Stockholms universitet

1-step-ahead forecast evaluation of conditional volatility models applied to Brent Oil log returns

Jesper Cedergren

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2016:19
Matematisk statistik
Juni 2016

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Mathematical Statistics
Stockholm University
Bachelor Thesis **2016:19**
http://www.math.su.se

1-step-ahead forecast evaluation of conditional volatility models applied to Brent Oil log returns

Jesper Cedergren*

June 2016

Abstract

The optimal scenario for perhaps every actor in the financial world would be to be able to control risk. Knowing what scenarios and events that would take place in the future would probably due aswell. The best an analyst in the financial market could do is to model and forecast the risk. This translates to model and forecast volatility. One particular asset in the financial market that is regarded to have a great economical impact and has fluctuated a lot during the last ten years is the oil price. In this paper the GARCH(1,1) and IGARCH(1,1) models' ability to forecast the volatility of the Brent Oil one day ahead is evaluated. The forecasting performance is first evaluated with regards to unconditional coverage. It is concluded that the GARCH(1,1) model using a Student-t distribution is the only model that on average forecasts adequately. It is then evaluated whether this model's symmetric prediction interval is satisfactory, in particular for the extreme observations. Finally, the entire distribution for the 1-step-ahead forecast is examined. The result is that the GARCH(1,1) model using a Student-t distribution performs well in every aspect considered when producing 1-step-ahead forecasts.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: jesper.cedergren@hotmail.com. Supervisor: Joanna Tyrcha, Mathias Lindholm and Filip Lindskog.

Preface and acknowledgement.

This is a Bachelor thesis to an extent of 15 ECTS, which will lead to a Bachelor's Degree in Mathematical Statistics at the Department of Mathematics at Stockholm University.

I sincerely would like to thank my supervisors Joanna Tyrcha, Mathias Lindholm and Filip Lindskog for their encouraging help and suggestions. Their inspiring discussions have motivated me along my work on this paper. I would also like to thank my fellow students for discussing and elaborating ideas and questions.

Contents

1 Introduction

In the financial world, the concept of volatility and ways of modelling volatility are highly stressed. Here volatility means the conditional standard deviation of the underlying asset. Volatility has many applications, for instance it can be used in options trading and risk management.

A common way to price options is to use the famous Black-Scholes model. In this model the volatility is assumed to be constant which is a simplification of reality since in financial time series "large changes tend to be followed by large changes-of either sign-and small changes tend to be followed by small changes" (Mandelbrot, 1963, p.418), a phenomenon referred to as volatility clustering. In other words, the volatility evolves over time and violates the assumption of homoscedasticity. However, this evolution can be modelled by using models that allow for heteroscedasticity, referred to as conditional heteroscedastic models.

According to Tsay (2010, p.113) conditional heteroscedastic models can be classified into two different categories, those which describe the evolution of the volatility by an exact function and those which use a stochastic equation. This paper will only consider models from the first category.

The fact that the volatility is not directly observable makes the evaluation of the forecasting performance a challenge. A lot of research has been done to try to create and evaluate different proxies for the volatility. It has been suggested that an estimate of the daily volatility can be obtained by using high-frequency data, such as 10-minute returns. Other proxies could be deduced from option prices by assuming that the prices are governed by an econometric model such as the Black-Scholes formula. These obtained proxies are then refereed to as implied volatility, implied by the price and the econometric model. This paper use a different approach outlined in section 4.2.

This thesis will in Section 2 shortly present previous studies including their results and models used where the volatility have been forecasted based on daily Brent Oil spot prices. Section 3 provides the theoretical framework needed. Section 4 will shortly describe how a time series should be dealt with when modelling and the evaluation procedures used in this paper will also be described. The initial data analysis of the log return series is done in Section 5, appropriate models are fitted in Section 6 and the results from the different evaluation methods are presented in Section 7. In Section 8 the results and the weaknesses of the tests will be discussed and the conclusions will be summarized in Section 9. Finally, Section 10 will suggest interesting ways to extend the paper with other tests and methods.

1.1 Background

The challenge with heteroscedasticity will always be regarded as needle in the eye of a statistician. However for statistcians and analysts modelling volatility the struggle was clearly alleviated when the American economist Robert F. Engle in 1982 introduced the autoregressive conditional heteroscedastic model, the ARCH model. As the name implies it uses past values in order to predict the future volatility. Shortly, the variance for the error term is assumed to be described as a linear function of its past squared values. In 2003 Engle together with Granger received the Swedish Riksbank's prize in Economics in Memory of Alfred Nobel for their methods modelling economical time series.

Four years later, in 1986 Tim Bollerslev extended the ARCH model to a more general form by letting the variance term not only be a linear function of its past squared values but also of its past conditional variances, giving birth to the generalized autoregressive conditional heteroscedastic model, the GARCH model.

A special case of the GARCH model is when the autoregressive polynomial in the GARCH process has a unit root, meaning that the coefficients in the GARCH model sum to one. The model is then referred to as an integrated GARCH model, IGARCH for short.

A weakness when modelling financial data with these models are that they respond equally to positive and negative shocks as seen in (3), (4) and (5). It is known that financial data in practice does not, see Tsay (2010, p.119). In particular, to allow for asymmetric effects between volatility and returns Nelson introduced the exponential GARCH, EGARCH.

There are many other extensions of the GARCH model which has different implications and capture different anomalies but the only ones discussed in this thesis is ARCH, GARCH, IGARCH and EGARCH.

1.2 Aim and purpose

The aim with this thesis is to fit appropriate time series models by inspecting the log return series of the Brent Oil price from January 4 2000 to January 2016 and evaluate the models' ability to forecast the volatility 1-step-ahead. The evaluation methods used are the back testing procedure described in section 3.3.8 and density forecast evaluation described in section 3.3.11.

2 Previous work

In order to get a sense of how the subject of this paper can be approached this section gives a general overview of what techniques that can be used and presents what has been done on the subject earlier including the results and different models used. All different techniques and models mentioned in this section that can and have been used are not further explained, they are presented to get a quick overview and hopefully to engage the interested reader to read up on and apply for him or herself.

According to Behmiri and Pires Manso (2013) the methods used to forecast the volatility based on the crude oil price can be either quantitative or qualitative, where the quantitative methods can be divided into econometric methods and non standard methods. This paper is solely using time series models which is an econometric method. Other econometric methods used are financial models and structural models. The non standard or computational approaches used are Artificial Neural Networks and Support Vector Machines.

Moreover, the most frequently used techniques in descending order are time series econometrics, financial methods, structural models and non standard computational models and the least used is qualitative knowledge based methods (Behmiri and Pires Manso, 2013).

Turning to the econometric time series models which are used in this paper, when modelling stocks and different index returns using daily data the volatility clustering property of the time series usually suggests using autoregressive conditional heteroscedasticity models. Just as stocks and index returns, the oil returns are experience volatility clustering and different types of autoregressive conditional heteroscedasticity models seems to be common practice. Considering only the previous work where the crude Brent Oil spot price have been used there are three studies of interest.

Cheong (2009, cited in Behmiri and Pires Manso, 2013, p.32) uses daily spot prices for the period from 4th January 1993 to 31st december 2008 to estimate out-of-sample forecasts for horizons of 5, 10, 20 and 100 days. The models compared are GARCH, asymmetric power ARCH (APARCH), fractionally integrated GARCH (FIGARCH) and fractionally integrated asymmetric power ARCH (FIAPARCH), all using normal and Student-t distribution. The GARCH models using normal and Student-t distributions are best for the 5 and 20 day forecasting horizons while the APARCH model is best for the longer forecasting horizons 60 and 100 days.

Kang, Kang and Yoon (2009, cited in Behmiri and Pires Manso, 2013, p.33) uses daily spot prices for the period from 6th January 1992 to 31st December 2009 to perform out-of-sample forecasting analysis on 1,5 and 20 days forecasting horizons. The models compared are GARCH, component GARCH (CGARCH), IGARCH and FIGARCH. The FIGARCH model outperforms the other models on all forecasting horizons.

Wei, Wang, and Huang (2010, cited in Behmiri and Pires Manso, 2013, p.33) use daily spot prices for the period 9th January 1992 to 31st December 2009. The work is an extension from the work by Kang, Kang and Yoon (2009) and the out-of-sample forecasting analysis is also performed on 1,5 and 20 days forecasting horizons. The models included are RiskMetrics, GARCH, IGARCH, Glosten-Jagannathan-Runkle GARCH (GJR-GARCH), exponential GARCH (EGARCH), APARCH, FIGARCH, fractionally assymetry power ARCH (FIAPARCH) and hyperbolic GARCH (HYGARCH). With regards to six loss functions there is no evidence that the model performance differently. However the linear GARCH models perform better when forecasting on 1 day horizons and the nonlinear performs better for 5 and 20 day forecasting horizons.

These studies have been done for roughly the same periods and models for both symmetric and asymmetric effects have been used. Overall linear GARCH models not allowing for asymmetric effects tend to perform best when estimating forecasts on short horizons as is the aim of this paper.

3 Theory

In this section the theoretical framework for the succeeding sections are presented, including theory for the specific models and tests.

3.1 Return

The theory in this subsection is from Ruppert (2004, section 3.3).

When modelling a financial time series the log return series rather than the price series is usually modelled. Let the simple return be defined as

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

A common model is to assume that the simple returns are mutually independent, identically normally distributed. However this model has two problems. Firstly, the returns are greater than -1 since it's not possible to lose more money than invested. This is a problem since a normally distributed random variables can take any value on the interval $[-\infty; \infty]$. Secondly, the multiperiod gross return $1+R_t(k) = (1+R_t)(1+R_{t-1})\cdots(1+R_{t-k+1})1$ is a product of normally distributed random variables, which is not normally distributed.

In order to get around this problems the simple log return is defined as

$$r_t = ln(1 + R_t) = ln\left(\frac{P_t}{P_{t-1}}\right) \tag{1}$$

where $1+R_t$ is the simple gross return. The simple log return has appealing properties. Firstly, just as a normally distributed random variable it can take any value value on the interval $[-\infty; \infty]$. Secondly, the multiperiod log return $ln(1 + R_t(k)) = ln((1 + R_t)(1 + R_{t-1}) \cdots (1 + R_{t-k+1})) = ln(1 + R_t) + ln(1 + R_{t-1}) + ... + ln(1 + R_{t-k+1}) = r_t + r_{t-1} + ... + r_{t-k+1}$ is a sum of normal random variables which is normally distributed. Another accompanying property of the log return is that it is symmetric around zero if the mean of the return is zero.

3.2 Time series

The theory in the preceding sections regarding time series and tests are found in Tsay (2005, chapter 1-3).

The process of log returns can be defined as

$$r_t = \mu_t + a_t \tag{2}$$

where μ_t is the expected value of the process $\{r_t\}$ and $\{a_t\}$ is a sequence of independent and identically distributed random variables with mean zero and conditional variance σ_t^2. The sequence of $\{a_t\}$ is referred to as *shocks* or *innovations* at time t. The equations that describe μ_t and σ_t^2 separately are referred to as the mean and variance equations respectively.

A time series is said to be weakly stationary if two conditions are met. Firstly, the mean of the return series $\{r_t\}$ should be constant, $E(r_t) = \mu$. Secondly, the covariance between different values, the autocovariance, should only depend on the lag length l, $Cov(r_t, r_{t-l}) = \gamma_l$. The basic idea behind stationarity is that the series is time independent.

If $\{r_t\}$ is a sequence of independent and identically distributed random variables with finite mean and variance then the time series is called white noise. In particular, if the $\{r_t\}$ is normally distributed with mean 0 and variance σ_t^2 it is called Gaussian white noise.

3.2.1 AR

The autoregressive model is used when μ_t can be expressed as a linear function of the time series previous values. This linear function is often referred to as the mean equation. An AR model is useful for instance when there is a trend in the series. An AR(p) model has the form

$$\mu_t = \phi_0 + \sum_{l=1}^{p} \phi_l r_{t-l} + a_t,$$

where $\left| \sum_{l=1}^{p} \phi_l \right| < 1$.

3.2.2 ARCH

The idea of ARCH models are that the the shock a_t is serially uncorrelated but dependent where this dependence can be described by a quadratic function of its lagged values as in (3).

$$a_t = \sigma_t \epsilon_t \quad \sigma_t^2 = \alpha_0 + \sum_{i=1}^{m} \alpha_i a_{t-l}^2 \tag{3}$$

where $\{\epsilon_t\}$ is a sequence of random variables with mean 0 and variance 1, $\alpha_0 \geq 0$, $\alpha_i \geq 0$.

As seen from (3) large values of the squared previous shocks $\{a_{t-l}^2\}_{i=1}^{m}$ implies that a large conditional variance σ_t^2 for the shock a_t is probable. In other words, a large shock a_{t-1}^2 tend to be followed by another large shock a_t^2 generating the behavior of volatility clustering.

The ARCH model responds slowly to large isolated shocks and tend to overpredict the volatility. It also responds equally to both to positive and negative shocks.

3.2.3 GARCH

The ARCH model often requires many lags to capture the behaviour of the volatility. However Bollerslev (1986, cited in Tsay, p.131) proposed an extension of the ARCH process, the generalized ARCH (GARCH).

The idea is the same as for an ARCH model and in addition the conditional variance σ_t^2 also depends on its previous values. Thus the a GARCH(m, s) model can be written as

$$a_t = \sigma_t \epsilon_t, \quad \sigma_t^2 = \alpha_0 + \sum_{i=1}^{m} \alpha_i a_{t-i}^2 + \sum_{j=1}^{s} \beta_j \sigma_{t-j}^2, \tag{4}$$

where $\{\epsilon_t\}$ again is a sequence of standardized random variables, $\alpha_0 \geq 0$, $\alpha_i \geq 0$, $\beta_j \geq 0$ and $\sum_{i=1}^{max(m,s)} (\alpha_i + \beta_i) < 1$ in order for the process to be stationary (Li, 2007, p.4). Following from the last condition that the sum of all parameters is less than one, a GARCH process is a mean reverting process.

In particular, if $m = s = 1$ which correspsond to a GARCH(1,1) process the variance equation in (4) is given by

$$\sigma_t^2 = \alpha_0 + \alpha_1 a_{t-1}^2 + \beta_1 \sigma_{t-1}^2.$$

The maximum likelihood estimation of the parameters in a GARCH(p, q) model and in particular in a GARCH(1,1) model is described in Appendix B.1.

3.2.4 IGARCH

If the process describing the variance equation in (4) is a unit root process, meaning that $\displaystyle\sum_{i=1}^{m} \alpha_i + \sum_{i=j}^{s} \beta_j = 1$, then it is referred to as an integrated GARCH process, IGARCH for short. In an IGARCH process the shocks are said to be persistent, meaning that the variance process is mean reverting slowly. An IGARCH(m, s) can be written as

$$a_t = \sigma_t \epsilon_t \quad \sigma_t^2 = \alpha_0 + \sum_{i=1}^{m=s} (1 - \beta_i) a_{t-i}^2 + \beta_i \sigma_{t-i}^2, \tag{5}$$

where $\{\epsilon\}$ is defined as earlier and $0 < \beta_i < 0$ as. Specifically, if $m = s = 1$ in (5) is given by

$$\sigma_t^2 = \alpha_0 + (1 - \beta_1) a_{t-1}^2 + \beta_1 \sigma_{t-1}^2,$$

3.2.5 Forecasting GARCH and IGARCH

The conditional variance of the GARCH model in (4) at time $h+1$ is given by

$$\sigma_{h+1}^2 = \alpha_0 + \sum_{i=1}^{m} \alpha_i a_{h+1-i}^2 + \sum_{j=1}^{s} \beta_j \sigma_{h+1-j}^2,$$

where at time h all the past values a_{h+1-i}^2 and σ_{h+1-j}^2 are known for all $i \leq m$ and $j \leq s$. Thus the 1-step ahead forecast from forecast origin h $\sigma_h^2(1)$ equals σ_{h+1}^2 and is given by

$$\sigma_h^2(1) = \alpha_0 + \sum_{i=1}^{m} \alpha_i a_{h+1-i}^2 + \sum_{j=1}^{s} \beta_j \sigma_{h+1-j}^2.$$

In particular for a GARCH(1,1) process with $\alpha_0 = 0$, the 1-step-ahead forecast from forecast origin h is

$$\sigma_h^2(1) = \alpha_1 a_h^2 + \beta_1 \sigma_h^2. \tag{6}$$

Specifically, for an IGARCH(1,1) process with $\alpha_0 = 0$, the 1-step-ahead forecast from forecast origin h is

$$\sigma_h^2(1) = \sigma_h^2 \tag{7}$$

3.3 Statistics and tests

3.3.1 T-test

A t-test is used to test whether a value of a parameter, usually the mean, is significantly different from a prespecified value under the null hypothesis on a given level of significance. It is assumed that the sample values used to estimate the mean is independent and normally distributed. The null hypothesis $H_0 : \mu = 0$ is tested against the alternative $H_1 : \mu \neq 0$. The test statistic is

$$T = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim t_{n-1} \tag{8}$$

where $\bar{x}$ and s is the sample mean and standard deviation respectively. The test statistic is t-distributed with $n-1$ degrees of freedom.

3.3.2 Autocorrelation, ACF

Autocorrelation is a generalization of the concept of correlation, it quantifies the linear dependence between r_t and its past values of lag l, r_{t-l}. Under the assumption of weak stationarity, the autocorrelation is a function of the lag length l only. In line with the definiton of the correlation coefficient of two random variables, the autocorrelation coefficient is defined accordingly

$$\rho_l = \frac{Cov(r_t, r_{t-l})}{\sqrt{Var(r_t)Var(r_{t-l})}} = \frac{Cov(r_t, r_{t-l})}{Var(r_t)} = \frac{\gamma_l}{\gamma_0}, \tag{9}$$

where $Var(r_{t-1}) = Var(r_t)$ since the series $\{r_t\}$ is weakly stationary. It should be noted that $\rho_0 = 1, \rho_l = \rho_{-l}$ and $-1 \leq \rho_l \leq 1$ by definition.

To test the null hypothesis $H_0 : \rho_l = 0$ versus the alternative $H_a : \rho_l \neq 0$ it can be used that if $\{r_t\}$ is independent and identically distributed then $\hat{\rho}_l \overset{asy}{\sim} N(0, \frac{1}{n})$ (Brockwell and Davis, 1991 cited in Tsay, 2005, p.31). The test statistic is

$$\mathrm{T} = \frac{\hat{\rho}_l}{\sqrt{1 + \frac{2\sum_{i=1}^{l-1} \hat{\rho}_i^2}{n}}} \sim t_{n-1}, \tag{10}$$

where n is the number of observations in the time series and $\hat{\rho}_l$ is the sample autocorrelation of lag length l.

3.3.3 Partial autocorrelation function, PACF

The partial autocorrelaton function is a function of its ACF. It describes the partial correlation of the time series with its own lagged values, controlling for the values of the time series at all shorter lags. It is useful to determine the lag order p of an AR(p) process. Considering the AR models:

$$\begin{aligned}
r_t &= \phi_{0,1} + \phi_{1,1} r_{t-1} + e_{1t} \\
r_t &= \phi_{0,2} + \phi_{1,2} r_{t-1} + \phi_{2,2} r_{t-1} + e_{2t} \\
&\vdots \\
r_t &= \phi_{0,k} + \phi_{1,k} r_{t-1} + \phi_{2,k} r_{t-2} + ... + \phi_{k,k} r_{t-k} + e_{kt}
\end{aligned}$$

where $\phi_{0,j}$, $\phi_{i,k}$ and $\{e_{jt}\}$ are respectively, the constant term, the coefficient of r_{t-i}, and the error term of an AR(j) model. These models are in the form of a multiple linear regression hence the coefficients can be estimated using the OLS method. The estimates $\hat{\phi}_{1,1}$, $\hat{\phi}_{2,2}$ and $\hat{\phi}_{k,k}$ of the respective equations are called the lag-1, lag-2, and lag-k sample PACF of r_t respectively. Thus the complete sample PACF describes the time series' serial correlation with its previous values of a specific lag controlling for the values of the time series at all shorter lags. The lag order is chosen where the PACF cuts off, where the lag-p sample is nonzero and where $\hat{\phi}_{j,j}$ is close to zero for all $j > p$.

3.3.4 Box Ljung

To jointly test whether several autocorrelations of r_t are zero, the null hypothesis $H_0 = \rho_1 = \rho_2 = ... = \rho_m = 0$ can be tested against the alternative that $\rho_l \neq 0$ for at least one l. The t Portmanteau statistic modified by Ljung and Box (1978, cited in Tsay, p.32) can be used,

$$Q(m) = n(n+2)\sum_{l=1}^{m}\frac{\hat{\rho}_l^2}{n-l} \sim \chi 2(m), \tag{11}$$

where m is the number of lags and n again is the number of observations in the time series.

3.3.5 Skewness

The third central moment measure the symmetry of a random variable X about its mean. The sample skewness $\hat{S}(x)$ is calculated according to

$$\hat{S}(x) = \frac{1}{(n-1)\sigma_x^3}\sum_{i=1}^{n}(x_i - \mu_x)^3. \tag{12}$$

A positive sample skewness is indicating that the tail on the right side of the probability density function is longer or fatter than the corresponding left and vice versa.

3.3.6 Kurtosis

The fourth central moment measure the tail behaviour of a random variable X. The sample kurtosis $\hat{K}(x)$ is calculated according to

$$\hat{K}(x) = \frac{1}{(n-1)\sigma_x^4}\sum_{i=1}^{n}(x_i - \mu_x)^4. \tag{13}$$

A sample value of the kurtosis greater than three is indicating that the probability density function is leptokurtic compared with a density of a normal distribution. A leptokurtic density function is characterized by a high, thin peak around its mean and heavy tails.

3.3.7 Kolmogorov Smirnov

The theory in this section is based on the theory in Bagdonavičius, Julius and Nikulin (2011, chapter 3).

It can be tested whether a sample from an unknown distribution F is equal to a particular distribution F_0 by formally testing the hypothesis $H_0 : F = F_0$ against the alternative $H_a : F \neq F_0$. This can be done with a Kolmogorov Smirnov test.

Let the empirical cumulative distribution function F_n for independent and identically distributed observations x_i from a sample of size n be defined as

$$F_n(x) = P(X \leq x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x),$$

where $I(X_i \leq x)$ is an indicator function taking value 1 if a given sample point is below or equal to x and 0 otherwise.

The law of large numbers implies that

$$\frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x) \to \mathrm{E}\left[I(X \leq x)\right] = P(X \leq x) \quad \text{as} \quad n \to \infty,$$

thus

$$F_n(x) \to F(x) \quad \text{as} \quad n \to \infty.$$

This means that the sample distribution will converge to the underlying distribution F as the sample size increases and will thus not approximate or depend on the reference distribution F_0. It follows that the distribution of D_n under H_0 can be tabulated for each n with no regards to the reference distribution. The test statistic used is the following

$$D_n = \sqrt{n} \sup_{x \in \mathbb{R}} |F_n(x) - F_0(x)|,$$

where n is the sample size and $\sup_{x \in \mathbb{R}} |F_n(x) - F_0(x)|$ is the largest distance from the sample c.d.f and the reference c.d.f for some observation x_i. The null hypothesis is rejected when the test statistic D_n exceeds a tabulated threshold value c_α for a given level of significance α, otherwise it is not rejected.

3.3.8 Back testing

The theory in this section is based on Christoffersen (1998).

In the back test, a given model is fitted for rolling windows over the entire log return series. From every window 1-step-ahead forecasts of the volatility is predicted with a belonging double sided 95 % prediction interval $PI_{0.95}$. When rolling the windows of a given length from start to end of the log return series it is of main interest to test whether the actual degree of coverage equals the degree of coverage implied by the prediction interval $PI_{0.95}$. The procedure of using rolling windows is further explained in section 4.2.

Let I_t be an indicator variable defined according to

$$I_t = \begin{cases} 1 & \text{if } r_t \in PI_{0.95} \\ 0 & \text{if } r_t \notin PI_{0.95}. \end{cases}$$

Since the GARCH model is conditioned on the past values, the successive 1-step-ahead forecasts are independent. So whether the observed log returns are inside its prediction interval can be seen as independent Bernoulli trials. The test can be formalized with the hypothesis

$$H_0 : p = 0.95$$
$$H_a : p \neq 0.95,$$

where under the null hypothesis $I_t \sim Be(n, p)$ and $\sum_{t=1}^{n} I_t \sim Bin(n, p)$ where n is the number of observations.

This could be tested with a likelihood ratio test, where the likelihood under the null hypothesis is

$$L(p; I_1, I_2, ..., I_T) = (1 - p)^{n_0} p^{n_1},$$

and under the alternative

$$L(\pi; I_1, I_2, ...I_T) = (1 - \pi)^{n_0} \pi^{n_1}.$$

So the likelihood ratio test is formulated according to

$$LR_{uc} = 2\log\left(L(\hat{\pi}; I_1, I_2, ..., I_n)/L(p; I_1, I_2, ..., I_n)\right) \sim \chi^2(1),$$

where the ML-estimate $\hat{\pi} = \frac{n_1}{n_0+n_1}$ and the degrees of freedom is one since there is one free parameter under the alternative hypothesis and zero under the null, thus the difference is one. The null hypothesis is rejected on 5% significance level if the LR_{uc} exceeds the critical value $\chi^2_{0.05}$, otherwise not. If the null hypothesis is rejected then the degree of coverage of the model is not what the prediction interval imply that it should be.

To obtain a confidence interval for the degree of coverage it is used that the sum of the indicator variable I_t are binomial distributed with parameters n and p. The likelihood for p is then given by

$$L(p) = \binom{n}{\sum_{i=1}^{n} x_i} p^{\sum_{i=1}^{n} x_i} (1 - p)^{n - \sum_{i=1}^{n} x_i}.$$

After some tedious steps and manipulations done in Appendix C the $(1 - \alpha)\%$ Wald confidence interval for p is given by

$$\left(\frac{\sum_{i=1}^{n} x_i}{n}\right) \pm \frac{z_{\alpha/2}}{n\sqrt{\left(\frac{1}{\sum_{i=1}^{n} x_i} + \frac{1}{n - \sum_{i=1}^{n} x_i}\right)}},$$

where $z_{\alpha/2} = \Phi^{-1}(1 - \alpha/2)$.

If $p = 0.95$ does not lie within this confidence interval then the null hypothesis that $p = 0.95$ is rejected, otherwise not.

3.3.9 Joint test of conditional coverage and independence

The test above tests the coverage on average, the unconditional coverage. However even if the model on average is right it is necessarily not the case that the degree of coverage is correct for smaller subperiods. Daily financial data is clustered and heteroscedastic volatility models try to account for that fact. So a model choice should not be regarded as successful if the outliers after trying to account for this effect still would be clustered. Christoffersen (1998) suggest that this could be tested with an appropriate likelihood ratio test. More specifically he suggests that the process describing an observation of going from being inside or outside the forecast interval to the preceding observation being inside or outside could be described as a Markov chain. The Markov chain has two states since the indicator variable is binary and the transition matrix is given by

$$\Pi = \begin{pmatrix} 1 - \pi_{01} & \pi_{01} \\ 1 - \pi_{11} & \pi_{11} \end{pmatrix}, \tag{14}$$

where the first and second row (column) represents starting (ending up) in a position inside and outside the forecast interval respectively.

If I_i is independent and identically distributed for all values of i then the probability of an observation being inside the forecast interval is the same no matter if the observation one step before were inside or outside the forecast interval, i.e. $P(I_t = 1|I_{t-1} = 0) = P(I_t = 1|I_{t-1} = 1)$. The test of conditional coverage independence is formalized with the hypothesis

$$H_0 : \pi_{01} = \pi_{11}$$
$$H_a : \pi_{01} \neq \pi_{11}$$

where $\pi_{ij} = P(I_t = j|I_{t-1} = i)$.

However, since financial data is assumed to respond differently to positive and negative shocks it would also be of great interest to test whether the coverage in the left tail and the right tail jointly performs as the model specifies. Testing for this could hint whether a model that allow for asymmetric effects might be called for. The idea of transitions between the two states being inside or outside the forecast interval can be refined to three states, being smaller than the lower forecasting bound, greater than the upper forecasting bound and being inside the forecasting interval. The transition matrix (14) could be extended to incorporate these three states accordingly

$$\Pi_a = \begin{pmatrix} \pi_{ll} & 1 - \pi_{ll} - \pi_{lu} & \pi_{lu} \\ \pi_{ml} & 1 - \pi_{ml} - \pi_{mu} & \pi_{mu} \\ \pi_{ul} & 1 - \pi_{ul} - \pi_{uu} & \pi_{uu} \end{pmatrix} \tag{15}$$

where the subindexes l, m and u denotes the states lower, middle and upper respectively.

Let S be the random variable that indicates what state the process situates in, then the approximate likelihood function for this process can be expressed as

$$L(\Pi_a; S_1, S_2, ..., S_n) = \pi_{ll}^{n_{ll}}(1 - \pi_{ll} - \pi_{lm})^{n_{ml}}...(1 - \pi_{ul} - \pi_{uu})^{n_{um}}\pi_{uu}^{n_{uu}}, \tag{16}$$

where n_{ij} is the number of observations that made the one step transition from state i to state j. It is approximate because it is conditioned on the first observation. Since this observation is the starting point a transition from a state in an earlier time period did not occur. The ML-estimates deduced in Appendix B.2 are given by

$$\hat{\pi}_{ij} = \frac{n_{ij}}{n_{il} + n_{im} + n_{iu}}, \tag{17}$$

for $i = l, m, u$, $j = l, u$.

Under the the assumption the state is independent of the previous state and that the tail probabilites are equal and prespecified according to some significance level $\alpha = 1 - p$ the transition matrix is specified according to

$$\Pi_0 = \begin{pmatrix} (1-p)/2 & p & (1-p)/2 \\ (1-p)/2 & p & (1-p)/2 \\ (1-p)/2 & p & (1-p)/2 \end{pmatrix}, \tag{18}$$

which has the likelihood

$$L(\Pi_0; S_1, S_2, ..., S_n) = (1-p)^{(n_{ll}+n_{lu}+n_{ml}+n_{mu}+n_{ul}+n_{uu})}p^{(n_{lm}+n_{mm}+n_{um})}. \tag{19}$$

Testing whether the matrix (15) equals the matrix (18), can be formalized by the hypothesis'

$$H_0 : (1-p)/2 = \pi_{ll} = \pi_{lu} = \pi_{ml} = \pi_{mu} = \pi_{ul} = \pi_{uu}$$
$$H_a : \text{at least one of the equalities under the null does not hold.}$$

The null hypothesis can be tested against the alternative using the likelihood ratio

$$LR = 2\log\Big(L(\hat{\Pi}_a; S_1, S_2, ..., S_n) - L(\Pi_0; S_1, S_2, ..., S_n)\Big) \sim \chi 2(6), \qquad (20)$$

since the difference in free parameters between the alternative and the null hypothesis is six. There are six free parameters under the alternative hypothesis and zero under the null.

The null hypothesis is rejected if the observed LR-statistic is greater than the critical value $\chi 2_\alpha(6)$, otherwise not.

3.3.10 Root mean squared error, RMSE

Even if observations are inside the models' prediction intervals it does not say anything regarding how far away the predictions are from the actual observations. Even if the model has the implied degree of coverage it perhaps underestimates the volatility completely when the prediction interval fail to include the observation or perhaps it overestimates the volatility and render wider prediction intervals than necessary. To measure the actual deviation from the predicted and the observed volatility the root mean squared error can be used.

Let $\hat{Y}$ be the predicted values of corresponding observed values Y, then the mean squared error of the predictor $\hat{Y}$ is estimated according to

$$RMSE(\hat{Y}) = \sqrt{MSE(\hat{Y})} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{Y} - Y)^2}, \qquad (21)$$

where n is the number of observations.

3.3.11 Density forecast

A more general approach to consider whether a model is appropriate would be to focus on how well the model forecast not only in the upper and lower quantiles but also how well it perform over all possible percentiles. In other words how well the model predict with regards to the entire distribution, thus evaluating the complete density forecast.

Let $Y = F_X(X)$. Then the probability integral transform states that for any continuous random variable X with distribution function F_X, Y is uniformly distributed on [0,1] (Held and Sabanés Bové, 2014, p.309-310). It follows since

$$F_Y(u) = P(Y \leq u) = P\left(F_X(X) \leq u\right) = P\left(X \leq F_X^{-1}(u)\right) = F_X\left(F_X^{-1}(u)\right) = u,$$

so $Y \sim U(0,1)$ by definition.

Specifically, if X_{t+1} denotes the distribution for the log returns in a GARCH(1,1) process at time point $t+1$ then the distribution for the random variable

$$Y_{t+1} = F_{X_{t+1|t}}(X_{t+1}) = P(X_{t+1} \leq X_{t+1}|F_{X_t}) = P(\sigma_{t+1}\epsilon_{t+1} \leq X_{t+1}|\sigma_{t+1}) \tag{22}$$

$$= P\left(\epsilon_{t+1} \leq \frac{X_{t+1}}{\sigma_{t+1}}\middle|\sigma_{t+1}\right) = F_{\epsilon_{t+1}}\left(\frac{X_{t+1}}{\sigma_{t+1}}\right) \sim U(0,1), \tag{23}$$

which shows that the cumulative distribution function for the standardized residuals ϵ_{t+1} evaluated in the observed points $\frac{x_{t+1}}{\sigma_{t+1}}$ should be uniformly distributed on the interval [0,1] for all t. To formally test the null hypothesis against the alternative accordingly

H_0: Y_{t+1} is uniformly distributed on [0,1].
H_a: Y_{t+1} is not uniformly distributed on [0,1],

the Kolmogorov-Smirnov test described in section 3.3.7 can be used.

4 Methodology

The data used is the Brent Oil spot price from January 4 2000 to January 22 2016 from which the log returns are calculated according to (1). This log return series is the time series which is subject to investigation.

4.1 Dealing with time series

A time series is usually not white noise, it is usually serially dependent meaning that the future values to some extent depend on the past values of different lags. In order to be able to forecast the future values of a time series behaving in a certain way this serial structure need to be identified and captured to some extent. It can be captured using more or less sophisticated mathematical models. However, these models are as all models a simplification of reality and should not be seen as the true model of the underlying time series, i.e. the price of a stock is not programmed to behave as a certain model. Every model is a simplification of reality so there exist no such model as a true model. The goal when dealing with time series is narrowed to successively navigate stepwise towards finding the best model at hand. In order to identify the serial structure it is suggested that the autocorrelation function is used. The autocorrelation is a generalisation of ordinary correlation which is nothing but a measure of linear correlation between the observations and its past values. The sample autocorrelations for all lags are computed according to (9) and plotted in correlograms including their 95 % confidence limits. Based on this the structure of the log return series is identified and appropriate models are fitted.

4.2 Forecast evaluation

In order to test how well the model predicts volatility it would be of interest to calculate the deviance from actual volatility, $\sigma^2_{forecast} - \sigma^2_{actual}$. However, the actual volatility cannot be observed. What actually can be observed is the log returns r_t which can be used as a proxy for σ_{actual}. Let F_{t-1} denote all available information up to and including time period $t-1$. Then

$$E(r_t|F_{t-1}) = E(\mu + \sigma_t\epsilon_t|\sigma_t) = \mu + \sigma_t E(\epsilon_t|\sigma_t) = \mu + \sigma_t E(\epsilon) = \mu$$

$$Var(r_t|F_{t-1}) = Var(\mu + \sigma_t\epsilon_t|\sigma_t) = \sigma_t^2 Var(\epsilon_t|\sigma_t) = \sigma_t^2 Var(\epsilon) = \sigma_t^2,$$

so the observations r_t should with $(1-\alpha)\%$ confidence be contained within the 1-step-ahead prediction intervals constructed according to

$$PI_t|F_{t-1} = \begin{cases} \mu \pm z_{\alpha/2}\sigma_{t-1}(1) & \text{using a normal distribution} \\ \mu \pm t_{\alpha/2}(df)\sigma_{t-1}(1) & \text{using a Student-t distribution,} \end{cases} \tag{24}$$

where $\sigma_{t-1}(1)$ is calculated according to (6) and (7) for GARCH(1,1) and IGARCH(1,1) models respectively.

In order to assess how well a given model perform the back testing procedure described in section 3.3.8 is used. The models are fitted on one hand with the assumption that the random variable ϵ in (4) and (5) are normally distributed and on the other assuming it is Student-t distributed. In sections 6.1 and 6.2 it is motivated that the models GARCH(1,1) and IGARCH(1,1) seem to resemble the underlying process of the log return series. Hence the models subject to 1-step-ahead forecast evaluation with accompanying distributions are:

- GARCH(1,1) using a normal distribution
- GARCH(1,1) using a Student-t distribution
- IGARCH(1,1) using a Normal distribution
- IGARCH(1,1) using a Student-t distribution.

For a given model rolling windows are used according to the following procedure. For instance when using a rolling window of length one year a model is fitted to the set of Brent Oil log returns using the observations for the first year in the time series. For this particular window the parameters in the model are estimated and a 1-step-ahead forecast $\sigma_h(1)$ is calculated according to (6) and (7) for the GARCH(1,1) and IGARCH(1,1) models respectively with the including 95% prediction intervals (24). Then the window is moved one step forward, corresponding to one day when using a daily time series. The new window leaves out the first observation and includes the first observation in the next year. So the window still includes log returns for a period of one year but is moved one day ahead. Again the model is fitted to the Brent Oil log returns that the window includes and a 1-step-ahead forecast is made and accompanied by a prediction interval. This procedure is repeated throughout the entire log return series until the window includes the most recent observation. So the window is rolling through the entire time series.
In order to get universal results the models are estimated using rolling windows of eleven different lengths, from six months, one year, two years and growing windows of one year up to and including ten years. Then for every combination of candidate in the list above and rolling window length the unconditonal coverage is tested as outlined in section 3.3.8. For appropriate models, based on the test of unconditional coverage, conditional coverage

and independence is jointly tested as described in section 3.3.9. Finally the density forecast is evaluated according to the procedure in section 3.3.11.

5 Data

5.1 Data analysis

In this section the price and log return series is plotted in Figure 1 and Figure 2 to detect specific patterns or anomalies such as volatility clustering, structural breaks or seasonal patterns. Further, some sample statistics are estimated and summarized in Table 1 accompanied with QQ-plots and histograms in Figure 3 and Figure 4 to get a sense of how the log return series are distributed. This could serve as a hint which distribution that would be the most appropriate when specify the likelihood function in order to estimate the parameters in the models.

To get a sense of how the time series is roughly behaving, the some initial plots and descriptive statistics can be inspected. As seen in Figure 1 the oil price seems to be hovering around some price level for specific time periods. That seems to be the case for the time period 2000-2005, then the price increases to a new level which it is lingering around for about one year. Following some volatile years including a big increase followed by an even larger decrease during 2008, from 2011 the price again seems to linger around some new level. The worst case scenario from a model selection perspective is that the time series would behave homogeneously within these different periods and at the same time heterogeneously with respect to these different periods. Then the different periods would call for different models.

Figure 1: Daily Brent Oil spot prices from January 4 2000 to January 22 2016

The spot price of the Brent Oil is not the time series that should be modelled rather it is the log return. Plotting the log returns and the absolute log returns could provide some useful insights, e.g. financial data tend to exhibit distinct changes in volatility over time. The log return and in particular the absolute log return series in Figure 2 is indicating that the magnitude of the volatility seem to decline over time. It is even clearer that some periods of large absolute log returns, e.g. after 2000, around 2005 and as expected also during the financial crisis 2008 are followed by periods of tranquility. This is an indication of volatility clustering. However the observed volatility clustering is more subtle than in most financial data where it is often very distinct.

Another observation is that the volatility appears to be smaller after the volatile period following the financial crisis than during the period 2005-2008, a period which itself seem to be somewhat different from the period 2000-2005. This arouses suspicion that there perhaps are structural breaks in volatility. The volatility clustering property and the fact that the variance is not constant suggest that autoregressive conditional heteroscedastic models would probably be good at mimicing the behaviour in the log return series.

In addition, the log return series does not seem to have any regularly repeating patterns so there is no signs of seasonality.

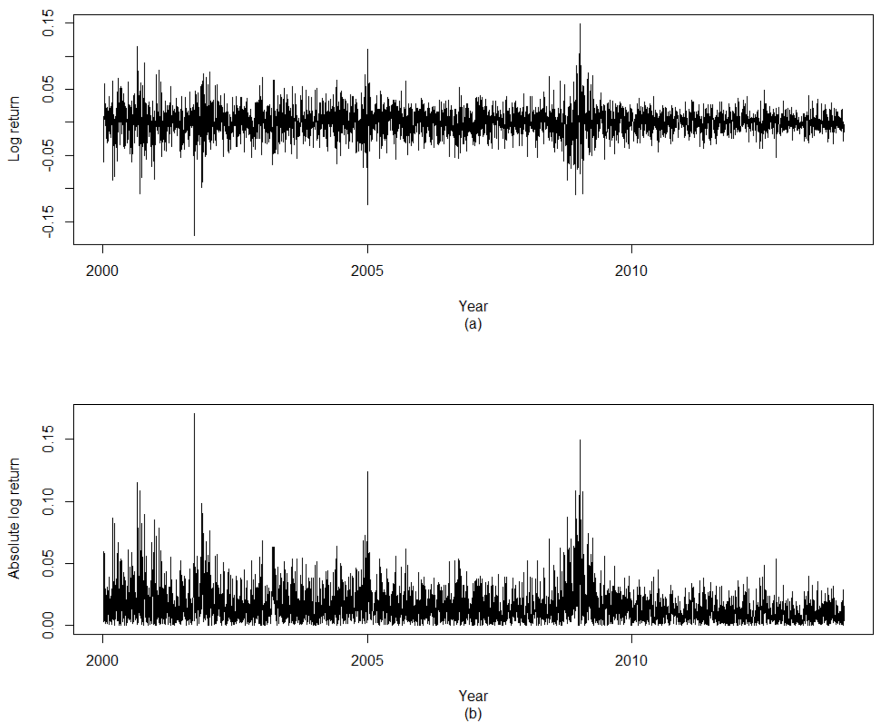

Figure 2: Daily Brent Oil log returns from January 4 2000 to January 22 2016: (a) log returns, (b) absolute log returns

By inspecting the descriptive statistics listed in Table 1 it is possible to get some indications how the data is distributed. As stated in subsubsection 3.3.5 and subsubsection 3.3.6 the sample skewness and kurtosis roughly signals how the log returns are distributed with respect to the tails. It is common for financial data that the sample kurtosis is greater than 3, thus being leptokurtic relatively a normal distribution.
From the the descriptive statistics in Table 1 it can be seen that the mean is very small. So not surprisingly, when testing the null hypothesis that the mean is zero against the alternative that it is different from zero using the t-test in (8) the null hypothesis is not rejected. So there is no evidence that the mean is non-zero and thus no indication of a trend in the data over the entire time span. Whether local trends exists is further discussed in section 6.3.1. It can be noted that the estimated standard deviation is roughly 0.021 so in order to reject the null hypothesis of zero mean with 5% level of significance the mean estimate need to be almost 900 times higher than currently estimated. However, when using a t-test in this case one should bear in mind that the test statistic is approximately t-distributed under the assumption that the observations are independent and identically

distributed. The log return series clearly violates the independency since it is heavily serially correlated as seen in Figure 5.
The negative sample skewness calculated according to (12) indicate that the tail on the left is longer or fatter than the right side. The sample kurtosis calculated according to (13) being greater than three indicates that as for most financial data Brent Oil log returns are leptokurtic, characterized by a high narrow peak and fatter tails compared to a normal distribution.

Mean	Standard deviation	Skewness	Kurtosis
0.00004739956	0.02132016	-0.1229448	6.863149

Table 1: Descriptive statistics for Brent Oil log returns from January 4 2000 to January 22 2016

As noted earlier in this section the distribution of the log return series could serve as a hint which distribution would be the most appropriate when specify the likelihood function in order to estimate the parameters in the models. The shortcoming of different distributions can easily be identified graphically by plotting a histogram or QQ-plot for the log returns. For comparison lines corresponding to the theoretical distributions are included in both the histogram and the QQ-plot. Large deviations between the data and these lines are signs of that the data is not distributed according to this particular distribution. The extreme observations in financial data are usually not captured by a normal distribution which usually calls for another distribution, e.g. a Student t distribution which has heavier tails. Having in mind that the the sample kurtosis for the log return is greater than three, it would not be surprising if a normal distribution does not fit the data well.

Together the QQ-plot and histogram in Figure 3 demonstrates what the sample kurtosis quantifies. From the QQ-plot in (a) it is clear that the tails are heavier than those of a normal distribution so as expected it does not capture the tail behaviour of the log returns. The histogram in (b) demonstrates the high and thin peak. The log return series is clearly leptokurtic and not normally distributed.

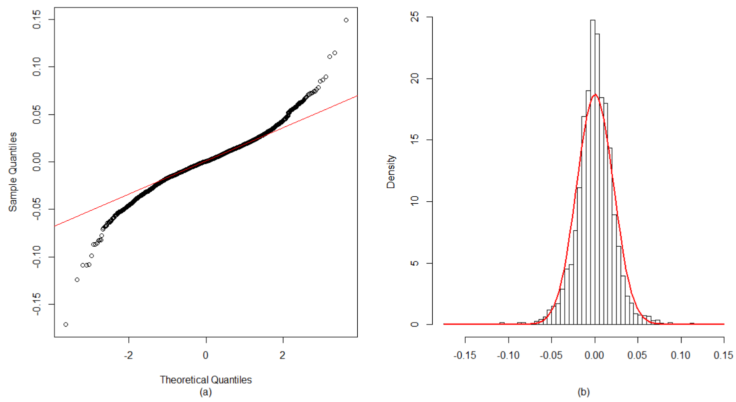

Figure 3: The distribution of the daily Brent Oil log returns from January 4 2000 to January 22 2016 illustrated in a: (a) normal QQ-plot, (b) histogram with a normal density curve

When choosing between rather simple distributions with few parameters such as the normal and Student-t distribution it is a trade off between having heavy tails and a high, thin peak around the mean. As seen in Figure 4 the tails of the Student-t distribution with six degree of freedom seem to be equally heavy as the tails of the log return series. Bearing in mind that the sample skewness is negative it is not surprising that the number of extreme observations in the lower part of the QQ-plot is greater than in the upper part. This further confirms that the left tail of the log return series is heavier than the right one. To conclude, a Student-t distribution with six degrees of freedom resembles the log return series and especially the tails of the log return series fairly well.

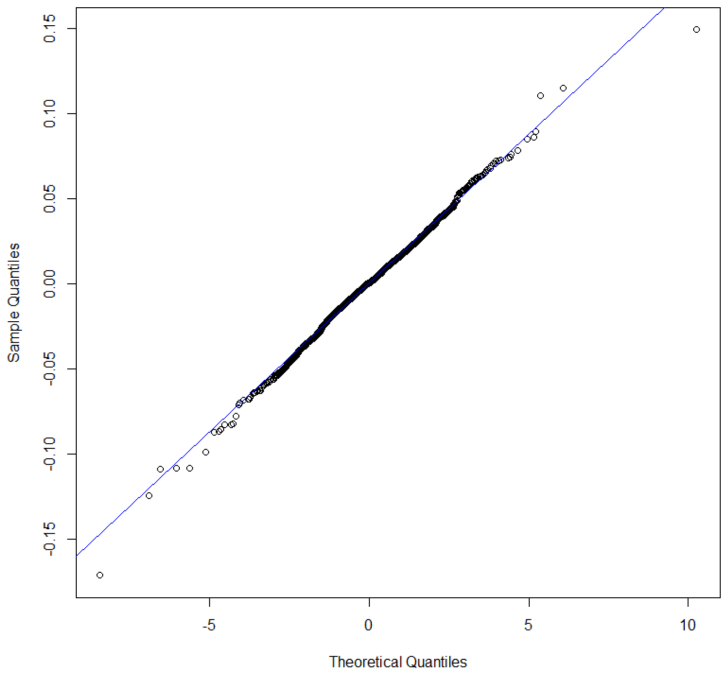

Figure 4: QQ-plot of the daily log returns of Brent Oil price from January 4 2000 to January 16 2016 against the theoretical quantiles for a Student-t distribution with six degress of freedom

To summarize, the log return series are experiencing volatility clustering and drastic changes in volatility, see Figure 2. This suggests using autoregressive conditional heteroscedastic models when modelling the time series. There are no indications of an existing trend with respect to the mean over the entire time period, see Table 1 and the text preceding it. The existence of trends will be further investigated in section 6.3.1. There are no signs of seasonality when inspecting Figure 2. Regarding the distribution, the series does not seem to be normally distributed and it is relatively leptokurtic, see Figure 3. The log return series is rather distributed according to a Student-t distribution with six degrees of freedom, see Figure 4. In addition the series is slightly negatively skewed, indicating that in the sample there are more negative log returns of a large magnitude compared to large positive ones, see Table 1 and Figure 4.

6 Modelling

This section progress with the next step in section 6.1, clarifying how the serial dependence in the data is structured using the sample ACF and PACF in Figure 5. Based on the model identification an appropriate model is fitted and its standardized residuals are graphically examined in section 6.2. Both with regards to how they are distributed in Figure 10 and further whether the standardized residuals have constant variance using Figure 11. In section 6.3 and section 6.3.1 it is further discussed whether it is appropriate to treat all the rolling windows based on the model selection of the entire period. In section 6.3.1 it is investigated whether there are local trends in the Brent Oil log return series using a simulation study.

6.1 Model identification

It is common that financial time series' suffer from structural breaks in volatility and volatility clustering. Different conditional heteroscedasticity models such as ARCH or GARCH models are frequently used to capture this. However in order to evaluate whether these or some other models are appropriate for the log return series at hand, the dependency structure in the log return series need to be assessed. This is done by studying how the log return values and its previous values cling together using different autocorrelation plots, correlograms, for different functions of the time series.

The correlograms for the untransformed time series show how a time series is correlated with its own previous values of different lag lengths. The autocorrelation plot for the squared and absolute time series indicates how the magnitude of the values of the time series is correlated with its own previous values for different lag lengths. Thus these plots describes the serial dependence in the time series. The partial autocorrelation plot for a time series depict the partial correlation of the time series with its own lagged values, controlling for the values of the time series at all shorter lags. Thus it describes time series serial correlation with its previous values of a specific lag. If partial correlation is significantly different from zero on a specific lag a trend is said to exist in the time series.

By inspecting the sample ACF and PACF plots (a) and (d) in Figure 5 the log returns appear to be white noise, the serial correlation is weak if any so there there is no indication of a trend in the data over the full time span. However, by inspecting the sample ACF plots for the (b) squared and (c) absolute log returns it appears that there indeed is a linear dependence in the log return series. This linear dependence does not seem to have any repeating patterns so there is still no signs of seasonality. These observations confirm the same observations made from Figure 2 regarding trend and seasonality. However the autocorrelation does not seem to decay at an exponential rate, rather at a slow hyperbolic rate. From the character of the squared and absolute log returns it should not be ruled out that the underlying process appear to be a long memory process. In such a process the squared and absolute log returns decay slowly even though the log return series does not exhibit serial correlation (Ding, Granger, and Engle, 1993 cited in Tsay, 2005, p.154).

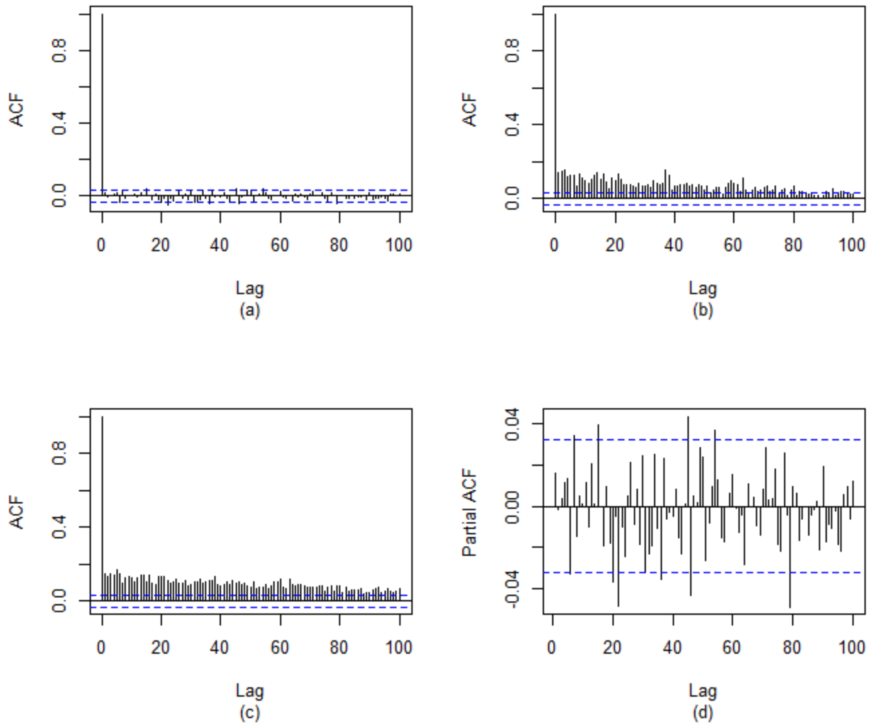

Figure 5: Sample ACF and PACF of various functions of Brent Oil log returns from January 4 2000 to January 22 2016: (a) ACF of the log returns, (b) ACF of the squared log returns, (c) ACF of the absolute log returns, (d) PACF of the log returns.

In order to identify what underlying process might give rise to this realisation simulations from theoretical processes can be done and it can then be evaluated whether the realisations of these simulations resemble the log return series with respect to the sample ACF and PACF plots. Log returns from ARCH(1) and GARCH(1,1) processes are simulated using the function *garchSim* in the package *fGarch* in R. The parameter values used are the ones obtained when fitting these models to the Brent Oil log return series using a Student-t distribution. All the sample ACF and PACF plots for the simulated GARCH(1,1) resemble the ones in Figure 5. It can be seen from (b) in Figure 6 that the sample ACF plot for the simulated GARCH(1,1) absolute log returns clearly resembles the sample ACF plot (c) for the absolute log return series in Figure 5. The same does not hold for the sample ACF (a) in Figure 6 for the simulated ARCH(1) series.

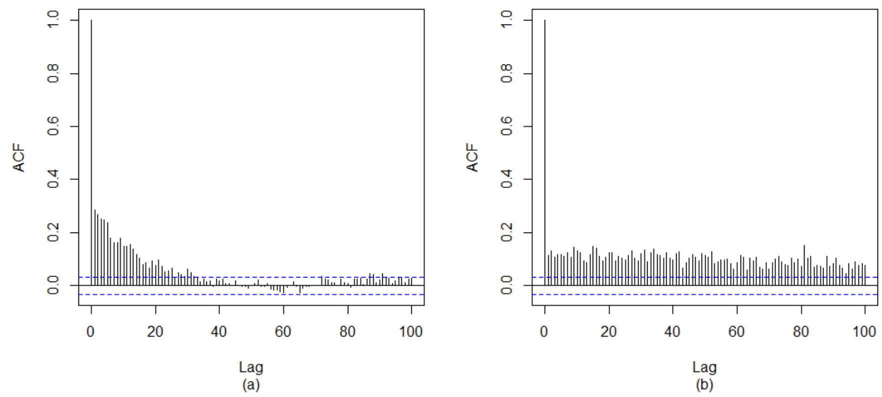

Figure 6: Sample ACF of 4000 simulated absolute log returns using parameter estimates obtained when fitting models using a Student-t distribution: (a) ARCH(1), (b) GARCH(1,1).

6.2 Model fitting

It has been established that a GARCH(1,1) process with Student-t distribution innovations seem to behave in a similar way as the log return series. However, since it is of interest to fit a simple model as possible a stepwise approach is applied. So far there is no indications that the mean equation describing μ_t in (2) need to be modelled in some way, the mean of the log return series is not significantly different from zero as seen in section 5.1. When fitting a randow walk model to the series the intercept is not significantly different from zero. This indicates that there is no drift in the mean equation that need to be modelled which further confirms that the mean in the series is not significantly different from zero. Further, the ACF and PACF plots in (a) and (d) in Figure 5 did not signal that there is a trend that need to be modelled either. So the mean equation is assumed to be zero. ARCH models of different lags are fitted using a normal distribution and the standardized residuals are obtained by dividing the innovations in the model by the estimated σ_t,

$$\epsilon_t = \frac{a_t}{\hat{\sigma}_t}.$$

As seen in the variance equation (3), in an ARCH(m) process the squared innovations constitutes an AR(m) process. So choosing the lag length of an ARCH process is done in a similar manner as when choosing the lag length in an AR process. The sample PACF plot for the squared log returns is used since in the ARCH model the squared shocks linearly depends on its past squared values. As seen in Figure 7 the serial dependence seem to cling off after lag six.

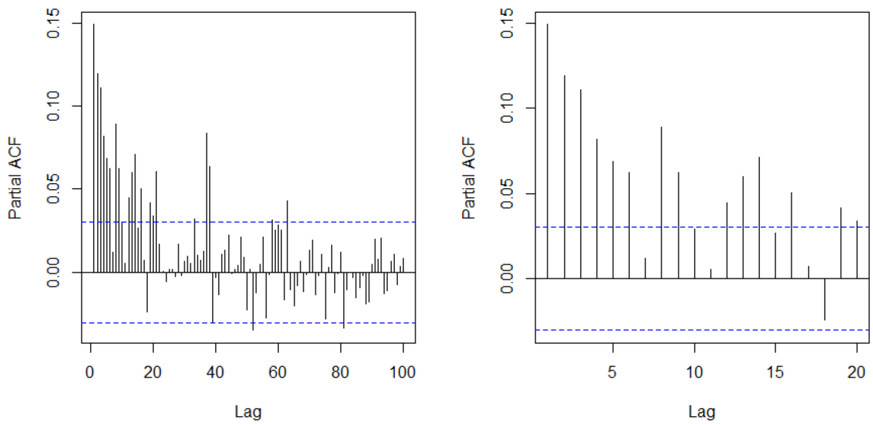

Figure 7: Sample PACF of the squared daily Brent Oil log returns from January 4 2000 to January 22 2016

It is confirmed by fitting models using both normal distribution and Student-t distribution that lags no less than six is needed in order for the standardized residuals from ARCH models to be white noise. However, when fitting models a simple model with as few lags as as possible is always preferable. So it would be of interest to fit a simpler model with regards to the number of parameters needed. In addition, according to Tsay (2010, p.119), the ability to capture excess kurtosis in ARCH models of higher orders is limited. Financial time series that mimic ARCH processes of higher orders can usually be resembled by GARCH processes of lower orders and as noted in previous sections the log return series at hand seem to mimic a GARCH(1,1) process. When fitting a GARCH(1,1) using a normal distribution the standardized residuals does not satisfy the normality assumption. It is illustrated in the QQ-plot in (a) in Figure 8 where the observations in the tails do not lie on the theoretical line. The same observation can be made from the histogram in (b) in the same figure, the normal density curve is below the bars in the tails. In addition, the normal density curve is also below the bars surrounding the mean in the center of the histogram. So as in the case of the log return series, the standardized residuals after fitting a GARCH(1,1) model using a normal distribution are leptokurtic.

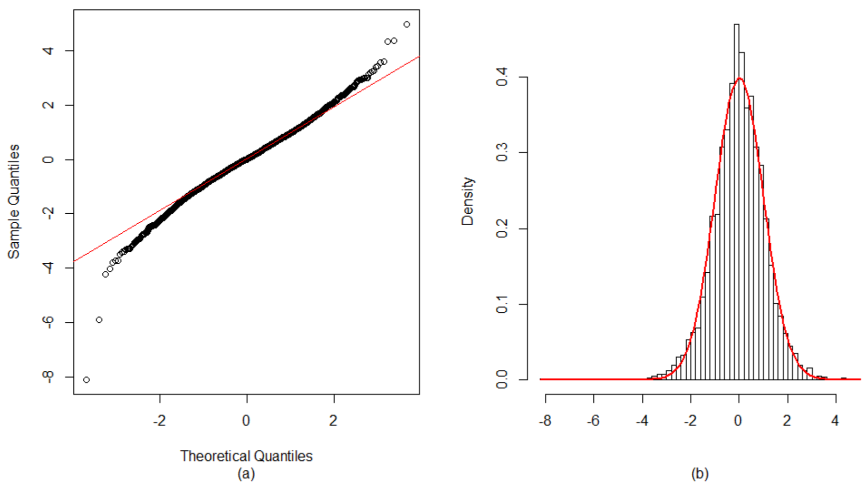

Figure 8: Plots of the standardized residuals when fitting a GARCH(1,1) model using a normal distribution to the Brent Oil log returns from January 4 2000 to January 16 2016: (a) QQ-plot, (b) Histogram with normal density curve.

It is concluded that the normality assumption does not hold. The tails of the distribution of the standardized residuals when fitting a GARCH(1,1) using a normal distribution were slightly heavier than the tails of a normal distribution suggesting that perhaps the use of the Student-t distribution is more appropriate. When fitting a GARCH(1,1) model with a Student-t distribution the standardized residuals are white noise, which is depicted in Figure 9. Testing the null hypothesis that the autocorrelations jointly are zero using the Box Ljung test (11) is not rejected for any lag length on 5 % level of significance. This confirms the visual conclusion.

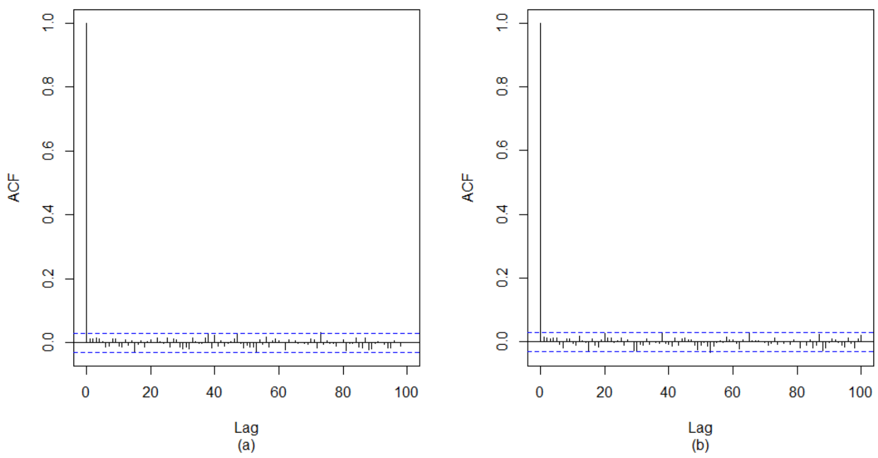

Figure 9: Sample ACF plots for various functions of the residuals when fitting a GARCH(1,1) model using a Student-t distribution to the daily Brent Oil log returns from January 2000 to January 2016: (a) ACF of the squared residuals, (b) ACF of the absolute residuals

In order to assess whether the model assumptions hold, that the standardized residuals are identically and independently Student-t distributed with zero mean and constant variance some model diagnostics is called for. The Student-t distribution assumption can be adressed with a QQ-plot and histogram. As seen in (a) in Figure 10 the standardized residuals seem to fit fairly well to a Student-t distribution with the exception for two extreme observations in the lower tail and that some observations in the upper tail do not lie on the line. In addition, the density curve of a Student-t distribution with 9.23 degrees of freedom in plot (b) follow the the bars in the histogram very closely. Since the normal distribution has a higher peak than an arbitrary Student-t distribution with finite number of degrees of freedom it is no surprise that the density curve for the Student-t distribution is below the bars around its mean just as in the case for the normal density curve in Figure 8. Even though the standardized residuals are leptokurtic, assuming that the standardized residuals are Student-t distributed seem justified.

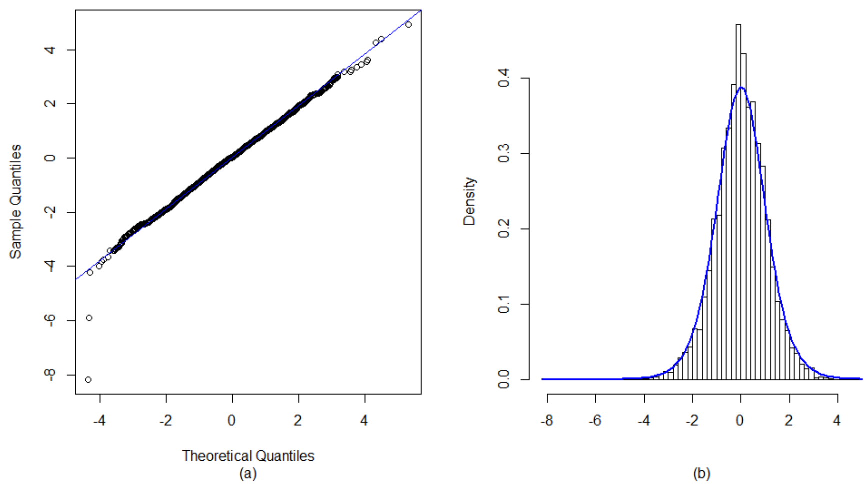

Figure 10: Plots of the standardized residuals when fitting a GARCH(1,1) model using a Student-t distribution to the Brent Oil log returns from January 4 2000 to January 16 2016: (a) QQ-plot, (b) Histogram with Student-t density curve

Regarding the assumption of constant variance and zero mean the model adequately captures the heteroscedasticity in the log returns depicted in (a) in Figure 11. This is illustrated clearly in (b) in Figure 11 where the standardized residuals are showing constant variance and seem to have mean zero. As expected after inspecting the QQ-plot in Figure 10 there are two outliers in the bottom of plot (b) in Figure 11. There are also some extreme observations in the top of the same plot.

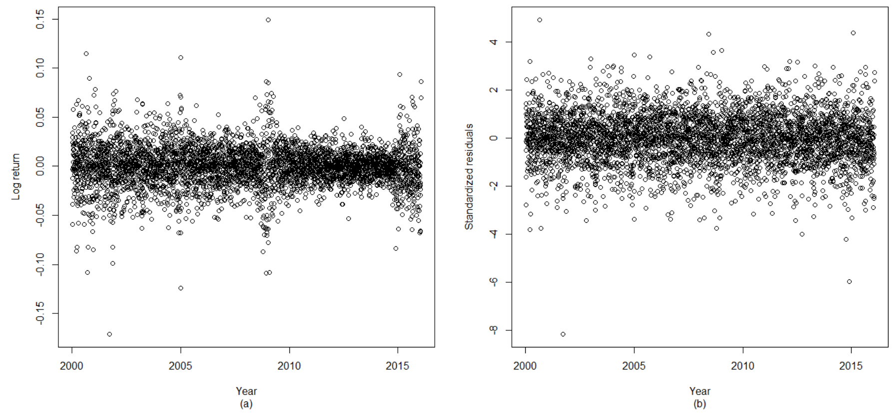

Figure 11: (a) Daily Brent Oil log returns from January 4 2000 to January 22 2016 (b) Standardized residuals when fitting a GARCH(1,1) using a Student-t distribution to the Brent Oil log returns for the same period

After inspecting these plots it is clear that the GARCH(1,1) model using a Student-t distribution seem to fit the log return series be fairly good. In addition the standardized residuals are even more skewed than the log return series with a sample skewness of -0.224 compared to -0.123 from Table 1. However the kurtosis is now 4.763 compared to 6.864 in the same table. This is indicating that the standardized residuals are more negatively skewed than the log return series and that they are leptokurtic but to a smaller extent than the log return series. All things considered, the Student-t distribution does not seem to capture the extreme observations nor does it manage to account for the skewness. It could be the case that a skewed t distribution would account for the skewness or maybe different distributions should be used for different subintervals. However, a Student-t distribution seems to be fairly good and is thus justified when proceeding. Even though the standardized residuals clearly are not normally distributed when fitting a GARCH(1,1) model to the log return series it does not necessarily imply that the forecasting performance of a GARCH(1,1) using a normal distribution is poor. It would be interesting to evaluate whether the distribution will affect the forecasting performance so model's forecasting performance using a normal distribution will also be evaluated.

When fitting a GARCH(1,1) using a Student-t distribution the mean equation is assumed to be zero. Neither the intercept term in the variance equation is significantly different from zero so the final model has the form

$$r_t = \sigma_t \epsilon_t, \quad \sigma_t^2 = \alpha_1 a_{t-1}^2 + \beta_1 \sigma_{t-1}^2$$

with the estimated variance equation

$$\hat{\sigma}_t^2 = 0.0348a_{t-1}^2 + 0.9642\sigma_{t-1}^2. \tag{25}$$

Since $\hat{\beta}$ in (25) is roughly 0.9642 it can be interpreted that the squared volatility at time t to a very large extent depend on the preceding squared volatility at time $t-1$ rather than the preceding shock.

It is now established that this model seem to fit the log return series well. The estimated parameters in the model sum to 0.999 so the IGARCH model is also a possible model candidate. The sum of the parameters using rolling windows is very close to 1 for the entire time span as seen in Figure 12. It should be noted that the ruotine rugarch used in R imposes the restriction that the sum of the parameters is less than or equal to 0.999 which explains the straight lines in the top of plot (a) and in (b) in Figure 12. As seen in plot (b) when using longer rolling windows such as five years the sum of the parameters is 0.999 for all windows throughout the entire time span.

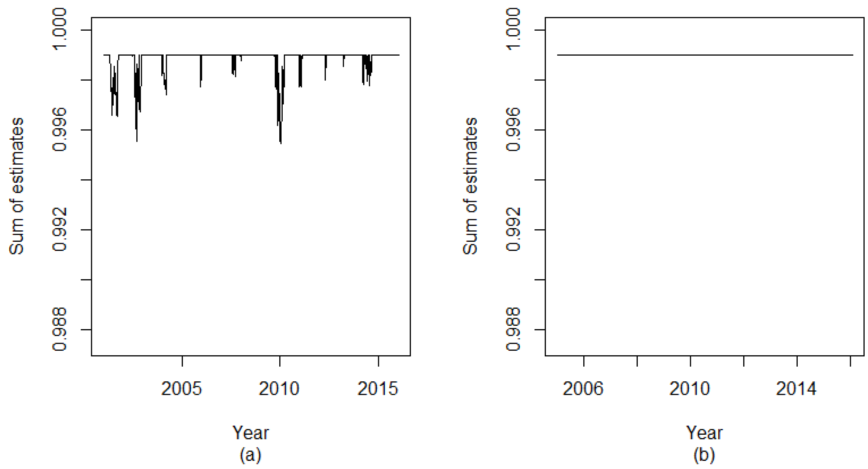

Figure 12: Sum of the estimated parameters α and β over time for the GARCH(1,1) model using a Student-t distribution and a rolling window length of: (a) one year (b) five years.

The fact that the sum of the parameters are so close to one is showing that the variance process in the log return series is mean reverting slowly, indicating that the shocks are persistent. This could explain why the log return series at hand resemble a long memory process as discussed in section 6.1. When fitting an IGARCH with a Student-t distribution the standardized residuals are still white noise. This is quite expected since the only change from the GARCH(1,1) is that the condition that the sum of the parameters α_1 and β_1 equals 1 is imposed.

What now is concluded is that the GARCH(1,1) and IGARCH(1,1) seem to mimic the log return series for the entire period from 2000-2016. No deeper evaluation regarding the fit of the models are done since the interesting question in this thesis is to assess whether these models are good at forecasting and if any of them are better than the other. However, when using one model for the entire time period it is assumed that the the entire time span is homogeneous which of course necessarily is not the case.

6.3 Structural breaks

As outlined in section 4.2 rolling windows should be used for which the models at hand should be estimated and in turn used to estimate the 1-step-ahead forecasts. It could be the case that different models are appropriate for different periods, i.e. that the log return series are dynamic regarding its behaviour during different circumstances or periods. That would clearly aggravate the model fitting in previous section. A crude approach to see if this is the case would be to find the best model for every rolling window or for different appropriate periods. Even the issue of defining what in this context would be appropriate periods would require considerably consideration. An even more naive approach would be to investigate how the parameters in the given models change over time, if they would change considerably that would be an indication that the given model at hand would not be appropriate. As seen in plot (a) in Figure 13 the parameters in a GARCH(1,1) using a Student-t distribution does not change considerably over time when using rolling windows of length one year. From the plot it can be seen that over the full time span, the parameter fluctuation is within a range of less than 0.1 except for 2005. The longer the rolling windows grow, the parameters fluctuates less. Plot (b) in Figure 13 depict the parameter estimates over the full time span for a GARCH(1,1) using a Student-t distribution for a rolling window length of five years. So based on this naive approach there is no evident signs that the GARCH(1,1) model is inappropriate when modelling the log return series. One interesting observation from plot (a) is that the parameter estimation change in absolute terms more and also more abruptly when the log returns from 2005 are included in the rolling window compared to when the financial crisis during 2008 is included.

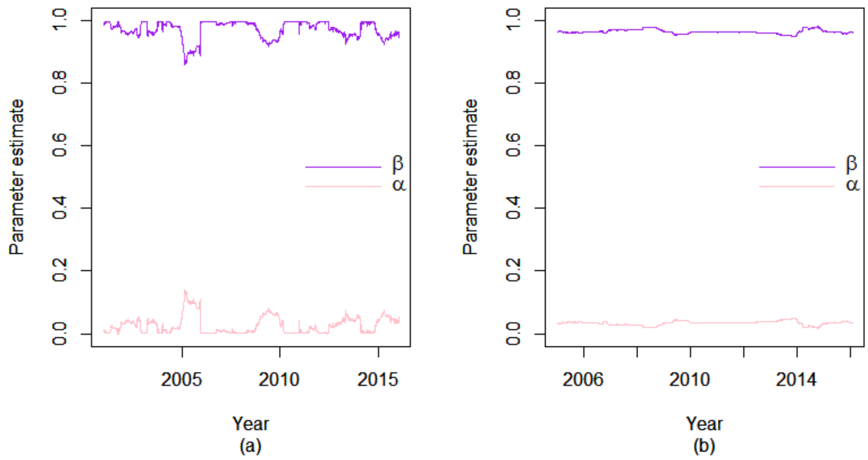

Figure 13: The parameter estimtaes of α and β over time when fitting a GARCH(1,1) model using a Student-t distribution. Using a rolling window of length: (a) 1 year, (b) 5 years.

6.3.1 Local trends

GARCH models are trying to capture the phenomenom that volatility is changing. Except from volatility shifts, a time series might exhibit some local trends. If there are local trends in the log return series then perhaps other models than GARCH(1,1) would have a better forecasting performance such as AR(1)-GARCH(1,1). One way to investigate whether there is a trend in form of an AR(1)-process would be to compare the distribution of estimated AR(1) components in the data at hand with a reference distribution of AR(1) estimates from a process where a trend is not present.

To obtain a reference distribution for the AR(1) components log returns are simulated from a GARCH(1,1) process using a Student-t distribution with the parameter estimates in (25). Then AR(1) coefficients are estimated for a rolling window through the entire simulated log return series.

This simulation and estimation process is repeated 1000 times and all AR(1) estimates together constitutes the reference distribution. From this reference distribution the empirical 0.025 and 0.975 quantiles are obtained. If there is no trend resembling an AR(1) process in a given log return series not more than 5% of the estimated AR(1) coefficients from this log return series should be smaller or greater than these quantile values.

To check if this is the case for the Brent Oil log return series AR(1) coefficients are estimated to the series using rolling windows of same length as when obtaining the reference distribution. One approach would be to count the AR(1) estimates from the Brent Oil log return series that are outside the

empirical 2.5% and 97.5% quantiles and then letting π denote the proportion of AR(1) estimates outside the empirical quantiles. Whether an estimate of an AR(1) coefficient is outside the empirical quantiles or not can be seen as a Bernoulli trial, where being outside is defined as a success. The sum of these Bernoulli trials is binomial distributed if the trials are independent and a binomial test can be performed. To test whether there is a greater number of AR(1) coefficients than can be explained by chance, the null hypothesis $H_0 : \pi = 0.05$ can be tested against the alternative $H_a : \pi \neq 0.05$ by constructing a Wald confidence interval according to (29) where x_i is 1 when the estimated AR(1) coefficient for a given window is outside the empirical 2.5% and 97.5% quantiles of the reference distribution. It should be noted that since the windows are rolling one step at a time the estimates in each window are clearly dependent which contradicts the assumption that the Bernoulli trials are independent. In order to have independent AR(1) estimates the log return series would have to be divided into non-overlapping windows. A problem with that is that the results could depend on where the windows are, as compared to when using rolling windows where every possible window is taken into account. In addition, the number of AR(1) estimates from the Brent Oil log return series would only be 16 when using a rolling window of one year and decrease towards one estimate when using a rolling window of ten years.
However, the reference distribution is obtained in the exact same way with rolling windows and will thus also render serially correlated estimates. So 5% of the AR(1) estimates for the Brent Oil log return series should lie outside the critical empirical quantiles under the assumption that there is no trend in the series.

The choice of the length of the rolling window would clearly affect the results since a local trend could vanish when using a longer window. On the other hand a local trend for very short windows would be rather unimportant and also hard to separate from sheer chance.

Without using a formal test it can be seen in plot (b) in Figure 14 where a rolling window of length five years has been used that after the window starts around 2009 (an thus starting with the window 2009-2014) the AR(1) estimates for almost all the rolling windows are above the empirical 97.5% quantile. Not only is there a lot more than 5% AR(1) estimates outside the critical quantiles, they are not randomly spread. The vast majority of the AR(1) estimates outside the quantiles are after 2009 and they are all above the 97.5% quantile, 14% to be precise. This is indication that there is a positive trend in the Brent Oil log return after 2010. From plot (a) in Figure 14 where a rolling window of length one year has been used it is clear that there probably is a local trend around 2010 to 2013 since the AR(1) estimates are above the 97.5% quantile for windows including this period. Almost 11% of all the AR(1) estimates in the plot are above the 97.5% quantile.

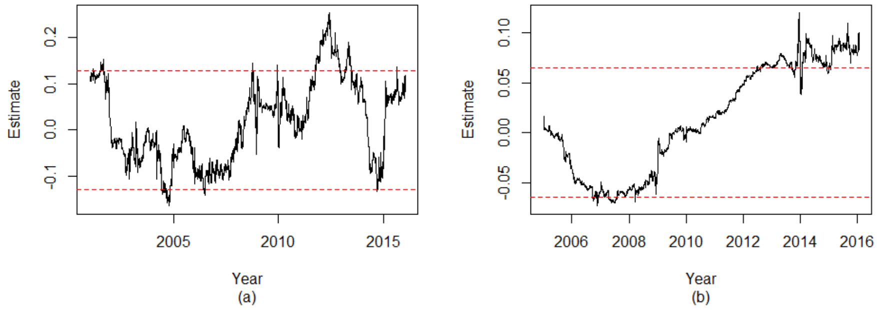

Figure 14: AR(1) estimates over time for the Brent Oil log return series using a rolling window of: (a) 1 year, (b) 5 years. The dotted lines are 2.5% and 97.5% quantiles of the reference distribution for AR(1) estimates obtained when estimating AR(1) estimates to a repeatedly simulated GARCH(1,1) process. The x-axis is indicating the last date in the rolling window.

It is suspected that there is a positive local trend in the Brent Oil log return series and that the underlying process at least for a period mimics the AR(1)-GARCH(1,1) process. For comparison it would also be interesting to see how such as simulated process would look like. An AR(1)-GARCH(1,1) model is fitted to the Brent Oil log return series and the parameters is used to simulate a process. Again, the function *garchSim* in the package *fGarch* in R is used for the simulation. In Figure 15 the AR(1) estimates estimated using a rolling window of one year to a realisation of such a process is plotted. The AR(1) estimates in plot(a) in Figure 15 over time is quite similar to the ones in plot (a) in Figure 14 considering that the estimates sometimes are below the empirical 2.5% quantile and when they are it is for a few subsequent windows. When the estimates are above the 97.5% quantile it is so for a greater number of subsequent windows. By inspecting plot (b) in Figure 15 it can be seen that the AR(1) estimates are not below the the 2.5% quantile and when it is above the 97.5% quantile it is so for very few subsequent windows. It does not resemble plot (b) in Figure 14 where the AR(1) estimates are above the 97.5% quantile for almost an entire span of rolling windows for three years. The conclusion that can be made from this is that the trend in the Brent Oil log return series seem to be a local one.

The AR(1)-GARCH(1,1) process has been simulated 1000 times and similar behaviour is seen in a sample of these simulated processes so it is not a special behaviour for the particular processes in Figure 15. It can be noted that the AR(1) estimate in the AR(1)-GARCH(1,1) model is modestly 0.02.

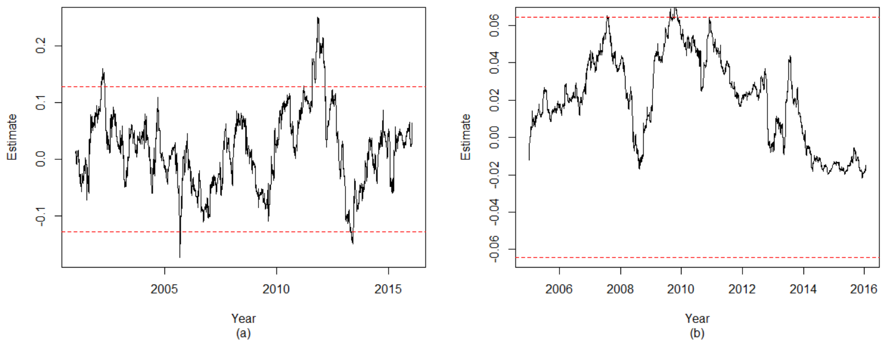

Figure 15: AR(1) estimates over time for a simulated AR(1)-GARCH(1,1) process using a rolling window of length: (a) 1 year, (b) 5 years. The dotted lines are 2.5% and 97.5% quantiles of the reference distribution for AR(1) estimates obtained when estimating AR(1) estimates to a repeatedly simulated GARCH(1,1) process. The x-axis is indicating the last date in the rolling window.

The estimates that gives rise to suspicion are positive. So the test of interest would be the one that tests the null hypothesis $H_0 : \pi_u = 0.025$ against the alternative $H_0 : \pi_u > 0.025$, where π_u is the proportion of the estimates greater than the empirical 97.5% quantile. The one-sided upper 95% Wald confidence bound for π_u are calculated according to:

$$\left(\frac{\sum_{i=1}^{n} x_i}{n}\right) - \frac{z_\alpha}{n\sqrt{\left(\frac{1}{\sum_{i=1}^{n} x_i} + \frac{1}{n - \sum_{i=1}^{n} x_i}\right)}}, \tag{26}$$

where n is the total number of estimates, x_i is the number of estimates greater than the 97.5% quantile and $z_\alpha = \Phi^{-1}(1-\alpha)$. For a rolling window of length one year the proportion of AR(1) estimates greater than the upper critical quantile is 0.108. The lower confidence bound obtained using (26) is 0.100 which is greater than 0.025 so the null hypothesis is rejected in favor of the alternative. The conclusion is that the high proportion of positive AR(1) estimates in the log return series greater than the empirical 97.5% quantile of the reference distribution cannot be explained by chance. The same is true when using rolling window lengths up to and including nine years. The fact that the proportion of AR(1) estimates outside of the the empirical 97.5% quantile is not greater than 0.025 when using a rolling window of length ten years is is further attesting that the trend is local. A trend that only is present for a shorter period will naturally not be discovered when looking at a longer time period where no other trends are present. The findings is in line with the observation made from Figure 14 that there seem to be a local positive trend in the log return series around 2010-2013.

Even though there is empirical evidence that there is a local trend in the log return series for simplicity and in order to somewhat limit the content in this paper it is assumed that there is not. In addition, in this paper it is of interest to evaluate given models for the entire time span from January 4 2000 to January 22 2016 while the discovered trend seem to be present only after 2010.

To summarize, the serial structure in the Brent Oil log return series is resembled by a GARCH(1,1) process. Having fitted the model using a Student-t distribution the standardized residuals are serially independent, see Figure 9. They are also showing constant variance, see Figure 11, so the standardized residuals seem to be white noise. In addition, assuming that the standardized residuals are Student-t distributed seem tenable, see Figure 10.
Since the sum of the estimates from the GARCH(1,1) model are close to 1, see (25), the structure in the Brent Oil log return series is possibly better resembled by an IGARCH(1,1). The process seem to large extent depend on the preceding volatility one day before rather than the shock one day before, see (25).
There is an indication that there is a trend in the log return series possibly around 2010-2013, see Figure 14. For simplicity when continuing it is assumed that there is no trend.

7 Analysis and results

This section will begin with presenting the results from the back testing procedure described in section 3.3.8 with including tests for unconditional coverage. In order to highlight the difference between the two models GARCH(1,1) and IGARCH(1,1) a graphical comparison between the volatility estimates conditioned on the distribution assumption is made in Figure 18. It will be followed by the results from the test of joint independent conditional coverage as described in section 3.3.9. Then finally the density forecast is evaluated.

7.1 Back testing

All the models independent of the length of the rolling windows that were used seem to be quite good in predicting the volatility.

The parameter estimates in the GARCH and the IGARCH model when using a given distribution is quite similar. Therefore it is not remarkable that the differences between plot (a) and (c) in Figure 16 or plot (b) and (d) is so small that they cannot visually be observed. However, when comparing the same models using different distributions it would be expected that the models using a Student-t distribution capture extreme observations to a larger extent than the ones using a normal distribution since the Student-t distribution's tails are heavier. By comparing plot (a) with (b) and plot (c) with (d) it is clear that that is the case, the models using Student-t distribution follow the extreme observations more closely.

Using same reasoning, since the log return series are leptokurtic the models using a normal distribution would be thought to follow the observations more closely when the log return are closer its mean zero since the peak of a normal distribution is higher than the corresponding of a Student-t distribution. This is not observable from the plots. Using a measure such as RMSE as described in section 3.3.10 could at best provide some information as to what extent a given model overall is wrong.

It should be noted that the plots using longer rolling windows look similar and the same reasoning regarding the comparisons above holds. However, the difference is that the models using the the longer windows does not seem to follow the extreme and sudden observations as closely. This illustrates the fact that even though the models depend on previous values of one lag, shorter window lengths and thus using more recent data when estimating the parameters makes the model more sensitive to sudden changes.

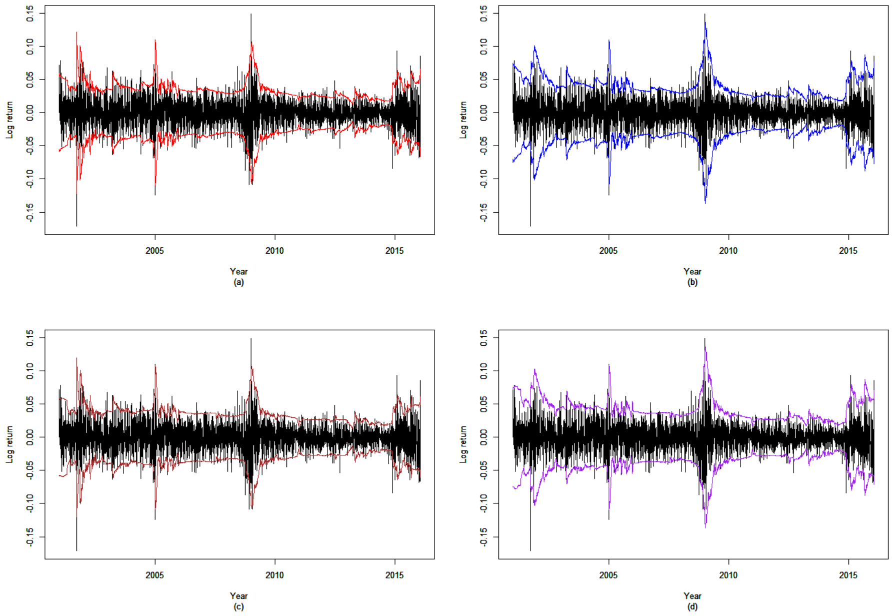

Figure 16: Brent Oil log returns from January 4 2000 to January 22 2016 in black with included forecasted 95% prediction intervals in color estimated using models: (a) GARCH(1,1) normal (b) GARCH(1,1) Student-t (c) IGARCH(1,1) normal (d) IGARCH(1,1) Student-t. Rolling windows of length one year is used.

7.1.1 Unconditional coverage

This coverage is on average expected to be 95% since the prediction intervals are created using 2.5 and 97.5% quantiles (24). The unconditional coverage including its corresponding 95% Wald confidence interval is calculated as described in section 3.3.9. Plotting these confidence interval for every window length for a given model in a plot gives a perspicuous picture of the unconditional coverage for a given model. As seen from (a) and (c) in Figure 17 the GARCH(1,1) and IGARCH(1,1) using normal distribution tend to underestimate the volatility since every Wald interval except one is below the desired degree of coverage 95%. The unconditional coverage for the GARCH(1,1) model using Student-t distribution is not significantly different from 95% for any window length as seen in plot (b). For an IGARCH(1,1) using Student-t distribution the coverage is on average not significantly different from 95% for window lengths greater than five years. The conclusions are that with regards to unconditional coverage there is no evidence that the assumption that the standardized residuals are conditionally normal distributed is tenable. Instead they seem to be conditionally Student-t distributed. This is in line with the the observation made in section 6.2. Assuming that the standardized residuals are conditionally Student-t distributed the GARCH(1,1) model performs better with regards to unconditional coverage than IGARCH(1,1) which tend to overestimate the volatility for shorter periods. Bearing in mind the observation made in the previous section, that using longer windows for estimation instead of using shorter windows that only includes the most recent observations made the model less sensitive to sudden changes. So the assumption of persistence in the IGARCH model is perhaps to strong when using shorter windows. Meaning that the combination of using short windows which cause the model to respond more heavily to shocks and the assumption that the shocks are persistent is causing the model to overpredict the volatility.

It could also be noted that for a given model the confidence intervals are getting wider the longer the used rolling windows are. This is the case since the variance estimates obtained from the observed Fisher information in Appendix C are consistent. The variances are inversely proportional to the sample size n. Because of that it is expected that the Wald intervals for a given level of significance would increase the longer the rolling window lengths are. Simply because when using longer windows the windows are rolling fewer times for a time series covering a given period.

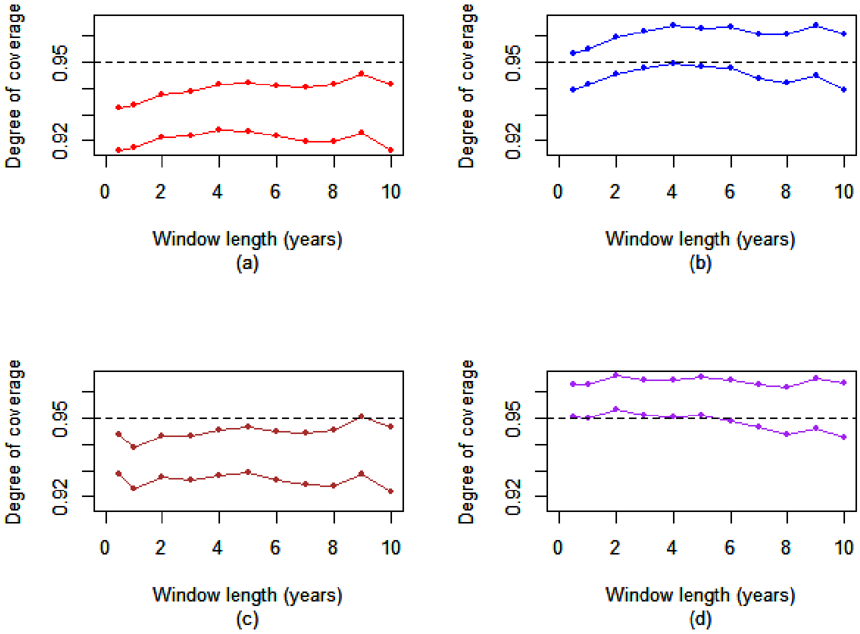

Figure 17: 95% Wald confidence intervals for the actual degree of coverage, plotted for different lengths of the rolling window for the model: (a) GARCH(1,1) normal, (b) GARCH(1,1) Student-t, (c) IGARCH(1,1) normal, (d) IGARCH(1,1) Student-t.

The fact that an IGARCH model imposes the assumption that shocks are persistent it would be expected that the IGARCH model would overestimate the volatility if the shocks in fact not are persistent. Consider the dotted lines in plot (a) and plot (b) in Figure 18. They correspond to the 1-step-ahead forecasted volatility for the IGARCH(1,1) models using normal and Student-t distributions respectively. These dotted lines are during tranquile periods above the whole lines in plot (a) and (b). These whole lines correspond to the 1-step-ahead forecasted volatility for the GARCH(1,1) models using normal and Student-t distributions respectively. So as expected the forecasted volatility one day ahead is for the most part greater for the IGARCH models than for the corresponding GARCH models. It can be noted that this difference narrows as the window length grows.

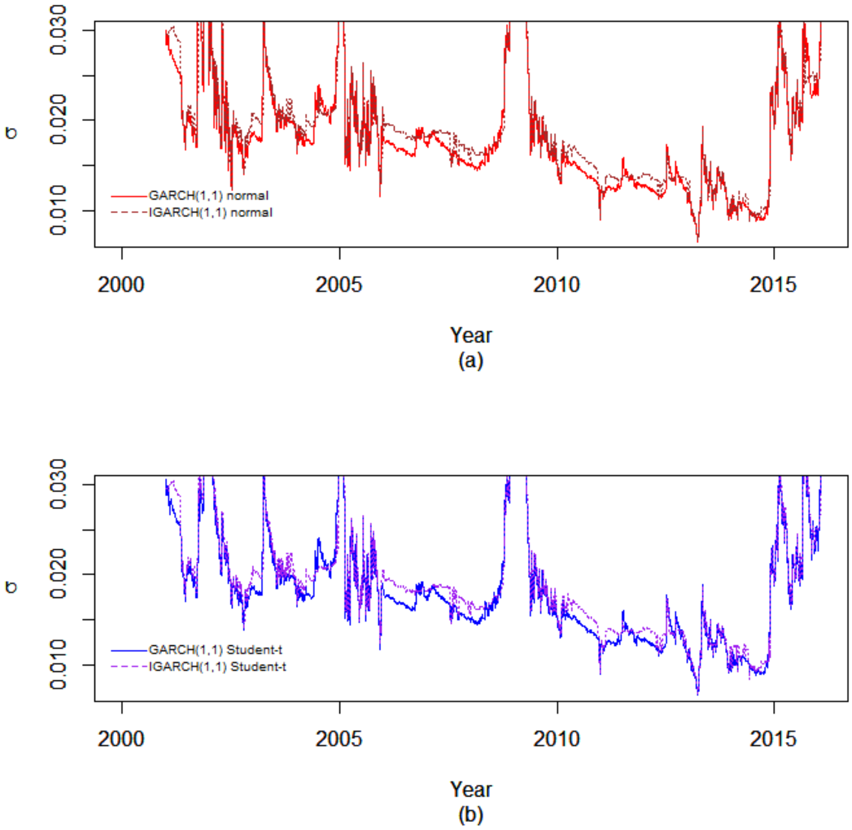

Figure 18: Plotting forecasted σ for all models when a rolling window length of one year is used. The plots are zoomed to highlight the difference during tranquile periods.

The observation made from Figure 16 that the the IGARCH(1,1) using a Student-t distribution performs well when longer rolling windows are used suggests that longer window lengths are favorable. This fact is further supported when inspecting Figure 19 from which it can be concluded that the forecasting error is smaller the longer rolling windows are used independent of the model. In addition it can also be seen that the RMSE is consistently higher for the IGARCH model than the GARCH model. Even though the differences in absolute numbers are small it speaks in favor of the use of the GARCH model. Another observation from the plot is that using the normal distribution gives a higher RMSE compared to when using a Student-t distribution conditioned on what model is used. This difference is in absolute terms very small.

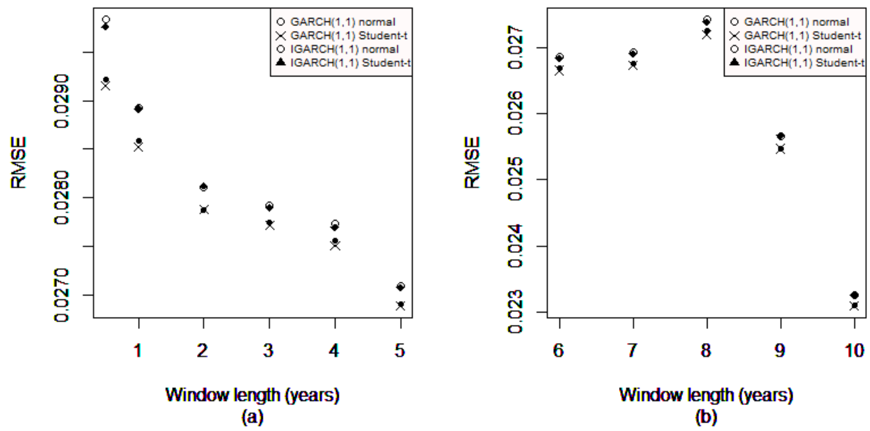

Figure 19: RMSE plotted for all different lengths of rolling windows for all combinations of models and distributions.

From the results it seems like the GARCH(1,1) model using a Student-t distribution is the best model. It is the only model that for all windows lengths has has the degree of coverage 95% on 95% confidence level. In addition the RMSE is also indicating that the GARCH(1,1) model using the Student-t distribution is the best candidate. Next step would be to further investigate this model with accompanying distribution.

7.1.2 Joint test of conditional coverage and independence

As discussed in section 3.3.9 there are two questions need that to be further investigated. Firstly, even if a model on average is right it is necessarily not the case that the degree of coverage is correct for smaller subperiods, bearing in mind that daily financial data is clustered. If the model fully accounts for the heteroscedasticity in the log return series then the outliers should not come in clusters. Secondly, it is known that financial data usually respond differently to positive and negative shocks. If the model consistently has more outliers in the lower than in the higher tail then this would be an indication that an asymmetric model would be preferable such as EGARCH. Or by bearing in mind the negative sample skewness perhaps using another distribution such as skewed-t would be motivated. It is assumed that the first order dependency as described in section 3.3.9 is an indication of clustering and that the two questions outlined above can be tested jointly using the LR-statistic (20). The p-values for the test is plotted in Figure 20. It should be noted that the p-values for longer rolling windows than seven years is not computable since there is not enough observations n_{ij} for all i and j. The same is the case for the rolling window length two years.

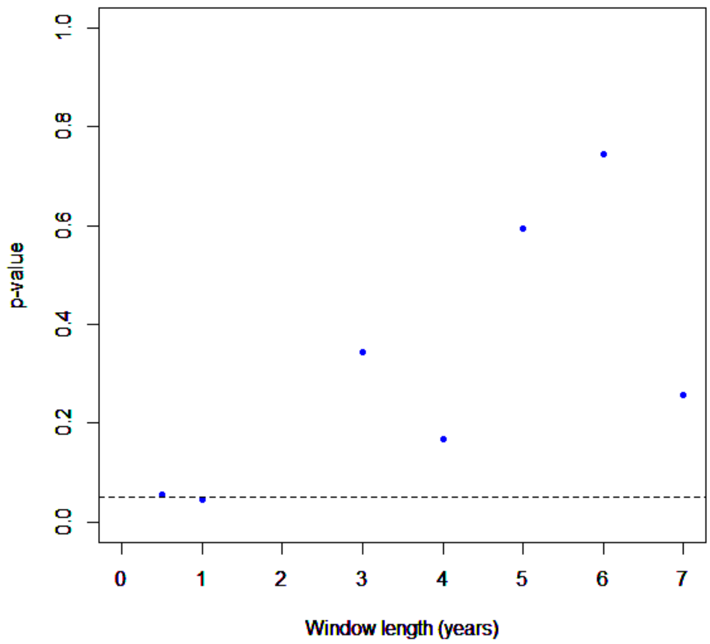

Figure 20: P-values when jointly testing for independence and conditional coverage of the GARCH(1,1) model using Student-t distribution. Plotted for rolling window lengths of 6 months to 7 years.

For the window lengths that the test is possible to perform the null hypothesis that the transition matrix is of the form

$$\Pi_0 = \begin{pmatrix} 0.025 & 0.95 & 0.025 \\ 0.025 & 0.95 & 0.025 \\ 0.025 & 0.95 & 0.025 \end{pmatrix} \tag{27}$$

is not rejected except for the window length of one year. In other words it cannot be rejected that whether the model will fail to predict the volatility 1-step-ahead at time origin h is independent of whether the model failed to predict the volatility 1-step-ahead at time point $h-1$. At the same time there is no empirical evidence that the probability that the volatility 1-step-ahead is lower or higher than implied by the prediction interval (24) is different from 0.025 and 0.025 respectively.

7.2 Density forecast

A more general approach to consider whether a model is appropriate would be to focus on how well the model forecast not only in the upper and lower quantiles but also how well it perform over all possible percentiles. In other words how well the model predict with regards to the entire distribution, thus evaluating the complete density forecast. As shown in section 3.3.11 the cumulative distribution function for the standardized residuals evaluated in the points $\frac{x_{t+1}}{\sigma_{t+1}}$ should be uniformly distributed on the interval [0,1] for all t. In Figure 21 Y_{t+1} obtained according to (22) for all t is plotted when a rolling window of one year has been used. Both plot (a) and (b) indicates that the cumulative distribution function for the standardized residuals evaluated in the points $\frac{x_{t+1}}{\sigma_{t+1}}$ for all t are uniformly distributed. The histograms and QQ-plots look almost identical no matter what length of the rolling window. The findings is further strengthened by the Kolmogorov-Smirnov test. The null hypothesis that Y_{t+1} is uniformly distributed is not rejected for any rolling window length for the GARCH(1,1) model using a Student-t distribution.

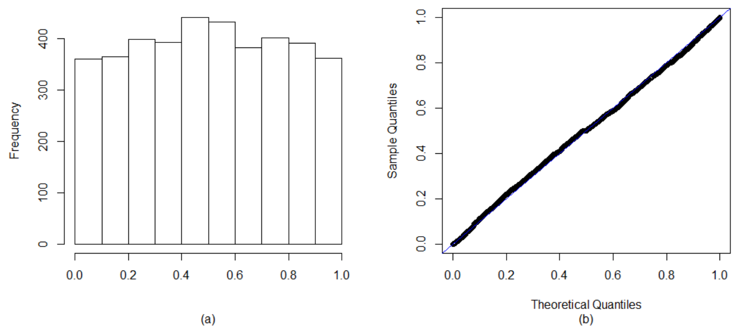

Figure 21: Distribution for the standardized residuals evaluated in the points $\frac{x_{t+1}}{\sigma_{t+1}}$ for the model GARCH(1,1) using a Student-t distribution when a rolling window length of one year has been used. Plotted in a: (a) Histogram, (b) QQ-plot.

To summarize the results, the GARCH(1,1) model using a Student-t distribution is the only candidate that for every rolling window length have unconditional degree of coverage not significantly different from 95% on 95% confidence level, see Figure 17. The RMSE indicates the same fact and in addition suggests that using longer windows when fitting the model would be preferable, see Figure 19.

There is hardly no empirical evidence against the fact that the model's prediction interval 1-step-ahead should be symmetric and independent on how the prediction interval performed in the previous time point, see Figure 20. There is also no empirical evidence against the fact that the entire distribution for the 1-step-ahead forecast is different from what the GARCH(1,1) using a Student-t distribution implies, see Figure 21.

8 Discussion

The method used in section 7.1.2 when evaluating whether the outliers come in clusters suffer from a great deal of limitation. The framework that is set up in section 3.3.9 is only considering independency of first order, the independency in one step. It is not considering whether the outliers are clustered for a longer period of time, in other words it is not considering independency of a more general form. Another limitation in the current test is that there are few observations that have made the so called transition from being outside the prediction interval to being outside the interval in the succeeding time period. This made the test impossible to perform when using longer rolling windows. It would be possible to extend the data set with more observations. It would also be possible to lower the confidence level for the prediction interval in (24) and test whether the model performs well on that level instead.

Regarding the window length used when estimating the parameters more observations is usually better since it leads to more precise estimates. However, the GARCH(1,1) and IGARCH(1,1) in this paper are used for 1-step-ahead forecasts and should to a large extent incorporate the most recent information in order to predict sudden changes. It is clear that when choosing the window length for estimation there is a trade off between parameter stability and given weight to the more recent information.

In Figure 17 it was observed that the unconditional coverage is significantly different from 95% for the IGARCH(1,1) model using a Student-t distribution for shorter rolling windows and not for longer. The lower Wald confidence bound was just above 95% degree of coverage when shorter periods were used and just below 95% for window lengths greater than five years. It cannot from this observation be concluded that the IGARCH(1,1) model using a Student-t distribution are better for longer rolling windows than for shorter ones. Simply because as noted in section 7.1.1 the Wald interval get wider for the longer windows at least partly because of that fewer observations are made.

However the forecasting error with regards to RMSE were almost consistently lower the longer estimation windows were used which suggests the use of longer windows is preferable. The RMSE were especially small for windows of length nine and ten years. Since the Brent Oil log return series in this paper started in the beginning of 2000, these windows are the only ones that include the financial crisis during 2008 for the first estimation windows. So the models using a rolling window length of nine and ten years do not forecast during the financial crisis, a period which was very volatile and probably resulted in high forecasting errors.

As seen in plot (a) and (b) in Figure 2 some periods during 2001, 2002 and 2005 were very volatile. So when the models are estimated using shorter windows these extremely volatile periods are rendering large forecasting errors. This is probably contributing to the fact that the RMSE when using shorter windows are higher than when using longer ones. So even though there is a lot of indications that longer rolling windows are prefereable it is necessarily not the case. To compare different window lengths perhaps the RMSE should be calculated for the exact same periods instead.

9 Conclusion

In section 5.1 it is concluded that the log return series are experiencing volatility clustering and drastic changes in volatility, see Figure 2. This suggests using autoregressive conditional heteroscedastic models when modelling the time series. There are no indications of an existing trend with respect to the mean over the entire time period, see Table 1 and the text preceding it. The existence of trends will be further investigated in section 6.3.1. There are no signs of seasonality when inspecting Figure 2. Regarding the distribution, the series does not seem to be normally distributed and it is relatively leptokurtic, see Figure 3. The log return series is rather distributed according to a Student-t distributed with six degrees of freedom when focusing on the tails, see Figure 4. In addition the series are slightly negatively skewed, indicating that in the sample there are more negative log returns of a large magnitude compared to large positive ones, see Table 1 and Figure 4.

In section 6 it is concluded that the serial structure in the Brent Oil log return series is resembled by a GARCH(1,1) process. Having fitted the model using a Student-t distribution the standardized residuals are serially independent, see Figure 9. They are also showing constant variance, see Figure 11, so the standardized residuals seem to be white noise.

Since the sum of the estimates from the GARCH(1,1) model are close to 1, see (25), the structure in the Brent Oil log return series is possibly better resembled by an IGARCH(1,1). The process seem to large extent depend on the preceding volatility one day before rather than the shock one day before, see (25).

There is an indication that there is a trend in the log return series possibly around 2010-2013, see Figure 14. For simplicity when continuing it is assumed that there is no trend and that the models GARCH(1,1) and IGARCH(1,1) are assumed to be appropriate for all rolling windows.

In section 7 it is concluded that the GARCH(1,1) model using a Student-t distribution is the only candidate that for every rolling window length have unconditional degree of coverage not significantly different from 95% on 95% confidence level, see Figure 17. The RMSE indicates the same fact and in addition suggests that using longer windows when fitting the model would be preferable, see Figure 19.

There is hardly no empirical evidence against the fact that the model's prediction interval 1-step-ahead should be symmetric and independent on how the prediction interval performed in the previous time point, see Figure 20. There is also no empirical evidence against the fact that the entire distribution for the 1-step-ahead forecast is different from what the GARCH(1,1) using a Student-t distribution implies, see Figure 21.

In section 8 the weaknesses of the joint test of conditional coverage and independence in section 7.1.2 is discussed. The test is only considering independency of first order, the independency in one step, and is not considering whether the outliers are clustered for a longer period of time. It is also stated that the test was not performable when longer rolling windows were used so data set covering a longer period could be used. Or the confidence level for the prediction interval in (24) could be lowered in order to get more observed transitions and test whether the model performs well on that level instead. It is also concluded that the even though there is indications that it would be preferable to use longer rolling windows, see plot (d) in Figure 17 and Figure 19, it cannot be ruled out that these indications can be explained by logical means. Concerning plot (a) in Figure 17, when longer rolling windows are used fewer observations are made. Thus the Wald confidence interval get wider as noted in section 7.1.1 and therefore includes the 95% level. Regarding the indication in Figure 19 this could be explained by the fact that the models using longer windows are not forecasting during some of the Brent Oil log returns' most volatile periods, the financial crisis especially.

This paper concluded that the GARCH(1,1) model using Student-t distribution performed well in every aspect considered when producing 1-step-ahead forecasts. The fact that the use of a model that allows for asymmetry such as an EGARCH model is not called for is in line with the findings of Wei, Wang, and Huang (2010) who found that linear GARCH models perform better when forecasting on 1 day horizons than nonlinear GARCH models do.

10 Further research

The standardized residuals when fitting the GARCH(1,1)-model using a Student-t distribution to the data were asymmetric. Combined with the fact that the fitted Student-t distribution does not seem to fit the standardized residuals perfectly in the QQ plot in Figure 10 it would be of interest to find a distribution which does and evaluate the forecasting performance of a GARCH(1,1) using this particular distribution. It would also be very interesting to incorporate models allowing for assymmetry such as EGARCH/TGARCH/APARCH for comparison.

Bearing in mind that there seem to be a positive local trend in the Brent log return it would be interesting to evaluate the forecasting performance of an AR(1)-GARCH(1,1). Since the trend is local the AR(1) component would be more or less redundant for the most part of the time span. So even though it would perform better during the period when the trend is present it is not sure that the performance over the entire time span would be better than the simpler GARCH(1,1) model.

This paper has only evaluated the different models performance using 1-step-ahead forecasts and a natural next step would be to produce forecasts on longer horizons and evaluate these. As noted in section 2 Kang, Kang and Yoon (2009, cited in Behmiri and Pires Manso, 2013, p.33), Wei, Wang, and Huang (2010, cited in Behmiri and Pires Manso, 2013, p.33) and Cheong (2009, cited in Behmiri and Pires Manso, 2013, p.32) found that other models than GARCH performed better on longer forecasting horizons. Even though other models are better it would be interesting to see how the performance of the models in this paper change on longer horizons.

The test for clustered outliers can be refined to test for independence of a more general form. The duration-based tests of independence that Christoffersen and Pelletier (2004) use can be used.

In light of the discussion that volatile periods render high forecasting error, it would also be of interest to evaluate and compare the given models specifically during the financial crisis.

As for the auther's own interest and development in the area fitting models such as Stochastic Volatility Models, Artificial Neural Networks and Support Vector Machines would be an inspiring challenge.

11 References

Bagdonavičius, V., Julius, K. and Nikulin, M. S., (2011). *Goodness-of-fit Tests Based on Empirical Processes, in Non-parametric Tests for Complete Data.* Hoboken, NJ, USA: John Wiley & Sons, Inc. Ch.3.

Behmiri, N.B. and Pires Manso, J.R., 2013. Crude oil price forecasting techniques: a comprehensive review of literature, *CAIA Alternative Investment Analyst Review*, 2(3), pp.30-48.

Christoffersen, P.F., 1998. Evaluating Interval Forecasts. *International Economic Review 39*, 4, pp.841–862.

Christofferson, P.F. and Pelletier, D., 2004. Backtesting Value-at-Risk: A Duration-Based Approach. *Journal of Financial Econometrics*, 2, p.84–10.

Held, L. and Sabanés Bové D., 2014. *Applied Statistical Inference.* New York: Springer. p.309-310.

Li, C., 2007. *On Estimation of GARCH Models with an Application to Nordea Stock Prices.* Department of Mathematics, Uppsala University.

Mandelbrot, B.B., 1963. The Variation of Certain Speculative Prices. *The Journal of Business*, 36(4), p.418.

Ruppert, D., 2004. *Statistics and Finance: An introduction.* New York, USA: Springer. Ch.3.

Tsay, R.S., 2005. *Analysis of Financial Time Series.* 2 edn. Hoboken, NJ, USA: John Wiley & Sons, Inc. Ch.1-3.

Appendices

A Distributions

A.1 Normal distribution

The density function for a normally distributed random variable X is

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(x-\mu)^2}{2\sigma^2}} \quad \text{for} \quad x \in (-\infty, \infty).$$

A.2 Student-t distribution

The density function for a Student-t distributed random variable X is

$$f_X(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\nu\pi}} \cdot \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} \quad \text{for} \quad x \in (-\infty, \infty),$$

where ν is the number of degrees of freedom.

B Maximum likelihood estimation

The theory in this section is from Li (2007, section 3.2).

The maximum likelihood estimate vector $\hat{\theta}_{MLE}$ is the estimate that is most likely given the sampla data, defined as

$$\hat{\theta} = \arg\max_{\theta \in \Theta} L(\theta).$$

Assuming that the vector of random variables $X = (x_1, x_2, ..., x_n)$ are independent and identical distributed $\hat{\theta}_{MLE}$ is vector that maximizes the joint likelihood

$$L(\theta; X) = f(X; \theta) = \prod_{i=1}^{n} f(x_i; \theta).$$

Taking log on both sides yields the log likelihood function

$$l(\theta; X) = \sum_{i=1}^{n} \log f(x_i; \theta).$$

The Score and Fischer information are given by

$$S(\theta; X) = \nabla l(\theta; X) = \frac{\delta l(\theta; X)}{\delta\theta}, \quad J(\theta; X) = \mathrm{E}\left(-\frac{\delta^2 l(\theta; X)}{\delta\theta\delta\theta^T}\right)$$

The estimates can be estimated using iterative methods. Let θ_j denote the parameter vector after the jth iteration. θ_{j+1} is then calculated using the Newton Raphson method according to

$$\theta_{j+1} = \theta_j + J(\theta)^{-1} S(\theta). \tag{28}$$

B.1 Time series

The theory in this section is from Tsay (2005, section 3.4) and Li (2007, section 3.2).

When specifying a likelihood function it is usually used that the observations are independent and identically distributed. Thus the likelihood equation is a product sum which often could be easy to deal with. In time series applications assuming that the log returns $\{r_t\}$ are independent is not tenable. This motivates an approach that uses the serial dependency by conditioning on the previous observation. More formally, using that

$$f(x_2|x_1) = \frac{f(x_2, x_1)}{f(x_1)} \Leftrightarrow f(x_2, x_1) = f(x_2|x_1)f(x_1),$$

can be generalized to

$$\begin{aligned} f(x_t, ..., x_1) &= f(x_t|x_{t-1}, ..., x_1)f(x_{t-1}, ..., x_1) \\ &= f(x_t|x_{t-1}, ..., x_1)f(x_{t-1}|x_{t-2}, ..., x_1) \\ &= f(x_t|x_{t-1}, ..., x_1)f(x_{t-1}|x_{t-2}, ..., x_1)...f(x_2|x_1)f(x_1) \end{aligned}$$

So the general joint likelihood can be written as

$$L(\theta; X) = \prod_{i=1}^{n} f(x_i|x_{i-1}, ...x_1; \theta)$$

and the joint log likelihood as

$$l(\theta; X) = \sum_{i=1}^{n} \log f(x_i|x_{i-1}, ...x_1; \theta)$$

Specially, estimating a GARCH(p, q) defined as in (4) assuming that ϵ_t is standard normal then

$$f(\epsilon_t|\epsilon_{t-1}, ..., \epsilon_0) = \frac{1}{\sqrt{2\pi}} e^{-\frac{\epsilon_t^2}{2}}.$$

Using that $a_t = \sigma_t \epsilon_t$ the conditional likelihood function of a_t is

$$f(a_t|a_{t-1}, ..., a_0) = \frac{1}{\sqrt{2\pi\sigma_t^2}} e^{-\frac{a_t^2}{2\sigma_t^2}},$$

which gives the conditional log likelihood of the parameter vector θ

$$\begin{aligned} l(\theta; a_{t-1}, ..., a_0) &= \sum_{t=q+1}^{n} \log f(a_t|a_{t-1}, ..., a_0) = \sum_{t=q+1}^{n} \log \left\{ \frac{1}{\sqrt{2\pi\sigma_t^2}} e^{-\frac{a_t^2}{2\sigma_t^2}} \right\} \\ &= \sum_{t=q+1}^{n} \left(-\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma_t^2) - \frac{a_t^2}{2\sigma_t^2} \right), \end{aligned}$$

where $\theta = (\alpha_0, \alpha_1, ..., \alpha_q, \beta_1, ..., \beta_p)^T$. So

$$\frac{\delta l(\theta)}{\delta\theta} = \frac{1}{2\sigma_t^2}\left(\frac{a_t^2}{\sigma_t^2} - 1\right)\frac{\delta\sigma_t^2}{\delta\theta},$$

and

$$\frac{\delta^2 l_t(\theta)}{\delta\theta\delta\theta^T} = \frac{1}{2\sigma_t^2}\left(\frac{a_t^2}{\sigma_t^2} - 1)\right)\frac{\delta^2\sigma_t^2}{\delta\theta\delta\theta^T} + \frac{1}{\sigma_t^4}\left(\frac{1}{2} - \frac{a_t^2}{\sigma_t^2}\right)\frac{\delta\sigma_t^2}{\delta\theta}\frac{\delta\sigma_t^2}{\delta\theta^T},$$

where $\frac{\delta\sigma_t^2}{\delta\theta} = (1, a_t^2, ..., a_{t-q}^2, \sigma_t^2, ..., \sigma_{t-p}^2)^T + \sum_{i=1}^{p}\beta_i\frac{\delta\sigma_{t-i}^2}{\delta\theta}$.

Thus the Score vector

$$S(\theta) = \tfrac{1}{2}\sum_{t=q+1}^{n}\frac{1}{\sigma_t^2}\left(\frac{a_t^2}{\sigma_t^2} - 1\right)\frac{\delta\sigma_t^2}{\delta\theta},$$

and the Fisher information matrix

$$J(\theta) = \tfrac{1}{2}\sum_{t=q+1}^{n} E\left(\frac{1}{\sigma_t^4}\frac{\sigma_t^2}{\delta\theta}\frac{\delta\sigma_t^2}{\delta\theta^T}\right).$$

In the special case of GARCH(1,1) model specified according to Equation 4, the vector $\theta = (\alpha_0, \alpha_1, \beta_1)^T$ should be estimated. Then

$$S(\theta) = \tfrac{1}{2}\sum_{t=2}^{n}\frac{1}{\sigma_t^2}\left(\frac{a_t^2}{\sigma_t^2} - 1\right)\frac{\delta\sigma_t^2}{\delta\theta},$$

and

$$J(\theta) = \frac{1}{2}\sum_{t=2}^{n} E\left(\frac{1}{\sigma_t^4}\frac{\sigma_t^2}{\delta\theta}\frac{\delta\sigma_t^2}{\delta\theta^T}\right).$$

The Score vector and Fisher information matrix are then used in the Newton Raphson method in (28) to obtain the vector $\hat{\theta}$ with maximum likelihood estimates.

If instead ϵ_t is assumed to be standardized Student-t distributed then the probability density function of ϵ_t for a given number of degrees of freedom ν is

$$f(\epsilon_t|\nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{(\nu-2)\pi}}\left(1 + \frac{\epsilon_t^2}{\nu - 2}\right)^{-\frac{\nu+1}{2}} \quad \text{for} \quad \nu > 2.$$

Using again that $a_t = \sigma_t\epsilon$ gives the conditional likelihood function of a_t is

$$f(a_t|a_{t-1}, ..., a_0) = \prod_{t=q+1}^{n}\frac{\Gamma(\frac{v+1}{2})}{\Gamma(\frac{v}{2})\sqrt{(v-2)\pi}}\frac{1}{\sigma_t}\left(1 + \frac{\epsilon_t^2}{(v-2)\sigma_t^2}\right)^{-\frac{v+1}{2}}.$$

Let $A_{t-1} = a_{t-1}, ..., a_0$, then the conditional log likelihood of θ is

$$l(\theta; A_{t-1}) = -\sum_{t=q+1}^{n} \left[\log \left(\frac{\Gamma(\frac{v+1}{2})}{\Gamma(\frac{v}{2})\sqrt{(v-2)\pi}} \right) + \frac{1}{2}\log(\sigma_t^2) + \left(\frac{v+1}{2} \right) \log \left(1 + \frac{a_t^2}{(v-2)\sigma_t^2} \right) \right].$$

The Score vector, Fisher information matrix and is obtained using the same procedure as when ϵ_t was assumed to be standard normal. Then again the Newton Rapshon method in (28) is used to obtain the vector $\hat{\theta}$ with maximum likelihood estimates.

B.2 Markov process

Let a Markov process have the transition matrix

$$\Pi_a = \begin{pmatrix} \pi_{ll} & 1-\pi_{ll}-\pi_{lu} & \pi_{lu} \\ \pi_{ml} & 1-\pi_{ml}-\pi_{mu} & \pi_{mu} \\ \pi_{ul} & 1-\pi_{ul}-\pi_{uu} & \pi_{uu} \end{pmatrix},$$

where the subindexes l, m and u denotes the states lower, middle and upper respectively.

The approximate likelihood function for this process is

$$\begin{aligned} L(\Pi_a; S_1, S_2, ..., S_n) &= \pi_{ll}^{n_{ll}}(1-\pi_{ll}-\pi_{lu})^{n_{lm}}...(1-\pi_{ul}-\pi_{uu})^{n_{um}}\pi_{uu}^{n_{uu}} \\ &= \prod_{i=l,m,u} \pi_{il}^{n_{il}}(1-\pi_{il}-\pi_{iu})^{n_{im}}\pi_{iu}^{n_{iu}}, \end{aligned}$$

where n_{ij} is the number of observations with value i followed by j. It is approximate because it is conditioned on the first observation since this observation is the starting point so a transition from a state in an earlier time period did not occur.

The log likelihood becomes

$$l(\Pi_a; S_1, S_2, ..., S_n) = \sum_{i=l,m,u} n_{il}\log(\pi_{il}) + n_{im}\log(1-\pi_{il}-\pi_{iu}) + n_{iu}\log(\pi_{iu}).$$

So the Score vector $S(\Pi_a) = \frac{\delta l(\Pi_a)}{\delta \pi_{ij}}$ for $i = l$ and $j = l, u$ is

$$\left(\frac{\delta l(\Pi_a)}{\delta \pi_{ll}}, \frac{\delta l(\Pi_a)}{\delta \pi_{lu}} \right)^T = \left(\frac{n_{ll}}{\pi_{ll}} - \frac{n_{lm}}{1-\pi_{ll}-\pi_{lu}}, \frac{n_{lu}}{\pi_{lu}} - \frac{n_{lm}}{1-\pi_{ll}-\pi_{lu}} \right)^T.$$

Putting this Score vector to zero yields after some tedious manipulation

$$\pi_{ll} = \frac{n_{ll}(1-\pi_{lu})}{n_{lm}+n_{ll}}, \quad \pi_{lu} = \frac{n_{lu}(1-\pi_{ll})}{n_{lm}+n_{lu}}.$$

.

Substituting π_{lu} into the expression for π_{ll} and π_{ll} into the expression for π_{lu} yields after tedious manipulation

$$\pi_{ll} = \frac{n_{ll}}{n_{ll} + n_{lm} + n_{lu}}, \quad \pi_{lu} = \frac{n_{ll}}{n_{ll} + n_{lm} + n_{lu}}.$$

By obtaining the Score vectors for $i = m, u$ the corresponding estimates for $\pi_{ml}, \pi_{mu}, \pi_{ul}$ and π_{uu} are obtained in the same way. The maximum likelihood estimates for Π_a are thus given by

$$\hat{\pi}_{ij} = \frac{n_{ij}}{n_{il} + n_{im} + n_{iu}} \text{for } i = l, m, u \text{ and } j = l, u.$$

.

C Wald interval for a binomial test

To obtain a confidence interval for the degree of coverage it is used that the sum of the indicator variable I_t are binomial distributed with parameters n and p. The likelihood for p is then given by

$$L(p) = \binom{n}{\sum_{i=1}^{n} x_i} p^{\sum_{i=1}^{n} x_i} (1-p)^{n - \sum_{i=1}^{n} x_i}.$$

.

The log likelihood is

$$l(p) \propto \left(\sum_{i=1}^{n} x_i\right) \log(p) + \left(n - \sum_{i=1}^{n} x_i\right) \log(1-p) + C,$$

and the Score function

$$S(p) = \frac{\delta l(p)}{\delta p} = \frac{\sum_{i=1}^{n} x_i}{p} - \frac{n - \sum_{i=1}^{n} x_i}{1-p}.$$

Putting the Score function to zero and solving for p through some tedious manipulation yields $\hat{p}_{MLE}$

$$\hat{p}_{MLE} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

The expected Fisher information and estimated expected Fisher information is after tedious manipulation given by

$$J(p) = -E\left(\frac{\delta^2 l(p)}{\delta p^2}\right) = n\left(\frac{1}{p} + \frac{1}{1-p}\right),$$

and

$$J(\hat{p}) = n^2\left(\frac{1}{\sum_{i=1}^{n} x_i} + \frac{1}{n - \sum_{i=1}^{n} x_i}\right).$$

The Wald statistic, T_W, which under the null hypothesis follow a standard normal distribution can be used to create a confidence interval for the degree of coverage p.

$$T_W = \frac{(\hat{p} - p)}{\mathrm{se}(\hat{p})} \sim N(0,1)$$

$$\begin{aligned} \implies T_W &= (\hat{p} - p) J(\hat{p})^{\frac{1}{2}} \\ &= \left(\frac{\sum_{i=1}^{n} x_i}{n} - p \right) n \sqrt{\left(\frac{1}{\sum_{i=1}^{n} x_i} + \frac{1}{n - \sum_{i=1}^{n} x_i} \right)} \sim N(0,1). \end{aligned}$$

Thus the Wald confidence interval CI_p for p is given by

$$\left(\frac{\sum_{i=1}^{n} x_i}{n} \right) \pm \frac{z_{\alpha/2}}{n \sqrt{\left(\frac{1}{\sum_{i=1}^{n} x_i} + \frac{1}{n - \sum_{i=1}^{n} x_i} \right)}}. \tag{29}$$

where $z_{\alpha/2} = \Phi^{-1}(1 - \alpha/2)$.