Automatized GARCH parameter estimation

Dennis Sundström

dennissu@kth.se
Supervisor at KTH: Boualem Djehiche
Supervisors at Cinnober Financial Technology AB: Mikael Öhman, Oskar Kallur

Department of Mathematics
Royal Institute of Technology
Stockholm, Sweden
August 2017

Abstract

This paper is about automatizing parameter estimation of GARCH type conditional volatility models for the sake of using it in an automated risk monitoring system. Many challenges arise with this task such as guaranteeing convergence, being able to yield reasonable results regardless of the quality of the data, accuracy versus speed of the algorithm to name a few. These problems are investigated and a robust framework for an algorithm is proposed, containing dimension reducing and constraint relaxing parameter space transformations with robust initial values. The algorithm is implemented in *java* with two models, namely the GARCH and gjr-GARCH model. By using real market data, performance of the algorithm are tested with various in-sample and out-of-sample measures, including backtesting of the widely used risk measure Value-at-Risk. The empirical studies conclude that the more complex gjr-sGARCH model with the conditional student's t distribution was found to yield the most accurate results. However for the purpose of this paper the GARCH or gjr-GARCH seems more appropriate.

Automatiserad approximation av GARCH parametrar

Sammanfattning

Denna uppsats undersöker möjligheten att automatisera approximationen av GARCH parametrar, där syftet är att använda algoritmen till ett automatiserat riskhanteringssystem. Med detta uppstår flera utmaningar som att garantera konvergens, kunna erhålla rimliga resultat oavsett datakvalitet, avvägning mellan algoritmens snabbhet och precision för att nämna några. Uppsatsen undersöker dessa problem och föreslår ett robust ramverk för en algoritm som innehåller transformationer av parameterrymden. Där dessa transformationer reducerar dimensionen av problemet samt reducerar antalet randvillkor. Algoritmen är implementerad i *java* med två modeller, GARCH och gjr-GARCH. Vidare så är algoritmen testad genom att använda riktig marknadsdata, där olika metoder använts för att utvärdera algoritmen. Modellerna som används backtestas på historisk data och det empiriska resultatet av detta talar för att gjr-sGARCH modellen med student's t fördelning levererar noggrannast resultat. Det är dock den mest komplexa modellen som används i denna uppsats och för denna uppsats ändamål anses GARCH eller gjr-GARCH modellerna mer passande.

Contents

1 Introduction

Ever since Engle [4] introduced the autoregressive conditional heteroscedasticity model or commonly referred to as the ARCH model, it has been generalized and modified in various ways for modeling financial time series. Where the ARCH model tries to capture the volatility clustering that financial time series often exhibit, that is, large volatility tends to be followed by large volatility and small volatility tends to be followed by small volatility. This is captured in the model by letting the variance of the current residuals be a function of the past squared residuals. Thus the model allows for time varying conditional variance, which arguably is the case of financial time series.

The model was then generalized by Bollerslev in [2], extending it by letting the current conditional variance be a weighted combination of past squared residuals and conditional variances allowing for longer memory. This extended model is commonly referred to as GARCH and has become very popular in financial applications. The GARCH model has then received several extensions to capture various types of "stylized facts" that financial time series exhibit. Some of those are the gjr-GARCH, EGARCH, IGARCH, TGARCH, to name a few. Using different types of GARCH models to account for volatility clustering in historical price data has been found to increase the accuracy of Value-at-Risk models. With main applications of the algorithm constructed in this paper being risk management, the GARCH models play a key role.

The purpose of this paper is to create a robust algorithm, capable of automatizing the parameter estimation of various GARCH type models, to be implemented in an automated risk monitoring system. GARCH parameter estimation can be solved by various methods. In this paper the most common one, being the maximum likelihood estimation, is covered, which is a non-linear optimization problem prone to spurious solutions. Typically in literature on the subject of GARCH models the technicalities and difficulties arising in the optimization problem when using the maximum likelihood approach are not covered. Some of the issues was addressed by Zumbach in [10] where he proposed a parameter space transformation reducing the dimension and relaxing the constrained optimization problem. This paper provides an extensive study of the optimization problem, covering the additional challenges that arise with automatizing the parameter estimation of the GARCH process, such as the trade-off problem between accuracy and time, guaranteeing convergence and handle different types of data sets. The trade-off problem lies not only in the algorithm itself, it is also a problem on how complicated models to employ, where the potential accuracy gained with more complicated models may not justify the use of it in an automated risk monitoring system since it takes too long time to estimate the parameters. The problems of convergence are addressed in this paper by constructing a method of choosing initial values for the maximum likelihood estimation and a fallback solution for the worst case scenario, being the case when the algorithm does not converge.

Furthermore the problem on handling various types of data sets does not only depend on the sample size of the data, also other problems arise such as data sets consisting of illiquid financial instruments.

The paper is organized as following. First the implemented models are defined in Section 2. Then the method for estimating the parameters, covering all the details of the maximum likelihood approach and a parameter space transformation together with a algorithm of choosing initial values is included, all of this are provided in section 3. Followed by section 4 and 5 with a topic on the assumption of i.i.d residuals and defining the methods to evaluate the forecast accuracy of each employed model. Finally a empirical study is provided in section 6, including two data sets with real market data.

2 Model specification

In this section the conditional volatility models used in this paper are specified. Starting with the standard GARCH model and proceeding with the more complex gjr-GARCH model. Followed by a specification of the conditional distributions considered. All data used in this paper are mean adjusted to contain zero mean, thus the models specified omits the mean. Through out this paper GARCH will refer to GARCH$(1,1)$ and small case, s, indicates the use of student's t distributed residuals.

2.1 GARCH

The generalized autoregressive conditional heteroscedasticity or commonly referred to as GARCH(p,q) is a parametric conditional volatility model and may be defined by

$$\begin{aligned} r_t &= \sigma_t \epsilon_t, \\ \sigma_t^2 &= \omega + \sum_{i=1}^{n} \alpha_i r_{t-i}^2 + \sum_{i=1}^{n} \beta_i \sigma_{t-i}^2, \end{aligned} \tag{1}$$

where $\{\epsilon_t\}$ is a set of i.i.d. random variables with zero mean and unit variance, $\omega > 0$, $\alpha \geq 0$, $\beta \geq 0$ are the parameters of the process. Although through out this paper the GARCH$(1,1)$ model, defined by

$$\sigma_t^2 = \omega + \alpha r_{t-1}^2 + \beta \sigma_{t-1}^2, \tag{2}$$

will be used. The process is further assumed to be stationary with finite variance i.e. $\mathbb{E}[\sigma_t^2] < \infty$ which implies $\alpha + \beta < 1$.

2.2 gjr-GARCH

The gjr-GARCH model proposed by Glosten, Jagannathan and Runkle in [5], extends the GARCH model by capturing the stylized fact that negative shocks at time $t-1$ impacts the volatility at time t more than the positive shocks. This is captured by defining the model

$$\begin{aligned} r_t &= \sigma_t \epsilon_t, \\ \sigma_t^2 &= \omega + (\alpha + \gamma \mathrm{I}_{t-1}) r_{t-1}^2 + \beta \sigma_{t-1}^2, \end{aligned} \tag{3}$$

where $\{\epsilon_t\}$ is a set of i.i.d. random variables with zero mean and unit variance. Note that with the introduced parameter $\gamma = 0$ we obtain the standard GARCH model. I_{t-1} is the indicator function defined by

$$\mathrm{I}_{t-1} = \begin{cases} 0, & r_{t-1} \geq 0, \\ 1, & r_{t-1} < 0, \end{cases} \tag{4}$$

thus the model captures negative shocks with coefficient $\alpha + \gamma$ and α for positive shocks. The process is further assumed to be stationary.

2.3 Conditional distributions

This section covers the distributions considered in this paper. The standard assumption is that the observed returns are conditional normally distributed. However returns observed in real market data often exhibit the stylized fact of fat tails, that is, the distribution exhibits large skewness or kurtosis. Thus the student's t distribution is widely used in financial applications since it is a leptokurtic distribution that captures the excess kurtosis in the distribution, meaning that it has fatter tails.

2.3.1 Normal distribution

Assume that $\{\epsilon_t\}$ is a set of i.i.d. random variables drawn from the standard normal distribution. Let $\mathcal{F}_t = \sigma\{r_s : s \leq t\}$ denote the sigma algebra representing the information of the process up to time t so that $\{\mathcal{F}_t\}_{t \in \mathbb{Z}_{>0}}$ is the natural filtration. Then the conditional distribution of r_t in equation 2 follows the normal distribution as following.

$$\begin{aligned} &r_t \mid \mathcal{F}_{t-1} \sim \mathcal{N}(0, \sigma_t^2), \\ &f(r_t \mid \mathcal{F}_{t-1}) = \frac{1}{\sqrt{2\pi\sigma_t^2}} e^{-\frac{r_t^2}{2\sigma_t^2}}, \qquad t = 1, 2, ..., n, \end{aligned} \tag{5}$$

where $\mathcal{F}_{t-1}$ is the natural filtration containing the information of the process up through time $t-1$ and let the parameter space be denoted by $\Theta = \{\omega, \alpha, \beta\}$.

2.3.2 Student's t distribution

Now let $\{\epsilon_t\}$ be a set of i.i.d. random variables drawn from the student's t distribution. First we need to scale ϵ_t with a factor to preserve the property $\text{Var}(r_t|\mathcal{F}_{t-1}) = \sigma_t^2$,

$$\begin{aligned} \eta_t &= \sqrt{\frac{\nu-2}{\nu}}\epsilon_t, \\ r_t &= \sigma_t \eta_t, \end{aligned} \tag{6}$$

where ν is the shape or degrees of freedom parameter, now the conditional distribution of r_t follows the standardized student's t distribution with variance σ_t^2

$$f_\nu(r_t \mid \mathcal{F}_{t-1}) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\sigma_t^2(\nu-2)\pi}\Gamma(\frac{\nu}{2})}\left(1 + \frac{r_t^2}{\sigma_t^2(\nu-2)}\right)^{-\frac{\nu+1}{2}}, \qquad \begin{aligned} &t = 1, 2, ..., n \\ &\nu > 2 \end{aligned} \tag{7}$$

and the parameter space is denoted by $\Theta = \{\omega, \alpha, \beta, \nu\}$.

3 Estimation

In order to estimate the parameters of the conditional volatility models in this paper maximum likelihood estimation is used, which is a constrained non-linear optimization problem. There exists many different algorithms to solve this problem. The focus in this paper lies on a line search algorithm, that is the non-linear conjugate gradient method, see [9]. Since the aim is to automate the process of estimating GARCH parameters, the algorithm is implemented in *java*. However before maximum likelihood estimation some preparatory work is done including parameter transformations and study of initial values in sections 3.1, 3.2 respectively to construct a robust framework for the algorithm.

3.1 Parameter transformation

In line with the findings of Zumbach in [10], some parameter transformations are done to reduce the dimension of the parameter space Θ and thereby reducing the number of

parameters $\theta \in \Theta$ to estimate by maximum likelihood estimation. Also by doing this transformations the original constrained optimization problem transforms to a unconstrained optimization problem to some extent.

3.1.1 GARCH

By rewriting the conditional variance in equation 2 to

$$\begin{aligned} \sigma_t^2 &= \sigma^2(1-\mu_{corr}) + \mu_{corr}(\mu_{ema}\sigma_{t-1}^2 + (1-\mu_{ema})r_{t-1}^2) \\ &= \sigma^2 + \mu_{corr}(\mu_{ema}\sigma_{t-1}^2 + (1-\mu_{ema})r_{t-1}^2 - \sigma^2), \end{aligned} \tag{8}$$

with the corresponding parameter transformation

$$\begin{aligned} \sigma^2 &= \frac{\omega}{1-\alpha-\beta}, \qquad \sigma^2 > 0, \\ \mu_{corr} &= \alpha + \beta, \qquad 0 < \mu_{corr} < 1, \\ \mu_{ema} &= \frac{\beta}{\alpha+\beta}, \qquad 0 < \mu_{ema} < 1. \end{aligned} \tag{9}$$

With some algebra and using the fact that $\mathbb{E}[\epsilon_t] = 0$, $\mathrm{Var}(\epsilon_t) = 1$ and that the process is stationary to calculate the unconditional variance of r_t, the advantages of this parameter transform starts to show

$$\begin{aligned} \mathrm{Var}(r_t) &= \mathbb{E}[r_t^2] - \mathbb{E}[r_t]^2 \\ &= \mathbb{E}[\sigma_t^2\epsilon_t^2] - 0 \\ &= \mathbb{E}[\sigma_t^2] \cdot 1 \\ &= \mathbb{E}[\omega + \alpha r_{t-1}^2 + \beta\sigma_{t-1}^2] \\ &= \omega + (\alpha+\beta)\mathrm{Var}(r_t), \end{aligned} \tag{10}$$

and thus the unconditional variance is obtained as

$$\mathrm{Var}(r_t) = \frac{\omega}{1-(\alpha+\beta)} = \sigma^2, \tag{11}$$

and now the unconditional variance of r_t appears explicit as the mean of the conditional variance in the process. By estimating the unconditional variance σ^2 with the unbiased

sample variance of the returns the parameters to estimate by maximum likelihood are reduced to (μ_{corr}, μ_{ema}) and also used as an initializer of the process i.e. $\sigma_0^2 = \sigma^2$. Furthermore μ_{corr} acts as a decay parameter that determines how fast the conditional variance converges to the unconditional variance. This may be seen by looking at the $t + h$ step forecast

$$\begin{aligned}
\widehat{\sigma}_{t+1}^2 &= \sigma^2 + \mu_{corr}(\mu_{ema}\sigma_t^2 + (1 - \mu_{ema})r_t^2 - \sigma^2) \\
&= \sigma^2 + \mu_{corr}(\mu_{ema}\sigma_t^2 + (1 - \mu_{ema})\mathbb{E}[r_t^2|\mathcal{F}_{t-1}] - \sigma^2) \\
&= \sigma^2 + \mu_{corr}(\sigma_t^2 - \sigma^2) \\
&\vdots \\
\widehat{\sigma}_{t+h}^2 &= \sigma^2 + \mu_{corr}^h(\sigma_t^2 - \sigma^2).
\end{aligned} \tag{12}$$

Hence for large h

$$\lim_{h\to\infty} \widehat{\sigma}_{t+h}^2 = \sigma^2, \tag{13}$$

which is why the unconditional variance are commonly referred to as the long term variance. and the last parameter μ_{ema} acts like a weight between the past conditional variance and squared return, similarly as the exponential moving average. With this parameters we still have a constrained optimization problem, thus another transformation is proposed as

$$\left.\begin{aligned} \mu_{corr} &= e^{-e^{-z_{corr}}} \\ \mu_{ema} &= e^{-e^{-z_{ema}}} \end{aligned}\right\} \iff \begin{aligned} z_{corr} &= \log\left(-\tfrac{1}{\log(\mu_{corr})}\right) \\ z_{ema} &= \log\left(-\tfrac{1}{\log(\mu_{ema})}\right), \end{aligned} \tag{14}$$

resulting in an unconstrained optimization problem. For completeness equation 8 is rewritten in the z coordinates

$$\sigma_t^2 = \sigma^2 + e^{-e^{-z_{corr}}}(e^{-e^{-z_{ema}}}\sigma_{t-1}^2 + (1 - e^{-e^{-z_{ema}}})r_{t-1}^2 - \sigma^2). \tag{15}$$

For student's t distributed residuals a transformation of ν is also done to relax the constraint $\nu > 2$ and for the sake of avoiding the possibility of dividing by zero

$$z_\nu = \log(\nu - 2). \tag{16}$$

3.1.2 gjr-GARCH

Rewrite equation 3 in similar way as for the GARCH model to obtain

$$\sigma_t^2 = \sigma^2 + \mu_{corr}\left((1-\mu_{ema})r_{t-1}^2 + 2(\mu_{ema}-\mu_{asy})\mathrm{I}_{t-1}r_{t-1}^2 + \mu_{asy}\sigma_{t-1}^2 - \sigma^2\right), \tag{17}$$

with the corresponding parameter transformation

$$\begin{aligned}
\sigma^2 &= \frac{\omega}{1-(\alpha+\beta+\frac{\gamma}{2})}, \qquad \sigma^2 > 0,\\
\mu_{corr} &= \alpha+\beta+\frac{\gamma}{2}, \qquad 0 < \mu_{corr} < 1,\\
\mu_{ema} &= \frac{\beta+\frac{\gamma}{2}}{\alpha+\beta+\frac{\gamma}{2}}, \qquad 0 < \mu_{ema} < 1,\\
\mu_{asy} &= \frac{\beta}{\alpha+\beta+\frac{\gamma}{2}}, \qquad 0 < \mu_{asy} < 1.
\end{aligned} \tag{18}$$

In the same way as in the previous section, it can be shown that unconditional variance σ^2 may be approximated by the unbiased sample variance of the returns as

$$\begin{aligned}
\mathrm{Var}(r_t) &= \mathbb{E}[r_t^2] - \mathbb{E}[r_t]^2\\
&= \mathbb{E}[\sigma_t^2\epsilon_t^2]\\
&= \mathbb{E}[\omega + (\alpha+\mathrm{I}_{t-1})r_{t-1}^2 + \beta\sigma_{t-1}^2]\\
&= \omega + (\alpha+\beta)\mathrm{Var}(r_t) + \gamma\mathbb{E}[I_{t-1}]\mathrm{Var}(r_t)\\
&= \omega + (\alpha+\beta+\frac{\gamma}{2})\mathrm{Var}(r_t),
\end{aligned} \tag{19}$$

and thus the unconditional variance is obtained as

$$\mathrm{Var}(r_t) = \frac{\omega}{1-(\alpha+\beta+\frac{\gamma}{2})} = \sigma^2, \tag{20}$$

assuming that the conditional distribution of the returns are symmetric around zero i.e. $P(r_t > 0) = P(r_t < 0) = \frac{1}{2}$. That works in the same matter as before, reducing the dimension and serving as an initializer of the process. The parameter μ_{corr} acts the same as in the case of GARCH model and the parameters μ_{ema}, μ_{asy} are somewhat more entangled and does not provide a natural explanation. Proceeding by transforming to the $\boldsymbol{z}$-space is a straightforward matter and does not add anything new except the asymmetric parameter, thus it is left out in this section.

3.2 Initial values

Since the algorithm used in this paper to solve the non-linear optimization problem of maximum likelihood estimation are based on a line search method the initial values have a big impact on speed and convergence. This is of great importance in an automated risk monitoring system. Therefore this part is about finding a good universal initial point for the algorithm, in the sense that the initial point should not only be good for specific data sets. By generating a sparse grid in the $\boldsymbol{z}$-space and evaluating the likelihood function in this grid a first maximum value is obtain. Then an iterative process is suggested by creating a new refined grid with the same grid size around the area of the recent point in the $\boldsymbol{z}$-space found to maximize the likelihood function and re-valuate the likelihood function in the new refined grid. Hence the iterations of refining the grid around the previous found maximum works as a "zoom". By iterating this procedure for some given number of iterations the $\boldsymbol{z}$ parameters found maximizing the likelihood function are passed as initial values to the optimization algorithm. Furthermore with empirical applications of GARCH volatility models the optimal $\boldsymbol{z}$ parameters is typically found around the region $1 \leq z_{corr}, z_{ema} \leq 4$ see [10], thus the first generated grid is created with this region. This method is also useful to use as a fallback solution in the worst case scenario of convergence failure, that is the case if the algorithm fails to converge after feeding it with the initial values, which should never occur but in an automatized system it is important to be on the safe side. This procedure is perhaps a bit naive since the number of evaluations of the likelihood function is growing exponentially with the number of parameters, k, with fixed number of iterations, I, and grid size N as $I \cdot N^k$. Hence this method is not suitable for large models containing many parameters. Let $\boldsymbol{\mu}$ denote the vector of parameters in the $\boldsymbol{\mu}$-space, then the outline of the procedure can be summarized to

- Create the first grid with intervals $\{\boldsymbol{\mu} \in \mathbb{R}^k | e^{-e^{-1}} \leq \boldsymbol{\mu} \leq e^{-e^{-4}}\}$ and grid size N. Evaluate the likelihood function in this grid to find the parameters maximizing the likelihood function.
- Let $\boldsymbol{\mu}^*$ denote the parameters found to maximize the likelihood function and create the refined grid around $\boldsymbol{\mu}^*$ as $[\boldsymbol{\mu}^* - \Delta, \boldsymbol{\mu}^* + \Delta]$ with the same grid size N. Where Δ is chosen to minimize the distance to the border of the $\boldsymbol{\mu}$-space i.e. $\Delta = \min(\boldsymbol{\mu}^*, 1 - \boldsymbol{\mu}^*)$.
- Evaluate the likelihood function in the new refined grid, $[\boldsymbol{\mu}^* - \Delta, \boldsymbol{\mu}^* + \Delta]$, to obtain a new maximum. Iterate the second and this step for a set number of iterations.

where likelihood evaluations are done in the $\boldsymbol{z}$-space and the grids are constructed in the $\boldsymbol{\mu}$-space to prevent under and overflows.

3.3 Maximum likelihood estimation

With the preparatory work covered in previous sections 3.1, 3.2 this section provides the mathematical background of the maximum likelihood estimation in the new parameter space $\boldsymbol{z}$. Consider the returns $r_1, ..., r_n$ to be observations of independent and identically distributed random variables $R_1, ..., R_n$ with the density function $f(R|\theta_0)$, where $\theta_0 \in \Theta$ denotes the set of true unknown parameters. Thus the random distribution would be completely characterized if we knew the set of parameters θ_0. Maximum likelihood estimation provides an estimate of the unknown parameters θ_0, as the parameter values $\theta \in \Theta$ that maximizes the probability of the observed data. Through out this paper $L(\cdot)$ and $LL(\cdot)$ denotes the likelihood function and the log-likelihood function respectively.

3.3.1 GARCH

Assume that ϵ_t in section 2.1 is normally distributed with zero mean and unit variance, then by parameter transformations and estimating σ^2 by the unbiased sample variance of the returns corresponding to the preparatory work in section 3.1.1. The parameter space is defined by

$$\Theta = \{\omega, \alpha, \beta\} \mapsto \Theta = \left\{\sigma^2, z_{corr}, z_{ema}\right\}.$$

Leaving the maximum likelihood estimation to the set of parameters, $\theta = (z_{corr}, z_{ema})$. Given the observations $r_1, ..., r_n$ the conditional distribution of r_t is defined as in equation 5, and by writing the joint distribution by the chain rule

$$f_{r_1,r_2,...,r_n}(\theta) = f_{r_n|r_1,r_2,...,r_{n-1}}(\theta) f_{r_{n-1}|r_1,r_2,...,r_{n-2}}(\theta) \cdots f_{r_1}(\theta), \tag{21}$$

the likelihood function is defined by

$$L(\theta|r_1, r_2, ..., r_n) = \prod_{t=1}^{n} \frac{1}{\sqrt{2\pi\sigma_t^2}} e^{-\frac{r_t^2}{2\sigma_t^2}}. \tag{22}$$

Since it is more convenient to work with sums than products the logarithm is applied to form the log-likelihood

$$\begin{aligned} LL(\theta|r_1, r_2, ..., r_n) &= \sum_{t=1}^{n} \left[\log\left(\frac{1}{\sqrt{2\pi\sigma_t^2}}\right) - \frac{r_t^2}{2\sigma_t^2} \right] \\ &= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{t=1}^{n}\log(\sigma_t^2) - \frac{1}{2}\sum_{t=1}^{n}\frac{r_t^2}{\sigma_t^2}, \end{aligned} \tag{23}$$

where the optimization problem that is finding the maximum of the log-likelihood function may be summarized to

$$\underset{\theta \in \Theta}{\operatorname{argmax}}\, LL(\theta | r_1, r_2, ..., r_n). \tag{24}$$

In order to solve this maximization problem the algorithm implemented in java is fed with the analytic derivatives. Where the gradient and a scheme of chain rules for the partial derivatives are obtained as

$$\nabla LL = \left(\frac{\partial LL}{\partial z_{corr}}, \frac{\partial LL}{\partial z_{ema}} \right), \tag{25}$$

$$\begin{aligned} \frac{\partial LL}{\partial z_{corr}} &= \frac{\partial LL}{\partial \sigma_t^2} \frac{\partial \sigma_t^2}{\partial z_{corr}}, \quad \frac{\partial \sigma_t^2}{\partial z_{corr}} = \frac{\partial \sigma_t^2}{\partial \mu_{corr}} \frac{\partial \mu_{corr}}{\partial z_{corr}} \\ \frac{\partial LL}{\partial z_{ema}} &= \frac{\partial LL}{\partial \sigma_t^2} \frac{\partial \sigma_t^2}{\partial z_{ema}}, \quad \frac{\partial \sigma_t^2}{\partial z_{ema}} = \frac{\partial \sigma_t^2}{\partial \mu_{ema}} \frac{\partial \mu_{ema}}{\partial z_{ema}}, \end{aligned} \tag{26}$$

where the algebra of calculating the gradient according to above are left out since it is a bit tedious and does not add anything important, although the following partial derivatives are obtained

$$\begin{aligned} \frac{\partial LL}{\partial z_{corr}} &= -\frac{1}{2} \sum_{t=1}^{n} \left[\left(\frac{1}{\sigma_t^2} - \frac{r_t^2}{\sigma_t^4} \right) \left(\mu_{ema} (\sigma_{t-1}^2 + \mu_{corr} \frac{\partial \sigma_{t-1}^2}{\partial \mu_{corr}}) + (1 - \mu_{ema}) r_{t-1}^2 - \sigma^2 \right) \mu_{corr} e^{-z_{corr}} \right], \\ \frac{\partial LL}{\partial z_{ema}} &= -\frac{1}{2} \sum_{t=1}^{n} \left[\left(\frac{1}{\sigma_t^2} - \frac{r_t^2}{\sigma_t^4} \right) \left(\sigma_{t-1}^2 + \mu_{ema} \frac{\partial \sigma_{t-1}^2}{\partial \mu_{ema}} - r_{t-1}^2 \right) \mu_{corr} \mu_{ema} e^{-z_{ema}} \right]. \end{aligned} \tag{27}$$

Now let η_t in equation 6 be scaled student's t distributed with zero mean and unit variance. Then by following the same procedure as above for the normally distributed innovations, we have the parameter space defined by

$$\Theta = \{\omega, \alpha, \beta, \nu\} \mapsto \Theta = \left\{ \sigma^2, z_{corr}, z_{ema}, z_\nu \right\}.$$

Again σ^2 is estimated by the unbiased sample variance of the returns, resulting in the set of parameters to estimate, $\theta = (z_{corr}, z_{ema}, z_\nu)$. Although a transformation to $z_\nu = \log(\nu - 2)$

is done with the sole reason to avoid the potential of dividing by zero, the z_ν values are transformed back to ν when evaluating log-likelihood and derivatives in the algorithm, therefore this transformation is ignored in the following equations. Using equation 7 with $\nu > 2$, the likelihood function is defined by

$$L(\theta|r_1, r_2, ..., r_n) = \prod_{t=1}^{n} \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\sigma_t^2(\nu-2)\pi}\Gamma(\frac{\nu}{2})} \left(1 + \frac{r_t^2}{\sigma_t^2(\nu-2)}\right)^{-\frac{\nu+1}{2}}, \tag{28}$$

where $\Gamma(\cdot)$ denotes the gamma function. Then by applying the logarithm the log-likelihood function is obtained as

$$\begin{aligned} LL(\theta|r_1, r_2, ..., r_n) &= \sum_{t=1}^{n} \log\left(\frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\sigma_t^2(\nu-2)\pi}\Gamma(\frac{\nu}{2})} \left(1 + \frac{r_t^2}{\sigma_t^2(\nu-2)}\right)^{-\frac{\nu+1}{2}}\right) \\ &= n \log\left(\frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\pi(\nu-2)}\Gamma(\frac{\nu}{2})}\right) \\ &\quad - \frac{1}{2}\sum_{t=1}^{n} \log(\sigma_t^2) + (\nu+1)\log\left(1 + \frac{r_t^2}{\sigma_t^2(\nu-2)}\right). \end{aligned} \tag{29}$$

Thus the maximization problem may be summarized to

$$\underset{\theta \in \Theta}{\operatorname{argmax}}\, LL(\theta|r_1, r_2, ..., r_n), \tag{30}$$

where the gradient is constructed by the partial derivatives

$$\nabla LL = \left(\frac{\partial LL}{\partial z_{corr}}, \frac{\partial LL}{\partial z_{ema}}, \frac{\partial LL}{\partial \nu}\right), \tag{31}$$

and the partial derivative scheme of chain rules are the same as in equation 26 except we have one more partial derivative to account for. Where we note that the conditional volatility does not depend on the degrees of freedom, ν, and thus the partial derivatives is the same as defined in equation 27 up to the factor of $\frac{\partial LL}{\partial \sigma_t^2}$. With some algebra the following partial derivatives are obtained

$$\frac{\partial LL}{\partial z_{corr}} = -\frac{1}{2}\sum_{t=1}^{n}\left[\left(\frac{1}{\sigma_t^2} - \frac{(\nu+1)r_t^2}{\sigma_t^4(\nu-2)+r_t^2\sigma_t^2}\right)\frac{\partial \sigma_t^2}{\partial z_{corr}}\right],$$
$$\frac{\partial \sigma_t^2}{\partial z_{corr}} = \left(\mu_{ema}(\sigma_{t-1}^2 + \mu_{corr}\frac{\partial \sigma_{t-1}^2}{\partial \mu_{corr}}) + (1-\mu_{ema})r_{t-1}^2 - \sigma^2\right)\mu_{corr}e^{-z_{corr}},$$

$$\frac{\partial LL}{\partial z_{ema}} = -\frac{1}{2}\sum_{t=1}^{n}\left[\left(\frac{1}{\sigma_t^2} - \frac{(\nu+1)r_t^2}{\sigma_t^4(\nu-2)+r_t^2\sigma_t^2}\right)\frac{\partial \sigma_t^2}{\partial z_{ema}}\right],$$
$$\frac{\partial \sigma_t^2}{\partial z_{ema}} = \left(\sigma_{t-1} + \mu_{ema}\frac{\partial \sigma_{t-1}^2}{\partial \mu_{ema}} - r_{t-1}^2\right)\mu_{corr}\mu_{ema}e^{-z_{ema}},$$

$$\frac{\partial LL}{\partial \nu} = \frac{n}{2}\left(\Psi\left(\frac{\nu+1}{2}\right) - \Psi\left(\frac{\nu}{2}\right) - \frac{1}{\nu-2}\right)$$
$$-\frac{1}{2}\sum_{t=1}^{n}\left[\log\left(1+\frac{r_t^2}{\sigma_t^2(\nu-2)}\right) - \frac{(\nu+1)r_t^2}{\sigma_t^2(\nu-2)^2 + r_t^2(\nu-2)}\right],$$

where $\Psi(\cdot)$ denotes the digamma function.

3.3.2 gjr-GARCH

Similarly as the procedure in section 3.3.1, provided by the log-likelihood function in equation 23 and 29 from that section and the conditional volatility defined as in equation 3, the log-likelihood function for the gjr-GARCH and gjr-sGARCH are defined. Then by following the parameter space transformation in section 3.1, the parameter space for gjr-GARCH with normally distributed residuals is defined by

$$\Theta = \{\omega, \alpha, \beta, \gamma\} \mapsto \Theta = \left\{\sigma^2, z_{corr}, z_{ema}, z_{asy}\right\}.$$

With the set of parameters to estimate, $\theta = (z_{corr}, z_{ema}, z_{asy})$. Moreover, the gradient is defined by

$$\nabla LL = (\frac{\partial LL}{\partial z_{corr}}, \frac{\partial LL}{\partial z_{ema}}, \frac{\partial LL}{\partial z_{asy}}), \tag{32}$$

with the partial derivatives obtained from following the partial derivative scheme defined in equation 26 to

$$\frac{\partial LL}{\partial z_{corr}} = -\frac{1}{2}\sum_{t=1}^{n}\left[\left(\frac{1}{\sigma_t^2} - \frac{r_t^2}{\sigma_t^4}\right)\frac{\partial \sigma_t^2}{\partial z_{corr}}\right],$$
$$\frac{\partial \sigma_t^2}{\partial z_{corr}} = \left((1-\mu_{ema})r_{t-1}^2 + 2(\mu_{ema}-\mu_{asy})\mathrm{I}_{t-1}r_{t-1}^2 + \mu_{asy}\sigma_{t-1}^2 + \mu_{corr}\mu_{asy}\frac{\partial \sigma_{t-1}^2}{\partial \mu_{corr}} - \sigma^2\right)\mu_{corr}e^{-z_{corr}},$$

$$\frac{\partial LL}{\partial z_{ema}} = -\frac{1}{2}\sum_{t=1}^{n}\left[\left(\frac{1}{\sigma_t^2} - \frac{r_t^2}{\sigma_t^4}\right)\left((2\mathrm{I}_{t-1}-1)r_{t-1}^2 + \mu_{asy}\frac{\partial \sigma_{t-1}^2}{\partial \mu_{ema}}\right)\mu_{corr}\mu_{ema}e^{-z_{ema}}\right],$$

$$\frac{\partial LL}{\partial z_{asy}} = -\frac{1}{2}\sum_{t=1}^{n}\left[\left(\frac{1}{\sigma_t^2} - \frac{r_t^2}{\sigma_t^4}\right)\left(\sigma_{t-1}^2 - 2\mathrm{I}_{t-1}r_{t-1}^2 + \mu_{asy}\frac{\partial \sigma_{t-1}^2}{\partial \mu_{asy}}\right)\mu_{corr}\mu_{asy}e^{-z_{asy}}\right].$$

For the student's t distributed residuals we have the addition of the parameter ν to the parameter space, which is then defined by

$$\Theta = \{\omega, \alpha, \beta, \nu, \gamma\} \mapsto \Theta = \left\{\sigma^2, z_{corr}, z_{ema}, z_\nu, z_{asy}\right\}.$$

Thereby the set of parameters to estimate, $\theta = (z_{corr}, z_{ema}, z_\nu, z_{asy})$. Furthermore the only difference from the partial derivatives obtained with normally distributed residuals except for the addition of a new parameter, is the factor $\frac{\partial LL}{\partial \sigma_t^2}$ which is given by

$$\frac{\partial LL}{\partial \sigma_t^2} = \left(\frac{1}{\sigma_t^2} - \frac{(\nu+1)r_t^2}{\sigma_t^4(\nu-2) + r_t^2\sigma_t^2}\right), \tag{33}$$

and the partial derivative of the log-likelihood function w.r.t the degrees of freedom ν is given by

$$\frac{\partial LL}{\partial \nu} = \frac{n}{2}\left(\Psi\left(\frac{\nu+1}{2}\right) - \Psi\left(\frac{\nu}{2}\right) - \frac{1}{\nu-2}\right) - \frac{1}{2}\sum_{t=1}^{n}\left[\log\left(1 + \frac{r_t^2}{\sigma_t^2(\nu-2)}\right) - \frac{(\nu+1)r_t^2}{\sigma_t^2(\nu-2)^2 + r_t^2(\nu-2)}\right].$$

3.3.3 Conjugate gradient method

There are various ways to solve the optimization problem of maximum likelihood estimation, in this paper the optimization problem is solved by using the non-linear conjugate gradient method. With the advantage over Newton-based methods of only requiring gradient computations, while Newton-based methods also requires the computational expensive Hessian. The optimization problem may be described by letting $f(x) = -LL(x)$ as

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in \mathbb{R}^n, \end{array} \tag{34}$$

and the non-linear conjugate gradient method is then defined by the iterative scheme,

$$x^{k+1} = x^k + \alpha^k d^k, \tag{35}$$

where α^k is the step size and d^k is the search direction. α^k is obtained by a line minimization, that is

$$f(x^k + \alpha^k d^k) = \underset{\alpha}{\text{minimize}} f(x^k + \alpha^k d^k), \tag{36}$$

and d^k by the following

$$d^k = -\nabla f(x^k) + \beta^k d^{k-1}. \tag{37}$$

There are various ways of determine β^k but the most common ones are Polak–Ribière

$$\beta^k = \frac{\nabla f(x^k)'(\nabla f(x^k) - \nabla f(x^{k-1}))}{\nabla f(x^{k-1})'\nabla f(x^{k-1})}, \tag{38}$$

and Fletcher–Reeves

$$\beta^k = \frac{\nabla f(x^k)'\nabla f(x^k)}{\nabla f(x^{k-1})'\nabla f(x^{k-1})}. \tag{39}$$

4 On the assumption of i.i.d. residuals

The work done in this paper heavily relies on the assumption that the residuals are i.i.d. random variables. In order to test this assumption the sample autocorrelation function, referred to as ACF, is used through out this paper. Let $\{x_1, ..., x_n\}$ be observations of a time series with zero mean. Then by [3] the sample autocorrelation function is defined by

$$\begin{aligned} \widehat{\gamma}(h) &= \frac{1}{n}\sum_{t=1}^{n-|h|} x_{t+|h|}x_t, \qquad -n < h < n, \\ \widehat{\rho}(h) &= \frac{\widehat{\gamma}(h)}{\widehat{\gamma}(0)}, \qquad -n < h < n, \end{aligned} \tag{40}$$

where h is the lag, $\widehat{\gamma}(h)$ is the sample autocovariance function and $\widehat{\rho}(h)$ is the sample autocorrelation function. the sample autocorrelations of a large sample of an i.i.d. sequence are approximately i.i.d. $N(0, 1/n)$. Thereby, the 95% confidence interval of the sample autocorrelations can then be computed as $\left[\frac{-1.96}{\sqrt{n}}; \frac{+1.96}{\sqrt{n}}\right]$. Consequently, for any sample of an i.i.d. sequence, approximately 95% of the sample autocorrelations should reside within the confidence interval, or conversely, no more than 5% should exceed the interval boundaries. Hence, the null hypothesis that the sample are i.i.d. is rejected if more than 5% of the sample autocorrelations fall outside the confidence interval. In this paper the sample ACF is used to check for independence in the residuals and typically normally distributed residuals are assumed, which can be tested with the normal Q-Q plot. If the assumption of normally distributed residuals is appropriate the normal Q-Q plot should approximate a straight line through the origin with slope one.

5 Forecast evaluation

To measure the forecast accuracy of the employed conditional volatility models some kind of out-of-sample testing is needed. In-sample performance are all fine and dandy but in reality it is the out-of-sample performance that are of most interest. Thus the out-of-sample performance of the models are obtained by constructing a backtest. This is done by using a rolling window. Let n be the sample size and create a window of size $w < n$. Now by letting the first window contain the first w data points of the sample and use this data to estimate the parameters of the GARCH model, a h-day-ahead forecast can be made. Proceeding by letting this window of size w "roll" through the sample, by letting the second window contain the second data point up through $w + 1$ and so on. Thus using

one-day-ahead forecast, i.e $h = 1$, the data set will be partitioned into $n_s = n - w + 1$ subsets although the last subset is redundant since we do not have market observations to measure the forecast accuracy, hence $n - w$ subsets are used. This procedure makes it possible to construct different forecast accuracy measures since the out-of-sample data are known but not used in the parameter estimation of the models. The procedure may be summarized as

For each subset:

- estimate model parameters based on data of the current subset
- estimate the h-day-ahead forecast
- calculate forecast measures

Where the various forecast measures are defined in the sections below.

5.1 Latent volatility

A mayor issue with volatility or variance forecast evaluation is the fact that the true underlying volatility is not observable, hence to construct a forecast accuracy measure a proxy of the latent volatility has to be estimated. Historically the latent volatility has been derived by using squared daily returns or other estimates where daily prices are used, however this has lead to the assumption that GARCH type models does not produce accurate volatility forecasts. In [1], Torben G.andersen and Tim Bollerslev argued and demonstrated that this is not the case and that the poor volatility forecast accuracy was due to the method of extracting latent volatility, where the proxies for the latent volatility was too noisy, thus they propose the use of high frequency data to extract the latent volatility. The realized variance using intraday returns, assuming there are m number of observations per trade day, are defined by

$$\widehat{\sigma}^2_{t+1} = RV^{(m)}_{t+1} = \sum_{i=1}^{m} r^2_{i,m,t+1}, \tag{41}$$

however using high frequency data can be computational heavy and complex market microstructure effects arise. Also high frequency data is costly and hence not accessible for this paper, however the intraday high low prices are typically accessible. Therefore the unbiased proxy proposed by Parkinson in [7], based on intraday high low prices are used. The estimate of latent variance based on high low prices is defined by

$$\widehat{\sigma}^2_{t+1} = \frac{1}{4\log 2}(\log\left(H_{t+1}/L_{t+1}\right))^2. \tag{42}$$

Now to evaluate the forecast accuracy of the conditional volatility models, the robust and homogeneous loss functions, MSE and QLIKE, proposed by Patton in [8] is employed. That is,

$$\begin{aligned} \text{MSE}: \quad & L(\widehat{\sigma}^2_{t+1}, \widehat{h}^2_{t+1}) = (\widehat{\sigma}^2_{t+1} - \widehat{h}^2_{t+1})^2, \\ \text{QLIKE}: \quad & L(\widehat{\sigma}^2_{t+1}, \widehat{h}^2_{t+1}) = \frac{\widehat{\sigma}^2_{t+1}}{\widehat{h}^2_{t+1}} - \log \frac{\widehat{\sigma}^2_{t+1}}{\widehat{h}^2_{t+1}} - 1, \end{aligned} \tag{43}$$

where $\widehat{\sigma}^2_{t+1}$ denotes the latent variance proxy defined by equation 42 and $\widehat{h}^2_{t+1}$ denotes the conditional variance forecast. Furthermore since the MSE loss function will have a variance proportional to the square of the variance of returns, i.e. σ^4, it is somewhat sensitive to extreme observations while the QLIKE loss function does not inherit this property due to the use of standardized error $(\frac{\widehat{\sigma}^2_{t+1}}{\widehat{h}^2_{t+1}})$.

5.2 Value-at-Risk

Value-at-Risk, referred to as VaR in this paper, is a widely used risk measure for financial applications. Although there are some shortcomings of this risk measure, such as hiding large potential losses in the tails and it does not encourage diversification since it is not a subadditive risk measure, it has the strength of being intuitive and backtestable. In this paper the VaR is calculated on losses to keep it intuitive. Now given the distribution of the potential loss L, Value-at-Risk at level p is simply the $(1-p)$ quantile of the potential loss, see [6], that is

$$\text{VaR}_p(X) = F_L^{-1}(1-p), \qquad p \in (0,1), \tag{44}$$

where X is the value of a portfolio, $L = -X/R_0$ with the natural interpretation as the discounted loss where R_0 is the percentage return of a risk-free asset. The interpretation of VaR is that the probability of a loss greater than $\text{VaR}_p(X)$ is p. Further the assumed distributions in previous section 3.3.1 are only used to estimate the parameters of the GARCH models and for the VaR calculations in this paper the empirical counter part are used, defined in section 5.3, and the percentage return of a risk-free asset is for simplicity assumed to be zero i.e. $R_0 = 1$. Furthermore a violation or breach of a one-day-ahead VaR forecast, meaning that the actual loss is larger than the one-day-ahead VaR forecast, may be described by the indicator function defined by

$$\text{I}_{t+1} = \text{I}(l_{t+1} > \text{VaR}_p^{t+1|t}), \tag{45}$$

where l_{t+1} is the loss and the probability of a loss exceeding $\text{VaR}^p_{t+1|t}$ is p, that is

$$P(\text{I}_{t+1}|\mathcal{F}_t) = p, \tag{46}$$

and thus the expected number of breaches is defined by

$$\mathbb{E}[\text{I}_{t+1}|\mathcal{F}_t] = p \cdot 1 + (1-p) \cdot 0 = p. \tag{47}$$

With this the expected number of breaches for a backtest with n numbers of VaR forecasts is therefore given by

$$\sum_{i=1}^{n} \mathbb{E}[\text{I}_{t+1}|\mathcal{F}_t] = np. \tag{48}$$

5.3 Empirical distribution

Let $\{X_1, X_2, ..., X_n\}$ be a sample of i.i.d. random variables with some unknown distribution function F_X, then the empirical distribution $F_{n,X}$ function may be approximated by assigning a probability weight of $1/n$ to each of the $X_k \leq x$, that is

$$\widehat{F}_{n,X} = \frac{1}{n} \sum_{i=1}^{n} \text{I}\{X_k \leq x\}. \tag{49}$$

Furthermore by ordering the sample of $\{X_1, X_2, ..., X_n\}$ such that $X_{1,n} \geq \cdots \geq X_{n,n}$ the empirical quantile estimate can be shown to be given by

$$\widehat{F}^{-1}_{n,X}(p) = X_{\lfloor n(1-p) \rfloor + 1, n}, \tag{50}$$

where $\lfloor \cdot \rfloor$ denotes the floor function. Hence the empirical counterpart of VaR in equation 44, given a sample $\{L_1, ..., L_n\}$ of independent copies of L, ordered as $L_{1,n} \geq \cdots \geq L_{n,n}$ is thus given by

$$\widehat{\text{VaR}}_p(X) = \widehat{F}^{-1}_{n,L}(1-p) = L_{\lfloor np \rfloor + 1, n}, \tag{51}$$

and therefore the $\widehat{\text{VaR}}_p(X)$ is simply the $\lfloor np \rfloor + 1$ worst outcome of the loss sample.

6 Empirical studies

This section covers the empirical studies, consisting of two different data sets, where the data sets are chosen for different reasons. The first data set contains the Brent crude oil spot price and the second contains front month lumber futures price. Front month refers to the shortest duration contract that can be purchased.

6.1 Brent crude oil

For the sake of testing the algorithm the first data set containing the ten year daily spot price of Brent crude oil, presented by the left plot in Figure 1, was chosen on the basis that crude oil is one of the more volatile financial instruments on the market. The data set consist of a total number of 2524 daily historical spot prices that are converted to log returns by $r_t = \log P_t/P_{t-1}$, furthermore the returns are mean adjusted to contain zero mean. By inspection it is evident that the data exhibit volatility clustering, that is the time series exhibit conditional heteroscedasticity and by inspection of the autocorrelation function in Figure 10a in Appendix A.3.1, a significant autocorrelation in the squared log returns are evident, indicating that volatility clustering is present.

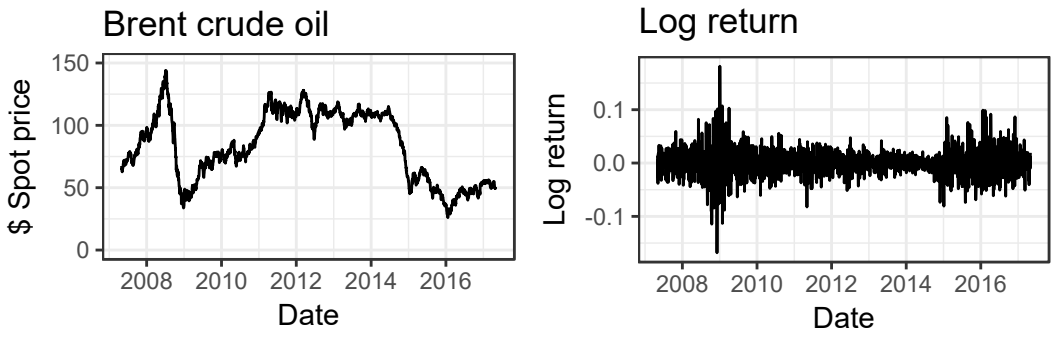
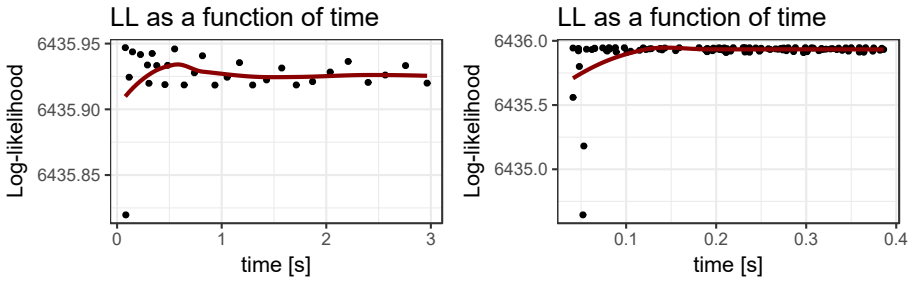


Figure 1: The plot to the left depicts ten year Brent crude oil daily spot price and to the right the log return, dating from 2007-05-01 to 2017-05-01.

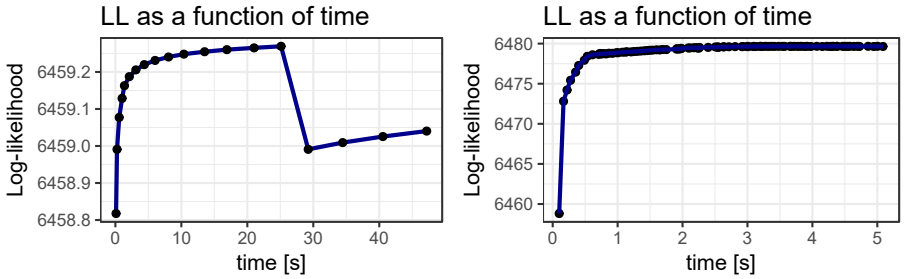
6.1.1 Robustness of solution

It is of no secret that optimization of GARCH parameters by maximum likelihood estimation is numerically difficult due to spurious solutions leading to sub-optimal solutions. Thus the robustness of the solution is examined by using different initial values according to 3.2, where a GARCH model is fed with the initial values and fitted to the data set with normally distributed and student's t distributed residuals. Figures 2a and 2b depicts the increase in accuracy in the sense of log-likelihood values by using the method in 3.2 for choosing the initial values. Worth mentioning is that without using the method for initial values the fit with normally distributed residuals a log-likelihood of 6435.51 is achieved while with student's t distributed residuals the algorithm struggles to converge.

The left hand plots in Figures 2a and 2b corresponds to log-likelihood as a function of time where each point represents a grid size increasing as $(10 \times 10,\ 20 \times 20,\ ...)$ for the two dimensional space (z_{corr}, z_{ema}) with normally distributed residuals, and $(10 \times 10 \times 10,\ 15 \times 15 \times 15,\ ...)$ for the three dimensional space $(z_{corr}, z_{ema}, z_\nu)$ with student's t distributed residuals and no iterations of refining grid around previous found maximum are done in these plots. The right hand plots corresponds to holding a constant grid size of (10×10) and $(10 \times 10 \times 10)$ for respectively parameter spaces and each point represents increased number of iterations of refining grid around previous found maximum, starting with no iteration and increasing with one for each point. The Figures shows that the solution are dependent of the initial values, meaning that both increasing grid size and number of iterations introduces new maximums. Where the iteration method clearly outperforms increasing grid size both in time and log-likelihood values, although they don't differ that much in the case of normally distributed residuals, they differ by a significant amount in the case of student's t residuals. Due to the trade-of problem of accuracy vs speed, a grid size of (20×20) for (z_{corr}, z_{ema}) and $(20 \times 20 \times 20)$ for $(z_{corr}, z_{ema}, z_\nu)$ with 20 iterations is proposed for this data set in order to preserve speed and obtaining good enough accuracy.



(a)



(b)

Figure 2: (a) Log-likelihood as a function of time with normally distributed residuals, Left plot: increasing grid size and no iterations, right plot: increasing number of iterations while holding grid size constant. (b) same as plot (a) for student's t distributed residuals.

6.1.2 Sensitivity analysis

Let the vector $\boldsymbol{\theta^*}$ denote the optimal solution found by using the method proposed in the section above. Since the global optimal solution is most certainly not going to be obtained by the algorithm a sensitivity analysis of $\boldsymbol{\theta^*}$ is conducted. By looking at the normalized sensitivity describing the percentage difference in log-likelihood value per percentage difference in parameter variable defined by

$$\frac{\frac{\Delta LL}{LL}}{\frac{\Delta \theta_i}{LL_i}} = \frac{\theta_i^*}{LL(\boldsymbol{\theta^*})} \frac{\partial LL}{\partial \theta_i}\bigg|_{\boldsymbol{\theta^*}}, \tag{52}$$

in order to capture relative sensitivity. Figure 3 depicts the relative sensitivities for each employed model, showing that the log-likelihood value is more sensitive to the parameter z_{ema} for the GARCH model with normally and student's distributed residuals and z_{asy} for the gjr-GARCH model. Further by looking at the gradient of the log-likelihood function evaluated at the optimal solution $\boldsymbol{\theta^*}$ presented in Table 5 in Appendix A.4, it is obvious that the gradient does not equal to zero in any of the employed optimizations, indicating that a spurious solution is found. There is various reasons of why spurious optimal solutions may arise, and to name a few:

- the log-likelihood function is almost flat around the optimal solution $\boldsymbol{\theta^*}$, resulting in algorithm stopping before reaching global optimum due to stopping criteria.
- the log-likelihood function is in fact not "well behaved" around the optimal solution $\boldsymbol{\theta^*}$, meaning that the function is not approximately quadratic near the optimum and thus the gradient are in fact not equal to zero at the optimal solution.
- the step size of the algorithm is to large, meaning that the global optimum is not reachable.

where the first case is the most probable since the percentage differences in log-likelihood value per percentage difference in parameter variable are small, see Figure 3. Since the number of parameters of GARCH models tend to be high it is hard to visualize the log-likelihood function, although with normally distributed residuals we can visualize it in the z-space. In Figure 4a and 4b the log-likelihood surface around the found optimal solution $\boldsymbol{\theta^*}$ is presented, where an almost flat "ridge" in the z_{corr} direction is apparent, containing spurious solutions making it complicated to locate the global optimum. In order to understand how big impact these spurious solutions has on the conditional volatility process, the partial derivative of the sample average of the conditional variance σ_t^2 with respect to the parameters, given the information $\mathcal{F}_{t-1}$, are defined by

$$\frac{\partial < \sigma_t^2 >}{\partial \theta_i} = \frac{1}{n-1} \sum_{t=1}^{n} \frac{\partial \sigma_t^2}{\partial \theta_i}, \tag{53}$$

where $< \cdot >$ denotes the sample average and the resulting partial derivatives evaluated at the found optimal solution $\boldsymbol{\theta}^*$ are presented in Table 6 in Appendix A.4. These results indicate that the conditional volatility process does not vary much within this "ridge" and the found solution are for the applications of this paper considered as good enough.

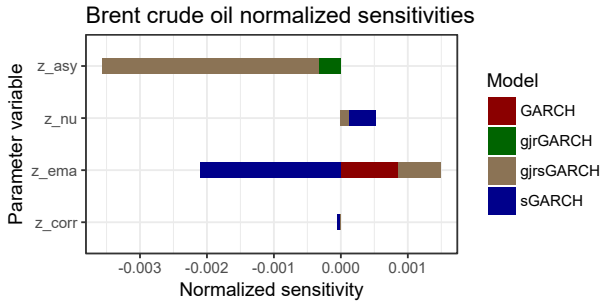
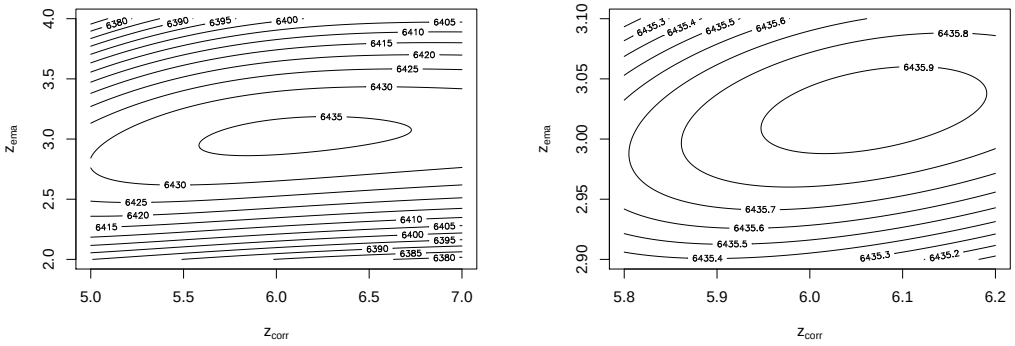


Figure 3: Normalized sensitivities for each employed conditional volatility model.



(a) (b)

Figure 4: Log-likelihood as a function of z_{corr} and z_{ema} for the GARCH model with normally distributed residuals.

6.1.3 In-sample fit

In Table 1 the parameter estimates are presented, where also for a soundness check the solutions obtained when using the *Rugarch* package in *R* are included. The *Rugarch* package provides various optimization algorithms and in this paper, *"solnp"* are used, that is based on the augmented Lagrangian method. Although the values obtained in *R* is of no guarantee to be the optimal solution and the definition of the likelihood function may vary.

For in-sample fit the Akaike information criterion (AIC), defined in Appendix A.1, is provided for model selection purposes and accordingly the gjr-GARCH model with student's t distributed residuals are preferred by (AIC) due to the lowest value of -5.141. Thus the data set seems to contain the stylized facts of fat tails and that bad news impacts the volatility more than good news. The fat tails can be seen in the normal Q-Q plot presented in Figure 11a in Appendix A.3.1 and the estimated positive values of γ indicates asymmetry between negative and positive shocks of same magnitude.

Let $< \cdot >$ denote the sample average, the difference between the sample variance of the data set and the conditional variance process are presented as $< r_t^2 > - < \sigma_t^2 >$, where the GARCH model yields the lowest absolute value of $1.882E-6$ corresponding to a difference of 0.3% between them. Figure 5 depicts the conditional volatility process with superimposed absolute value of the log returns for each employed conditional volatility model, where we see that the employed conditional volatility models capture the changes in the returns.

Furthermore the log returns are filtered with the conditional volatility process σ_t of the employed models to obtain the residuals $\{\frac{r_t}{\sigma_t}\}$ and investigate the assumption off i.i.d. residuals. Figure 13a through 13d in section A.3.1 presents the ACF for the residuals obtained with the filters. By inspection the resulting autocorrelation is not zero for all lag $h \neq 0$, however it is reasonably close with 3 sample autocorrelations falling outside the confidence bounds where the expected number of sample autocorrelations to fall outside the confidence boundaries would be $35 * 0.05 \approx 2$. Thus the null hypothesis that the residuals are i.i.d. are theoretically rejected. Although in this context the sample autocorrelations are considered to be reasonably close to zero and therefore the residuals are considered to be independent. Where the normal Q-Q plot in Figure 11a in Appendix A.3.1 showed the need of a more heavy-tailed distribution than the normal distribution.

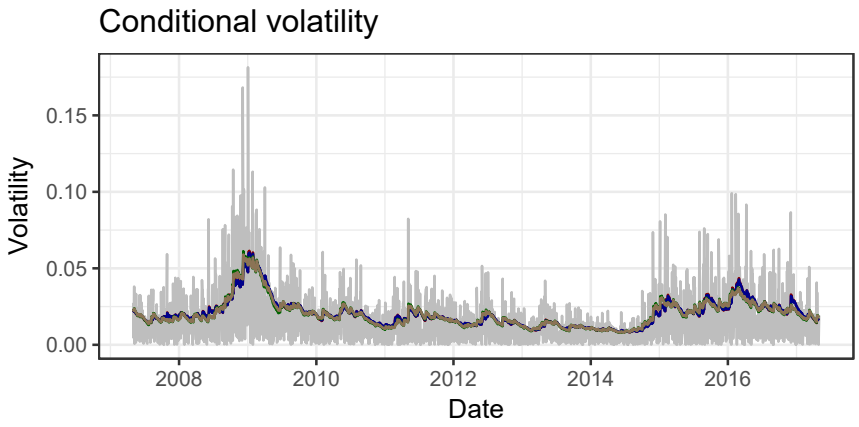


Figure 5: Conditional volatility with superimposed absolute value of log returns in grey. Although it is hard to distinguish the processes, red line corresponds to GARCH, blue for sGARCH, green for gjr-GARCH and brown for gjr-sGARCH.

			$\boldsymbol{\theta^*}$						
	ω	α	β	ν	γ	LL	AIC	$< r_t^2 > - < \sigma_t^2 >$	t_{run} (s)
GARCH	1.121e-6	0.04722	0.9505	-	-	6435.946	-5.097	-1.882e-6	0.608
	(1.042e-6)	*(0.04910)*	*(0.9500)*	-	-	*(6436.265)*	-	-	-
sGARCH	1.054e-6	0.04239	0.9555	7.209	-	6479.662	-5.131	-2.044e-6	7.347
	(1.005e-6)	*(0.04404)*	*(0.9550)*	*(6.997)*	-	*(6479.809)*	-	-	-
gjr-GARCH	9.690e-7	0.01361	0.9561	-	0.05656	6459.568	-5.115	4.743e-6	7.179
	(9.272e-7)	*(0.01408)*	*(0.9555)*	-	*(0.05876)*	*(6459.908)*	-	-	-
gjr-sGARCH	8.438e-7	0.01248	0.9615	7.480	0.04848	6492.929	-5.141	4.387e-6	61.93
	(7.755e-07)	*(0.01265)*	*(0.9612)*	*7.877*	*(0.05029)*	*(6493.237)*	-	-	-

Table 1: Parameter estimates of each conditional volatility model employed, the letter, s, in the model names indicates student's t distributed residuals and the Akaike information criterion (AIC) is also provided. Values in parentheses are values obtained by fitting the conditional volatility models in R. The run time of the algorithm are provided as t_{run}.

6.1.4 Forecast accuracy

Since the spot price of Brent crude oil is provided daily by *U.S. Energy Information Administration* and are used as a benchmark for other financial instruments such as futures, high-frequency data or high low prices are not available.

Hence the perhaps more intuitive approach by backtesting Value-at-risk are conducted as described in section 5. Now consider a portfolio consisting of a single contract on a barrel of crude oil with spot price P_t at time t, with a portfolio value tomorrow of $P_{t+1} - P_t$.

Then the empirical estimate of the one-day-ahead VaR forecast for the subsets with size w of the partitioned data set are estimated by creating independent copies of the potential future loss L at time $t+1$, defined by

$$L_{t+1} = -(P_{t+1} - P_t) = -P_t(e^{R_{t+1}} - 1) = g(R_{t+1}). \tag{54}$$

Thus, under the assumption that the observations r_{t-j}, for $j = 0, ..., w-1$ are independent copies of R_{t+1}, the subsets of independent copies of L are created by

$$l_{t-j} = -P_t(e^{r_{t-j}} - 1), \qquad \text{for} \quad j = 0, ..., w-1, \tag{55}$$

and the losses are further assumed to follow the model

$$l_t = \sigma_t z_t, \tag{56}$$

where the set $\{z_t\}$ is assumed to be i.i.d. with zero mean and unit variance with the unknown distribution approximated by the empirical distribution $\widehat{F}_{n,Z}$. Given the information $\mathcal{F}_t$, the conditional volatility process σ_t is estimated by the GARCH models for each of the subsets of historical losses and the potential loss L at time $t+1$ is then given by

$$L_{t+1|t} = \sigma_{t+1|t} Z, \qquad \text{for} \quad t = w, ..., n-1. \tag{57}$$

Thus by using equation 51 and filtering the historical losses with the conditional volatility process σ_t (to obtain the residuals $\{\frac{l_t}{\sigma_t}\}$) and ordering them such that $Z_{1,n} \geq \cdots \geq Z_{n,n}$, the empirical estimate of the one-day-ahead VaR forecast for the subsets are obtained by equation 51 and Proposition 1 in Appendix A.2 as

$$\widehat{\mathrm{VaR}}_p^{t+1|t} = \sigma_{t+1|t} \widehat{F}_{n,Z}^{-1}(1-p) = \sigma_{t+1|t} Z_{\lfloor np \rfloor + 1, n}, \qquad \text{for} \quad t = w, ..., n-1. \tag{58}$$

Although the empirical cdf and the quantile of the filtered losses are re-evaluated for each subset, the right tail of the empirical cdf corresponding to the full data set of filtered losses, filtered with the employed conditional volatility models are presented in Figure 12 in Appendix A.3.1 to get a less cluttered visualization of the empirical cdf.

The backtest is then conducted by evaluating $\mathrm{VaR}_{0.01}^{t+1|t}$ by the procedure described above, and the resulting one-day-ahead, 1% VaR forecast for each employed conditional volatility model is presented in Figure 6 with superimposed losses. Clearly, by inspection, each employed model yielded fairly accurate VaR forecasts. With a sample size of $n = 2524$ and

a window size of $w = 1000$, the number of expected breaches of the backtest according to equation 48, are obtained as $n_s p = 1524 \cdot 0.01 \approx 15$. Let ϵ denote the difference between the number of breaches acquired from the backtest and the expected number of breaches, that is

$$\epsilon = \sum_{t=1}^{n_s} I_t - n_s p. \tag{59}$$

Which is summarized in Table 2, with the gjr-GARCH model yielding the lowest of 14 breaches. Furthermore the sample ACF of the squared losses and the filtered losses are presented in Figures 10b and 14a through 14d in Appendix A.3.1. With the interpretation that the losses exhibit volatility clustering and that the assumption of i.i.d. filtered losses can not be rejected.

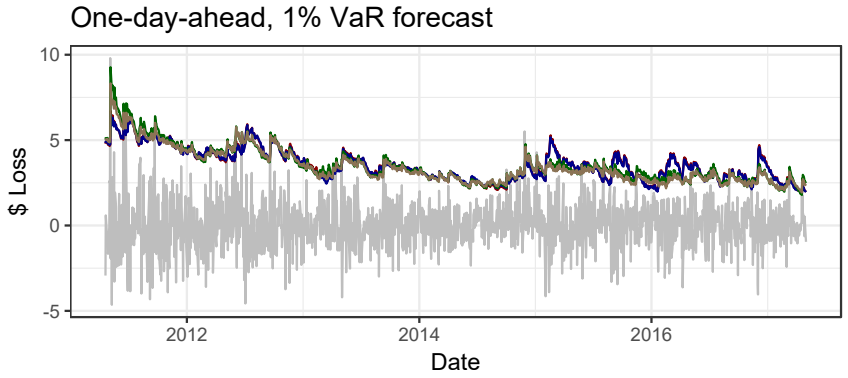


Figure 6: One-day-ahead, 1 % VaR forecast for each employed conditional volatility model with superimposed losses in grey. Red line for GARCH, blue for sGARCH, green for gjr-GARCH and brown for gjr-sGARCH.

	#	ϵ
GARCH	15	0
sGARCH	15	0
gjr-GARCH	14	-1
gjr-sGARCH	15	0

Table 2: Number of breaches are denoted by # and ϵ is the difference from the expected breaches

6.2 Lumber

The lumber data set contains the two year front month daily futures price of lumber, where left plot in Figure 7 depicts the historical daily futures price and the right plot depicts the mean adjusted log returns. Lumber futures market is highly illiquid and the average daily trading volume of the data set is 375 contracts with a minimum as low as 6 traded contracts. illiquid markets are subject to large bid-ask spreads, high volatility etc. Hence this data set is chosen to evaluate how the algorithm perform on illiquid instruments. Furthermore the data set consists of 505 historical daily futures prices. By inspection of Figure 7 it is not easily seen if the time series exhibit volatility clustering and Figure 15a in section A.3.2 shows that the sample autocorrelations of the squared log returns are not significantly different from zero. Hence this data set may not be the ideal for GARCH models, however the algorithm should yield reasonable results anyway.

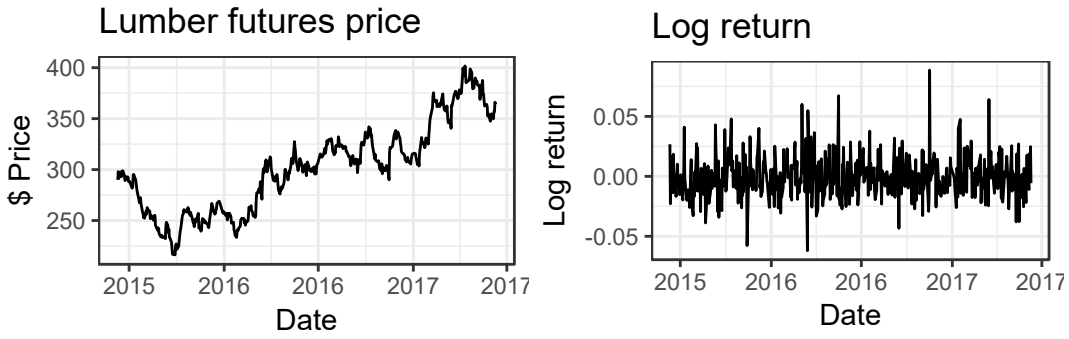


Figure 7: The left plot depicts the two year lumber daily futures price and the right plot depicts the log returns, dating from 2015-06-09 to 2017-06-09.

6.2.1 In-sample fit

Figure 8 depicts the conditional volatility process of each GARCH model with superimposed absolute value of the log returns. The spikes in the data set seen in Figure 8 is probably caused by large bid-ask spreads where the models seems to capture this pretty well. Otherwise the volatility process is almost constant due to the lack of clustering. Furthermore the parameter estimates of each model are presented in Table 3, where the sGARCH obtained the lowest AIC value of -5.327 and gjr-sGARCH yielded the lowest difference in sample variance of $4.546e-7$, corresponding to 0.15% of the data sample

variance. Thus this data set also seems to exhibit fat tails, where the normal Q-Q plot in Figure 16a found in Appendix A.3.2 confirms this.

The i.i.d. assumption of the residuals are tested similarly as in the crude oil section. The ACF for the various filters are presented in Figure 18a through 18d in section A.3.2. Where the expected number of atoucorrelations to fall outside the confidence bounds are $27*0.05 \approx 1$ and the obtained autocorrelations to fall outside the boundaries is 2. Thus the null hypothesis that the residuals are i.i.d. is theoretically rejected. Although in this context the sample autocorrelations are considered to be reasonably close to zero and therefore the null hypothesis is not rejected and the residuals are considered as approximately i.i.d. random variables.

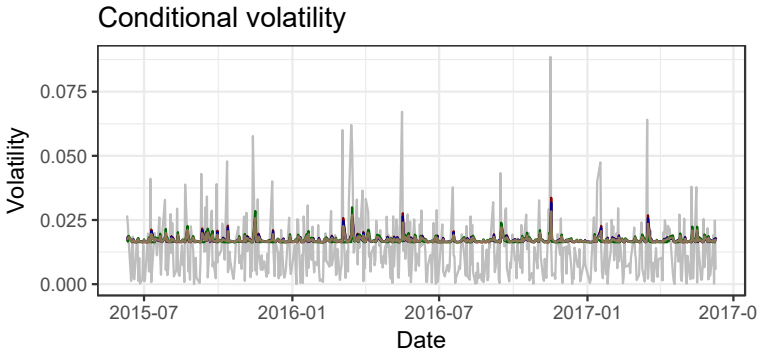


Figure 8: Conditional volatility with superimposed absolute value of log returns in grey. Where red line corresponds to GARCH, blue for sGARCH, green for gjr-GARCH and brown for gjr-sGARCH.

	$\boldsymbol{\theta^*}$								
	ω	α	β	ν	γ	LL	AIC	$< r_t^2 > - < \sigma_t^2 >$	t_{run} (s)
GARCH	2.629e-4	0.1104	0.01064	-	-	1333.445	-5.279	-5.870e-7	0.309
sGARCH	2.709e-4	0.09433	9.433e-7	6.350	-	1346.655	-5.327	-5.888e-7	2.244
gjr-GARCH	2.014e-4	3.268e-6	0.2449	-	0.1635	1333.701	-5.276	3.581e-6	2.269
gjr-sGARCH	2.541e-4	0.06806	0.05700	6.150	0.05078	1346.697	-5.324	4.546e-7	12.70

Table 3: Parameter estimates of all conditional volatility models employed, the letter, s, in the model names indicates student's t distributed residuals and the Akaike information criterion (AIC) is also provided. The run time of the algorithm are provided as t_{run}.

6.2.2 Forecast accuracy

Evaluation of the one-day-ahead forecast accuracy of the models are conducted with both the Parkinson estimate of the latent variance and VaR at level $p = 0.01$. The backtest is constructed by using a window size of $w = 200$, thus the backtest consists of 304 one-day-ahead forecasts. Figure 9 depicts the VaR forecasts and Table 4 presents the sample average of the loss functions defined in section 5.1 and the obtained number of VaR breaches. Also the right tail of the empirical cdf corresponding to the full historical set of filtered losses are presented in Figure 17 in Appendix A.3.2. A check on the independence of the residuals are provided by the sample ACF in Figure 19a through 19d which are considered as not significantly different from zero and therefore considered as independent. The expected number of breaches by using Equation 48 is 3, thus each employed model yields less breaches than the expected. This was however an "easy" loss sample for the VaR forecast since the loss sample does not contain large fluctuations. For the variance forecast measures we see that the gjr-sGARCH performs best both in the sense of MSE and QLIKE.

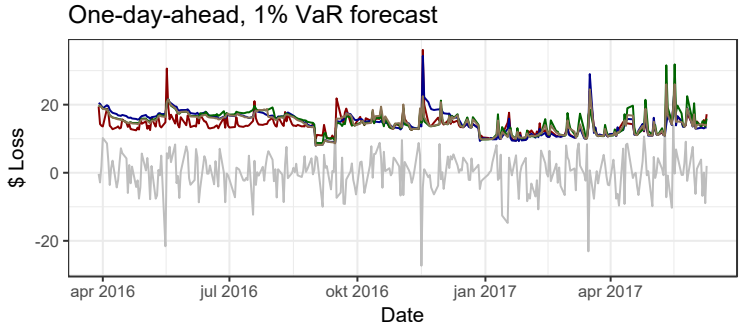


Figure 9: One-day-ahead, 1 % VaR forecast for each employed conditional volatility models with superimposed losses in grey. Red line for GARCH, blue for sGARCH, green for gjr-GARCH and brown for gjr-sGARCH.

	MSE	QLIKE	#	ϵ
GARCH	6.719e-08	0.4426	2	-1
sGARCH	6.611e-08	0.4453	2	-1
gjr-GARCH	5.358e-08	0.4440	2	-1
gjr-sGARCH	4.675e-08	0.4350	2	-1

Table 4: Sample average of the loss functions MSE and QLIKE for each employed conditional volatility model. Also the number of VaR breaches are provided as # and the difference between obtained number of breaches and the expected noted by ϵ.

7 Conclusions

This paper provides the framework for a robust algorithm for automatizing the GARCH parameter estimation. Where a parameter space transformation from the original parameter space to the $\boldsymbol{z}$-space is conducted, resulting in a more robust algorithm. Meaning that the dimension of the estimation problem is reduced, relaxed constraints and also the mean of the conditional variance process σ_t^2 is approximated by the sample variance of the data, thus the long term conditional variance is bounded by the sample variance of the data. Then to deal with the potential problem of convergence, a method to obtain initial values for the algorithm is proposed in Section 3.2, that is also used as a fallback solution in the worst case scenario if the algorithm fails to converge. The robustness of the solution is investigated in Section 6.1.1, where it is evident that the accuracy of the solution obtained when solving the maximum likelihood optimization problem with the nonlinear conjugate gradient method depends on the initial values. The causes for this was found to be the complex surface of the maximum likelihood, see Section 6.1.2. The approach in this paper is to find a "good enough" solution for the applications intended for the algorithm. Hence the grid size of the method for finding initial values to k number of parameters in the $\boldsymbol{z}$-space, was proposed as 20^k and 20 iterations of refining the grid. Where this scheme yielded no convergence failures and good enough accuracy versus speed of the algorithm.

This paper includes testing the algorithm on two different data sets and the algorithm was further tested on various types of data sets not included in this paper. In-sample fit and out-of-sample forecast performance are evaluated on both data sets. Where the conclusion is that there is no clear winner, however AIC seems to prefer the models with student's t distributed residuals. One can also see that the gjr-sGARCH model overall yielded the best results. Indicating that the data sets considered in this paper exhibits the stylized facts of fat tails and that bad news impacts the volatility more than good news. However it is also the most complex model employed and is notably slower, given it has the most parameters to estimate. This is due to the exponentially increasing number of evaluations of the log-likelihood function that is performed by the method of finding initial values. Noteworthy is that in the backtest of the one-day-ahead VaR all the employed models kept the number of VaR breaches to the expected or less. Hence it is more tempting to use the GARCH or gjr-GARCH model in this context to preserve speed of the algorithm. Also both Polak–Ribi‘ere and Fletcher–Reeves formula for the conjugate gradient method was tested with no notably difference in accuracy or speed for the data sets considered in this paper. Furthermore the algorithm managed to handle illiquid assets and smaller data sets, in the sense of yielding reasonable results.

For further research on this topic, the method of finding initial values can be improved. Since the method included in this paper is quite naive and is not a good approach for

large models. Of course there exist lots of algorithms to solve the optimization problem and other algorithms may outshine the nonlinear conjugate method in this context, which was chosen due to only requiring first order derivatives. In the perfect world it would be preferred to construct a robust optimization algorithm that does not depend on the initial values. Also there is the possibility of using other faster computer languages, such as *C*.

As a final comment, the method used to construct the algorithm capable of automatizing GARCH parameter estimation in this paper resulted in a powerful *java* implementation of it. Showing great ability of volatility forecasting while handling the problems that arise with automatizing the estimation.

A Appendix

A.1 Akaike information criterion

$$AIC = \frac{2k}{n} - \frac{2LL}{n} \tag{60}$$

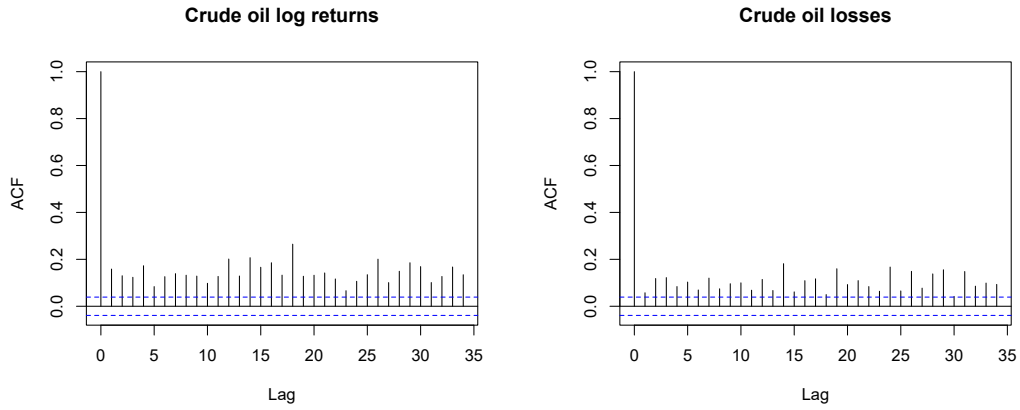
where k is the number of parameters of the model and n is the sample size.

A.2 Propositions

Proposition 1
If $g : \mathcal{R} \to \mathcal{R}$ is non-decreasing and left continuous, then for any random variable Z it holds that $F^{-1}_{g(Z)}(p) = g(F^{-1}_Z(p))$ for all $p \in (0, 1)$.
Proposition 2
For any random variable X, $F^{-1}_{-X}(p) = -F^{-1}_X((1-p)+)$ for all $p \in (0, 1)$. In particular, if F_X is continuous and strictly increasing, then $F^{-1}_{-X}(p) = -F^{-1}_X(1-p)$.

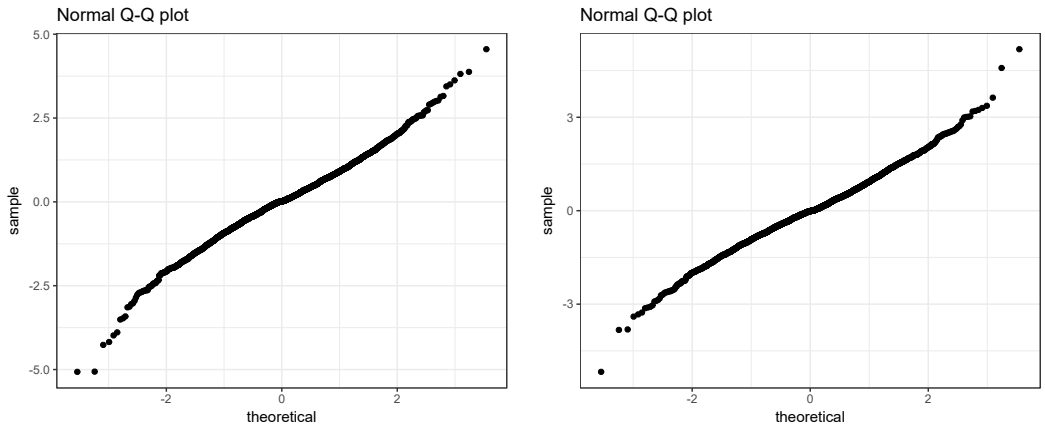
A.3 Figures

A.3.1 Brent crude oil



(a) (b)

Figure 10: Plot (a) depicts sample ACF of the squared log returns and plot (b) depicts the sample ACF of the squared losses.



(a) (b)

Figure 11: Plot (a) depicts the normal Q-Q plot of Brent crude oil filtered log returns (b) depicts the normal Q-Q plot of Brent crude oil filtered losses. Where the filtering is done with GARCH.

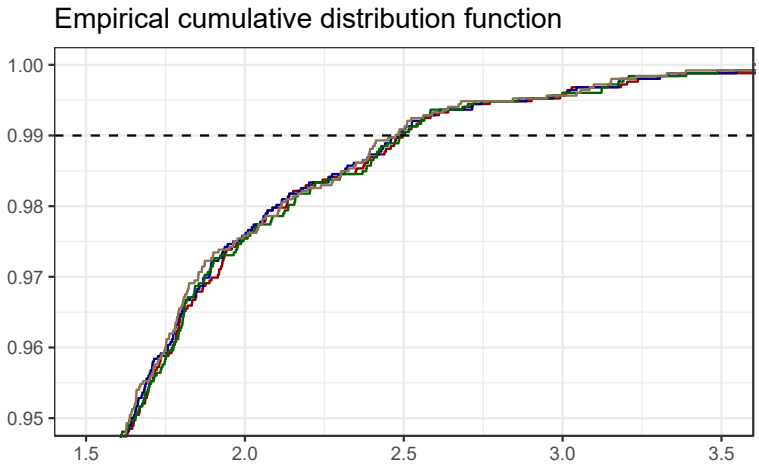
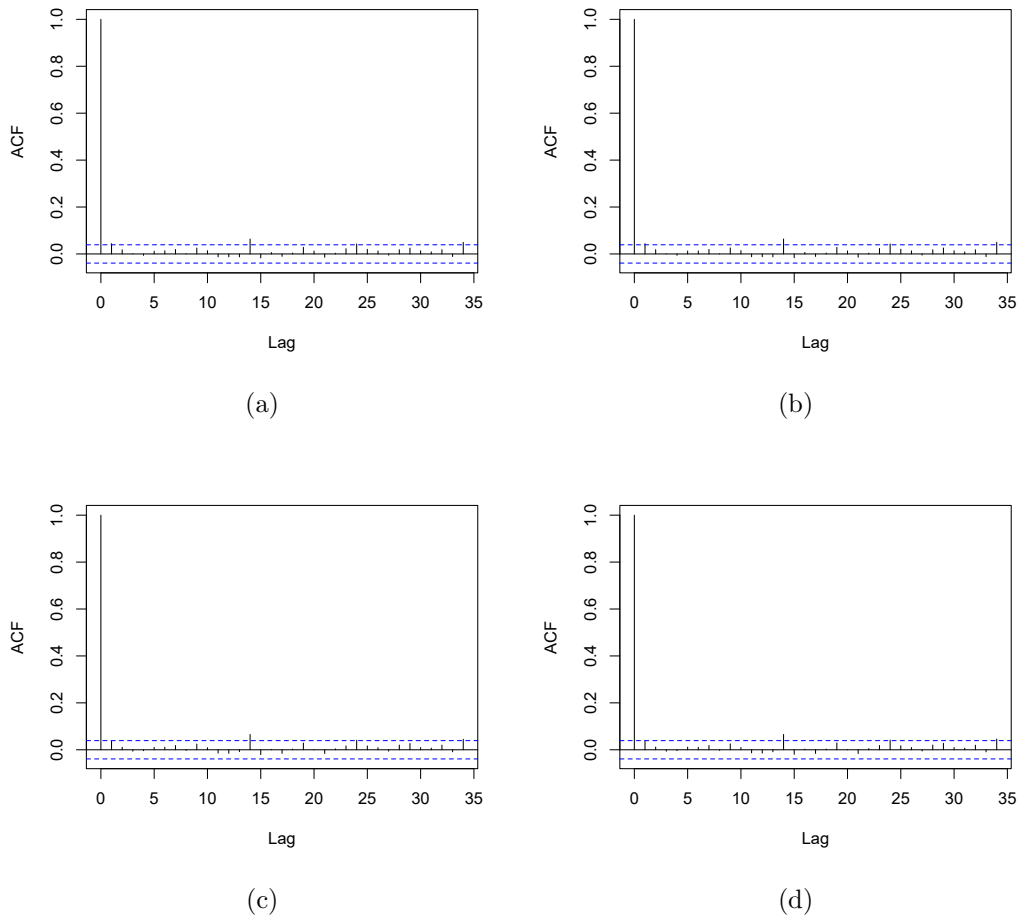


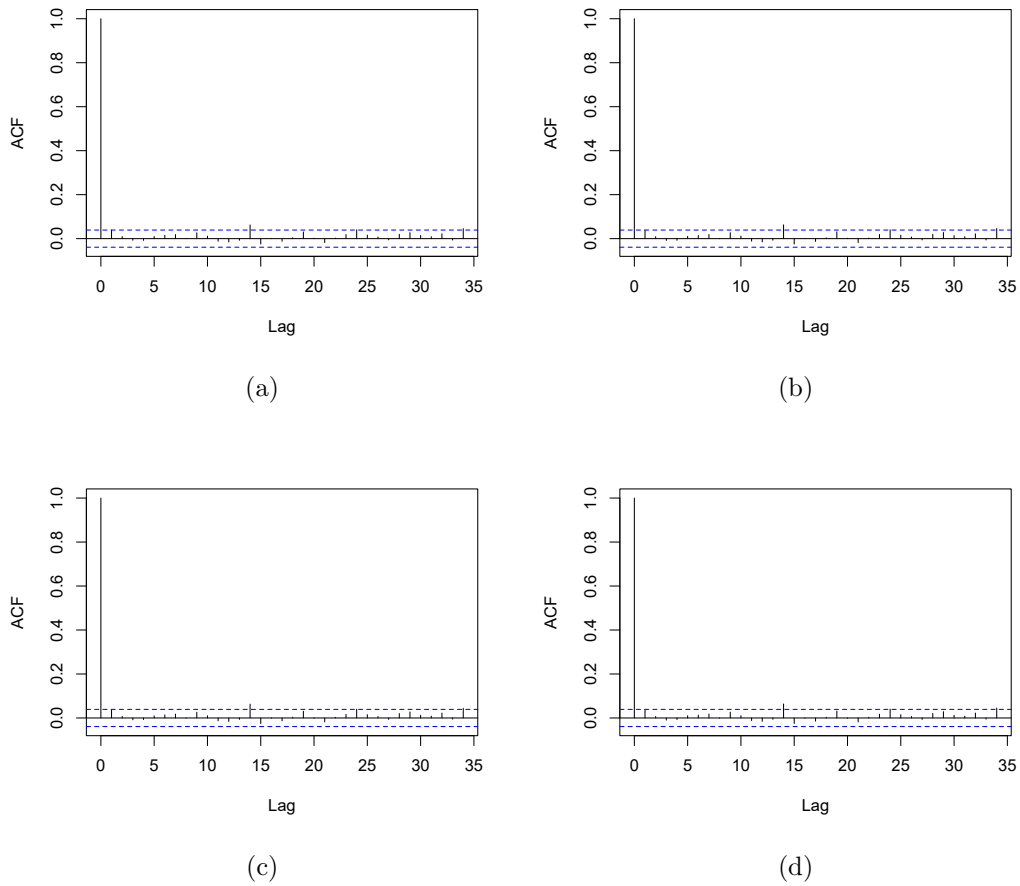
Figure 12: Right tail of the empirical cdf of Brent crude oil filtered losses (residuals), where the dashed line marks the 0.99 level. Red line for GARCH, blue for sGARCH, green for gjr-GARCH and brown for gjr-sGARCH.



(a) (b)

(c) (d)

Figure 13: Sample ACF of crude oil filtered log returns (residuals), plot (a) filtered with GARCH, plot (b) filtered with sGARCH, plot (c) filtered with gjr-GARCH and plot (d) is filtered with gjr-sGARCH.

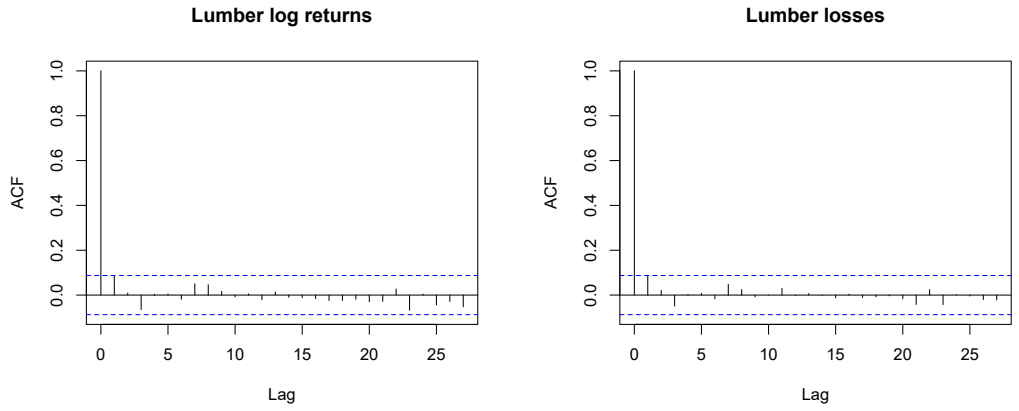


(a) (b)

(c) (d)

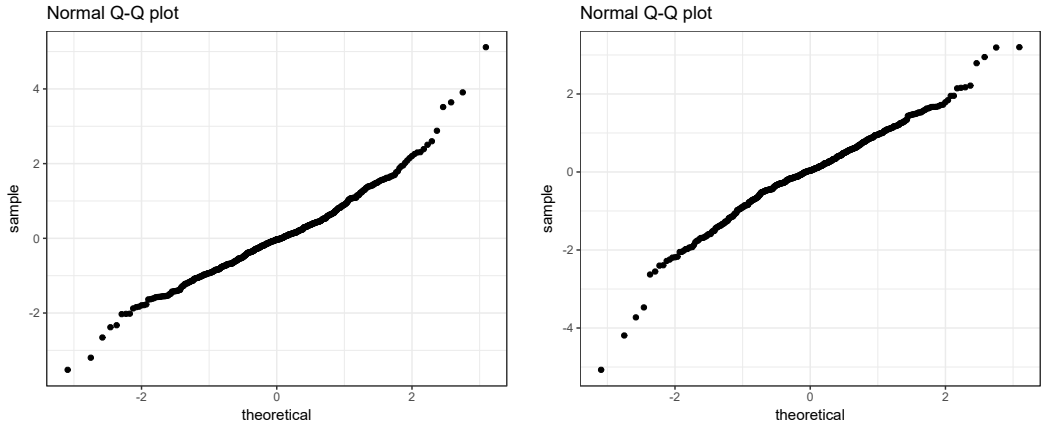
Figure 14: Sample ACF of crude oil filtered losses (residuals), plot (a) filtered with GARCH, plot (b) filtered with sGARCH, plot (c) filtered with gjr-GARCH and plot (d) is filtered with gjr-sGARCH.

A.3.2 Lumber



(a) (b)

Figure 15: Plot (a) depicts sample ACF of the squared log returns and plot (b) depicts sample ACF of the squared losses.



(a) (b)

Figure 16: Plot (a) depicts the normal Q-Q plot of lumber futures filtered log returns (b) depicts the normal Q-Q plot of lumber futures filtered losses. Where the filtering is done with GARCH.

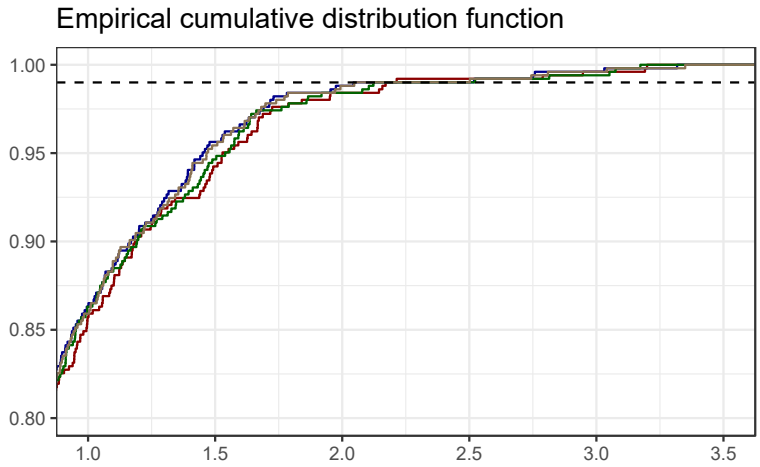


Figure 17: Right tail of the empirical cdf of Lumber futures filtered losses, where the dashed line marks the 0.99 level. Red line for GARCH, blue for sGARCH, green for gjr-GARCH and brown for gjr-sGARCH.

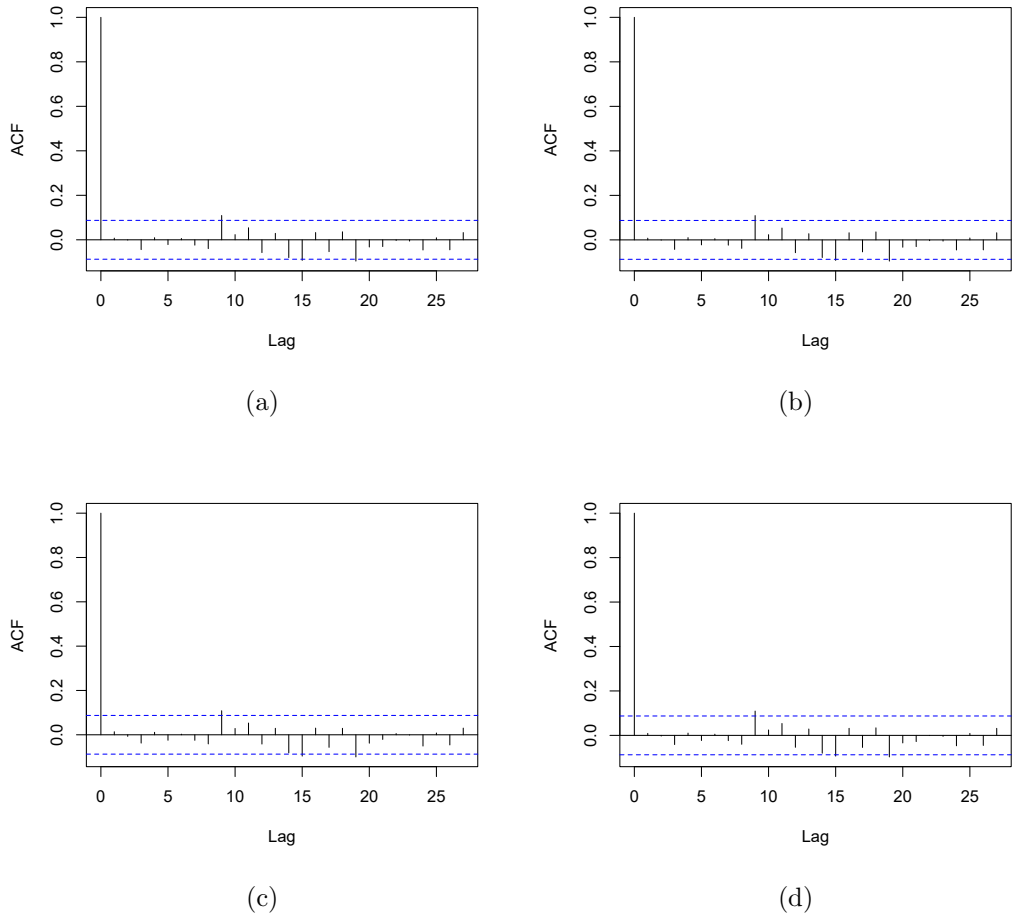


Figure 18: Sample ACF of lumber futures filtered log returns (residuals), plot (a) filtered with GARCH, plot (b) filtered with sGARCH, plot (c) filtered with gjr-GARCH and plot (d) is filtered with gjr-sGARCH.

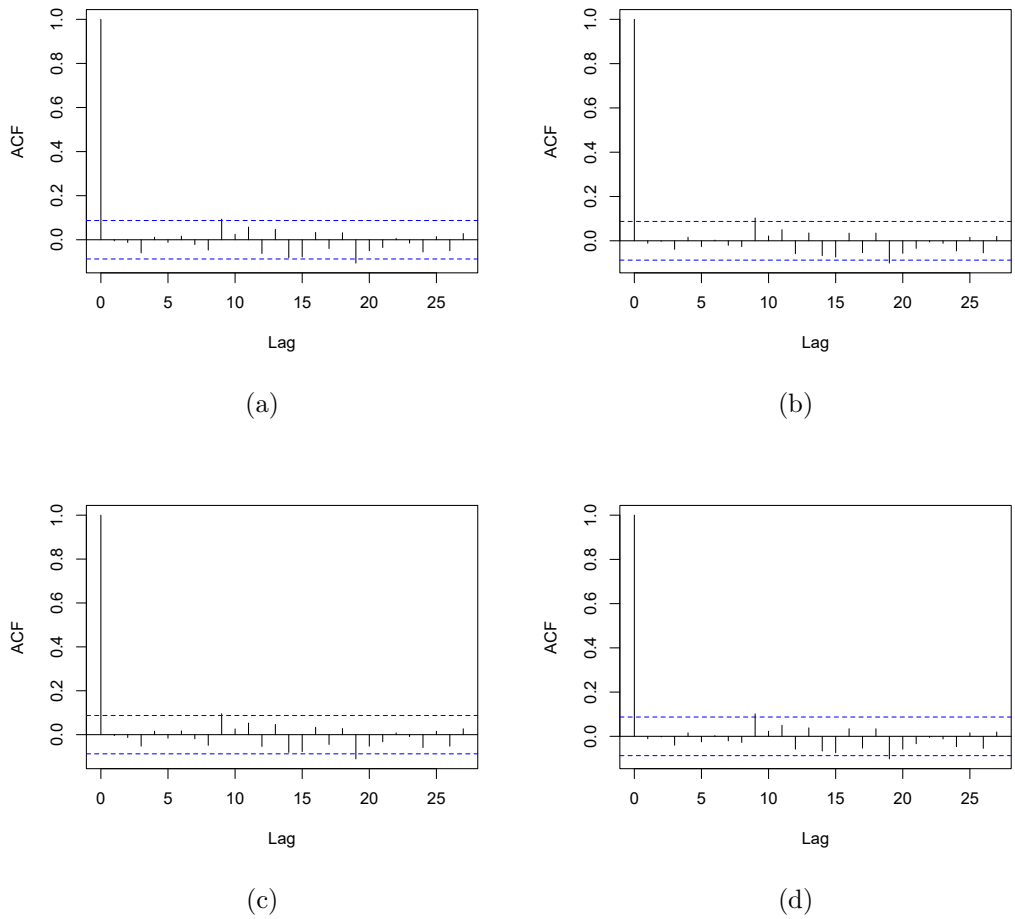


Figure 19: Sample ACF of lumber futures filtered losses (residuals), plot (a) filtered with GARCH, plot (b) filtered with sGARCH, plot (c) filtered with gjr-GARCH and plot (d) is filtered with gjr-sGARCH.

A.4 Tables

	∇LL_i			
	$\frac{\partial LL}{\partial z_{corr}}$	$\frac{\partial LL}{\partial z_{ema}}$	$\frac{\partial LL}{\partial z_{\nu}}$	$\frac{\partial LL}{\partial z_{asy}}$
GARCH	-0.006006	1.804	-	-
sGARCH	-0.007995	-4.426	-0.07938	-
gjr-GARCH	-0.002261	-0.01445	-	-0.1653

Table 5: Gradient of the log-likelihood function evaluated at the optimal solution $\boldsymbol{\theta^*}$.

	$\frac{\partial <\sigma_t^2>}{\partial z_{corr}}$	$\frac{\partial <\sigma_t^2>}{\partial z_{ema}}$	$\frac{\partial <\sigma_t^2>}{\partial z_{asy}}$	$\boldsymbol{\theta^*} = (z_{corr}, z_{ema}, z_{\nu}, z_{asy})$
GARCH	3.638E-9	8.796E-8	-	(6.084, 3.026, -, -)
sGARCH	4.146E-9	9.376E-8	-	(6.146, 3.139, 1.650, -)
gjr-GARCH	-1.028E-8	-1.341E-7	-1.341E-7	(6.230, 4.288, -, 3.148)

Table 6: Partial derivatives of sample variance w.r.t parameters.

References

[1] Torben G. Andersen and Tim Bollerslev. "Answering the Skeptics: Yes, Standard Volatility Models do Provide Accurate Forecasts". In: *International Economic Review* 39 (1998).

[2] Tim Bollerslev. "GENERALIZED AUTOREGRESSIVE CONDITIONAL HETEROSKEDASTICITY". In: *Journal of Econometrics 31* 31 (1986).

[3] P.J. Brockwell and R.A. Davis. *Introduction to Time Series and Forecasting*. Third edition. Springer, 2002. ISBN: 9780387953519.

[4] Robert F Engle. "Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation". In: *Econometrica* 50 (1982).

[5] Lawrence R. Glosten, Ravi Jagannathan, and David E. Runkle. "On the Relation between the Expected Value and the Volatility of the Nominal Excess Return on Stocks". In: *The journal of finance* 48 (1993).

[6] H. Hult, F. Lindskog, O. Hammarlid, and C.J. Rehn. *Risk and Portfolio Analysis: Principles and Methods*. Springer Series in Operational Research and Financial Engineering. Springer Science+Business Media New York, 2012. DOI: 10.1007/978-1-4614-4103-8_1.

[7] Michael Parkinson. "The Extreme Value Method for Estimating the Variance of the Rate of Return". In: *The Journal of Business* 83 (1980).

[8] Andrew J. Patton. "Volatility forecast comparison using imperfect volatility proxies". In: *Journal of Econometrics* 160 (2011).

[9] Dimitri P.Bertekas. *NONLINEAR PROGRAMMING*. Athena Scientific, Belmont, Massachusetts, 2016. ISBN: 978-1-886529-05-2.

[10] Gilles Zumbach. "The pitfalls in fitting GARCH processes". In: *Dunis C.L. (eds) Advances in Quantitative Asset Management. Studies in Computational Finance* 1 (2000).