For this reason one, would probably, need a completely new model in order to handle outliers in holding data.

Other possible extension would be related to dependence between assets. For instance, if there is an extreme movement (positive or negative) in the market, then the models might "detect" outliers even though these points are true (false positive error). If a large fraction of the assets are detected to have outliers then it is likely that these are false positives, so it could probably be possible to remove some of the false positive errors in a case such as the one described.

Other possible methods more related to "machine learning" could be interesting to investigate, such as a neural network or a support vector machine approach, see [21] & [25] respectively.

Bibliography

[1] C.C. Aggarwal. *Outlier Analysis.* Springer International Publishing, 2016.

[2] N. S. Altman. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185, 1992.

[3] L. Andrews. Beta function. In *Field Guide to Special Functions for Engineers*, pages 26–26. SPIE Press, 2011.

[4] A. Berchtold and A. Raftery. The mixture transition distribution model for high-order markov chains and non-gaussian time series. *Statistical Science*, 17(3):328–356, 2002.

[5] T. Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31:307–327, 1986.

[6] T. Bollerslev. A conditionally heteroskedastic time series model for speculative prices and rates of return. *The review of economics and statistics*, 69(3):542–547, 1987.

[7] G.E.P. Box and G.M. Jenkins. *Time series analysis : forecasting and control.* Holden-Day series in time series analysis. Holden-Day, San Francisco, rev. ed.. edition, 1976.

[8] G.E.P. Box, G.M. Jenkins, and G.C. Reinsel. *Time series analysis forecasting and control.* Wiley Series in Probability and Statistics. John Wiley, 4th ed.. edition, 2008.

[9] T.J. Brailsford, J.H.W. Penm, and R.D. Terrell. Selecting the forgetting factor in subset autoregressive modelling. *Journal of Time Series Analysis*, 23(6):629–649, 2002.

[10] P.J. Brockwell and R.A. Davis. *Introduction to Time Series and Forecasting.* Number v.1 in Introduction to Time Series and Forecasting. Springer, 2002.

[11] A. Charles and O. Darné. Outliers and garch models in financial data. *Economics Letters*, 86(3):347–352, 2005.

[12] A. DasGupta. *Probability for Statistics and Machine Learning: Fundamentals and Advanced Topics.* Springer Texts in Statistics. Springer New York, 2011.

[13] J. Fan and Q. Yao. *Nonlinear Time Series Nonparametric and Parametric Methods.* Springer Series in Statistics. Springer, 2005.

[14] C. Francq and J.M. Zakoïan. *GARCH Models: Structure, Statistical Inference and Financial Applications.* Wiley, 2011.

[15] P.H. Franses and H. Ghijsels. Additive outliers, garch and forecasting volatility. *International Journal of Forecasting*, 15(1):1–9, 1999.

[16] D. Freedman and P. Diaconis. On the histogram as a density estimator: L^2 theory. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 57(4):453–476, 1981.

[17] G.K. Grunwald, R.J. Hyndman, L. Tedesco, and R.L. Tweedie. Non-gaussian conditional linear ar(1) models. *Australian and New Zealand Journal of Statistics*, 42(4):479–495, 2000.

[18] M. Gupta, J. Gao, C. C. Aggarwal, and J Han. Outlier detection for temporal data: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 29(6):2250–2267, 2014.

[19] Allan Gut. *Probability: A Graduate Course.* Springer Texts in Statistics. Springer New York, 2013.

[20] Richard H. *Numerical Methods for Scientists and Engineers.* Dover Books on Mathematics. Dover Publications, Newburyport, 2012.

[21] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition.* Springer Series in Statistics. Springer New York, 2009.

[22] D.M. Hawkins. *Identification of Outliers.* Monographs on applied probability and statistics. Chapman and Hall, 1980.

[23] M.S. Heracleous. *Volatility Modeling Using the Student's t Distribution.* PhD thesis, Virginia Polytechnic Institute and State University, 2003.

[24] H. Hult, F. Lindskog, O. Hammarlid, and C.J. Rehn. *Risk and Portfolio Analysis: Principles and Methods.* Springer Series in Operations Research and Financial Engineering. Springer New York, New York, NY, 2012.

[25] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning: with Applications in R.* Springer Texts in Statistics. Springer New York, 2013.

[26] E. Jondeau, S.H. Poon, and M. Rockinger. *Financial Modeling Under Non-Gaussian Distributions.* Springer Finance. Springer, 2007.

[27] H. Kamranfar, R. Chinipardaz, and B. Mansouri. Detecting outliers in garch(p,q) models. *Communications in Statistics - Simulation and Computation*, 46(10):7844–7854, 2017.

[28] R. Karp. On-line algorithms versus off-line algorithms: How much is it worth to know the future? In *International Computer Science Institute*, 1992.

[29] E.M. Knorr and R.T. Ng. Algorithms for mining distance-based outliers in large datasets. In *Proceedings of the 24rd International Conference on Very Large Data Bases*, VLDB '98, pages 392–403. Morgan Kaufmann Publishers Inc., 1998.

[30] L.J. Latecki, A. Lazarevic, and D. Pokrajac. Outlier detection with kernel density functions. In *Machine Learning and Data Mining in Pattern Recognition*, volume 4571, pages 61–75, 2007.

[31] C.K.-S. Leung, R.K. Thulasiram, and D.A. Bondarenko. An efficient system for detecting outliers from financial time series. In *Flexible and Efficient Information Handling*, pages 190–198. Springer Berlin Heidelberg, 2006.

[32] C. Leys, C. Ley, O. Klein, P. Bernard, and L. Licata. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, 49(4):764–766, 2013.

[33] D. Peña, G.C. Tiao, and R.S. Tsay. *A course in time series analysis.* Wiley series in probability and statistics: Probability and statistics. Wiley, 2001.

[34] F. Provost and T. Fawcett. Robust classification for imprecise environments. *Machine Learning*, 42(3):203–231, 2001.

[35] J. Qu. Outlier detection in financial data based on voronoi diagram. In *2008 4th International Conference on Wireless Communications, Networking and Mobile Computing*, pages 1–4, Oct 2008.

[36] D. Ruppert and D.S. Matteson. *Statistics and Data Analysis for Financial Engineering: with R examples.* Springer Texts in Statistics. Springer New York, 2015.

[37] Md.S. Sadik and L. Gruenwald. Dbod-ds: Distance based outlier detection for data streams. In *Database and Expert Systems Applications*, pages 122–136. Springer Berlin Heidelberg, 2010.

[38] D.W. Scott. On optimal and data-based histograms. *Biometrika*, 66(3):605–610, 1979.

[39] D.W. Scott. *Multivariate density estimation : theory, practice, and visualization.* Wiley Series in Probability and Statistics. Wiley, 2nd ed.. edition, 2015.

[40] B.W. Silverman. *Density estimation for statistics and data analysis.* Monographs on statistics and applied probability [26]. Chapman and Hall, 1986.

[41] F. Javier Trívez and Beatriz Catalán. Detecting level shifts in arma-garch (1,1) models. *Journal of Applied Statistics*, 36(6):679–697, 2009.

[42] R.S. Tsay. *Analysis of financial time series.* Wiley series in probability and statistics. Wiley, 3rd ed. edition, 2010.

[43] J. Zhang and X. Wang. Robust normal reference bandwidth for kernel density estimation. *Statistica Neerlandica*, 63(1):13–23, 2009.

Appendices

Appendix A

Mathematics

Conditional Distribution of GARCH Models

The PDF of $Z_t|\mathcal{F}_{t-1}$ when $e_t \sim \mathcal{N}(0,1)$ is as follows

$$f_{Z_t|\mathcal{F}_{t-1}}(x) = \frac{1}{\sqrt{2\pi h_t}} \exp\left(-\frac{x^2}{2h_t}\right),$$

and when the $e_t \sim \sqrt{\frac{\nu-2}{\nu}} t_\nu$ the PDF of $Z_t|\mathcal{F}_{t-1}$ becomes

$$f_{Z_t|\mathcal{F}_{t-1}}(x) = \frac{\Gamma(\frac{\nu+2}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi(\nu-2)h_t}} \left(1 + \frac{x^2}{h_t(\nu-2)}\right)^{-\frac{\nu+1}{2}} \qquad \nu - 2 > 0.$$

For the normal distribution the PDF is fairly simple to obtain, given that $\mathrm{Var}[Z_t|\mathcal{F}_{t-1}] = h_t$. For the Student's t-distribution it is a bit trickier but consider the PDF of a Standard Student's t-r.v. X,

$$f_X(x) = \frac{\Gamma(\frac{\nu+2}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} \qquad \nu > 0.$$

Here $\mathrm{Var}[X] = \frac{\nu}{\nu-2}$. Then consider the r.v. $Z_t|\mathcal{F}_{t-1} = \sqrt{h_t \frac{\nu-2}{\nu}} X$. Then $\mathrm{Var}\left[Z_t|\mathcal{F}_{t-1}\right] = h_t$. So given the PDF of X it is easy to obtain the PDF for $Z_t|\mathcal{F}_{t-1}$ with the help of the proposition below.

Proposition. If a r.v. X has the PDF $f_X(x)$ then the r.v. aX, $a > 0$, has the PDF $\frac{1}{a} f_X(\frac{x}{a})$.

Proof. Let the r.v. X have CDF $F_X(x) = \mathbb{P}(X \leq x)$. Consider the r.v. aX, $a > 0$, and its CDF, then: $F_{aX}(x) = \mathbb{P}(aX \leq x) = \mathbb{P}(X \leq \frac{x}{a}) = F_X(\frac{x}{a})$. If $\frac{d}{dx}F_X(x) = f_X(x)$ then $\frac{d}{dx}F_{aX}(x) = \frac{d}{dx}F_X(\frac{x}{a}) = \frac{1}{a} f_X(\frac{x}{a})$. □

Appendix B

Implementation Details

Python Libraries

The `Python` libraries that have been used for this thesis are listed below.

- `Matplotlib`[1]
- `Pandas`[2]
- `NumPy`[3]
- `SciPy`[4]
- `StatsModels`[5]
- `Spectrum`[6]
- `ARCH`[7]

Yule-Walker Equations versus Burg's Algorithm

Based on data from the Swedish stock index OMXS30[8], using daily closing prices from 2000-01-03 to 2017-12-29, in total 4528 price points. An AR(1) model was fitted to the mean corrected daily log returns (in percent), i.e. the model was fitted to

$$R_t^* = R_t - \bar{R}_t, \quad R_t = 100 \log \frac{P_t}{P_{t-1}}.$$

The functions `aryule()` and `arburg()`, from the library `Spectrum` was used to approximate the coefficients, $\hat{\phi}_1$ and $\hat{\sigma}^2$ for the model

$$R_t^* = \phi_1 R_{t-1}^* + \epsilon_t, \quad \epsilon_t \sim \text{IID}(0, \sigma^2).$$

The returns, R_t, where divided up into sections of $n = 100$ observations and then an AR(1) was fitted to each of these sections. In Figure B.1 the result is presented. As the scatter plot

[1] matplotlib 2.2.2

[2] pandas 0.22.0

[3] numpy 1.14.2

[4] SciPy 1.0.1

[5] statsmodels 0.8.0

[6] Spectrum Analysis Tools.

[7] arch 4.3.1

[8] The data can be fetched from Nasdaq Nordic.

indicates there are no significant difference between the Yule-Walker estimates and the estimates from Burg's algorithm, so Yule-Walker is chosen due to the easier implementation.

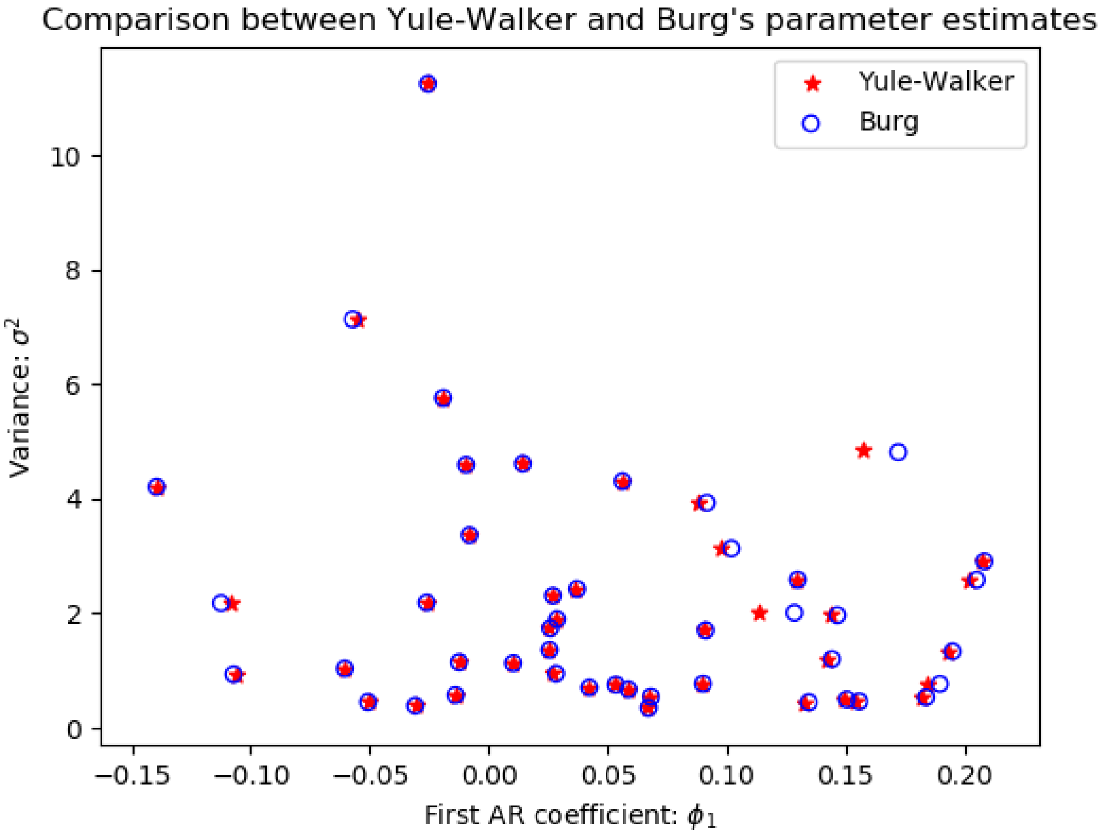


Figure B.1: AR(1) coefficients from the Yule-Walker equations vs. Burg's algorithm.

Implementation Details for Distance Based Outlier Detection

What one essentially wants to compute is the probability

$$p(\tilde{x}, r) = \int_{\tilde{x}-r}^{\tilde{x}+r} \frac{1}{h} \frac{\sum_{i=1}^{n} \gamma^{n-i} K\left(\frac{x_i - x}{h}\right)}{\sum_{i=1}^{n} \gamma^{n-i}} dx = \frac{1}{h} \frac{1}{\sum_{i=1}^{n} \gamma^{n-i}} \int_{\tilde{x}-r}^{\tilde{x}+r} \sum_{i=1}^{n} \gamma^{n-i} K\left(\frac{x_i - x}{h}\right) dx. \quad \text{(B.1)}$$

Here the points $\{x_i\}_{i=1}^{n}$ are the n logarithmic returns and $K(\cdot)$ is the Gaussian kernel function. The difficult part with this computation is that the integral cannot be solved explicitly, so it has to be approximated numerically. This can be done with the trapezoidal rule, see [20] for details. In general, given the closed interval $x \in [a, b]$, this interval is split into T sample points: $a = x_0 < x_1 < \cdots < x_{T-1} < x_T = b$. With the distance between the samples $\Delta x_j = x_j - x_{j-1}$ then it is possible to do the approximation

$$\int_{a}^{b} f(x)dx \approx \sum_{j=1}^{T} \frac{f(x_j) + f(x_{j-1})}{2} \Delta x_j.$$

With the help of the trapezoidal rule it is possible to evaluate the probability in equation (B.1) and thus implement the outlier detection algorithm.

Appendix C

Additional Investigation

C.1 Largest Daily Movement in Index Price

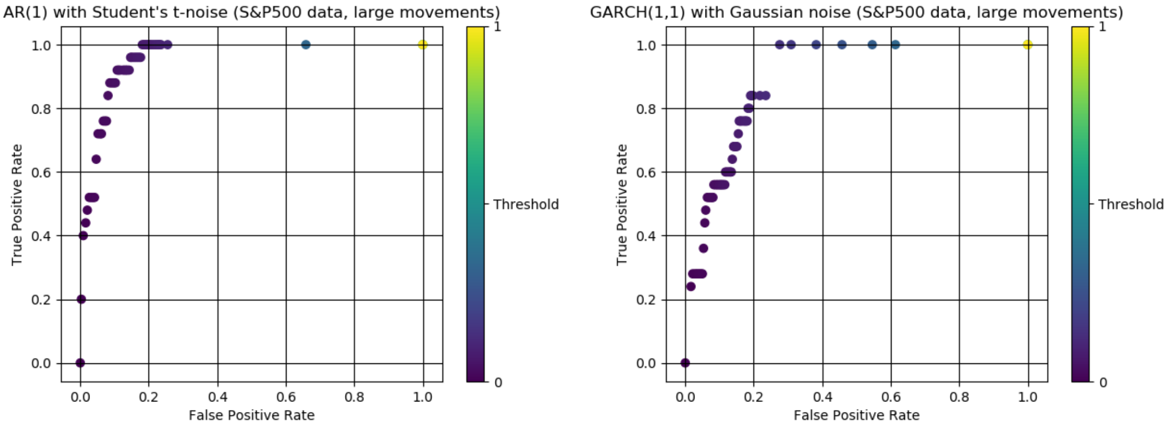
In this section yet another definition of outliers has been investigated. Here the definition of an outlier is not as previously, a false data point which is extreme, but rather “extreme” movements in the index price between two consequent business days. All events marked as “outliers” had a price change of either $> +5\%$ or $- < 5\%$. The percentage change in price where computed as follows

$$R_t = 100\frac{P_t - P_{t-1}}{P_{t-1}},$$

where P_t is the price of the index at day t.

S&P 500 Index

This investigation was based on the S&P 500 index between 2000-01-03 to 2018-04-09, a total of 4595 data points, here 25 movements where found to be larger than 5% or smaller than −5%. In Figure C.1a & C.1b one can see the performance for the AR(1) model with Student’s t-noise and the GARCH(1,1) model with Gaussian noise respectively. It is rather clear, by eye inspection that the AR(1) model “dominates” the GARCH(1,1) model, which is also confirmed by the ROC AUC in Table C.1. A possible explanation why this is the case when outliers are defined as “extreme” events rather than false data points is provided in the next section.



(a) AR(1) model with Student’s t-noise. **(b)** GARCH(1,1) model with Gaussian noise.

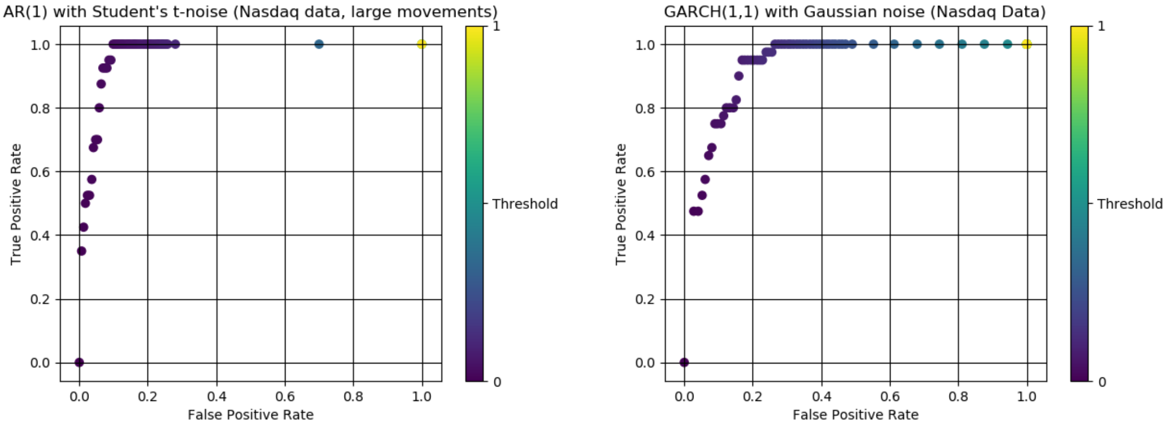
Figure C.1: Performance based on the S&P 500 data with outliers defined as “extreme events”.

Model	ROC AUC
AR(1) with Student's t-noise	0.9577
GARCH(1,1) with Gaussian noise	0.8969

Table C.1: ROC AUC for two models based on data from the S&P 500 index.

Nasdaq Composite Index

Price data for the Nasdaq Composite stock index[1] between 1971-02-05 and 2018-03-27 was used, a total of 11890 data points. A total of 40 points where marked as outliers, all the outliers were extreme movements in the price, just as in the previous section. From Figure C.2a & C.2b it is rather clear that the AR(1) model performs better than the GARCH(1,1) model and this is also confirmed by the ROC AUC in Table C.2. The author of this thesis has come up with one possible explanation why this is the case, that the AR model outperforms the GARCH model. The outliers are here defined to be "extreme" events and they often occur in a more volatile market, e.g. during the last financial crisis in 2008. A GARCH model is designed to be able to handle the volatility clustering and the heteroskedasticity of financial data, which is why the GARCH model does not label these "extreme" events as outliers. The AR model on the outlier hand is a very simple time series model and cannot handle such "extreme" events, hence the AR model label "extreme" points as outliers.



(a) AR(1) model with Student's t-noise.

(b) GARCH(1,1) model with Gaussian noise.

Figure C.2: Performance based on the Nasdaq Composite data with outliers defined as "extreme events".

Model	ROC AUC
AR(1) with Student's t-noise	0.9694
GARCH(1,1) with Gaussian noise	0.9334

Table C.2: ROC AUC for two models based on data from the Nasdaq Composite index.

[1] The data can be fetched from Yahoo Finance.

C.2 Simulated Time Series Models

In this section one AR(1) model with Student's t-noise and one GARCH(1,1) model with Gaussian noise have been simulated. This is done in order to see how well the models detect outliers when the data truly comes from the time series model used. Additionally the models are also run on data that was not generated by the same model. Firstly the AR(1) is simulated as follows

$$R_t = \phi_1 R_{t-1} + \epsilon_t, \quad \epsilon_t \sim t_5,$$

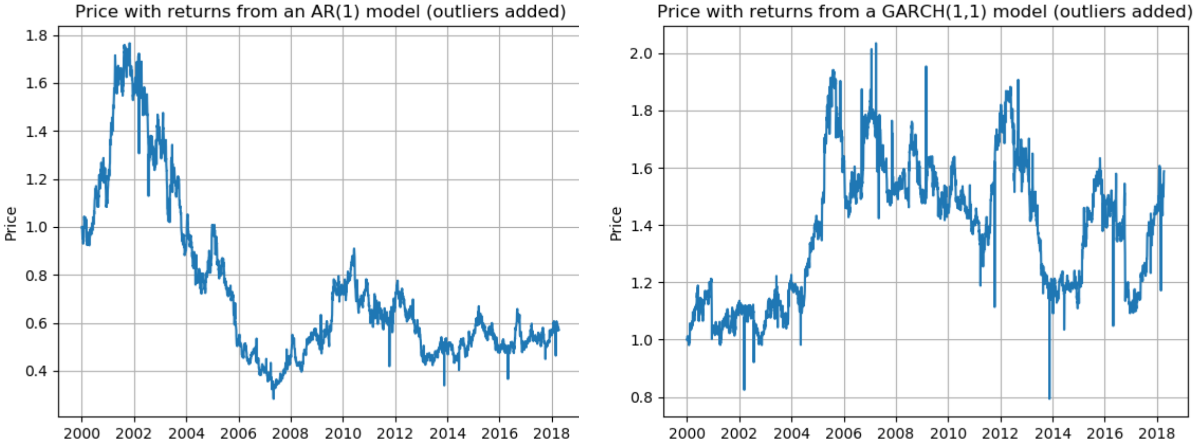
where the parameter ϕ_1 has been chosen to 0.25. This is done such that the time series contain 4595 samples (same as the S&P 500 training set). Furthermore $\{R_t\}_{i=1}^n$ is considered to be the percentage logarithmic return of the price (see equation (2.3)). Then the price, $\{P_t\}_{i=1}^n$, is computed as follows

$$P_t = P_{t-1} \exp\left\{\frac{R_t}{100}\right\}, \quad t = 1, \dots, n, \;\; P_0 = 1. \tag{C.1}$$

Similarly with the GARCH(1,1) model but with the parameters as follows

$$R_t = \sqrt{h_t} e_t, \quad e_t \sim \mathcal{N}(0,1),$$
$$h_t = 0.2 + 0.5R_{t-1}^2 + 0.3h_{t-1}.$$

Where the price then is computed according to equation (C.1). Outliers are then added to the price, P_t, according to the description in section 5.4.3, with the parameters: $\alpha_1 = 0.05$, $\alpha_2 = 0.5$, $\beta_1 = 0.5$ and $\beta_2 = 0.95$. In Figure C.3 the two price series with outliers added are shown. The outliers looks to be of smaller magnitude in Figure C.3a than in Figure C.3b but they are generated in the same way. The reason for this is that the price series in the model with AR(1) returns is below one which makes the magnitude of the outliers smaller.

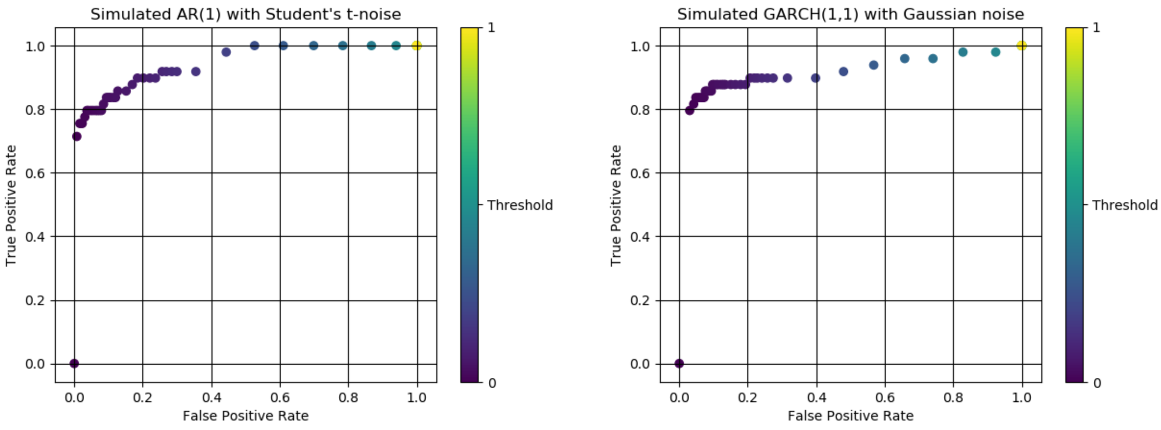


(a) Returns simulated from an AR(1) model with Student's t-noise.

(b) Return simulated from a GARCH(1,1) model with Gaussian noise.

Figure C.3: Price series based on simulated time series models with outliers added.

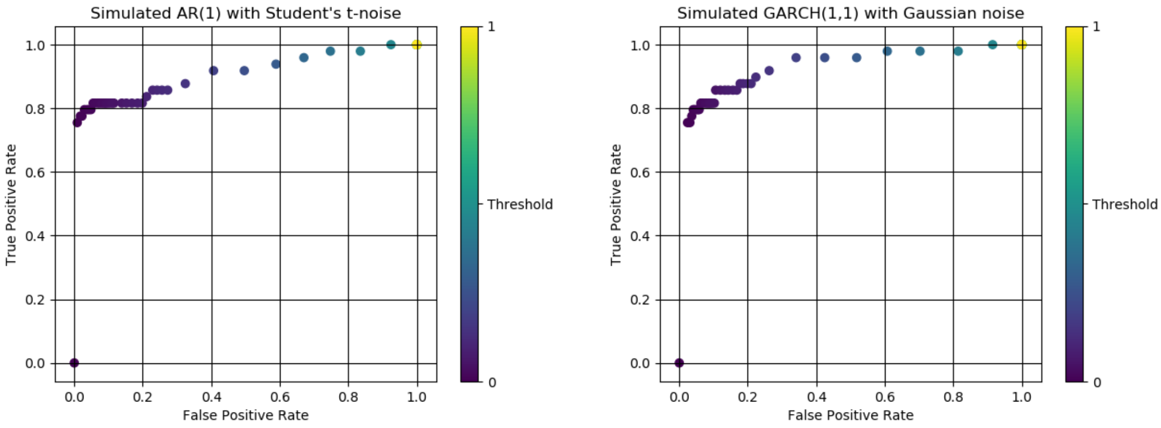
Two different performance tests have been done. Firstly, in Figure C.4a & C.4b the models were run on the data that they were simulated from, i.e. the AR model were run on data that was generated from an AR model, similarly with the GARCH model. Secondly, in Figure C.5a & C.5b there is the opposite. In the former figure the underlying data is from Figure C.3b and in the latter figure the underlying data is from Figure C.3a. The ROC AUC for these four runs can be seen in Table C.3. Here one can see that the AR model performs close to 4% better (in terms of ROC AUC) on the AR data compared to on the GARCH data. For the GARCH model the performance is the opposite, it performs 2% worse (in terms of ROC AUC) on the GARCH data compared to the AR data. One should remember that these are small differences and in order to make any proper inferences the investigation should be more thorough.



(a) AR(1) model with Student's t-noise.

(b) GARCH(1,1) model with Gaussian noise.

Figure C.4: Returns simulated from the same model that is used for outlier detection.



(a) AR(1) model with Student's t-noise with GARCH(1,1) as underlying returns.

(b) GARCH(1,1) model with Gaussian noise with AR(1) as underlying returns.

Figure C.5: Returns simulated from another model compared to the outlier detection model used.

Model	**ROC AUC**
AR(1) with t-noise (underlying data: AR)	0.9433
AR(1) with t-noise (underlying data: GARCH)	0.9088
GARCH(1,1) with Gaussian noise (underlying data: GARCH)	0.9109
GARCH(1,1) with Gaussian noise (underlying data: AR)	0.9296

Table C.3: ROC AUC for the four simulated runs.

TRITA -SCI-GRU 2018:071

www.kth.se