Master Thesis

Copenhagen Business School

Cand.Merc(mat.)

Modeling Structural Changes in Volatility Using Markov Switching GARH Models

Authors:
Jeppe Brauer
Niels Kristensen

Supervisor:
Anders Rønn-Nielsen

Number of pages: 108

May 15, 2020

Abstract

Knowing the risk associated with a financial investment is relevant for everyone in the financial world. Determining a way to estimate this risk accurately, has been a topic of discussion for decades. Knowledge regarding the behaviour of financial asset returns, has lead to a lot of interesting discoveries, such as the heavy tailed nature of the returns as well as the heteroscedastic behavior of the volatility. These findings have been accounted for in different ways throughout the years.
A type of behaviour that is rarely accounted for in risk modeling is structural changes in the modeling, over time. These changes can be seen when the financial markets enter a crisis, where the volatility usually skyrocket. It is therefore interesting to model these periods differently, than periods where the financial markets are stable, and the volatility is generally low.

This thesis applies the theory behind hidden Markov models to expand upon the GARCH model, such that it can account for periods of structurally different volatility. This model is called a Markov switching GARCH model, and it will be used to explain the volatility of the return process for a selected part of the S&P 500 index. In doing so the model parameters will be estimated based on a historical time period, that include times of financial crisis, as well as more stable times. When estimating the parameters of the MS-GARCH model based on this period, the variance of the financially stable periods converge to being constant, whereas the variance in the financial crises are modelled well by the GARCH model.

Estimating the parameters in the MS-GARCH is made difficult since the structure of the model introduces a path dependence in the conditional variance. This is overcome by using a Bayesian estimation procedure, instead of maximum likelihood, to determine the parameter estimates. The method used for the estimation of the parameters is a Gibbs sampler, which is used due to its effectiveness when working with high dimensional estimation.

The capabilities of the MS-GARCH model as a risk model are also examined. Here it is found to produce good risk estimates on historical observations, however the risk estimates generated from day to day, are not ideal.

Resumé

At kende risikoen forbundet med en finansiel investering, er relevant for enhver person i den finansielle verden. Dog har måden hvorpå denne risiko estimation bliver lavet, blevet diskuteret i årtier. Kendskab til opførslen af finansielle afkast har ledt til mange interessante opdagelser, såsom den tung-halede fordeling af de finansielle afkast, samt den heteroskedastiske opførsel af volatiliteten. Disse opdagelser er gennem årene også blevet implementeret i diverse risikomodeller.
En type af risikomodellering som der sjældent bliver taget højde for, er strukturelle ændringer i risikomodelleringen gennem tid. Disse ændringer kan ses når de finansielle markeder går ind i en krise, hvor volatiliteten normalt eksploderer. Det er derfor interessant at modellerer disse perioder anderledes, end de perioder hvor markederne er stabile og volatiliteten er lav.

Denne afvikling anvender teorien bag Hidden Markov modeller til at udvide GARCH modellen, således at der kan tages højde for de strukturelle forskelle i volatilitets strukturen. Denne type model kaldes en Markov Switching GARCH-model, og vil blive brugt til at beskrive variansen i afkast processen for en udvalgt periode af S&P 500 indekset. For at gøre dette, vil modellens parametre blive baseret på en historisk periode som inkluderer både finansielle kriser, og finansielt stabile perioder. I estimationen af parametrene i MS-GARCH-modellen, baseret på denne periode, bliver det fundet at variansen konvergerer til at være konstant i de stabile perioder, samt at GARCH-modeller forklarer volatiliteten godt i de finansielle kriser.

Estimationen af parametrene i MS-GARCH-modellen er svær at lave, grunden strukturen af modellen, som introducerer en løbende afhængighed af den betingede varians proces. Dette overkommes ved at anvende en Bayesiansk estimations procedure i stedet for Maximum likelihood estimation. Den brugte metode til estimation af parametrene er Gibbs sampling, som bruges grundet dens effektivitet når der arbejdes med højdimensional estimation.

MS-GARCH modelles egenskaber som risiko model undersøges også. Her ses det at der produceres gode risiko estimater på historiske observationer, men risikoestimaterne der bliver lavet dag til dag, er ikke ideelle.

Acknowledgement

We want to thank our supervisor Anders Rønn-Nielsen for all his help and his big engagement in the this project. Anders has always been easy to get a hold of, and the various questions we have had, have always been answered almost instantly. Our weekly meetings have been a huge help, and without Anders' expertise, this project would not have been possible.

Table of contents

1 Introduction

The behaviour financial asset returns can change drastically from day to day. Modelling this behaviour can be very difficult, since the expectation of the future will have to be derived from the available historical information.
Modelling financial asset returns as draws from a fixed distribution, often leads to issues, as the underlying variance, referred to as volatility, is rarely constant. This has been a focus point in financial risk modelling for decades, and has led to numerous models which attempt to account for this.
Examples of such models include the ARCH model developed in Engel (1982) as well as the GARCH model developed in Bollerslev (1986).
Both of these models account for a non-constant volatility, also called a heteroscedastic volatility, which is often seen in financial asset returns.

Instead of only considering the volatility as heteroscedastic, the volatility can also be seen as behaving structurally different over time. These structurally different periods could for instance be seen as times of financial crisis versus times where the financial markets are stable. It could be assumed that these periods would behave differently, and neither the ARCH nor the GARCH model are able to account for this.
A type of model that can account for these structural changes in the underlying process, across time, are called regime shift models.
We wish to implement a model which combines the GARCH model and a regime shift model. Such a model will possibly be able to capture both the heteroscedasticity from the GARCH model, as well as accounting for the structural changes in the underlying process. This model determines the underlying regimes by assuming the existence of an unobservable Markov chain, which determines the type of regime.

Such a model is called a Markov switching GARCH model, and was first implemented in Gray (1996), and further developed in Bauwens et al. (2010). Estimating the parameters of this model, will also be a big focus point, as some alternative measures have to be implemented, because of the existence of a path-dependence in the underlying variance process. In Bauwens et al. (2010) it is suggested that a Bayesian approach could be used to estimate the parameters, instead of maximum likelihood estimation.

With the implementation of the Markov switching GARCH model, we attempt to model

the volatility of the S&P 500 index, in times of financial crisis and times of market stability. One of the goals with this estimation is to clearly separate these structurally different periods, in order to model them differently.
Further investigation of the estimated model will determine how well it estimates risk, compared to a standard GARCH model. Furthermore the usefulness of the regime shifting capabilities, in a risk modelling framework, will also be examined.

1.1 Thesis statement

The main objective of this thesis is to examine how a regime-shift model, in combination with a GARCH process, can be used to model the structural changes in the volatility of financial asset returns. We will further investigate how the parameters in such a model can be estimated, as well as examining the performance of it, as a risk model, compared to other volatility models.

- What is the MS-GARCH model, and which attributes makes it useful in volatility modelling?
- How are the parameters and the regimes of the MS-GARCH model estimated?
- How accurate is the estimation procedure for the MS-GARCH model?
- How are financial times of crisis, and times of stability identified and modelled when using the MS-GARCH model?
- Is it possible to use the MS-GARCH model work as risk model, and how does it work compared to other volatility models?

2 Stochastic processes

A stochastic process is a sequence of random variables iterated by time. These random variables all have values in a state space $\mathcal{X}$. The state space can be defined as subsets $\mathcal{X} \subseteq \mathbb{N}^d$ or $\mathcal{X} \subseteq \mathbb{R}^d$, depending on whether it is discrete or continuous, where d denotes the number of dimensions. In this section we will cover stochastic processes with both discrete and continuous state space, as these are both used throughout the thesis.
Another characteristic of a stochastic processes is the frequency of which the variables are collected. This can either be done in discrete or continuous time, meaning that there can either exist observations at every point in time, or at a discrete set of points.

Financial time series, can be considered in either discrete or continuous time. This is due to the fact that prices on financial assets theoretically exist at every point in time. This way of defining financial time series is used in several instances, one of which being option pricing. The theory of option pricing rely greatly on the concepts of stochastic integrals, which assume that the stochastic process is defined in continuous time.
However since the prices can only be observed discretely, and because it allows for some simplifying conditions, this thesis will consider the financial time series as discrete time stochastic processes.

2.1 Markov chains

There exist several different types of stochastic processes, which meet certain conditions. A particularly useful type of stochastic process is called a Markov chain, which defines a stochastic processes that satisfies the Markov property. The Markov property states that each observation is only dependent on the previous observation, and not the entire preceding chain of observations. The theory behind Markov Chains is covered in Lawler (2006), where the mathematical definition of a Markov chain X in discrete time and with discrete state space is given as

$$\mathbb{P}(X_t = x_t \mid X_1 = x_1, ..., X_{t-1} = x_{t-1}) = \mathbb{P}(X_t = x_t \mid X_n = x_n) \ \forall \ x_1, ..., x_t \in \mathcal{X}.$$

Markov chains are also defined with continuous state space, where the process is required to follow a similar definition, with $X_t = x_t$ replaced by $X_t \in A$ for a set $A \subseteq \mathcal{X}$.
Markov chains are also usually assumed to be time-homogeneous, meaning that the probability of transferring from one state to another is determined by a constant probability,

which does not depend on time. This condition can be expressed mathematically as

$$\mathbb{P}(X_t = x_t \mid X_1 = x_1, ..., X_{t-1} = x_{t-1}) = \eta_{x_{t-1},x_t}$$

where η_{x_{t-1},x_t} does not depend on time.
For a Markov chain in a discrete state space, the time homogeneity makes it possible to collect all transition probabilities in a transition matrix. In a transition matrix each row describes the probability of transferring from the state, corresponding to the row, to each of the other states, such that each row sum is equal to 1. Given a state space of $\mathcal{X} = 1, 2, ..., k$ the transition matrix will be

$$\Gamma = \begin{bmatrix} \eta_{1,1} & \eta_{1,2} & \cdots & \eta_{1,k} \\ \eta_{2,1} & \eta_{2,2} & \cdots & \eta_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ \eta_{k,1} & \eta_{k,2} & \cdots & \eta_{k,k} \end{bmatrix},$$

where $\eta_{i,j}$ defines the probability of transferring from regime i to regime j.
For the time homogeneous Markov chains the sojourn times, which defines the amount of time it takes to leave a state, will be geometrically distributed

$$\mathbb{P}(X_1 = i, ..., X_{t-1} = i, X_t \neq i \mid X_1 = i) = \eta_{i,i}^{t-1} \cdot (1 - \eta_{i,i}).$$

Since the sojourn times in Markov chains are geometrically distributed, it is also possible to calculate the amount of time the process is expected to stay in a state. Defining $T_i = \min\{t \in \mathbb{N} \mid X_t \neq i\}$ we have

$$\mathbb{E}(T_i \mid X_1 = i) = \frac{1}{1 - \eta_{i,i}}$$

The transition matrix Γ can contain multiple communication classes, that indicates which states have a positive probability of transferring between them. These classes can be divided into two groups, recurrent and transient. The type indicates how frequently a state is visited in the limit.
Given an infinite number of observations a state is recurrent, if it is visited an infinite number of times, whereas a transient state will only be visited a finite number of times.

Definition 2.1. *Let ζ_i define the total number of times the Markov chain X visits state i, $\zeta_i = \sum_{n=0}^{\infty} I\{X_n = i\}$, then state i is recurrent if $\mathbb{E}(\zeta_i | X_0 = i) = \infty$*

If $\mathbb{E}(\zeta_i | X_0 = i) < \infty$, the state is called transient. If a communication class is recurrent, the states in that communication class are called recurrent states.

In this thesis we only consider Markov chains with one communication class, in which case the chain is called *irreducible*. If the Markov chain was not irreducible, it would not be independent of the starting distribution ϕ_0 in the long run, and it will be uncertain if the chain would leave the state or return if it left the state.

Definition 2.2. *A Markov chain is called irreducible if there for all i,j exist a* $n > 0$ *such that* $\mathbb{P}(X_n = i \mid X_0 = j) > 0$.

For an irreducible Markov chain, we define the *period*, $d(i)$ for state i, as the greatest common divisor of J, where J contain the number of steps the Markov chain can use in order to return to state i.

Definition 2.3. *A state* i *is called aperiodic if* $d(i) = 1$ *i.e it holds that* $\eta_{i,i} > 0 \ \forall \ i$

Since Markov chains are time-homogeneous with transition matrix Γ, it is possible to determine an invariant probability distribution of the states. The invariant probability distribution denotes how much the process is in each of the states in $\mathcal{X}$. The invariant probability distribution is defined as π where $\pi\Gamma = \pi$ and $\pi 1^T = 1$, and it can be found by taking the limit of the n-step transition probabilities, i.e.

$$\lim_{n \to \infty} \phi_0 \Gamma^n = \begin{bmatrix} \pi_1 \\ \vdots \\ \vdots \\ \vdots \\ \pi_n \end{bmatrix},$$

where ϕ_0 is the initial probability distributions.
for a transition matrix in two dimensions the invariant probability distribution can be found as

$$\pi = \begin{bmatrix} \pi_1 \\ \pi_2 \end{bmatrix} = \begin{bmatrix} \frac{\eta_{1,2}}{\eta_{1,2}+\eta_{2,1}}, \\ \frac{\eta_{2,1}}{\eta_{1,2}+\eta_{2,1}} \end{bmatrix}. \tag{1}$$

2.2 Continous state space

The definitions for Markov chains with continuous state space is described in Hahn (2013-2014). Most of the definitions in continuous state space stay the same, expect for some cases. These special cases will be examined further in this section.
It is not possible to define the transition matrix Γ in continuous state space. Instead the transition matrix will be considered as a transition kernel. Let $\mathcal{X}$ be a measurable space, then for a time homogeneous Markov chain the transition kernel can be written as

$$\mathbb{P}(X_{n+1} \in B | X_n = x) = P(x, B) \ \forall \ n \in \mathbb{N}, x \in \mathcal{X} \ \text{ and } \ B \subseteq \mathcal{X}$$

For $P(.,.)$ to be a transition kernel on $\mathcal{X}$ it must be satisfied that any fixed $x \in \mathcal{X}$, $P(x,.)$ is a probability measure.
As the definitions for Markov chains with continuous state space are generally the same as in discrete state space, we will just apply this change in the notation.
In continuous state space a measure ψ on $\mathcal{X}$ is used to describe some properties. A Markov chain X is said to be ψ-irreducible if there exists a measure ψ such that

$$\forall \ B : \psi(B) > 0 \Rightarrow \exists n : P^n(x, B) > 0 \ \forall x \in \mathcal{X},$$

This is a general notation in continuous state space, where it is necessary to consider all measurable subsets of $\mathcal{X}$ in order to determine a probability measure.

In continuous state space, recurrence will be defined as Harris recurrence, which is defined as

Definition 2.4. *A Markov chain is Harris recurrent if there exists a measure ψ, such that the Markov chain is ψ-irreducible and $\forall B$ with $\psi(B) > 0$ it holds that, $P(h_B = \infty | X = x) = 1 \ \forall x \in B$*

The notation of h_B indicates the number of passages of the set B. Furthermore, if a Markov chain is Harris recurrent and aperiodic, then it is also ergodic.
An ergodic Markov chain ensures that an invariant distribution exist, no matter where it is initialized.

Definition 2.5. *A Markov chain X_n on $\mathcal{X}$ with transition kernel $P(.,.)$ and invariant distribution $\pi(.)$ is ergodic, if $\forall x \in \mathcal{X}$,*

$$||P^n(x,.) - \pi(.)||_{TV} \xrightarrow[n \to \infty]{} 0.$$

The notation TV stands for *total variation norm* and can be written as

$$||P - \pi||_{TV} = \frac{1}{2} \int_{\mathcal{X}} |P(dx) - \pi(dx)|.$$

Thereby we are ensured that the Markov chain approaches an invariant distribution
Using Definition 2.5 we can approximate the expectation of the invariant distribution

$$\mathbb{E}_\pi h(x) := \int_{\mathcal{X}} h(x)\pi(x)dx$$

by the partial mean,

$$S_n(h) := \frac{1}{n} \sum_{i=1}^{n} h(X_i).$$

The strong law of large number gives the condition under which the mean converges. This is referred to as the Ergodic Theorem.

Theorem 2.1. *If X_n has a σ-finite invariant measure π, then*

$$\mathbb{P}\Big(\lim_{n \to \infty} S_n(h) = \mathbb{E}_\pi h(x) | X_0 = x \Big) = 1 \quad \forall x \in \mathcal{X}$$

if and only if X_n is Harris recurrent.

2.3 Stationarity

Given the randomness of a stochastic process it is difficult to determine how the process will behave over a large period of time. A way of evaluating this large time behavior of the stochastic process, is by evaluating the stationarity of the process. The concepts behind stationarity are covered in Ruppert and Matteson (2015), which is also the basis for the following definitions in this section.
A stochastic process can be stationary at different levels. The stationarity requirement with the strictest assumptions is called a strictly stationary process. A strictly stationary process is a process where all aspects are unchanged over time.

Definition 2.6. *A process is strictly stationary if* $(X_1, ..., X_n) = (X_{1+m}, ..., X_{n+m}) \quad \forall \ n, m.$

This also means that the distribution of each set of observations is the same, and is therefore not dependent on the time origin nor the number of observations.

A stochastic process can also be weakly stationary, which lessens the assumptions of the process. For a process to be weakly stationary it must have a finite and constant unconditional mean and variance. Furthermore the covariance between two observations must only depend on the time distance between them.

Definition 2.7. *A process is weakly stationary if* $\mathbb{E}(Y_t) = \mu \ \forall \ t$, $Var(Y_t) = \sigma^2 \ \forall \ t$ *and* $Cov(Y_t, Y_s) = \gamma(| \ t - s \ |) \ \forall \ t, s$, *for some function* γ.

Finally it is possible for a stochastic process to be higher order stationary, which indicates the existence of a higher order unconditional moment.

Definition 2.8. *A process is m-order stationary if the* m^{th} *moment of* Y_t *is finite and constant.*

3 Value at risk

In the financial world it is important to have an idea about how risky an investment is. There are several ways of quantifying risk in the financial world, and one of the more well know methods is called Value at Risk (VaR). VaR aims to determine the maximum loss that is expected to occur within a given interval of time, at a certain confidence level. In this thesis we will cover single day VaR, and the general theory behind VaR is described in Röman (2017).

The confidence level used in VaR is called α and VaR_α is the VaR estimate associated with the confidence level. As α is a confidence level it can be seen as the classification of how often the VaR estimate is too low. $VaR_{99\%}$ can therefore be seen as the level of the loss, which will only be expected to be exceeded every 100^{th} day.
The VaR_α estimate can therefore also be written as

$$VaR_\alpha(L) = \inf\{c : \mathbb{P}(L > c) \leq 1 - \alpha\},$$

where L denotes the loss.
Calculating VaR is mainly done using either the parametric or non-parametric methods. This thesis will only cover the parametric method, however the difference between the two, is that the parametric method assumes an underlying distribution whereas the non-parametric method does not.
Since the parametric VaR method assumes an underlying distribution, it is possible to calculate the VaR estimate from Equation 3 as

$$VaR_\alpha(L) = F^{-1}(\alpha; \theta),$$

where F^{-1} is the quantile function of the underlying distribution, and θ is the parameter set for that distribution.
the $VaR_{95\%}$ and $VaR_{99\%}$ are illustrated in Figure 1 using a standard normal distribution. Here it is seen that VaR_α becomes more negative when the confidence level increases.

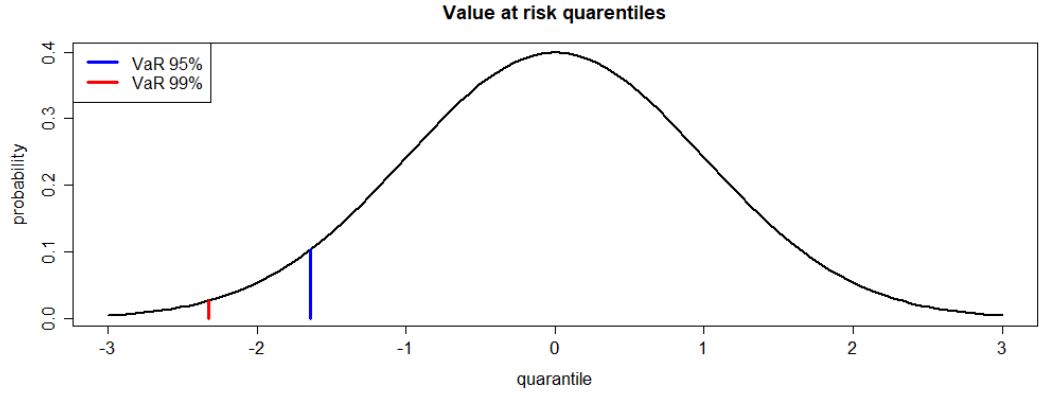


Figure 1: Illustration of the VaR estimate using a normal distribution.

Since the parametric VaR method uses a underlying distribution in order to calculate the risk, it is necessary to consider which distribution is used, as-well as how to parameters of the distribution are determined. This is not straightforward as the empirical distributions of financial asset returns usually have much heavier tails than the normal distribution, i.e. a positive excess kurtosis.
One possible way of accounting for the heavy tails in the financial asset returns could therefore be to use a more heavy tailed distribution. However finding a distribution with as heavy tails as financial asset returns can be a task in itself, and working with such distributions can be even more challenging, and therefore this is rarely a feasible solution. Alternatively a normal distribution could be used, and the heavy tails could instead be accounted for by changing the variance of the distribution continuously, this can help increase the weight of the tail probabilities.

In the next section we will investigate how such a changing variance can be calculated in order to account for the distribution of financial asset returns as best as possible.

4 Volatility processes

How to model the variance of financial asset returns, often referred to as volatility, has been a topic of discussion for decades. The main reason why this metric is difficult to model, is that it can not be found as a snapshot of how the world is today. Instead it has to be derived from the available historical information. This creates an issue, as we will not expect history to repeat itself in the same exact way, and we therefore have to construct a model which turns history into reality.

The simplest way to construct a model that describes the changing variance, is to use a running variance with a fixed window. This makes it possible to estimate confidence intervals where new observations would be expected to appear, and therefore an estimation of the risk, however with this model is far from optimal.
One issue with this method is determining how long an observation is relevant in the estimation of the risk, i.e. how large should the fixed window be? The size of the window can have a very significant effect on the variance level, and it is therefore possible to get two completely different risk estimations solely based on this parameter. Since there is no theoretically correct window size, the risk is going to be based on a subjective choice. Therefore it is necessary to construct a more complicated model in order to explain the movements in financial time series, such that the risk attached with an investment can be estimated accurately.

One option could be to use an auto-regressive model, described in Ruppert and Matteson (2015), to model the returns. Auto-regressive models assume that the value of a future return can be seen as a function of past returns. This means that if there was a large negative return one day, the probability of seeing another large negative return the next day would be higher. This might seem like a good solution, as it sounds reasonable that a bad day for the market, will increase the chance of seeing another bad day. However financial returns rarely exhibit any auto-correlation, and this type of behaviour is therefore not observed in financial time-series.

Even though financial time series do not exhibit significant auto correlation, it does not mean that the idea behind auto-regressive models cannot be used. The problem with using auto-regressive models to model financial asset returns, is that the direction of the movement, or the sign of the return, is expected to be the same between two days.

What is actually observed in financial returns, is that the magnitude of the return one day affects the magnitude of the future returns. This is usually detected in financial returns as the absolute value of the returns, or the squared returns, exhibit a lot of auto-correlation. When modelling the magnitude of observations in an auto-regressive manner, it is in fact the same as modelling the variance. Since the conditional variance moves based on previous observations, it is also assumed that it is non-constant, and this property is usually referred to as heteroscedasticity.

Heteroscedasticity can occur in many different ways, however the first model which accounted for it in an auto-regressive framework was introduced in Engel (1982), where the ARCH model was created. This model can be split into three parts: AR (Auto-regressive) which means that the present return is conditioned on a set of previous returns, C (Conditional) which means that the conditional variance is modeled, and H (heteroscedastic) which means that the conditional variance in the model is not constant.
In the ARCH model the conditional variance of the process is modeled from the size of the squared value of the lagged observations, such that large observations will increase the variance of the future observations.

4.1 GARCH model

The ARCH model was further developed in Bollerslev (1986) in which the GARCH model was created. This model added the G (Generalized) term which changes the conditional variance, such that it will not only be based on the size of the squared value of the lagged returns, but also on the level of the previous conditional variances.
The advantage of the GARCH model over the ARCH models is the possibility of having a higher persistence in the conditional variance. This makes it easier to account for volatility clustering, which is also very prominent in financial time series. Volatility clustering means that the volatility is high in certain limited periods, and relatively low outside of these periods.

The GARCH model is constructed to account for both the observations and conditional variances that lie before the previous observations. Such a model is defined as a GARCH(p, q) model, where p is the number of lagged squared returns (the ARCH term), and q is the number of lagged conditional variances. In this thesis, we will only consider the simplest model, the GARCH(1,1) where only the previous, squared return,

and previous conditional variance is accounted for. This model can be written as

$$\begin{aligned} y_t &= \mu + \sigma_t u_t \\ \sigma_t^2 &= \omega + \alpha \epsilon_{t-1}^2 + \beta \sigma_{t-1}^2 \\ \epsilon_t &= y_t - \mu \\ \theta &= \{\omega, \alpha, \beta\} \end{aligned} \tag{2}$$

Where u_t is a draw from a standard-normal distribution, drawn independently of σ_t.

In this model ω is the lowest possible standard deviation of the model, and it is therefore clear that we must have $\omega > 0$, since it would otherwise be possible to get a variance of zero, and if this was to happen the process would become continuously constant. Further we must have that $\alpha \geq 0$ and $\beta \geq 0$, since the variance cannot assume a negative value. Lastly $\alpha + \beta < 1$ ensures that the process is stationary. The stationarity of the GARCH model is explained in Bollerslev (1986) p. 310, where it is described that the process is weakly stationary if $\alpha + \beta < 1$. However using this result, as well as the strict stationarity requirement found in Francq and Zakoïan (2010) p. 24, it can easily be found that $\alpha + \beta < 1$ also ensures strict stationarity, by using Jensens inequality

$$\begin{aligned} & \mathbb{E}[\log(\alpha u_t^2 + \beta)] < 0 \\ \implies & \mathbb{E}[\log(\alpha u_t^2 + \beta)] \underbrace{\leq}_{\text{Jensens inequality}} \log(\mathbb{E}[\alpha u_t^2 + \beta]) = \log(\alpha + \beta) < 0 \\ \implies & \alpha + \beta < 1. \end{aligned}$$

The parameters α and β have two different effects on the GARCH process. α can be seen as the sensitivity to the level of the previous observations, and β can be seen as the persistence of the shocks. In Figure 2 four simulated GARCH processes are shown, with different parameter choices. We see that α makes the process more explosive, and β increases the time it takes for the process to return to a normal level.

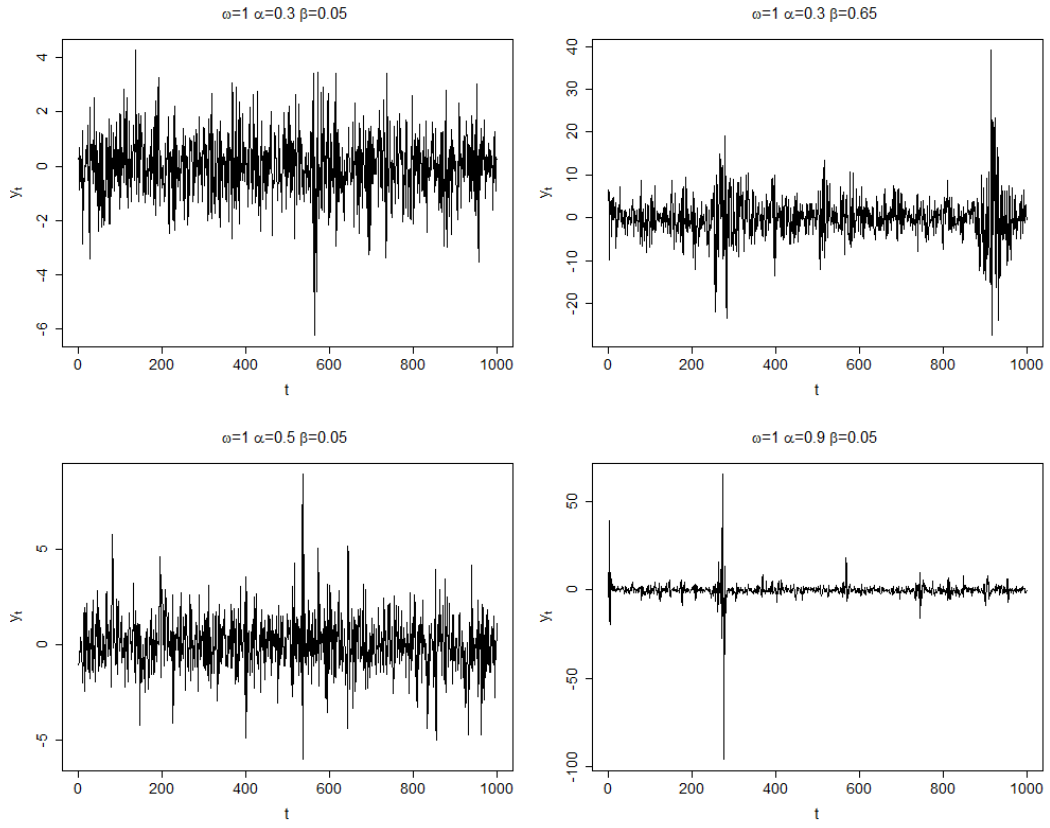


Figure 2: Simulated sample path with 1000 samples of a GARCH-model described in equation 2. the two upper plots shows the effect of the β parameter and the two on the bottom show the effect of the α parameter in the GARCH model.

4.1.1 Central moments of the GARCH-model

In order to get a better grasp of how a GARCH time-series behave, it is useful to have knowledge of the unconditional central moments of the process. The derivations of the results will follow Posedel (2005), that originally investigated these principals, as well as Francq and Zakoïan (2010). The third moment, or the skewness, is not considered here. This is due to the fact that the noise term in the GARCH model is a normal distribution, i.e. a symmetric distribution, and therefore the GARCH-model will not involve skewness.

Unconditional mean

The unconditional mean is simply found as the mean of the GARCH process. That is

$$\begin{aligned} \mathbb{E}[y_t] = \mathbb{E}[\mu + \sigma_t u_t] &= \mathbb{E}[\mu] + \mathbb{E}[\sigma_t u_t] \\ &= \mathbb{E}[\mu] + \mathbb{E}[\sigma_t]\mathbb{E}[u_t] \\ &= \mathbb{E}[\mu] + \mathbb{E}[\sigma_t] \cdot 0 \\ &= \mu, \end{aligned} \tag{3}$$

due to the independence of σ and μ.

Unconditional variance

In order to calculate the unconditional variance, the expected value of the squared process is needed, which can be found as

$$\begin{aligned}\mathbb{E}[y_t^2] &= \mathbb{E}[(\mu + \sigma_t u_t)^2] \\ &= \mu^2 + 2\mathbb{E}[\mu\sigma_t u_t] + \mathbb{E}[\sigma_t^2]\mathbb{E}[u_t^2] \\ &= \mu^2 + \mathbb{E}[\sigma_t^2].\end{aligned}$$

Using Equation 3 and 4.1.1 and also $\mathbb{E}[\epsilon_t^2] = \mathbb{E}[(y_t - \mu)^2] = \mathbb{E}[\sigma_t^2]$ the unconditional variance can be calculated using the variance formula

$$\begin{aligned}Var[y_t] = \mathbb{E}[y_t^2] - \mathbb{E}[y_t]^2 = \mathbb{E}[\sigma_t^2] \\ &= \mathbb{E}[\omega + \alpha\epsilon_{t-1}^2 + \beta\sigma_{t-1}^2] \\ &= \omega + \alpha\mathbb{E}[\epsilon_{t-1}^2] + \beta\mathbb{E}[\epsilon_{t-1}^2] \\ &= \omega + (\alpha + \beta)\mathbb{E}[\epsilon_{t-1}^2].\end{aligned} \tag{4}$$

Due to strict stationarity of the GARCH process (achieved by $\alpha + \beta < 1$) the distribution of y_t and y_{t-1} are identical, which leads to $Var[y_t] = Var[y_{t-1}] = \mathbb{E}[\epsilon_{t-1}^2]$. Using this it is possible to re-write the unconditional variance as

$$\begin{aligned}&Var[y_t] = \omega + (\alpha + \beta)\mathbb{E}[\epsilon_{t-1}^2] \iff \\ &Var[y_t] = \omega + (\alpha + \beta)Var[y_t] \iff \\ &Var[y_t] = \frac{\omega}{1 - (\alpha + \beta)}.\end{aligned} \tag{5}$$

Unconditional kurtosis

In Posedel (2005) it is noted that in order for the existence of a stationary kurtosis the following condition must be met

$$\beta^2 + 2\beta\alpha + 3\alpha^2 < 1. \tag{6}$$

This condition has also been illustrated in Figure 3, where it is clearly seen that this requirement is more strict than the regular stationary requirement of the GARCH process.

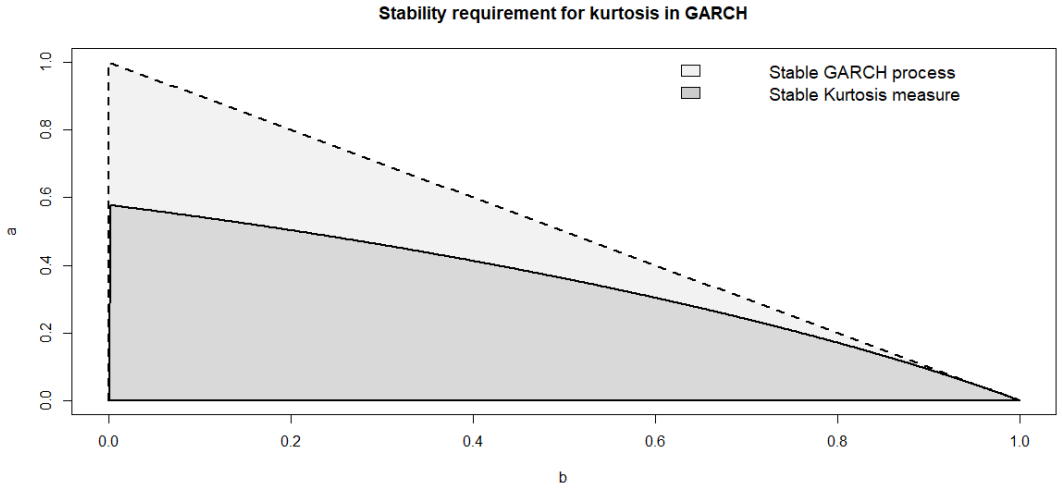


Figure 3: Area of the $\{\alpha, \beta\}$ space in which there exist a stationary kurtosis.

The kurtosis can be found as the fourth moment of the centralized process, or more explicitly as

$$k = \frac{\mathbb{E}[(y_t - \mu)^4]}{(\mathbb{E}[(y_t - \mu)^2])^2} = \frac{\mathbb{E}[(y_t - \mu)^4]}{(Var[y_t])^2}.$$

Since $Var[y_t]$ is already known, as the unconditional variance, only the expression $\mathbb{E}[(y_t - \mu)^4]$ has to be determined

$$\mathbb{E}[(y_t - \mu)^4] = \mathbb{E}[\epsilon_t^4] = \mathbb{E}[(\sigma_t u_t)^4] = \mathbb{E}[\sigma_t^4 u_t^4] = 3\mathbb{E}[\sigma_t^4]$$

This can then be further evaluated

$$\begin{aligned} 3\mathbb{E}[\sigma_t^4] &= 3\mathbb{E}[(\omega + \alpha\epsilon_{t-1}^2 + \beta\sigma_{t-1}^2)^2)] \\ &= 3(\omega^2 + 2\omega(\alpha+\beta)\mathbb{E}[\epsilon_{t-1}^2] + \alpha^2\mathbb{E}[\epsilon_{t-1}^4] + \beta^2\mathbb{E}[\sigma_{t-1}^4] + 2\alpha\beta\mathbb{E}[\epsilon_{t-1}^2\sigma_{t-1}^2]) \\ &= 3(\omega^2 + 2\omega(\alpha+\beta)\mathbb{E}[\epsilon_{t-1}^2] + 3\alpha^2\mathbb{E}[\sigma_{t-1}^4] + \beta^2\mathbb{E}[\sigma_{t-1}^4] + 2\alpha\beta\mathbb{E}[\sigma_{t-1}^2 u_{t-1}^2\sigma_{t-1}^2]), \end{aligned}$$

Which can be rewritten as

$$\begin{aligned} 3\mathbb{E}[\sigma_t^4] &= 3\frac{\omega^2 + 2\omega(\alpha+\beta)\mathbb{E}[\epsilon_{t-1}^2]}{(1 - 3\alpha^2 - \beta^2 - 2\alpha\beta)} \\ &= \frac{3\omega^2(1+\alpha+\beta)}{(1-\alpha-\beta)(1-\beta^2-2\alpha\beta-3\alpha^2)}. \end{aligned}$$

Given this, the kurtosis of the GARCH process can be found as:

$$\begin{aligned} k &= \left(\frac{3\omega^2(1+\alpha+\beta)}{(1-\alpha-\beta)(1-\beta^2-2\alpha\beta-3\alpha^2)}\right) \cdot \left(\frac{\omega}{1-\alpha-\beta}\right)^{-2} \\ &= \left(\frac{3(1+\alpha+\beta)(1-\alpha-\beta)}{1-(\beta^2+2\alpha\beta+3\alpha^2)}\right). \end{aligned} \tag{7}$$

Here it can be see why the condition in Equation 6 must be met, because it allows the denominator of the unconditional kurtosis expression to stay positive.

The expression of the kurtosis leads to an interesting discovery. If α is equal to zero, the kurtosis expression becomes

$$k = \left(\frac{3(1+\beta)(1-\beta)}{1-\beta^2}\right) = \left(3\frac{1-\beta^2}{1-\beta^2}\right) = 3.$$

This indicates that if α is close to zero the unconditional distribution of the GARCH process will be a normal distribution with variance equal to the unconditional variance in Equation 5. Investigating this further, we can evaluate the variance expression in Equation 4 with $\alpha = 0$

$$Var[y_t] = \mathbb{E}[\omega + \beta\sigma_{t-1}^2]$$

This expression is completely independent of y_t, and will therefore converge in a fixed, and predictable manner based on starting variance $Var[y_0]$, as well as the ω and β parameters. Figure 4 shows this convergence for three different β values, assuming $Var[y_0] = 1$ and $\omega = 1$. Here we see that the variance of the processes converge quickly to the unconditional variance, at which point the process will have converged to a white noise process.

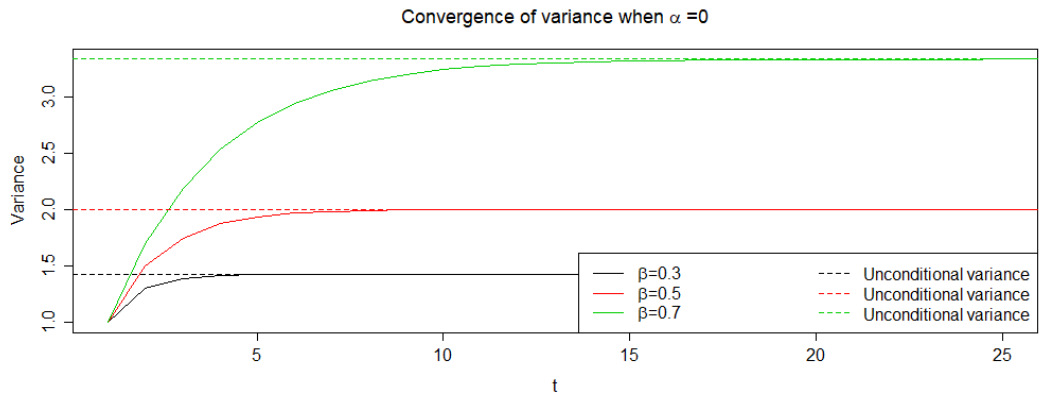


Figure 4: Convergence of the variance in a GARCH model with $\alpha = 0$

In general the GARCH model performs quite well in modelling financial time series, and it corrects most of the problems with using a running variance. Since it allows for much heavier tails it mostly resolves the kurtosis issue.

The GARCH model also removes the need of assuming a window size, instead it is necessary to determine the parameters p and a q parameters. The task of determining a value of p and q is however more feasible than determining a window size of a running variance, and usually a GARCH(1,1) model performs quite well. In the case where a GARCH(1,1) cannot be used, it is also possible to estimate the correct q and p terms. Usually this is done by using AIC or BIC tests.

The GARCH process furthermore captures some of the volatility clustering in the financial

time series, however this is also where the model has its greatest shortcoming. Since β determines the levels of persistence in the volatility shocks produced by α and, $\alpha + \beta < 1$, there is a limit to the amount of persistence in the volatility shocks.
This can be an issue in financial modeling, as structural changes in the financial market, caused by financial crises, can cause longer periods of structurally different volatility.

One way of overcoming this issue is to assume that these structurally different financial periods do in fact exists, and that they should be modelled with different parameters. These kinds of models are called regime shifting models.

4.2 Regime Shift Models

Regime shift models try to explain structural changes, or non-liniarities, in time series models, by assuming the existence of a number of different regimes with different parameters, or even different underlying models. Regime shift models are often divided into two different types, namely threshold models and Markov-Switching models. These models differ in the way that the regimes are defined, and thereby how they are found.
Threshold models assume that shifts in regimes can be determined based on the level of an observed variable, in relation to an unobserved threshold. Threshold models were first introduced in Tong (1978), and include models such as the SETAR (Self-Exciting Threshold AutoRegressive) model and the STAR (Smooth Transition AutoRegressive) model. We will not be working with threshold models further in this thesis, however models such as the one described in this thesis has been created using the threshold framework in Brooks (2001).
Markov-Switching models were first introduced in Goldfeld and Quandt (1973), and in this framework it is assumed that there exists an underlying Markov-Chain which governs which regime the observable variable exists in at any given time. The underlying Markov chain is not observable, as the regimes themselves are assumed to exist, but are not directly observable. A Markov-Chain which is not directly observable, but can be implied through connected observations is called a Hidden Markov model.

4.2.1 Hidden Markov models

A Hidden Markov models denotes a set of two (or more) stochastic processes S and Y, where S is an unobservable Markov chain, that contain information regarding Y which is observable.

The applications for Hidden Markov models are numerous, as it is rarely possible to observe every determining variable, in an empirical study. In Yang (2010) the general theory is covered very well, and this will also be the basis of the theory which is covered in the following section.

Since the unobservable process S contain information regarding the Y variable, the distribution of Y will be dependent on S. Given this the conditional probability of observing Y_i with respect to the set of unobservable S_i can therefore be written as

$$\mathbb{P}(Y_n = y_n \mid S_1 = s_1, ..., S_n = s_n).$$

Here it is essential to clarify that since S is a Markov chain it fulfills the Markov property, however when Y is conditioned by S the process does not necessarily fulfill a Markov like property. That is

$$\mathbb{P}(S_t = s_t \mid S_1 = s_1, ..., S_{t-1} = s_{t-1}) = \mathbb{P}(S_t = s_t \mid S_{t-1} = s_{t-1})$$
$$\mathbb{P}(Y_t = y_t \mid S_1 = s_1, ..., S_{t-1} = s_{t-1}) \neq \mathbb{P}(Y_t = y_t \mid S_{t-1} = s_{t-1}).$$

Using the expression in Equation 4.2.1 it is possible to use maximum-likelihood to determine the most probable values of $S_i \ \forall \ i \in 1, ..., n$, and thereby derive the Hidden Markov chain through the observable variable Y.
One thing to note when calculating the maximum likelihood values of S given Y, is the dependence of Y on S. Using Bayes rule on Equation 4.2.1 gives

$$\mathbb{P}(Y_n = y_n \mid S_1 = s_1, ..., S_n = s_n) = \frac{\mathbb{P}(S_1 = s_1, ..., S_n = s_n \mid Y_n = y_n)\mathbb{P}(Y_n = y_n)}{\mathbb{P}(S_1 = s_1, ..., S_n = s_n)}.$$

Here the term $\mathbb{P}(S_1 = s_1, ..., S_n = s_n \mid Y_n = y_n)$ shows that each state s_i will be conditioned on $y_i, ..., y_T \ \forall \ i \in \{1, ..., T\}$.

In order to account for this dependence a smoothing algorithm is commonly used when maximizing the likelihood expression for a Hidden Markov model. Such an algorithm is described in Hamilton (1994), and the idea is to re-calculate each probability such that it is dependent on the full information set, and not just the previous observation. The probabilities of each state is therefore changed, such that

$$\mathbb{P}(S_t = s_t \mid Y_t) \xrightarrow{Smooth} \mathbb{P}(S_t = s_t \mid Y_T)$$

In order to perform the smoothing, a so called 'forward-backward' algorithm is applied. The 'forward-backward' algorithm is comprised of two steps, first the regular Markov-chain probabilities are calculated, from the probability $\mathbb{P}(S_t = s_t \mid Y_t, S_{t-1} = s_{t-1})$, this

is the forward part of the algorithm. After the forward part has been calculated, each probability will be re-calculated, with a dependence on the future values state variable. In order to do this it is proposed in Hamilton (1994) that the smoothing can be found as

$$\hat{\xi}_{t|T} = \hat{\xi}_{t|t} \odot (\Gamma(\hat{\xi}_{t+1|T} \oslash \Gamma^T \hat{\xi}_{t|t})) \tag{8}$$

where $\odot$ denotes the element-wise product, $\oslash$ denotes the element-wise division and

$$\hat{\xi}_{t|t} = \begin{pmatrix} p(s_t = 1 \mid \Omega_t, \theta_1, \Gamma) \\ \vdots \\ p(s_t = k \mid \Omega_t, \theta_k, \Gamma) \end{pmatrix}$$

$$\Omega_t = \{Y_1, ..., Y_t\}$$

$$\theta_i = \text{parameter space of regime } i.$$

This is the backwards part of the algorithm, since it is calculated iteratively from the last observation to the first. The reason why this is calculated from the last observation to the first is because $\hat{\xi}_{T|T}$ and $\hat{\xi}_{T-1|T-1}$ are known from the forward part of the algorithm. These probabilities can then be used to calculate $\hat{\xi}_{T-1|T}$ using Equation 8, and this procedure can then be continued, until every probability has been smoothed.

4.3 Example of a regime-shifting model

When using Hidden Markov models in regime shifting models it is normal to assume a state space of the state variable $S \in \{1, ..., k\}$ where k is the total number of regimes. In this way each regime will be referenced to by its corresponding integer in the state space.

Using this definition of S, and assuming two regimes, $S \in \{1, 2\}$, a simple regime shift model can be created, such that it has a positive drift in one regime, and a negative drift in another. Such a model can be written as

$$Y_t = Y_{t-1} + \theta_{s_t} + u_t$$

$$\theta_{s_t} = \begin{cases} \mu & \text{if } s_t = 1 \\ -\mu & \text{if } s_t = 2 \end{cases}, \tag{9}$$

where μ defines the drift-rate and u_t is an i.i.d. draw of a standard normal distribution. In Figure 5 and 6 simulated sample paths of the process described in Equation 9 are shown with $\mu = 1$ and $\mu = 0.1$ respectively, and using the transition matrix

$$\Gamma = \begin{bmatrix} 0.99 & 0.01 \\ 0.01 & 0.99 \end{bmatrix}$$

With these transition probabilities the expected length of each regime will be

$$\frac{1}{1-0.99} = 100,$$

using the expression in Equation 2.1. Furthermore the invariant probabilities will be 50% since the transition matrix is symmetrical.

In Figure 5 it is very easy to visually distinguish between the two regimes, since the drift outweighs the error term, the growth rate is almost strictly positive for regime 1, and almost strictly negative for regime 2. Since there is such a clear distinction between the two states in the growth rate, it is almost possible to determine the regimes from the sign of the growth rate alone.

In Figure 6 the error term has a substantially higher effect on the process, as the drift is lower, this makes it much harder to distinguish between the two regimes. It is, however, still clear that there exist two regimes, as there are upward trending periods and downward trending periods.

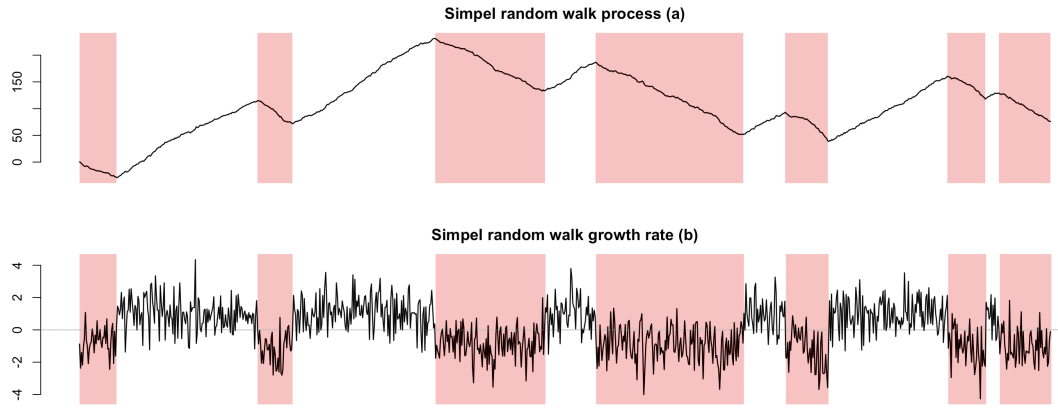


Figure 5: Simulated sample path with 1000 samples of a simple regime shift model described in Equation 9 with $\mu = 1$. The background color illustrates the regime state which the process is in. A white background color indicates regime one (positive drift), and a red background indicates regime two (negative drift) Figure (a) shows the sample path, and Figure (b) shows the growth rate, which is found as the integrated process of (a).

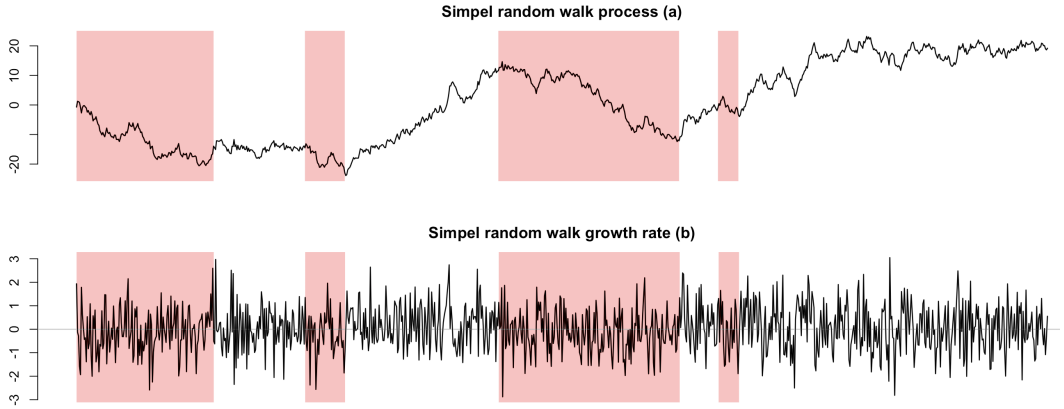


Figure 6: Simulated sample path with 1000 samples of a simple regime shift model described in Equation 9 with $\mu = 0.1$. A white background color indicates regime one (positive drift), and a red background indicates regime two (negative drift) Figure (a) shows the sample path, and Figure (b) shows the growth rate, which is found as the integrated process of (a).

Using regime shifting properties in conjunction with the GARCH model, it may be possible to account for the structurally different volatility, as well as the heteroscedasticity, seen in the financial asset returns, and thereby predicting the volatility more precisely.

5 Markov Switching GARCH model

The original idea of applying Markow Switching models to autoregressive processes, was created in Hamilton (1989), in which a Markov Switching ARIMA model was constructed. This was further developed in Cai (1994) and Hamilton and Susmel (1994) which introduced the Markov Switching ARCH model. The reason why the Markov Switching model was originally only applied to the ARCH model, and not the GARCH model, is due to the fact that the GARCH model introduces a path dependence of the conditional variance. In Gray (1996) it is proposed to use to expected value of the variance, instead of the actual variance in the GARCH model, as a way of overcoming this issue. However this issue has later been completely overcome in Bauwens et al. (2010), where a Gibbs sampler is used to approximate the solution of the maximum likelihood expression.
The theory behind Bayesian inference and MCMC methods, such as the Gibbs sampler, will be explored further in section 6, and how these methods are used in the MS-GARCH model will be explored in section 7.

In the Markov Shifting GARCH (MS-GARCH) model, the framework from the standard GARCH model shown in Equation 2, is reused and expanded. The expansion of the model is done such that each of the parameters of the model at time t is conditioned on the regime state s_t, which is generated from a hidden Markov chain S_t. The MS-GARCH(1,1) model can be written as

$$\begin{aligned} y_t &= \mu_{s_t} + \sigma_t u_t \\ \sigma_t &= \omega_{s_t} + \alpha_{s_t}\epsilon_{t-1}^2 + \beta_{s_t}\sigma_{s_t}^2 \\ \epsilon_t &= y_t - \mu_{s_t} \\ \theta_{s_t} &= \{\omega_{s_t}, \alpha_{s_t}, \beta_{s_t}\} \end{aligned} \tag{10}$$

As with the standard GARCH model we must have that $\omega_{s_t} > 0$, $\alpha_{s_t} \geq 0$ and $\beta_{s_t} \geq 0$, for each state $s_t \in S$ due to the same argument presented in section 4.1.

As the thesis, only covers the GARCH(1,1) model as well as a two regime models, the model parameters can be written as

$$\theta = \begin{bmatrix} \omega_1 & \alpha_1 & \beta_1 \\ \omega_2 & \alpha_2 & \beta_2 \end{bmatrix} \quad \mu^T = \begin{bmatrix} \mu_1, \mu_2 \end{bmatrix} \quad \Gamma = \begin{bmatrix} \eta_{1,1} & \eta_{1,2} \\ \eta_{2,1} & \eta_{2,2} \end{bmatrix}.$$

Using these model parameters the invariant probabilities can be calculated as described in section 2.1

$$\pi = \begin{bmatrix} \pi_1 \\ \pi_2 \end{bmatrix} = \begin{bmatrix} \frac{\eta_{1,2}}{\eta_{1,2}+\eta_{2,1}} \\ \frac{\eta_{2,1}}{\eta_{1,2}+\eta_{2,1}}. \end{bmatrix}$$

In order to show the application of the MS-GARCH model, a simulated MS-GARCH process is shown in Figure 7 (a) which has been simulated using the following parameters

$$\theta = \begin{bmatrix} 0.3 & 0.05 & 0.2 \\ 0.05 & 0.1 & 0.88 \end{bmatrix} \quad \mu^T = \begin{bmatrix} 0.05 & -0.05 \end{bmatrix} \quad \Gamma = \begin{bmatrix} 0.995 & 0.005 \\ 0.005 & 0.995 \end{bmatrix}.$$

This simulation shows a significant resemblance to financial returns, as there are periods of high variance where the asset returns behave in an explosive manner, and periods of low variance.

Figure 7 (b) shows how an asset would have behaved if the asset returns had been the same as the simulated process. In this figure it can be seen that the price movements generated by the MS-GARCH model also resemble real movements in asset prices.

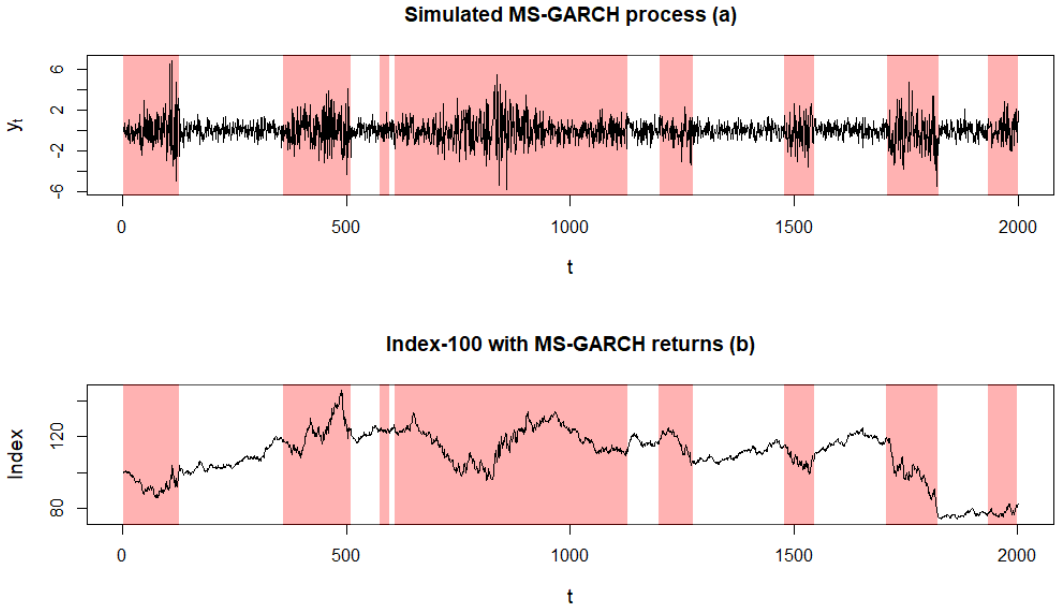


Figure 7: Figure (a) show a simulated sample path with 2000 samples of a MS-GARCH model described in Equation 10. Figure (b) shows how the price of an asset, indexed to 100, would move if the asset returns followed the simulated sample path. The background color indicates the regimes, where white is regime one and red is regime two.

5.1 Stationarity of the MS-GARCH model

Since the MS-GARCH model includes multiple different GARCH processes, the stationarity requirement in Bollerslev (1986), can no be applied directly. However in Bauwens et al. (2010), the stationarity requirements for the MS-GARCH model are given.
Theorem 2.1 in Bauwens et al. (2010) states that a process Y_t is geometrically ergodic, and strictly stationary, if it is initiated from its stationary distribution and follow these three assumptions:

Assumption 1 The error term u_t is i.i.d. with a density function that centered on zero and is positive and continuous everywhere on the real line, and
$\mathbb{E}[\mid u_t^2 \mid^{\delta}] < \infty$ for some $\delta > 0$.

Assumption 2 $\alpha_i > 0$, $\beta_i > 0$ and $\eta_{i,i} \in (0; 1)$ for all $i, j \in 1, ..., m$.

Assumption 3 $\sum_{i=1}^{n} \pi_i \mathbb{E}[log(\alpha_i u_t^2 + \beta_i)] < 0$.

Since we will only consider a standard normal distributed error term in this thesis, ***Assumption 1*** will be satisfied.
Assumption 2 follows non-negativity requirement from stationarity assumptions in Bauwens et al. (2010), however it is more strict, as $\alpha = 0$ and $\beta = 0$ are not allowed.
Assumption 3 follows the strict stationarity requirement from Francq and Zakoïan (2010), where the stationarity of each regime is weighted by the invariate probability of that regime. This assumption therefore state that if all regimes are strictly stationary, the MS-GARCH process will also be strictly stationary. However it also says that for the MS-GARCH process to be strictly stationary all regimes *need* not be strictly stationary, as long as there is enough stationarity in the remainder of the regimes.

In Bauwens et al. (2010) it is also argued that under ***Assumption 1 - 2*** the moments of order k will exist if the following conditions are met

$$\sum_{i=1}^{m} \pi_i \mathbb{E}\left[(\alpha_i u_t^2 + \beta_i)^k\right] < 1$$

$$\mathbb{E}[y_t^{2k}] < \infty.$$

Furthermore, moments of 2. order will exist when the conditions are met for k=1.

5.2 Central moments in the MS-GARCH model

Since the mean in the MS-GARCH process is not affected by the variance, it is possible to determine the unconditional mean as a weighted average of the drift coefficients, where

the weight is determined by the invariant probabilities, as described in Section 4.2. The expression for the unconditional mean of the MS-GARCH model will therefore be

$$\mu_{MS-GARCH} = \pi_1\mu_1 + \pi_2\mu_2.$$

Determining the remaining central moments of the MS-GARCH model is much more difficult, as they are affected by a path dependency of the variance process. This path dependency is described further in section 5.3.
It is however possible to estimate these parameters, by simulating a set of MS-GARCH paths, and then calculate the central moments based on these paths.

In Table 1 the central moments of the MS-GARCH model with the parameters shown in Equation 5 can be seen. These values are found as the average of the empirical moments from 10.000 different MS-GARCH sample paths, each with a length of 10.000. The moments from each of the regimes has been calculated using the theoretical expressions described in section 4.1. Here we see that the variance of the MS-GARCH model lie just below the average of the variances of the two underlying regimes.
We also see that the kurtosis is much higher than either of the two underlying regimes. We do also expect this as the high volatility regime pushes observations out in the tails, whereas the low volatility regime draw observations closer to 0. This results in a high number of observations close to zero, and a high number of observations far away from 0, which creates high kurtosis.

	MS-GARCH	Regime 1	Regime 2
Mean	0	0.5	-0.5
Variance	1.24	0.4	2.5
Kurtosis	7.33	3.02	6.06

Table 1: Comparison of the estimated central moments of the MS-GARCH model, and the central moments of the underlying regimes.

5.3 Maximum likelihood in MS-GARCH

When applying the MS-GARCH model to a data set, the only information available is the Y process it self, and therefore it is necessary to estimate both the parameters of the model, as well as the underlying regime state Markov chain. In order to do this, a maximum likelihood expression is needed. Since we are estimating two processes, namely

the process itself and the underlying regime states, the likelihood expression will be a joint probability of the two, or more precisely it will be

$$\begin{aligned} p(y_t, s_t \mid \mu, \theta, \Gamma, Y_{t-1}, S_{t-1}) &= p(y_t \mid \mu, \theta, \Gamma, Y_{t-1}, S_{t-1}) p(s_t \mid \Gamma, S_{t-1}) \\ &= p(y_t \mid \mu, \theta, \Gamma, Y_{t-1}, S_{t-1}) p(s_t \mid \Gamma, s_{t-1}) \end{aligned}$$

where $Y_t = \{y_1, ..., y_t\}$ and $S_t = \{s_1, ..., s_t\}$. The last equality arises due to the Markov property of S_t

Since the error term of the process is a normal distribution, the probability of observing y_t can be found from a normal distribution as well

$$p(y_t \mid s_t, \mu, \theta, Y_{t-1}, S_t) = \frac{1}{\sqrt{2\pi}\sigma_t} \exp\left[\frac{(y_t - \mu_{s_t})^2}{2\sigma_{s_t}^2}\right]. \tag{11}$$

Since S is a Markov chain, and therefore is time homogeneous the state probabilities are given as the probability of being in a specific state, given the previous state. Using the transition matrix, the state probabilities can therefore be found as

$$p(s_t \mid \Gamma, s_{t-1}) = \eta_{s_{t-1}, s_t}. \tag{12}$$

The maximum likelihood expression for the joint series of Y and S is therefore

$$\begin{aligned} p(Y, S \mid \mu, \theta, \Gamma) &\propto p(y_1, s_1 \mid \mu, \theta, \Gamma, Y_0, S_0) \cdot ... \cdot p(y_T, s_T \mid \mu, \theta, \Gamma, Y_{T-1}, S_{T-1}) \\ &\propto p(y_1 \mid \mu, \theta, \Gamma, Y_0, S_0) p(s_1 \mid \Gamma, s_0) \cdot ... \\ &\quad ... \cdot p(y_T \mid \mu, \theta, \Gamma, Y_{T-1}, S_{T-1}) p(s_T \mid \Gamma, s_{T-1}) \\ &\propto \prod_{t=1}^{t=T} \frac{1}{\sqrt{2\pi}\sigma_t} \exp\left[-\frac{(y_t - \mu_{s_t})^2}{2\sigma_t^2}\right] \eta_{s_{t-1}, s_t}. \end{aligned} \tag{13}$$

Maximizing the likelihood expression in Equation 13 turns out to be very difficult, because of the structure of the σ-parameter. The variance at each point in time is dependent on the previous variance, and this effect adds up, such that each variance will depend on all the previous variances. Given that there exists several regimes in the MS-GARCH model, the parameter set changes over time, which also changes the variance, and since each variance is dependent on the previous, a path dependency is created.

The generation of the path dependency is also illustrated in Figure 8, which shows how the variance behaves as time passes.

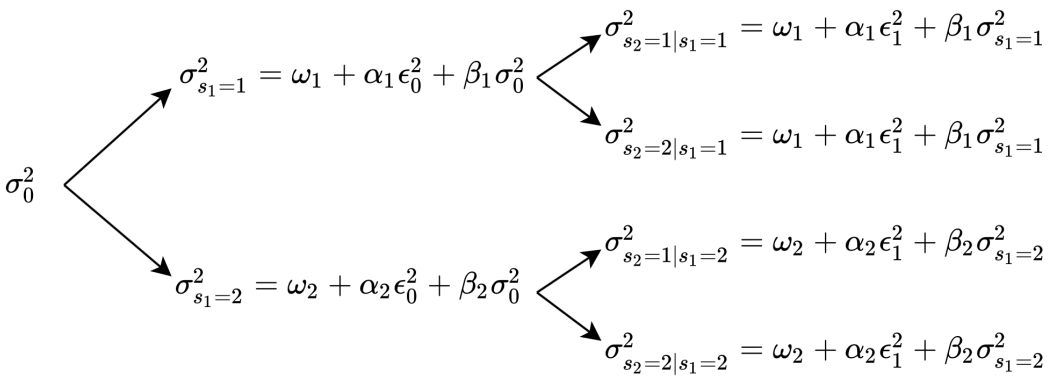


Figure 8: Illustration of path dependence of MS-GARCH variance, based on equivalent diagram in Bauwens et al. (2010).

The path dependence of the σ-parameter is transferred to the maximum likelihood in Equation 13, since it depends on the σ-parameter. In order to maximize the likelihood, via a regular frequentistic approach it would be necessary to test each combination of regime states, and assuming there is n observations and k regimes this result in k^n difference combinations, which quickly becomes impossible to compute with the hardware available today.

It is therefore necessary to use a different approach to the maximization, in Bauwens et al. (2010) a Bayesian approach is suggested, which utilizes MCMC methods in order to maximize the likelihood expression. The theory behind this is covered in more detail in section 6, and the estimation algorithm is described in section 7.

6 Bayesian statistics

The information regarding Bayesian inference and Monte Carlo methods are described widely in the literature. The literature used in this thesis has been found in Tsay (2005), Gelman et al. (2014) and Hahn (2013-2014). Furthermore Gamerman and Lopes (2006) gives a thorough explanation of Markov chain Monte Carlo.

In Bayesian inference the idea is that the parameter values θ can be explained from a probabilistic point of view. In this setup θ is conditioned on an observable value Y, which is expressed as $p(\theta|Y)$. In a Bayesian framework θ is considered to be a random variable spanning a parameter space. Since θ is a random variable it is necessary to use a probability distribution as a representation for it.

Another approach to statistic inference is frequentist inference where θ is assigned to a fixed value which optimizes the likelihood based on a sample of observable data Y. This is referred to as maximum likelihood estimation.
As described in Section 5.3 this method can be rather complex when there exists a path dependence in the Y-process. In such cases Bayesian inference is preferred.

To give a guidance for further notation, $p(\theta, y)$ denote the joint density, $p(.|.)$ the conditioned density and $p(.)$ is the marginal density. As mentioned earlier Bayesian statistic tries to determine $p(\theta|Y)$, but first of all to get an idea of $\theta|Y$ the joint density of θ and Y is introduced as $p(\theta, Y)$. This expression can be written as

$$p(\theta, Y) = p(\theta)p(Y|\theta), \tag{14}$$

where $p(\theta)$ is a prior density of the parameter values. This is an indicator of how θ is distributed before Y is collected/observed, and it contains the prior belief before collecting data. As an example, in Section 4.2 it was described how a Hidden Markov model changes state based on a unobserved transition probability. When estimating these probabilities a proper choice of prior densities could be a uniform or a beta distribution, which are both limited to the interval $[0; 1]$.
The prior density is therefore compliant with the space in which the parameter is defined. Since the the joint density is affected by the prior density, this can create issues and lead to skewed image of the joint density. Therefore the choice of prior density may affect posterior density. This problem is solved when the number of data points in Y is large.

$p(Y|\theta)$ can be seen as the "likelihood-density" of the observed Y conditioned by θ. It provides the likelihood for each value of θ having led to the observations of Y. In Frequentist inference estimating θ will be equivalent to maximize $p(Y|\theta)$ with respect to θ.
Bayesian inference is based on Bayes rule, which is found by normalizing the expression in equation 14 with the marginal density of Y

$$p(\theta|Y) = \frac{p(\theta, Y)}{p(Y)} = \frac{p(\theta)p(Y|\theta)}{p(Y)}, \tag{15}$$

which is the normalized posterior density. Since the denominator does not include θ, it can simply be seen as a constant. By normalizing with $p(y)$, each value of θ is therefore scaled with the same $p(y)$. Thus it can be left out and the un-normalized posterior density can be written as

$$p(\theta|y) \propto p(\theta)p(y|\theta). \tag{16}$$

where the denominator $p(y)$ has been replaced with a proportional expression, which is an alternative way of denoting the scaling of $p(\theta)p(y|\theta)$ such that it is represented as a density.

6.1 Monte Carlo methods

A possible solution to problems with dependence structures is, to use a Markov Chain Monte Carlo (MCMC) method.

The MCMC method contain two components, namely the Markov chain and the Monte Carlo method. The Markov chains follows the theory described in section 2.1 both for discrete and continues state space. Since using an iterative method for estimating the parameters, a Markov chain for each of the parameters is created with the posterior distribution as invariant distribution.

The Monte Carlo part refers to random draws from a process or distribution in order to obtain a numerical result. This must be done a sufficient amount of times to ensure, that the underlying Markov chain has converged. Otherwise it is a uncertain it the Markov chain has reached a invariant distribution.

The goal for the MCMC procedure is that given a large n the invariant distribution, π, can be approximated for the posterior distribution. it is required that certain properties described in Section 2 must be satisfied, otherwise it is uncertain whether the Markov chain will converge to the invariant distribution.

It is required that the Markov chain is aperiodic and Harris recurrent, and thus also ergodic, see. Definition 2.2 - 2.5 and theorem 2.1. When the chain is Harris recurrent it will also be irreducible. If these conditions are met, it ensures that the Markov chain will converge and that the chain is independent of the initial starting point.

Numerous methods involving MCMC have been developed for solving high dimensional estimations. This could be the case for parameters in a high dimensional distribution. We will examine one of the most well-known in the following section.

6.2 Gibbs sampling

One MCMC method which is frequently used when dealing with high dimensional distribution, is the Gibbs sampler. This is a special case of the Metropolis-Hastings algorithm, which was originally introduced in Metropolis et al. (1953) and later extended in Hastings (1970).
The Gibbs sampler was introduced in Geman and Geman (1984), and its usefullness for handling multidimensional estimation problems was pointed out in Gamerman and Lopes (2006).
Other MCMC methods, like the Metropolis-Hastings algorithm, draws from the joint distribution of θ. This can be very difficult in with high dimensional distributions.
With the Gibbs sampling algorithm the parameter set θ is divided into d sub-vectors $\theta = \{\theta_1, ..., \theta_d\}$. This way sampling θ can be handled by sampling from the conditional distribution $p(\theta_j|\theta_{-j})$.
Furthermore the Gibbs sampler is applicable when the joint distribution is unknown or difficult to sample from, but the conditional distribution for all subset of the parameters are accessible to sample from.

The Gibbs sampling method samples the parameter space spanning $\theta = \{\theta_1, ..., \theta_d\}$, which is done by iterating over the full conditional distribution

$$\theta_j \sim p(\theta_j|\theta_{-j}),$$

where $\theta_{-j} = \{\theta_1, ..., \theta_{j-1}, \theta_{j+1}, ...\theta_d\}$. At iteration r the sampling of θ_j^r can be written as

$$\theta_j^r \sim p(\theta_j|\theta_{-j}^{r-1}),$$

where the superscript denotes the current iteration of the parameter, and
$\theta_{-j}^{r-1} = \{\theta_1^r, ..., \theta_{j-1}^r, \theta_{j+1}^{r-1}, ...\theta_d^{r-1}\}$.
Thus the current component of θ is updated conditioned on the past values of iteration r

as well as the past values in iteration $r-1$ which have not yet been updated.

Since the Gibbs sampler is a MCMC method, the sampling of $\theta_j \sim p(\theta_j|\theta_{-j})$ is repeated until each Markov chain of $\theta_1, ..., \theta_d$ have converged to the invariant distribution. Usually a burn-in period is removed, consisting of the first iterations of the Gibbs sampler, where the distribution has not yet converged. This is done such that all the samples have been drawn from the invariant distribution.
In Section 2.1 we described the properties of the Markov chains, which must be satisfied in order to ensure that a invariant distribution exists. This must be accounted for, for each of the parameters. Furthermore due to the Ergodic Theorem 2.1 we can approximate the mean of the posterior distribution and thereby get an estimate of the parameters for the MS-GARCH model.

The Gibbs sampling algorithm draws from the conditional distribution, the trajectories can be shown to be orthogonal. This is showed by the following example.
Suppose we have a single set of coordinates (y_1, y_2) that follow a bivariate normal distribution with unknown mean $\mu = (\mu_1, \mu_2)$ and known variance $\Sigma = \begin{pmatrix} 1 & 0.7 \\ 0.7 & 1 \end{pmatrix}$.
When the prior distribution for μ is uniformly distributed, the posterior distribution can be written as:

$$\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \Big| y \sim \mathcal{N}\left(\begin{pmatrix} y_1 \\ y_2 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right)$$

and the full conditional distribution is given by:

$$\mu_1|\mu_2, y \sim \mathcal{N}(y_1 + \rho(\mu_2 - y_1), 1 - \rho^2)$$
$$\mu_2|\mu_1, y \sim \mathcal{N}(y_2 + \rho(\mu_1 - y_2), 1 - \rho^2)$$

This was shown in Gelman et al. (2014) and gives great view of the aforementioned trajectory of the Gibbs sampler.

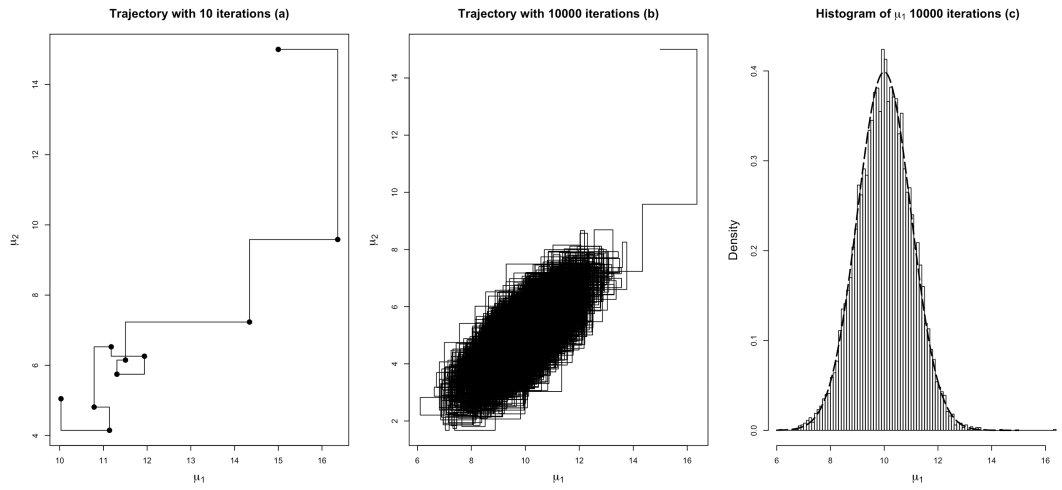


Figure 9: Figure (a) and (b) shows trajectories of the Gibbs sampling method with 10 and 10,000 iterations respectively. The filled black dots in Figure (a), indicates each iteration. The marginal density of μ_1 shown in Figure (c), where the black dashed line is a $\mathcal{N}(10, 1)$.

From the trajectories shown in Figure 9 (a) and (b), it is clear to see the sub-vector wise updating of μ_1 and μ_2, hence the orthogonal movement from updating each parameter.
However this can be cumbersome for a large number of parameters, yet in many cases it is still easier to compute than the Metropolis-Hastings, which requires sampling from the joint density i.e. sampling all parameters at once.
It is not always possible to work with conjugate distributions and it may therefore be necessary to use numerical integration to overcome this issue.

6.2.1 Griddy-gibbs sampling

Sometimes the true full conditional distribution $p(\theta_j|\theta_{-j})$ is unobtainable, making it impossible to draw sample directly from it. In estimation of the MS-GARCH model, this is case for θ and μ, as it will later be explained in Section 7.5.
In order to resolve this issue, Ritter and Tanner (1992) constructed the Griddy-gibbs sampler, which samples from an approximated conditional distribution, instead of the full conditional distribution.
The way this is done, is by numerically integrating the known conditional density, to get an approximated conditional distribution, or CDF. When the CDF has been approximated it is inverted, leaving the quantile function, which can be used to draw samples from.
More specifically the Griddy-Gibbs sampling algorithm will follow these steps:

1. Select a grid of m points $\{\theta_{i,1}, ..., \theta_{i,m}\} \in \Theta$ and evaluate the conditional posterior distribution $p_i(\theta_{i,j})$ for $j = 1, .., m$.

2. Construct an approximation of the inverse CDF of $p_i(\theta_{i,j})$, which belongs in the interval [0;1].

3. Draw a value from a uniform distribution and transform it by using the approximated inverse CDF in order to obtain a value for θ_i

There are several ways of iterating through step 1 to 3, however the inverted CDF must be in [0;1], such that the sampling of θ_i can be obtained by drawing from uniform distribution in that interval. The theory discussed in this section, will be applied to the MS-GARCH model in the following section. Here a Gibbs sampling method will be applied, which resolves the issue created by the path dependence of the variance structure.

7 Estimation of MS-GARCH using Gibbs sampling

Throughout this section we describe the estimation procedure of the MS-GARCH model, as well as some modifications which have been necessary to implement.
Recall Equation 10 the MS-GARCH model was written as

$$
\begin{aligned}
y_t &= \mu_{s_t} + \sigma_{s_t} u_t \\
\sigma_{s_t} &= \omega_{s_t} + \alpha_{s_t} \epsilon_{t-1}^2 + \beta_{s_t} \sigma_{s_{t-1}}^2 \\
\epsilon_t &= y_t - \mu_{s_t} \\
\theta_{s_t} &= \{\omega_{s_t}, \alpha_{s_t}, \beta_{s_t}\}
\end{aligned}
$$

For this model the parameters that need to be estimated are θ, μ, Γ and S.
The sampling procedure follows a Gibbs-sampling algorithm as described in section 6.2. However some of the concepts have been tweaked, in order to make the Gibbs-sampling work in the MS-GARCH framework. The exact sampling procedure for these parameters will be examined later in this section.
The estimation algorithm is written in *RCPP*, which is a *C++* implementation embedded in the statistical software *RStudio*. The choice of using *RCPP* was made because of the computational load associated with MCMC estimation. The implementation of the algorithm can be found in the appendix, along with an example code.

7.1 Choosing prior distributions

For the estimation of Markov-switching models using a Bayesian approach, a proper prior distribution must be chosen. A property concerning the parameters, that must be accounted for when choosing a prior distribution, is the restrictions regarding the parameter space of the variable.

The parameters α, β and Γ are all restricted to the interval [0;1], as described in Section 5. Knowing these limitations it would be appropriate to chose a distribution which also is limited to that interval. A distribution that lie in [0;1] could the beta distribution. When using the beta distribution, a set of shape parameters has to be selected. These shape parameter can either be chosen such that they introduce some prior beliefs, to the shape of the posterior distribution. Otherwise they can both be set to 1 which is equivalent to choosing a uniform distribution.

The minimal variance of the process, ω, is strictly positive. Because of this, the prior distribution of the parameter must also be strictly positive. Some examples of suitable prior distributions could therefore be the *log-normal* distribution, the χ^2 distribution, or another distribution which is truncated such that it is strictly positive.

The drift parameter μ is not limited. Therefore we are able to choose any distribution spanning the interval $]-\infty;\infty[$.
In financial time series μ often lie around 0, therefore choosing a distribution that is symmetric around 0 could be a suitable.

For all the parameters contained in θ, the uniform distribution with appropriate boundaries, could be a good choice. By choosing a prior distribution other than uniform two things must be taking into account. Firstly, if the coefficients in the distribution are chosen, what should they be in order to capture the prior belief of the shape of the posterior distribution. Secondly, if the coefficients are not chosen in advance, these hyper-parameters must estimated, which can be difficult.
A more feasible approach would be choosing a distribution which has equal probability across the entire parameter space, hence choosing a uniform distribution which span a pre-selected area.

A problem that can occur, if the parameter space of two regimes are too similar is that the Gibbs sampler may not be able to separate the regimes. This problem is described in James D. Hamilton and Zha (2007) as label-switching of the parameters. When label-switching is not accounted for, it can cause the two regimes to switch state numbers, throughout the estimation procedure, resulting in inaccurate estimates, and non-converging posterior densities.
There are several ways to avoid this issue, where one of the more simple ways is to restrict each parameter to a confined area, such that a wide overlap between regimes is avoided. However when choosing the boundaries of the parameter space, two issues must be accounted for, as described in Bauwens et al. (2010).

1. The truncation of the prior distribution is to narrow and thereby the parameter can be caught in the boundaries.

2. Choosing too wide boundaries, such that the prior distribution is multiplied by 0 for a large area of the parameter space, making the posterior computation inefficient.

7.2 Log transformation

Often it is more convenient to work with the log-likelihood expression instead of the normal likelihood expression written in equation 13. It has also been necessary to make this transformation in order to estimate the parameters. This is due to the computational capabilities for 64-bit computing, where (very) small number are rounded off to equal zero. As an example 2.225074e-308 is the smallest non-zero number, that R can handle. This can quickly become an issue for a large value of T in the likelihood expression, as this is a product of terms which are mostly between 0 and 1. This issue will further be referred to as numerical instability.

Due to proportionality of the posterior distribution we can easily handle the numerical instability by parallel shifting the log-likelihood before transforming it back. Doing this will keep the the proportionality of the likelihood expressions, and it will make it numerically stable.

For the log-likelihood expression adding a number is equivalent to multiplying in the likelihood expression. This way we can successfully transform to and back from the log-scale and avoiding numerical instability.

7.3 Sampling regime states s

When estimating the parameters in the MS-GARCH model, we will start by estimating the regime states. This procedure is shown in the orange part of the flowchart on page 43.

To do this, the likelihood expression for $p(y, s, \mid \mu, \theta, \Gamma)$ can be rewritten, to find the conditional posterior distribution of s_t given everything else, as

$$\begin{aligned} p(y, s, \mid \mu, \theta, \Gamma) &\propto \prod_{j=1}^{T} \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left[-\frac{(y_j - \mu_{s_j})^2}{2\sigma_j^2}\right] \eta_{s_{j-1}, s_j} \\ &\propto \prod_{j=1}^{T} \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left[-\frac{(y_j - \mu_{s_j})^2}{2\sigma_j^2}\right] \prod_{j=1}^{T} \eta_{s_{j-1}, s_j} \\ \propto p(s_t | S_{\neq t}, \mu, \theta, \Gamma, y) &\propto \prod_{j=1}^{T} \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left[-\frac{(y_j - \mu_{s_j})^2}{2\sigma_j^2}\right] \eta_{s_{t-1},1}^{2-s_t} \eta_{s_{t-1},2}^{s_t - 1} \eta_{s_t,1}^{2-s_{t+1}} \eta_{s_t,2}^{s_{t+1}-1}. \end{aligned}$$

We arrive at the last expression since $p(s_t | S_{\neq t}, \mu, \theta, \Gamma, y)$ is only dependent on terms which include s_t, as the rest of the terms are equivalent for all state values of s_t. Since it is a proportional expression, these values will evaluated as constants, and thereby have no effect on the likelihood.

The dependence of the transition probabilities from the previous state, and to the next

state, stem from the Markov property of S, which is why they have to be accounted for in this expression.

The expression can now be further rewritten, as

$$\begin{aligned} p(s_t|S_{\neq t},\mu,\theta,\Gamma,y) &\propto \eta_{s_{t-1},1}^{2-s_t}\eta_{s_{t-1},2}^{s_t-1}\eta_{s_t,1}^{2-s_{t+1}}\eta_{s_t,2}^{s_{t+1}-1}\prod_{j=1}^{T}\frac{1}{\sqrt{2\pi}\sigma_j}\exp\left[-\frac{(y_j-\mu_{s_j})^2}{2\sigma_j^2}\right] \\ &\propto \eta_{s_{t-1},1}^{2-s_t}\eta_{s_{t-1},2}^{s_t-1}\eta_{s_t,1}^{2-s_{t+1}}\eta_{s_t,2}^{s_{t+1}-1}\prod_{j=t}^{T}\frac{1}{\sqrt{2\pi}\sigma_j}\exp\left[-\frac{(y_j-\mu_{s_j})^2}{2\sigma_j^2}\right] \\ &\quad\cdot\prod_{j=1}^{t-1}\frac{1}{\sqrt{2\pi}\sigma_j}\exp\left[-\frac{(y_j-\mu_{s_j})^2}{2\sigma_j^2}\right] \\ &\propto \eta_{s_{t-1},1}^{2-s_t}\eta_{s_{t-1},2}^{s_t-1}\eta_{s_t,1}^{2-s_{t+1}}\eta_{s_t,2}^{s_{t+1}-1}\prod_{j=t}^{T}\frac{1}{\sqrt{2\pi}\sigma_j}\exp\left[-\frac{(y_j-\mu_{s_j})^2}{2\sigma_j^2}\right] \\ &\propto \eta_{s_{t-1},1}^{2-s_t}\eta_{s_{t-1},2}^{s_t-1}\eta_{s_t,1}^{2-s_{t+1}}\eta_{s_t,2}^{s_{t+1}-1}\prod_{j=t}^{T}\frac{1}{\sigma_j}\exp\left[-\frac{(y_j-\mu_{s_j})^2}{2\sigma_j^2}\right]. \end{aligned} \tag{17}$$

The reason why the product term only has to be evaluated from $t \to T$, follows the same argumentation as for the transition probabilities, since the product term from $1 \to t-1$ will also be evaluated as a constant.

The reason why there is a dependence on the terms after time t is due to the path dependency of σ_t which is affected by every s_i for $i \leq t$.

As described previously it is preferable to use the log-likelihood expression, Equation 17 will therefore be transformed as

$$q(s_t|S_{\neq t},\mu,\theta,\Gamma,y) \propto \log\left(\eta_{s_{t-1},1}^{2-s_t}\eta_{s_{t-1},2}^{s_t-1}\eta_{s_t,1}^{2-s_{t+1}}\eta_{s_t,2}^{s_{t+1}-1}\right) + \sum_{j=t}^{T} -\log(\sigma_j) - \frac{(y_j-\mu_{s_j})^2}{2\sigma_j^2} \tag{18}$$

Since s_t can only take on the two values 1 and 2, we can sample it by drawing from a Bernoulli distribution, with probability π, where π for regime 2 is calculated as the proportion of the likelihood

$$\pi_2 = \frac{e^{q(s_t=2|.)}}{e^{q(s_t=1|.)}+e^{q(s_t=2|.)}} = \frac{p(s_t=2|.)}{p(s_t=1|.)+p(s_t=2|.)},$$

and the probability of sampling regime 1 will be $1-\pi$. This procedure is then repeated for $t = 1, ..., T$, such that each regime state is estimated.

7.4 Sampling transition probabilities Γ

The next part of the sampling procedure will be the sampling of Γ. This can be seen in the blue part of the flowchart on page 43.

The the conditional posterior distribution can be derived as

$$\begin{aligned} p(y, s, \mid \mu, \theta, \Gamma) &\propto p(\Gamma) \prod_{j=1}^{T} \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left[-\frac{(y_j - \mu_{s_j})^2}{2\sigma_j^2}\right] \eta_{s_{j-1}, s_j} \\ &\propto p(\Gamma) \prod_{j=1}^{T} \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left[-\frac{(y_j - \mu_{s_j})^2}{2\sigma_j^2}\right] \prod_{j=1}^{T} \eta_{s_{j-1}, s_j} \\ \propto p(\Gamma|S) &\propto p(\Gamma) \prod_{t=1}^{T} \eta_{st-1, s_t}. \end{aligned} \tag{19}$$

Since Γ does not depend on μ, θ and y, the conditional posterior distribution will only depend on S.
Furthermore the product term can be split up as

$$p(\Gamma|S) \propto p(\Gamma) \cdot \eta_{1,1}^{n_{11}} \cdot \eta_{1,2}^{n12} \cdot \eta_{2,1}^{n21} \cdot \eta_{2,2}^{n22},$$

where $n_{i,j}$ denotes the number of times $s_{t-1} = i$ and $s_t = j$.
Since $\sum_{j=1}^{2} \eta_{i,j} = 1$ for $i = 1, 2$ the expression can be rewritten, by changing $\eta_{1,2}$ to $(1 - \eta_{1,1})$ and $\eta_{2,1}$ to $(1 - \eta_{2,2})$

$$p(\Gamma|S) \propto p(\Gamma)(\eta_{1,1}^{n_{11}}(1 - \eta_{1,1})^{n_{12}}(1 - \eta_{2,2})^{n_{21}}\eta_{2,2}^{n_{22}}).$$

By the independence of $\eta_{1,1}$ and $\eta_{2,2}$ the probabilities can be considered separately, which will simplify the problem substantially.
Because all elements in Γ belongs to the interval [0;1], a beta prior distribution can be used for each of them. In Equation 20 this is done for $\eta_{1,1}$

$$\begin{aligned} p(\eta_{1,1}|S) &\propto \frac{\eta_{1,1}^{a_{11}-1}(1 - \eta_{1,1})^{a_{12}-1}}{\Gamma(a_{11}, a_{12})}(\eta_{1,1}^{n_{11}}(1 - \eta_{1,1})^{n_{12}}) \\ &\propto \eta_{1,1}^{a_{11}-1}(1 - \eta_{1,1})^{a_{12}-1}(\eta_{1,1}^{n_{11}}(1 - \eta_{1,1})^{n_{12}}) \\ &\propto \eta_{1,1}^{a_{11}-1+n_{11}}(1 - \eta_{1,1})^{a_{12}-1+n_{12}}, \end{aligned} \tag{20}$$

where $\Gamma(.,.)$ denotes the gamma distribution.
The values a_{11}, a_{12} are the shape parameters belonging to the prior distribution. When $a_{11} = a_{12} = 1$ the prior distribution will be a uniform distribution. Therefore the samples from the posterior distribution concerning $\eta_{1,1}$, and equivalently for $\eta_{2,2}$, can be drawn from a beta distribution.

7.5 Sampling θ and μ

The conditional posterior distribution of θ and μ are found equivalently to the previous parameters, and the sampling procedure can be seen in the purple and green part of the

flowchart on page 43.
Firstly the conditional posterior distribution is found as

$$\begin{aligned} p(y, s, \mid \mu, \theta, \Gamma) &\propto p(\theta, \mu|S, y) \prod_{j=1}^{T} \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left[-\frac{(y_j - \mu_{s_j})^2}{2\sigma_j^2}\right] \eta_{s_{j-1}, s_j} \\ &\propto p(\theta, \mu) \prod_{j=1}^{T} \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left[-\frac{(y_j - \mu_{s_j})^2}{2\sigma_j^2}\right] \prod_{j=1}^{T} \eta_{s_{j-1}, s_j} \qquad (21) \\ \propto p(\theta, \mu|S, y) &\propto p(\theta, \mu) \prod_{j=1}^{T} \frac{1}{\sigma_j} \exp\left[-\frac{(y_j - \mu_{s_j})^2}{2\sigma_j^2}\right], \end{aligned}$$

where unnecessary constants have been left out, and $p(\theta, \mu)$ denotes the chosen prior distribution.

It is once again advantageous to log-transform Equation 21 because of the numerical instability. With a log transformation the expression becomes

$$q(\theta, \mu|S, y) \propto \log\left[p(\theta, \mu)\right] + \sum_{j=1}^{T} -\log(\sigma_j) + -\frac{(y_j - \mu_{s_j})^2}{2\sigma_j^2}, \qquad (22)$$

with $q(\theta, \mu|S, y) = \log(p(\theta, \mu|S, y))$.

Since this expression can not be simplified to a closed form solution, and in particular since it does not follow a standard distribution, sampling these parameters is not as is not as straightforward as for the other parameters.
It will be necessary to make an approximation of the posterior distribution for each iteration r that is different than the ones used in Section 7.3 and 7.4.
In Section 6.2.1 we described such a method for handling such an issue, namely the Griddy Gibbs method.
By applying this method we start by sampling ω_1

1. By using Equation 22 we can compute the kernel density of the posterior distribution, $\kappa_i(\omega_1^i|S^{(r)}, \beta_1^{(r-1)}, \alpha_1^{(r-1)}, \theta_2^{(r-1)}, \mu^{(r-1)}, y)$, where (r) denotes the iteration step. κ_i denotes evaluating the density in the i^{th} point of a fixed grid $(\omega_1^1, ..., \omega_1^G)$. By evaluating the posterior distribution over the chosen grid, the vector $\nu_\kappa = (\kappa_1, ..., \kappa_G)$ is obtained.

   Since a log-transformation of the posterior distribution is used, the points calculated for ν are on a log scale. In order to compute the CDF it is therefore necessary to make the transformation of ν back from the log scale, by taking the exponential

of ν.

We will denote the element-wise exponential transformation of ν by $\Psi = e^{\nu}$, where a parallel shift has been applied to the series of ν such that the numerical instability, as explained previously, is avoided.

2. To approximate the CDF, a Riemann sum over the chosen grid can be used. This can be calculated as

$$f_j = \sum_{i=2}^{j} \Psi_i(\omega_1^i - \omega_1^{i-1}) \text{for } j = 2, ..., G$$

To make f a proper distribution function, every point in f is normalised by $\frac{1}{f_G}$.
Later in this section we explain more thoroughly how to ensure an accurate approximation of the CDF.

3. Find the inverse of f, i.e. the quantile function, f^{-1}. Then draw
$\omega_1^r \sim p(\omega_1|S^{(r)}, \beta_1^{(r-1)}, \alpha_1^{(r-1)}, \theta_2^{(r-1)}, \mu^{(r-1)}, y)$ as follows:
First draw $u \sim uniform(0, 1)$, and then use the approximated quantile function to obtain ω_1^r.
In order to use the approximated quantile function, it is necessary to interpolate between the points generated by the numerical integration. There are several ways to interpolate between the points in f^{-1}, however a linear interpolation is the most simple, and through testing it has been found to be sufficient.

4. Continue similarly for the parameters $\alpha_1^r \to \beta_1^r \to \omega_2^r \to \alpha_2^r \to \beta_2^r \to \mu^r \to \omega_1^{r+1}$.

One of the issues encountered when implementing the estimation procedure of the parameters in the MS-GARCH model, was instability in the numerical integration of the posterior distributions for θ and μ.
In the following section we will give a more thorough description of the numerical integration procedure used in the Griddy Gibbs sampler.

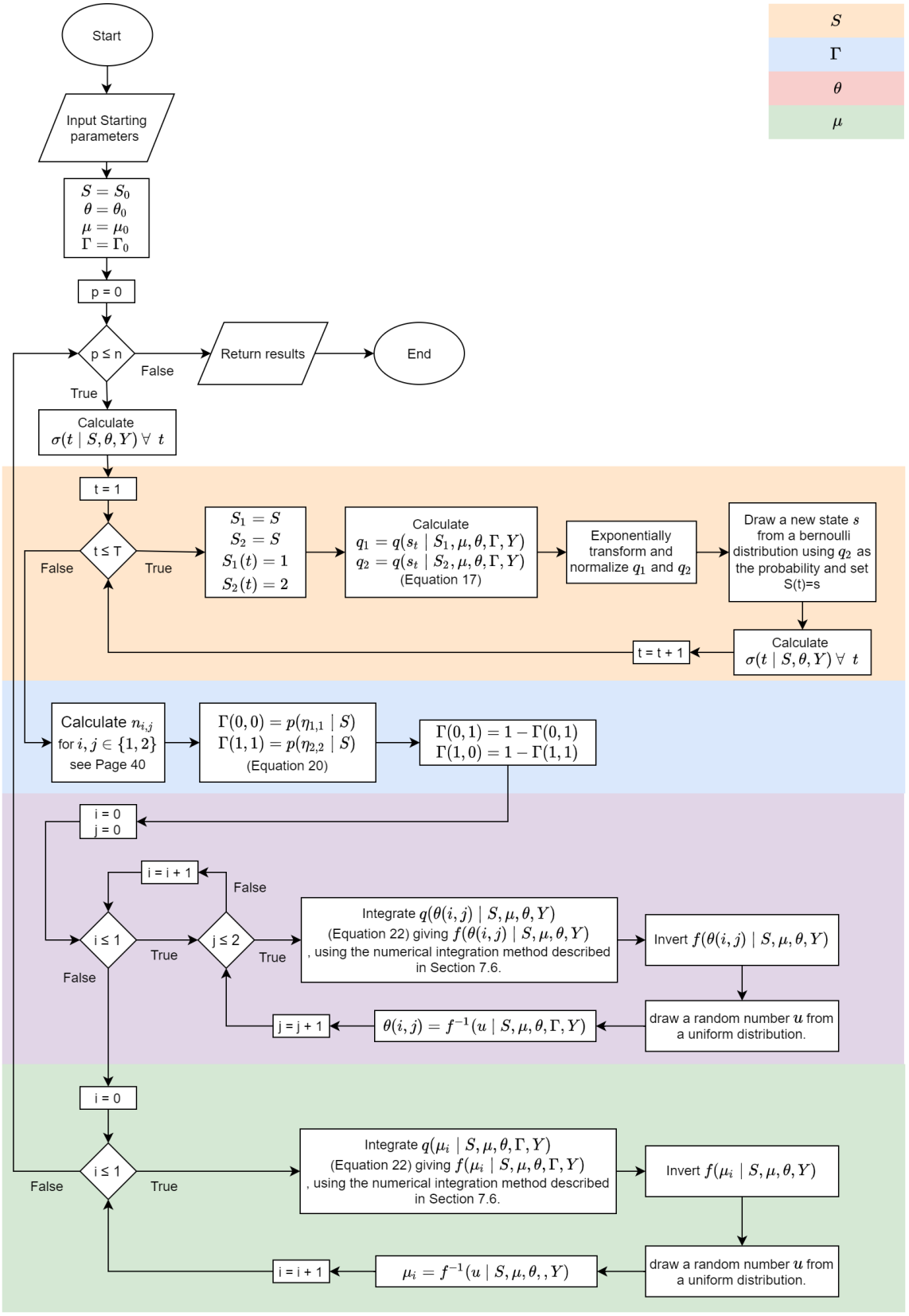


Figure 10: A flowchart showing the estimation procedure of the MS-GARCH model. A high quality version can be seen in the appendix.

7.6 Numerical integration algorithm

As explained previously the prior distributions are chosen such that they encapsulate the sample space of the posterior distributions as well as possible. However since it is unknown how the conditional posterior distributions look, and given the fact that the intervals of the probability mass of the conditional posterior density, used in the Gibbs-sampler, can change at each iteration, the numerical integration method has to be considered.
In Bauwens et al. (2010) it is suggested that M points should be used to estimate the integrated density. However since there is a chance that the probability mass lie on a very small area within the boundaries, chosen by the prior distribution, it is necessary that M is very large.
This is shown for a beta distribution in Figure 11 below, where a grid of 10 point seems to produce an acceptable approximation for a wide distribution such as the one in Figure 11 (c). However for a narrow distribution as the one in Figure 11 (a) it is notable that the same grid gives a poor approximation of the CDF. This is because only one of the points in the integration falls within the probability mass of the distribution.

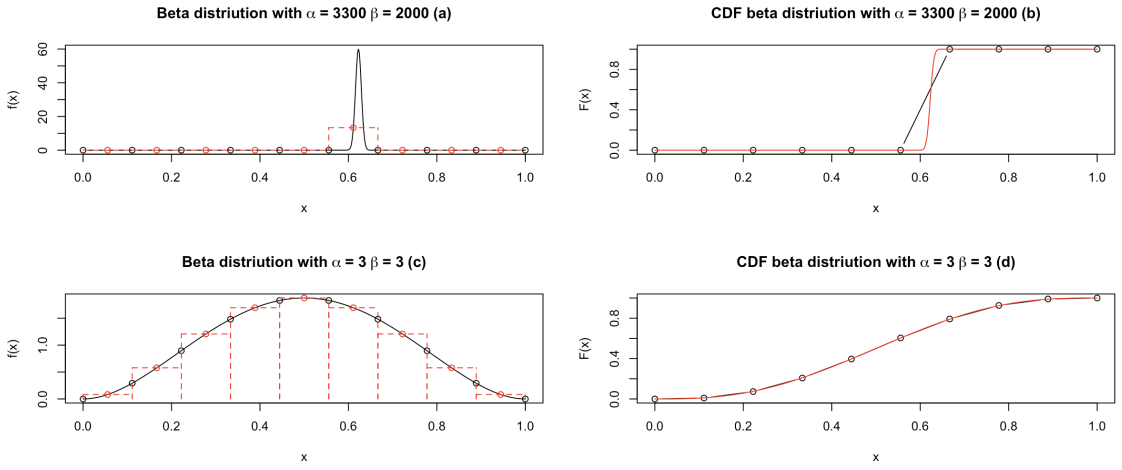


Figure 11

Under normal circumstances this is not a big issue, since the M can easily be increased. However the functions which are being integrated in the Gibbs sampler are computationally cumbersome, because the entire σ process has to be re-calculated with each calculation of the function, therefore it is not possible to increase M infinitely.

Instead of using a standard Riemann sum, we have developed an algorithm which determines where the probability mass of the density function lie, and uses that interval in the integration process. By doing so, we ensure that the probability mass is calculated accurately, no matter how slim the distribution is. Using this algorithm does require some

extra calculations. However, if a Riemann sum, with the same level of precision, was used, the computational load would be far greater.

A flowchart showing the steps of the integration algorithm can be seen on page 49. A more in-depth explanation will also be given in the following section.

The algorithm is split into three sections.

1. The mode parameter, or maximum, of the posterior density is found.
2. The boundaries of where the probability mass of the posterior density lie are found.
3. A Riemann sum is calculated between the boundaries found in step (2) and the remaining area between the boundaries of probability mass, and the boundaries of the parameter space are approximated using triangles.

The calculation of the mode parameter serves multiple purposes. Since the posterior distributions in the Gibbs sampler have been log transformed, they have to be transformed back to true probabilities again. As explained in section 7.2, this can be done by parallel shifting of the log transformed function before the exponential transformations, and by doing so the function will retain the same proportionality.
Using the mode parameter for the parallel shift of the posterior density, ensures that the function value in the mode will be 1 after the transformation. This way of parallel shifting the posterior distributions will also ensure that the scaling of the posterior distributions remain the same throughout the Gibbs sampling algorithm.
The mode parameter will also be used in the calculation of the boundaries of probability mass. How this is done will be explained further later in this section.

Estimation of the mode parameter

The mode parameter is found using a golden search approach, in which the algorithm starts at point x_{mid}, in the center of the parameter space. In x_{mid} the differential coefficient is then calculated (using the finite differences method). Given that the posterior density is unimodal we know that a positive differential coefficient ensures that the mode is on the right of x_{mid}, and the mode is on the left of x_{mid} if the differential coefficient is negative. Using this knowledge a process can be implemented to search for the mode, following this procedure

1. Define the bounds of the parameter space $x_1 = x_{min}$ and $x_2 = x_{max}$.
2. Define x_{mid} as $\frac{x_1+x_2}{2}$.

3. Calculate the differential coefficient in x_{mid}, using the finite differences method.

4. If the differential coefficient is negative set $x_1 = x_{mid}$ else set $x_2 = x_{mid}$.

5. Repeat step 2 - 4 until the mode has been found.

We define the number of times step 2-4 are repeated as τ, and the value of τ is discussed further in section 7.6.1.
When the procedure has been run τ times the mode parameter is found as x_{mid}, and the function value in this point, which is used in the parallel shift in the exponential transformations, will be defined as γ_{mid}.
Since the golden search approach continuously halves the maximum distance to the desired point, only a very small number of calculations are needed in order to determine a very accurate estimate of the mode. Instead of using the golden search approach for locating the mode, it could be argued that the Newton-Raphson method should be used, as this method is more widely used. However, since the interval in which the mode is found is limited, the Newton-Raphson method will not work.

Effective bounds of the probability mass

The next step of the algorithm seeks to determine the boundaries of where the probability mass of the posterior distribution lie, i.e. where it would be expected to draw samples from the distribution. This is done since only a small part of the sample space of a narrow posterior distribution, has a probability of sampling which is significantly higher than zero. An example can be seen in Figure 12, where a density is dawn, according to which the combined probability of drawing an observation between $[0; 0.4]$ or $[0.6; 1]$ is $1.684 \cdot 10^{-10}$. In this distribution it would therefore be a waste to accurately calculate the integral in these intervals, since the probability of drawing a sample in the intervals is approximately zero.

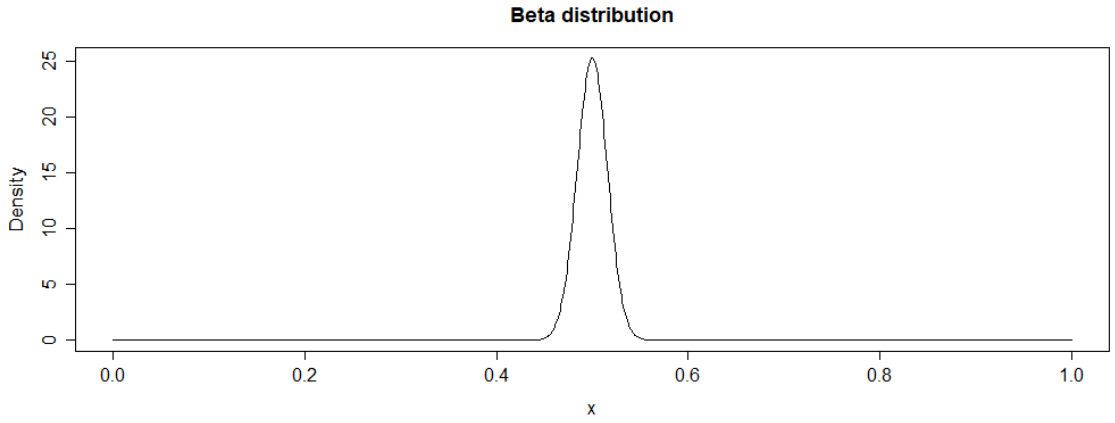


Figure 12: An example of a distribution with a slim probability mass. The distribution is a beta distribution with parameters $shape_1 = shape_2 = 500$

We propose a limit ϵ which indicates the maximum difference between the function value in the mode parameter, γ_{mid}, and the log transformed posterior density function, where it still remains relevant to calculate the integral correctly. This is equivalent to limiting the integral to only integrate, where the function value of the posterior density is above $e^{-\epsilon}$. In order to determine the boundaries, another version of a golden search has been implemented, since the sign of the differential coefficients are known on either side of the mode parameter.
Instead of the differential coefficient, the function value is used directly. The boundaries are moved using the difference between the tested point and γ_{mid} relative to the ϵ limit. Using this difference it is possible to determine if the boundary is to the left or the right of the tested point. This in turn leads to the following approach for finding the bound to the left of the mode

1. Define the bounderies using the prior distribution $x_1 = x_{min}$ and $x_2 = x_{mid}$.
2. Define x_{low} as $\frac{x_1+x_2}{2}$.
3. Define γ_{low} as the function value of x_{low}.
4. If $\gamma_{low} - \gamma_{mid} > \epsilon$ set $x_1 = x_{low}$ else set $x_2 = x_{low}$.
5. Repeat step 2 - 4 until the limit has been found.

As with the procedure to find the mode of the conditional density function, step 2-4 will be repeated τ times. The same number of repetitions is used here, as the in the previous procedure, since we expect a similar number of repetitions are needed in order accurately determine the two values, given the similarity of the two procedures.

After x_{low} has been found, the process is then performed similarly in order to calculate the limit on the right hand side of the mode. This leaves two limits for the integral x_{low} and x_{high}.

Integration method

The last step of the algorithm calculates the approximate integral using x_{low}, x_{high} and γ_{mid} found previously.

The area between x_{low} and x_{high} found in the previous step is calculated using a modified Riemann sum. It turned out that it was computationally easier to implement a Riemann sum, where the top of each column is a triangle, instead of as flat column. Since the posterior distributions are smooth this type of integral should yield the same result as a regular Riemann sum when enough grid points are used for the integration. How the integral is calculated using the alternative Riemann sum is illustrated in Figure 13.

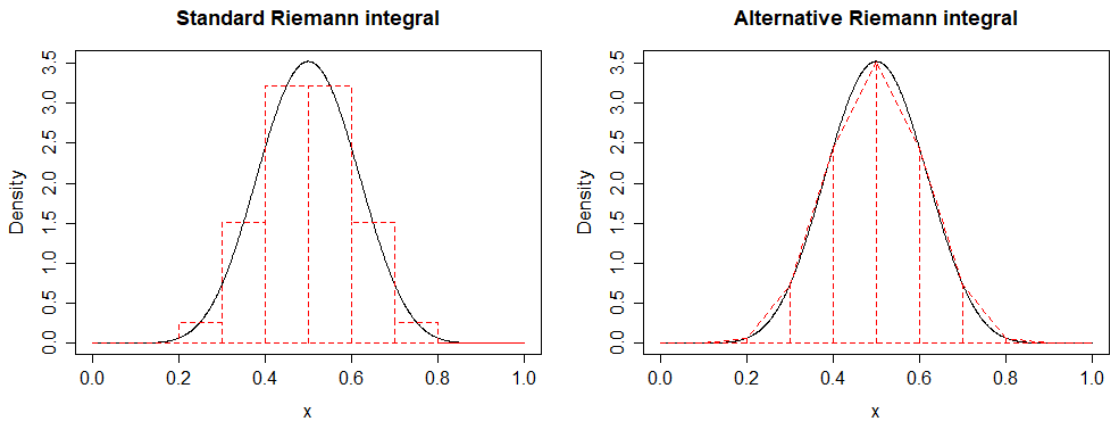


Figure 13: Illustaion of an alternative way of calculating the column area of a Riemann sum. Instead of using columns with flat tops, the columns will have a triangular shape, defined by the function values in the limits of the columns.

The areas between the boundaries of the probability mass and the boundaries of the parameter space are approximated using triangles. This will result in an overestimation of the tail probabilities, however if ϵ is large enough, the probability of drawing an observation in the tails is approximately zero. How large ϵ has to be is discussed further in section 7.6.1.

An alternative solution would be to set the probability of ending between these points to zero. However this could possibly prevent the posterior distributions from reaching certain points of the parameter space, which would break the ergodicity requirement of the Gibbs sampler. Because of this the tails will be estimated using triangles.

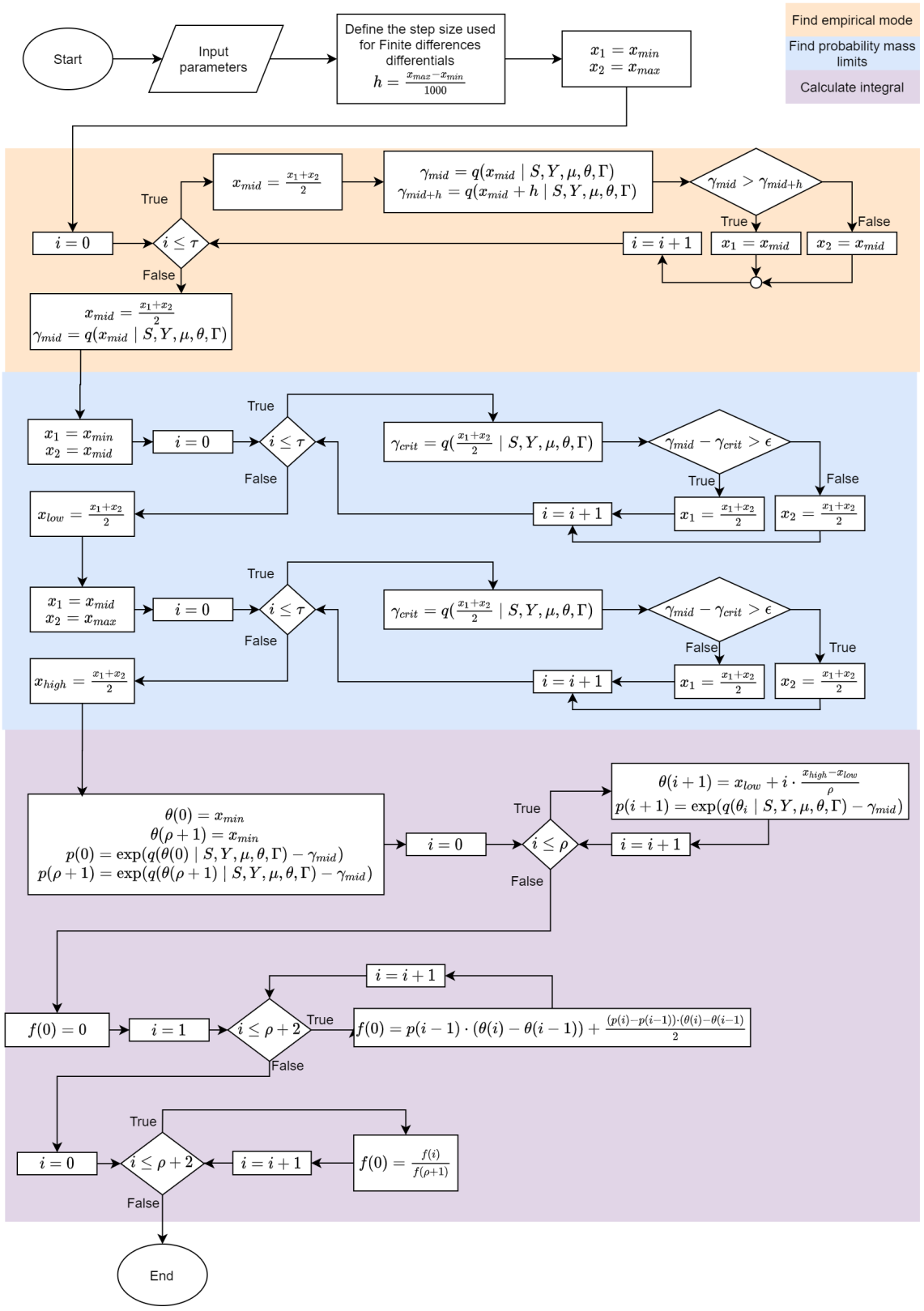


Figure 14: A flowchart showing the numerical integration procedure used for the estimation of θ and μ in the estimation procedure. A high quality version can be seen in the appendix.

7.6.1 Input parameters for the numerical integration

Since this integration method relies on different parameters, this section will evaluate what the values of these parameters need to be, in order for the integration algorithm to function properly. The parameters which will be evaluated are

- τ, the number of times the golden search is run in order to find the mode, and the boundaries of the probability mass, x_{low} and x_{high}.
- ϵ, the critical value of the boundaries.
- ρ, the number of grid points used for the integration withing the boundaries x_{low} and x_{high}.

The parameters will be choosen based on the performace of the algorithm when applied to a number of beta distributions with different parameters. This will test how the algorithm perform in different scenarios.

Choosing the number of iterations for the golden search τ
In order to keep the parameter set of the numerical integration algorithm as simple as possible, the same τ will be used for determining both the mode and the boundaries of the probability mass.

Figure 15 shows the convergence of the mode parameter for different beta distributions. It is notable that the convergence seems more or less unaffected by shape and placement of the distribution. This is good when having to pick a τ, since the chosen τ should work equally well for most distributions, according to Figure 15.
Investigating Figure 15 further, $\tau = 10$ seems suitable since it looks like it has converged for all distributions at this level.

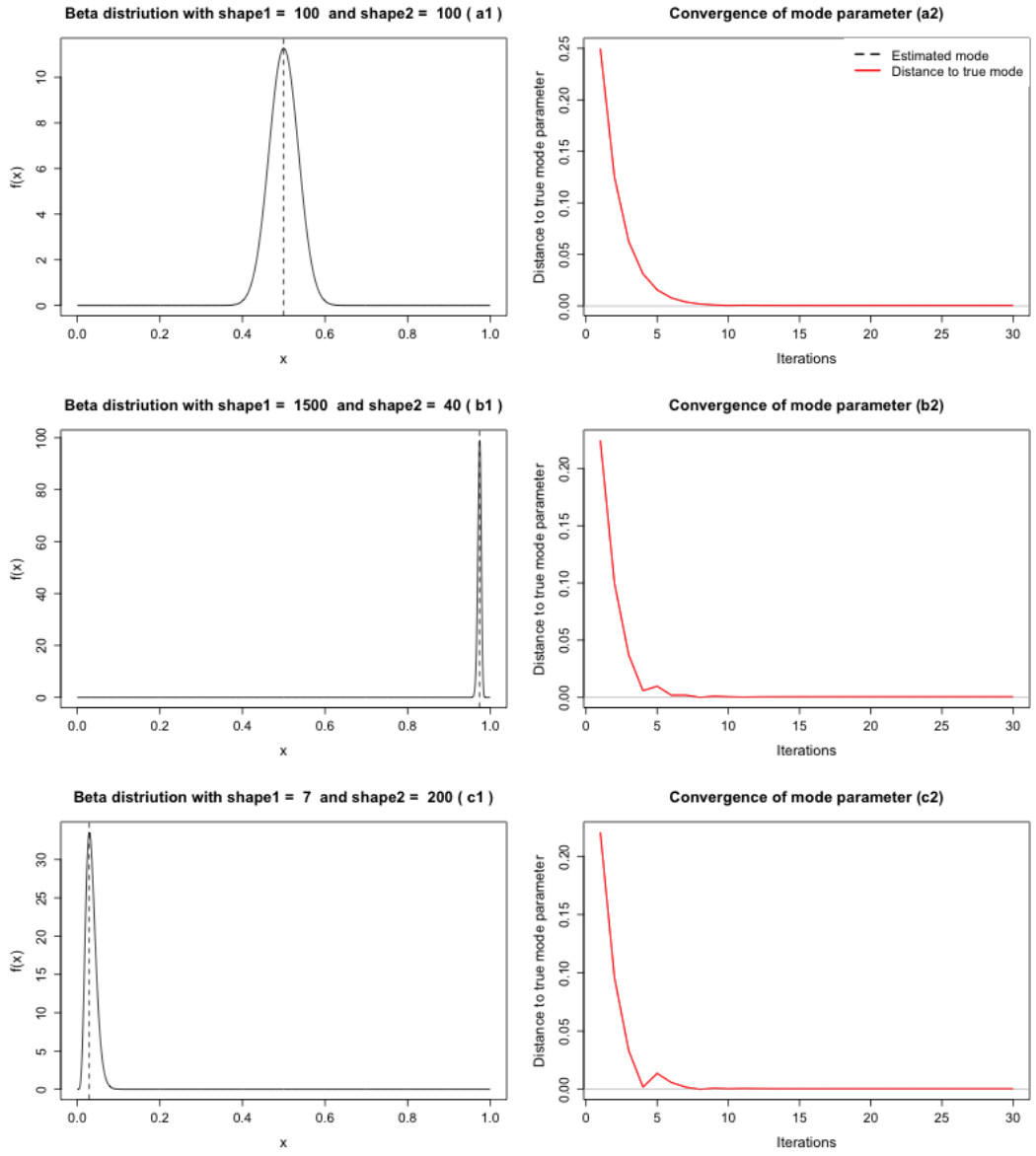


Figure 15: Convergence of the distance between the estimated mode parameter, and the true mode of the distributions using the golden search method.

Next, the limits, which capture the probability mass, will be determined. An equivalent figure to Figure 15, where limits are evaluated instead of the mode, can be seen in Figure 16.

The limits found in Figure 16 are determined using the true mode, instead of the mode found using an equivalent τ. This is done since its more simple, and it encapsulates the effect of the golden search method of the limits.

in Figure 16 we see that the algorithm also seems to have converged for $\tau = 10$.

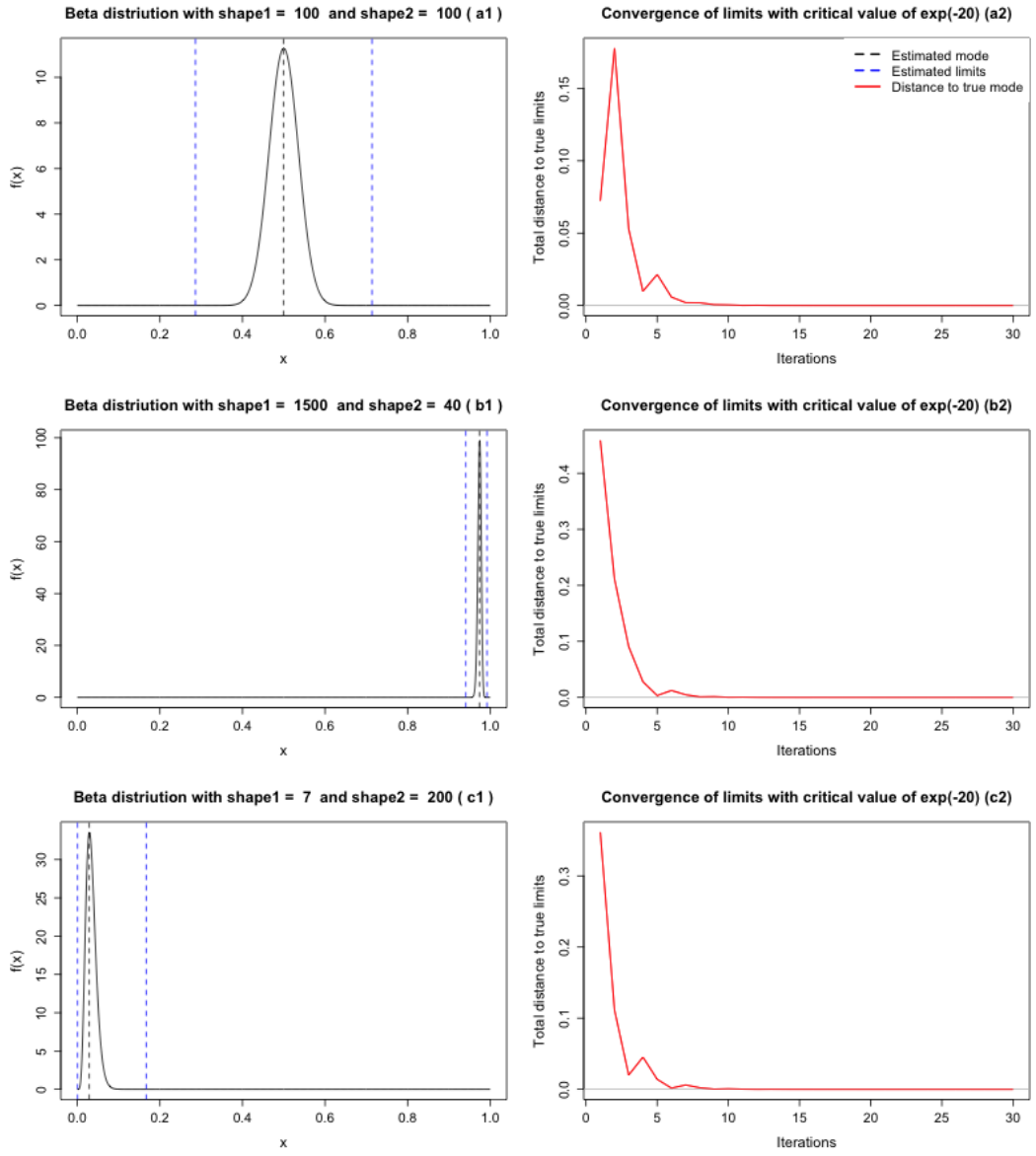


Figure 16: Convergence of the distance between the estimated bound, and the true bound, given by ϵ, of the distributions using the golden search method.

In Table 2 the absolute distances between the estimated and real mode, as well as the estimated and true boundaries are shown for a subset of τ values. This table also shows that $\tau = 10$ seems reasonable, as the precision is quite high.

A point worth noting when looking at the table is that the precision of the mode parameter does not increase significantly after 15 iterations. The reason for this is that the finite differences method is used for the differentiation process, in the golden search method for the mode parameter. Since the step size in this finite differences method has been chosen as $\frac{1}{1000}$ of the area of the parameter space, the precision of the golden search method will at some point surpass the precision of the finite difference method. This is one of the main

problems with numerical analytics. However since we has chosen $\tau = 10$, this should not affect our results much.

	Number of iteration				
	5	10	15	20	30
Figure a1					
Mode	1.56e-02	4.88e-04	4.58e-05	5.01e-05	5e-05
Limits	2.14e-02	9.19e-04	1.05e-05	4e-08	4.66e-10
Figure b1					
Mode	9.73e-03	5.21e-04	4.83e-05	4.97e-05	5e-05
Limits	3.86e-03	5.91e-04	1.40e-05	4.22e-07	4.66e-10
Figure c1					
Mode	1.36e-02	4.60e-04	4.77e-05	5e-05	5e-05
Limits	1.444e-02	3.35e-04	1.50e-05	7.59e-07	4.65e-10

Table 2: Absolute error of estimated limits

Choosing the critical value ϵ

Next we will consider the sufficient level for ϵ, such that the probability for being in interval from the boundaries of the parameter space to the chosen limits is acceptably small. If ϵ becomes too large the approximation of the tail probabilities will be very bad, due the triangular density estimation used in the tails. The probability of being on the edge of the parameter space will simply be too high. On the other hand, if ϵ is too small the area, which is integrated will become too large, which negates the purpose of the integration algorithm.

In order to determine a suitable ϵ, the tail probabilities for different values of ϵ have been estimated. To get an indication of what might be a sufficiently low probability outside the limits, this is compared with the true probability for the beta distribution.
In Figure 17 the aforementioned comparison of the tail probabilities, from the boundaries of the parameter space to the limits, are shown. These are shown for the same beta distribution, as used in Figure 16 (b1).

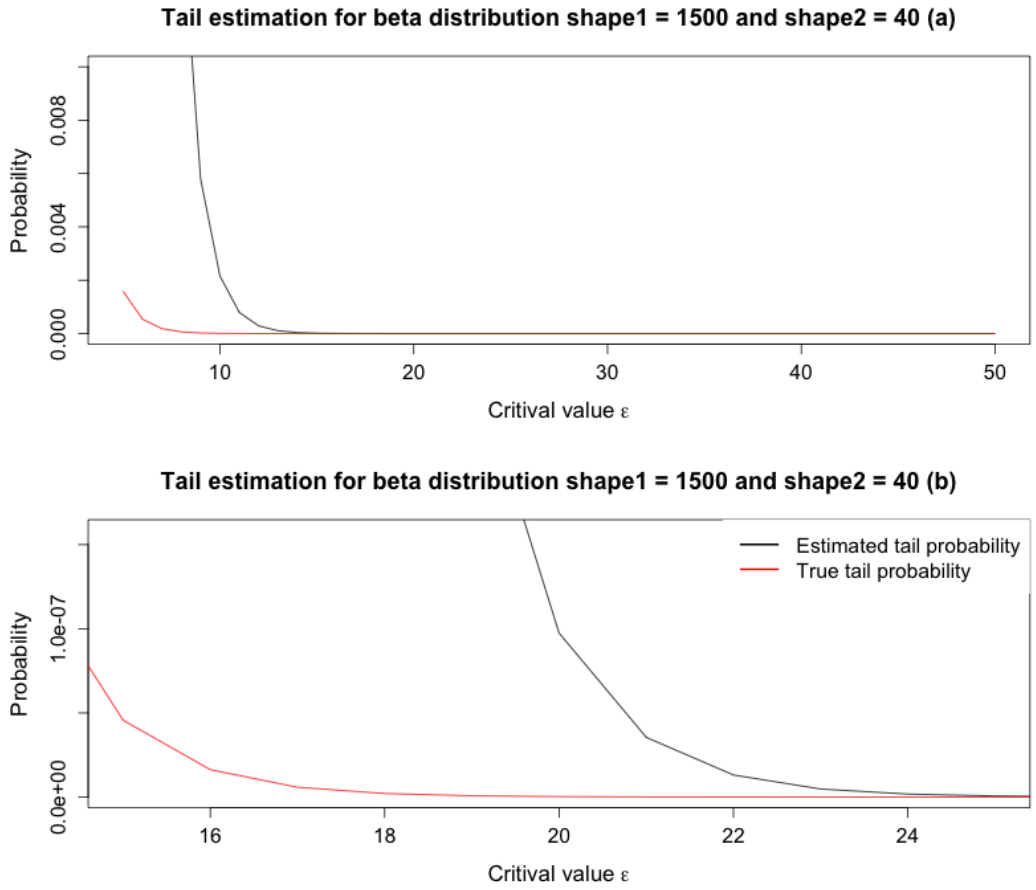


Figure 17: Tail probability for different value of ϵ. Figure (b) is a close up of figure (a).

It can be seen in Figure 17 (a) that the estimated tail probability quickly converges to the true value, when ϵ increases. However, it is difficult to see the difference between true and estimated probability as they both goes to 0.
Figure 17 (b) shows a closeup of Figure 17 (a) for $15 \leq \epsilon \leq 25$. Here it is clear, that for a sufficiently large ϵ, the estimated probability in the tails is equivalently small such that it will not affect the approximated CDF, in a significant manner.

	Critical value ϵ				
	5	20	30	40	50
Excess probabilities	0.2444	1.31e-08	7.71e-14	3.36e-18	8.52e-21

Table 3: Excess tail probabilities outside of boundaries for different values of ϵ.

In Table 3 the excess tail probabilities can be seen for selected values of ϵ. Here $\epsilon = 20$ will be small enough for it to have a very little effect on the integration. Using $\epsilon = 20$ will

give a one out of 100 million chance that an observation will end up in the tails, when it wouldn't have if the integration was conducted properly. Since we are not drawing close to 100 million samples, this estimation seems reasonable.

Finding a sufficient number of grid points ρ

The last thing to consider is the amount of grid points used for the integration, in order to have an accurate approximation of the CDF. Since the numerical integration algorithm, explained in Figure 14, ensures that we have determined where the probability mass is located, a sufficient number of grid points is still needed, within these boundaries.
In order to determine this we use the Kolmogorov–Smirnov statistic, $\mathcal{D}$, to compare the estimated CDF with the true CDF for a given distribution. The Kolmogorov–Smirnov statistic is the supremum of the set of distances between the empirical and theoretical CDF. This statistic is used as an indicator of how much the approximated CDF deviates from the true CDF. This information is sufficient since $\lim_{\rho \to \infty} \mathcal{D} = 0$. Therefore a sufficiently small $\mathcal{D}$ has to be chosen such that the approximation is accurate.

Figure 18 shows the CDF approximation for the beta distribution in Figure 15 (b1), using the alternative integration method with 10, 100 and 400 grid points.

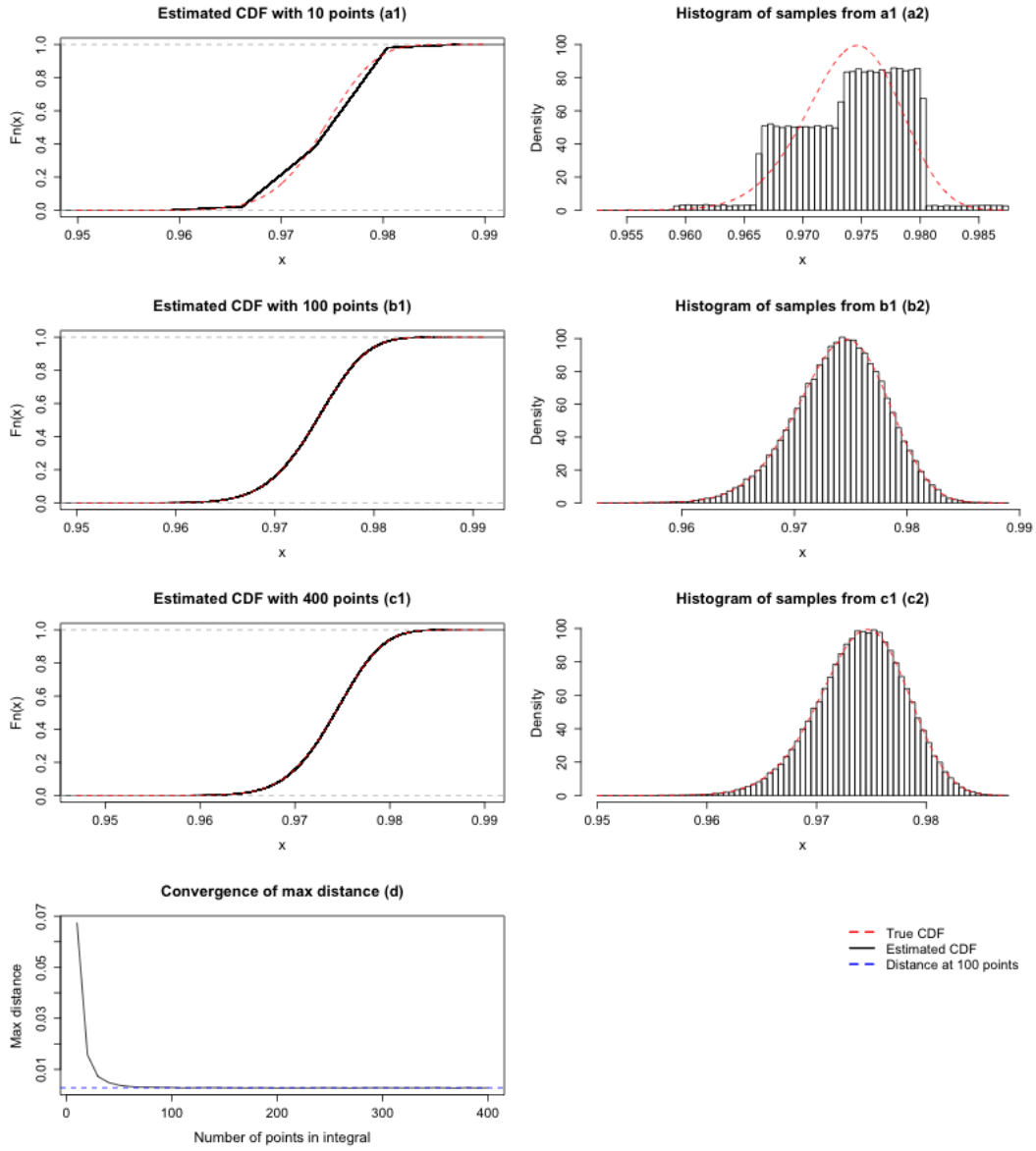


Figure 18: precision of the estimated integral given the number of points, ρ, used for the Riemann sum.

Figure 18 gives an indicator, of which ρ yields an acceptable approximation. With the chosen limits for the probability using 10 grid points is far from enough, as shown in Figure 18 (a1) and (a2). The histogram in (a1) accentuates the issue with not using enough grid points, since the estimated density will not be a smooth curve function, but instead falls in a set of uniformly distributes areas.

With 100 and 400 grid points the approximation of both the CDF and density are almost identical to the true beta distribution, shown in Figure 18 (b1)-(c2).

Using more than 100 grid points, does not seem necessary as the approximation does not seem to become more accurate by doing so. This is also supported by Figure 18 (d) where

the Kolmogorov–Smirnov statistic seems to have converged after using 100 grid points.

In order to evaluate the Kolmogorov–Smirnov statistic generated by the approximated CDF's, they will be compared to the expected maximal distance of an empirical distribution function, generated by drawing n observations from the true distribution. Using this measure it is then possible to evaluate whether the approximated CDF is sufficiently precise.
When performing this comparison we will use $n = 100.000$ draws, since we will not expect to draw more than 100.000 samples from any conditional posterior distribution when using the Gibbs-sampling algorithm. Furthermore we will perform the test 100.000 times, in order to get the expected Kolmogorov–Smirnov statistic for 100.000 draws using the different ρ values.
The result of this experiment is shown in Table 4 where the Kolmogorov–Smirnov statistic almost stays constant after $\rho = 100$. Therefore this seems to be an acceptable amount to use for the numerical ingratiation.

	Expected distance	ρ				
		10	50	100	200	400
Max distance	0.0027	0.067	0.0038	0.00278	0.00266	0.00278

Table 4: the average Kolmogorov-Smirnov test statistic, and p-value using the estimated CDF and the true CDF and a sample size of $n = 100.000$

7.6.2 Computational advantage

As described earlier, the general idea of encapsulating the probability mass is for computational advantages when integrating the posterior distribution in order to approximate the CDF. This is done instead of integrating over the entire parameter space.
We have showed that 100 grid points are sufficient for the numerical integral in order to ensure accuracy of the approximation. Since the area which encapsulates the probability mass can vary in size, this will affect the number of points that must be used in a standard Riemann sum.

In order to compare the standard Riemann sum with the integration method used in this thesis, the number of calculations needed for a given precision will be compared. Figure 19 shows the number of calculation needed when using the standard Riemann sum,

given that 100 grid points has to be included in a percentage of the integral area. Here it is also when where the standard Riemann sum will be disadvantageous and where it will be advantageous.

For computing the number of calculation points in Figure 19, the information provided from the earlier sections is used for the limited Riemann sum. The number of calculation points for an equivalent standard Riemann sum has been calculated as

$$\frac{\rho \cdot area\ of\ parameter\ space}{area\ of\ encapsulated\ probability\ mass},$$

where ρ is the number of grid points chosen for the integral.

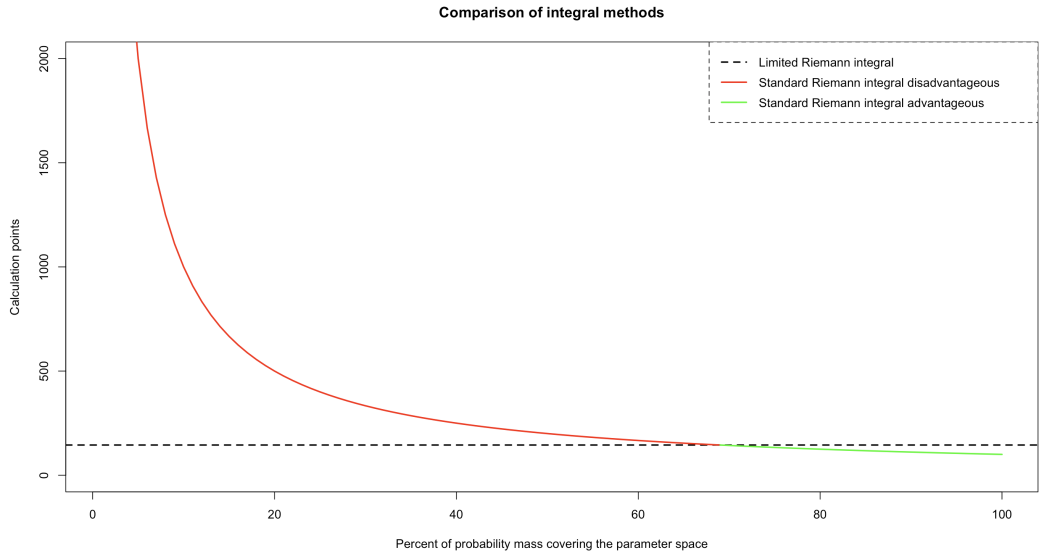


Figure 19: Number of calculation points needed when using the standard Riemann sum, given that 100 grid points has to be included in a percentage of the integral area.

Because the integration process used in this thesis requires a fixed number of calculations in order to determine the boundaries, it will be disadvantageous if the probability mass can be encapsulated on an area less than 70% of the parameter space. However we can not be sure that the encapsulated probability mass will only cover 70% of the parameter space, for each iteration of the Gibbs-sampler. Therefore it is necessary that number of points for the standard Riemann sum correspond to what might be expected.

8 Estimating simulated MS-GARCH

The following section will examine how the estimation algorithm, described in Section 7, performs when applied to a simulated MS-GARCH data set. By doing so, we will be able to test how well the estimation procedure, described in Section 7, performs.
The simulated MS-GARCH process used as basis for the estimation algorithm has been generated using the following parameter set, and the process can be seen in Figure 20.

$$\theta_0 = \begin{bmatrix} 0.5 & 0.35 & 0.15 \\ 1.5 & 0.15 & 0.65 \end{bmatrix} \quad \mu_0^T = \begin{bmatrix} 0.05 & -0.05 \end{bmatrix} \quad \Gamma_0 = \begin{bmatrix} 0.99 & 0.01 \\ 0.03 & 0.97 \end{bmatrix} \tag{23}$$

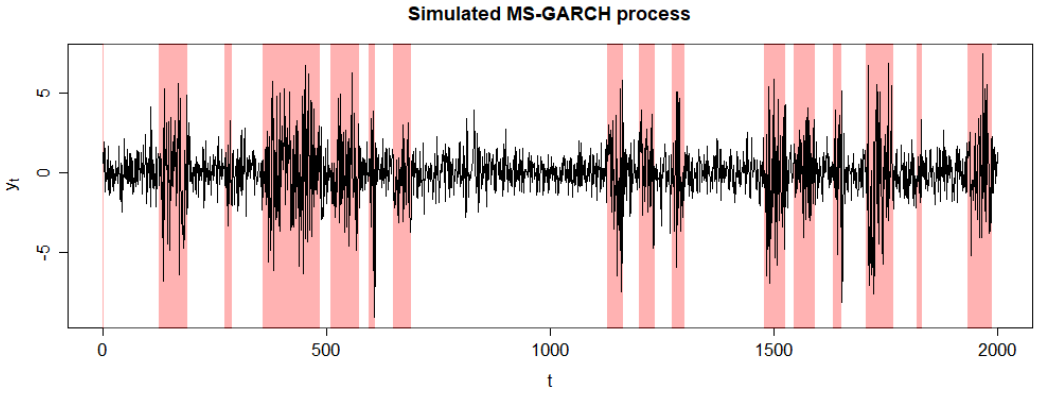


Figure 20: Simulated process of 2000 observations constructed using the MS-GARCH model described in Section 5, and the parameters shown in Equation 23. The White background indicated regime one, and the red background indicates regime two.

Before the estimation algorithm can be run, a set of inputs have to be defined. These inputs include the starting point of the regime states, the prior distributions for the parameters, as well as a starting value for the parameters.

8.1 Initial values of the estimation

The starting points of all the regime states have all been set to regime one, i.e. $s_t = 1 \ \forall \ t$. This choice of starting point should challenge the algorithm since it presents no prior knowledge of the regime states.

The starting points for the parameter set also have to be selected. The ones used for this estimation can be seen in Table 5. These starting points have been chosen such that

regime 1 is less explosive, and less persistent than regime 2.

Lastly the prior distributions has to be considered. The prior distribution for each parameters will be chosen to be uniform, as this will allow the model to be optimized solely on the information of the data.
In order to separate the regimes, such that the model does not switch the labels of the two regimes throughout the optimization process, the prior distributions for the parameters will be limited to certain boundaries, as explained in section 7.1.
Another thing that has to be considered when choosing the boundaries, is the stationarity of the model. This has to be considered, as we wish the final model to be stationary, and since the MS-GARCH model does not follow the same stationarity restrictions as a standard GARCH model, it is possible to set the boundaries such that the the sum of α and β exceeds one for certain regimes.
The variance stationarity requirement that exists in the MS-GARCH model is $\sum_{i=1}^{k} \pi_i(\alpha_i + \beta_i) < 1$, as explained in section 5.1. Since the transition matrix changes throughout the estimation process, the invariant probabilities π_i are also changing. Because of this it is only possible to be completely sure that every process is stationary if $\alpha_i + \beta_i < 1 \ \forall \ i \in \{1, ..., k\}$, however it can be useful to consider states which allows for explosive regimes.
In Table 5 the chosen intervals for the prior distributions can be seen.

	Parameters		Limits	
	Real	Starting	Min	Max
ω_1	0.50	0.10	0.01	1.00
α_1	0.35	0.10	0.10	0.50
β_1	0.15	0.10	0.01	0.25
μ_1	0.05	0.00	0.10	0.10
ω_2	1.50	0.50	0.50	3.00
α_2	0.15	0.20	0.01	0.25
β_2	0.65	0.70	0.30	0.90
μ_2	-0.05	0.00	0.20	0.10
$\eta_{1,1}$	0.99	0.90	0.00	1.00
$\eta_{2,2}$	0.97	0.90	0.00	1.00

Table 5: The parameters used in for simulating the MS-GARCH process used in the estimation, as well as the model parameters stating points and boundaries.

8.2 Parameter estimation

When determining the posterior distributions from the estimation algorithm, it is necessary to have a burn-in period, as described in section 6.2. Since it is not known how long the algorithm has to run before the posterior distributions have converged, the algorithm will be run 50.000 times, after which, it will be evaluated whether the process has converged. The way in which the convergence will be determined is by evaluating a trace-plot.

In Figure 21, the trace plot of the simulations can be seen, here we see that all the parameters converge quite quickly, and the burn-in period could therefore be equally small. To be certain that the process has converged, and since there are a lot of observations, the burn-in period will be set to 10.000.

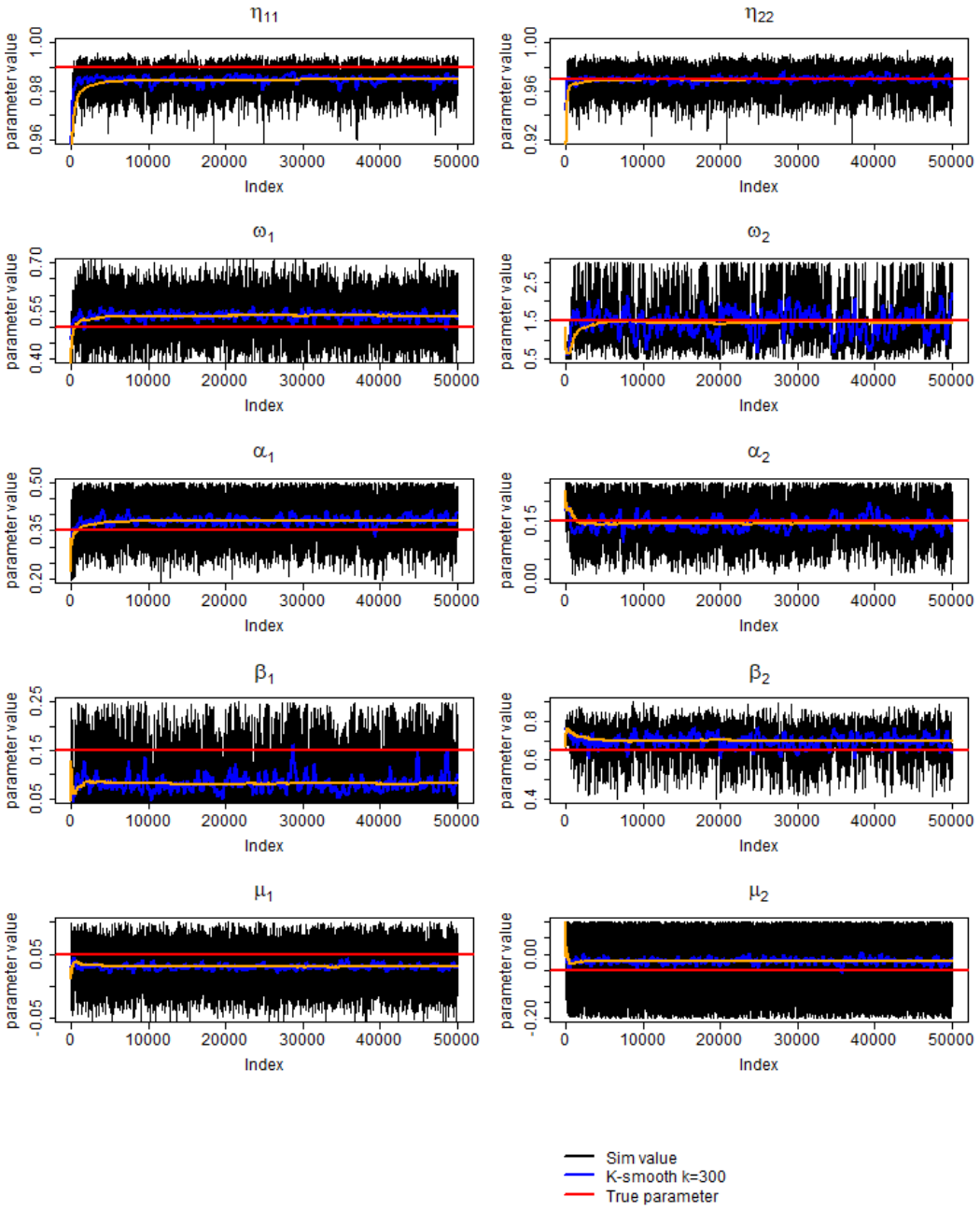


Figure 21: A trace plot showing the parameter estimates for each iteration of the Gibbs-sampler.

In the trace-plot it seems like some of the parameters, such as β_1 and μ_2 may have too tight boundaries, as they hit the boundaries quite frequently. In order to evaluate whether the boundaries are too tight, as well as investigating the posterior distributions, the histograms of the posterior distributions are shown in Figure 22.

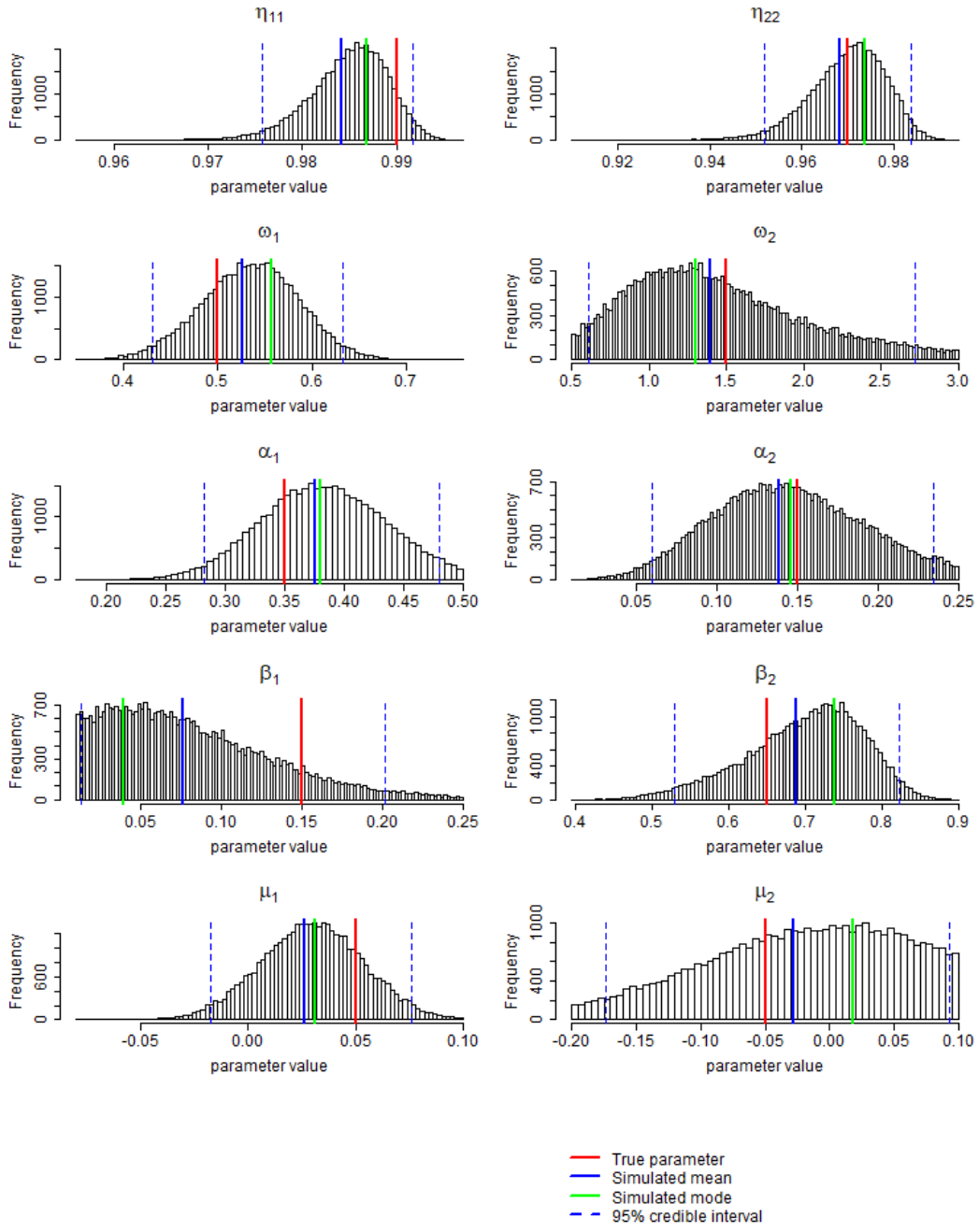


Figure 22: Histograms showing the posterior distributions, when a burn-in period of 10.000 observations are removed. The mode parameters have been calculated using a Gaussian kernel density estimate

In the histograms the same issues can be seen, as the probability mass of the β_1 and μ_2 parameters have been cut off. However for both parameters it seems that mode parameter is within the boundaries, and the estimate should therefore still be quite accurate. Another note to this point is that the boundary for the β_1 parameter can not be negative.

Using the posterior distributions the model parameters have to be chosen. There are two main options when choosing the model parameter. Either the mean, or the mode parameter of the posterior distribution can be chosen. The mean is the easiest choice, as it can easily be calculated, and it should also be a good estimate.
The issue with using mean is that it is not possible to take the average of the entire posterior distribution, because of the boundaries set by the prior distributions. This is not an issue for every parameter, as some of them stay well within their boundaries. It does however become an issue for parameters which are limited by the boundaries. This is especially prevalent for the μ_2 parameter, where the estimated mean will be lower than the actual mean, since there is accounted for more probability mass in the left hand side of the posterior distribution.
A way to resolve this issue is to use the mode parameter of the posterior distributions. Since the mode parameter does not depend on the entire distribution, but only on the most likely value, it is not affected by boundaries, as long at the mode of the posterior distribution lie within the boundary limits.
One issue with using the mode parameter, is that the posterior distributions are empirical with a continuous first-axis, and the mode parameter is therefore not possible to determine directly. There do exist several methods of determining mode parameters, however these methods will not necessarily produce the same results.
Because the mean is easier to calculate and easier to evaluate, this will be chosen as the parameter estimate. In Figure 22 it can also be seen that the two estimates are quite close in most cases, even for the μ_2 where the mean should produce the biggest issues.

The estimated model parameters found using the means of the posterior distributions can be seen in Table 6.

	Real parameter	Estimation	
		Mean	Std.
ω_1	0.50	0.5361	0.0513
α_1	0.35	0.3824	0.0510
β_1	0.15	0.0822	0.0503
μ_1	0.05	0.0286	0.0239
ω_2	1.50	1.4840	0.5526
α_2	0.15	0.1415	0.0453
β_2	0.65	0.6965	0.0758
μ_2	-0.05	-0.0188	0.0732
$\eta_{1,1}$	0.99	0.9858	0.0041
$\eta_{2,2}$	0.97	0.9700	0.0081

Table 6: Estimated parameters of the simulated MS-GARCH process in Figure 20, using the MS-GARCH model.

The estimated parameters of regime 2 are very close to the real parameters, where as the parameters of regime 1 are quite different. It mainly seems like the β_1 parameter is too high, and the ω_1 and α_1 parameters are too low. However if the unconditional variance and kurtosis of the regime 1 GARCH process are calculated using both the estimated parameters and the real parameter we get the following result.

	Estimated	Real
Unconditional variance	1.0016	1
Unconditional kurtosis	4.7858	4.4554

Table 7: Unconditional variance and kurtosis of the estimated regime 1 GARCH model parameters, and the real regime 1 GARCH model parameters.

This means that while there are some differences between the estimated and the real process, these are most likely minor, since the variance of regime 1 is very low. Furthermore the unconditional distribution of the two processes are almost identical. Since the two

processes are so similar it seems reasonable that the estimation algorithm converges to the wrong parameter estimates.

8.3 Estimation of the regime states

Another reason that could cause the estimation algorithm to converge to the wrong parameter estimates would be if the regime states are not correctly identified, and therefore this will also have to be investigated.
Optimally the regime states would be determined as the regime each observation is in the most, across the iterations of the gibbs sampler, similarly to how the rest of the model parameter have been chosen. However, there is one issue with this approach. Since it is not known how long the burn-in period will be before the estimation algorithm has run, the regime of every observation would have to be saved for each iteration of the estimation algorithm. This would make it possible to ignore the burn-in period, and use the remaining iterations to determine the estimated regime states. The issue with this approach is that having to save 2.000 observations for 50.000 iterations would require saving 100 million data-points each time the algorithm is run. This would significantly slow down the algorithm, and since no further analysis would be done using the additional information, this approach is quite excessive.
Alternatively the algorithm is run an additional 10.000 times after the initial 50.000 in order to have a 'clean' sample of the regime states.
This does mean that there is a difference between the underlying iterations for the parameter estimates, and the regime states. However the reason why the burn-in period is removed is exactly because the process is assumed to have converged after the burn-in period. Since the estimation algorithm has converged to the invariant distribution, the samples are drawn from the same posterior distribution. Given the ergodic theorem, see section 2 theorem 2.1, the end result should therefore be the same.

After the additional 10.000 iterations the probability of the regime states can be estimates, and the results of which can be seen in Figure 23. Here we see that the probabilities lie close to being in each regime in 50% of the iterations of the Gibbs sampler. Because of this, the estimated regimes of the observations will be defined as the most likely regime.

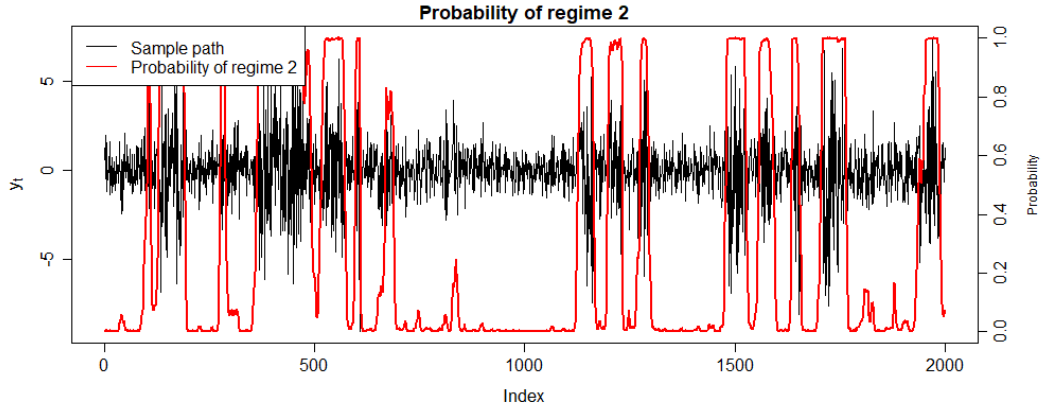


Figure 23: the probability of being in regime two for each observation in the sample-path.

In Figure 23 the real regimes of the model are compared to the ones estimated by the algorithm. Here it is clear that the estimated regimes are very similar to the real regimes.

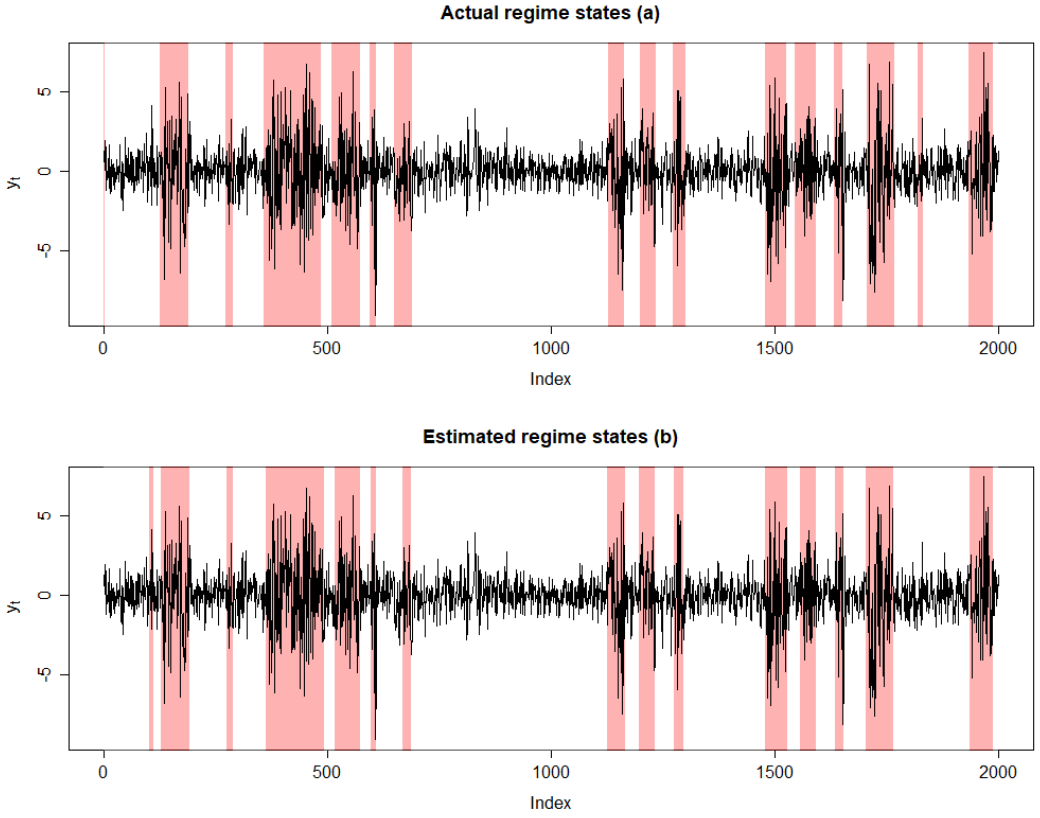


Figure 24: Figure (a) shows the real regimes used to generate the MS-GARCH process. Figure (b) shows the regimes estimated by the probabilities in Figure 23. The regimes are estimated by classifying regimes which have a probability higher than 50% of being in regime 2, as regime 2.

Having the model parameter estimates, as well as estimated regime states, it is possible to calculate the estimated conditional σ-process, which is done by using the formula in

Equation 10 on Page 24.

In Figure 25 the real conditional σ-process are compared with the ones calculated using the estimated parameters and regimes. In the comparison we see that σ is approximately the same for both models, except when the estimated and real regimes differ. Given this it is not expected that the difference in the parameters discussed previously will have much of an effect on the model, however the estimated regimes may.

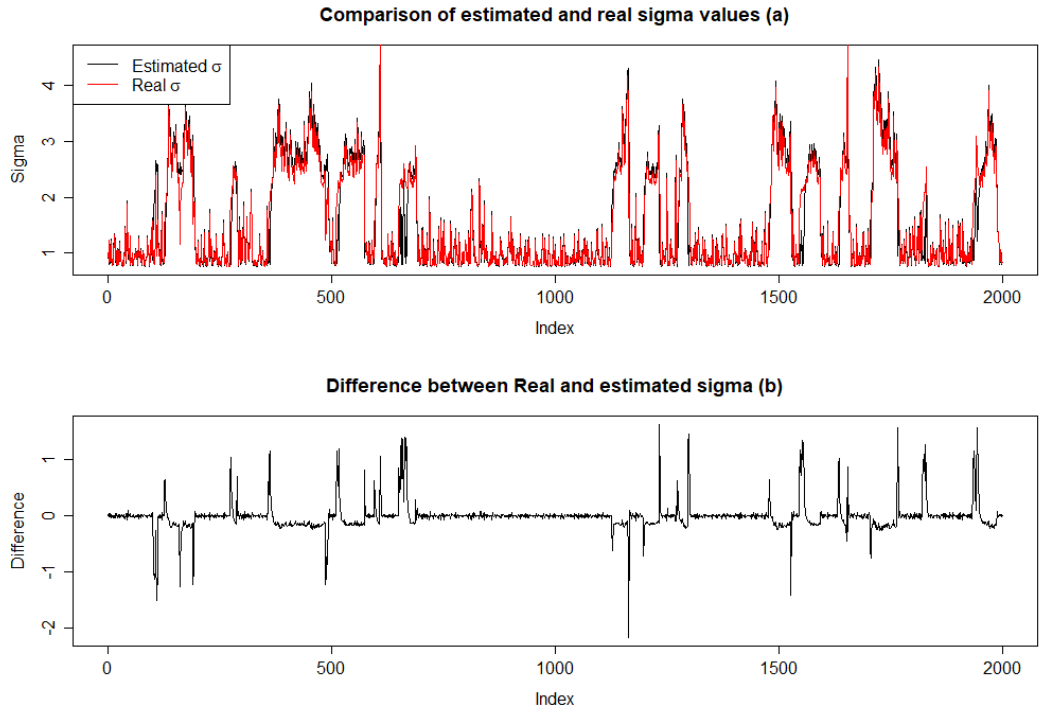


Figure 25: The σ-process generated using the estimated parameters and regime states, compared with the σ-process generated using the real parameters and regime states.

8.4 Estimated model evaluation

In order to evaluate the performance of the model further, the standardized residuals are calculated using the formula

$$\hat{r}_t = \frac{y_t - \hat{\mu}_{s_t}}{\hat{\sigma}_t}.$$

Both the standardized residuals as well as the squared standardized residuals, will be evaluated.

The standardized residuals should give insight into whether the mean of the white noise u_t, from Equation 10, is 0, and the squared standardized residuals tests if the variance of u_t is 1. These residual tests will therefore show if the process left after removing the MS-GARCH model is in fact white noise drawn from a standard normal distribution, as

defined by the model in Equation 10.

In Figure 26 the residuals, and the squared residuals are shown. In this figure the residuals lie around zero, which indicates that the mean of the residuals is zero. Furthermore we see that the squared residuals lie close to 1, this is also what is expected since a squared standard normal distribution is a χ^2-distribution with 1 degree of freedom, and the mean of such a distribution is 1. Since the mean of the squared residuals is approximately 1 and the mean of the residuals is 0, it indicates that the residuals are white noise, and that the model explains the variance of the process well.

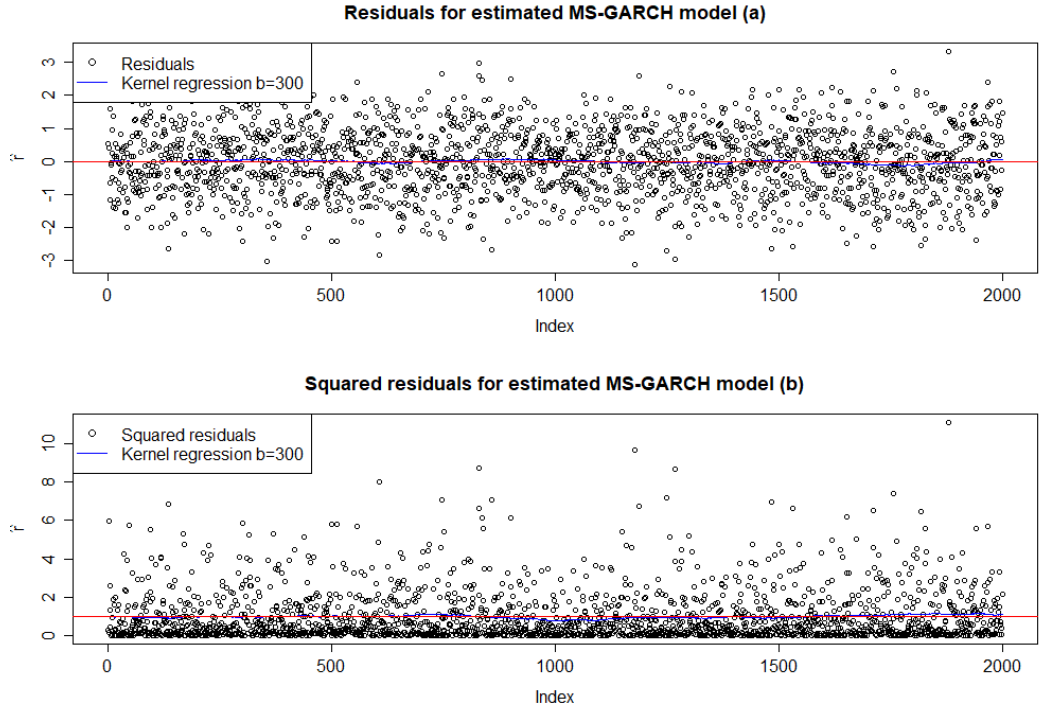


Figure 26: Residuals, and squared residuals, calculated using the estimated parameters and regime states.

Another way to evaluate the residuals is through a qq-plot, where the residuals are compared to a theoretical estimate of the residuals. The qq-plots of the residuals and the squared residuals are shown in Figure 27, here we see that the residuals are clearly normally distributed, however the squared residuals are not perfectly χ^2-distributed.

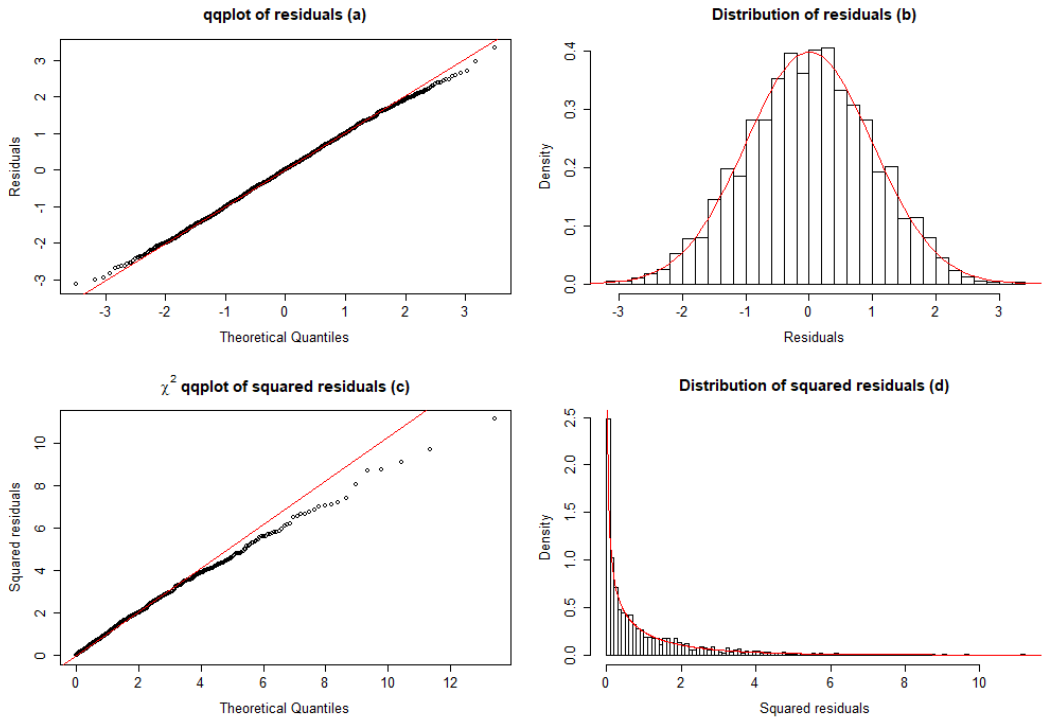


Figure 27: qq-plot of the residuals and squared residuals using the estimated parameters and regime states. The qq-plot in Figure (c) is generated using a χ^2 distribution.

As discussed previously the main difference between the real model and the estimated model, is the estimated regime states. In order to quantify how much of the inaccuracy of the residuals is caused by the wrongly estimated regime states, the qq-plots can also be seen in Figure 28 where the real regimes, but the estimated parameters are used. Here we see that the squared residuals become χ^2-distributed, which indicates that the biggest issue when estimating the model is estimating the regime states.

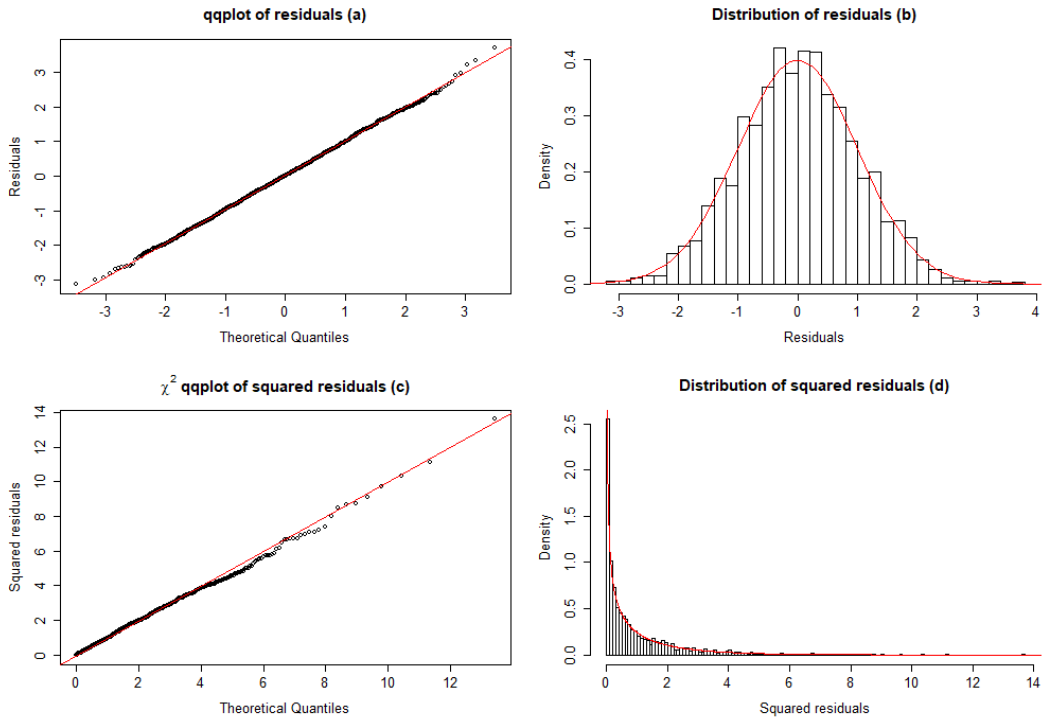


Figure 28: q-plot of the residuals and squared residuals using the estimated parameters and the real regime states used to generate the MS-GARCH sample-path. The q-plot in Figure (c) is generated using a χ^2 distribution.

Since the white noise in the MS-GARCH model is i.i.d. there can not be any auto-correlation in the residuals, since this would indicate that the noise terms are not independent of each other. In Figure 29 the auto-correlation of both the residuals and the squared residuals, using the estimated regime states and the estimated parameters, are shown. Here it is clear to see that the residuals exhibit no auto-correlation, which again indicates that the residuals are white noise.

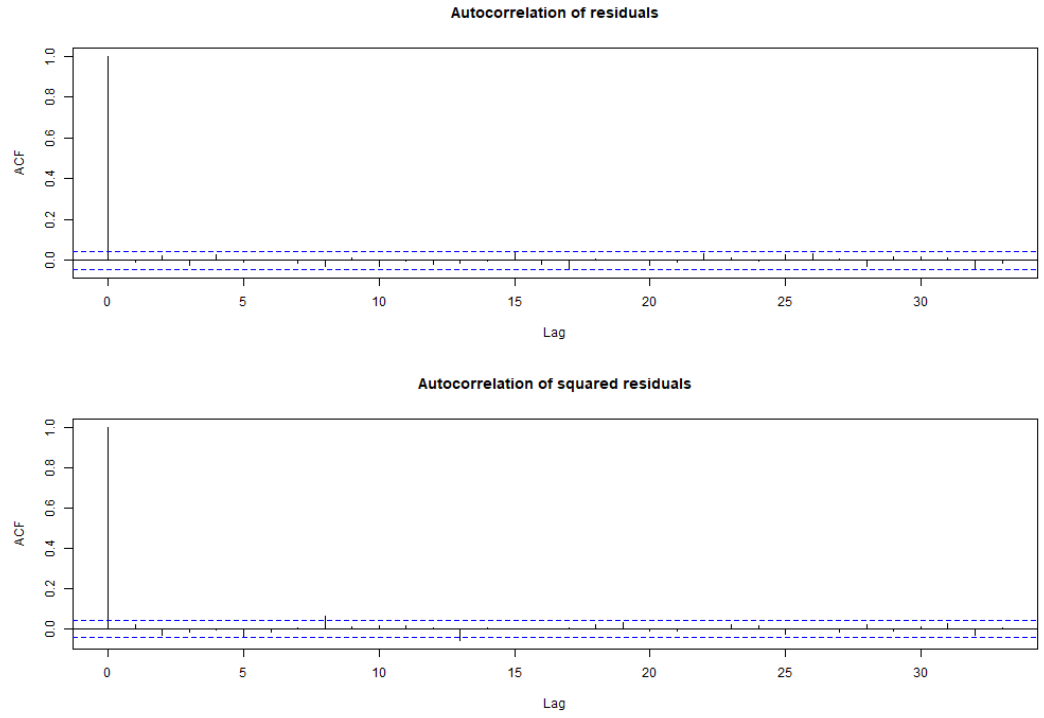


Figure 29: Auto correlation plots of the residuals and the squared residuals.

Lastly the likelihood estimates of the estimated, and the real model can be evaluated. This can be done using the AIC or BIC criterions, however since the number of observations, and number of parameters stay constant, these criterions will not provide more information than the likelihood alone. Therefore the likelihood will be evaluated by calculating the log likelihood. These log-likelihood estimated are calculated by taking the log of the expression in Equation 13 leading to

$$\begin{aligned}
\log(p(Y, S \mid \mu, \theta, \Gamma)) &= \log\left(\prod_{t=1}^{T} \frac{1}{\sqrt{2\pi}\sigma_{s_t}} \exp\left[-\frac{(y_t - \mu_{s_t})^2}{2\sigma_{s_t}^2}\right] \eta_{s_{t-1},s_t}\right) \\
&= \sum_{t=1}^{T}\left(\log\left(\frac{1}{\sqrt{2\pi}\sigma_{s_t}}\right) + \log\left(\exp\left[-\frac{(y_t - \mu_{s_t})^2}{2\sigma_{s_t}^2}\right]\right) + \log\left(\eta_{s_{t-1},s_t}\right)\right) \\
&= \sum_{t=1}^{T}\left(\log(\eta_{s_{t-1},s_t}) - \log(\sqrt{2\pi}\sigma_{s_t}) - \frac{(y_t - \mu_{s_t})^2}{2\sigma_{s_t}^2}\right).
\end{aligned} \tag{24}$$

Even though the Bayesian approach does not directly attempt to maximize the likelihood, it would still be expected that the estimated model provides a good log-likelihood estimate. The log-likelihood for the estimated and real models are therefore shown in Table 8. Here we see that the estimated model returns a log-likehood which is higher than the log-likelihood for the real model. This might help explain why the estimation algorithm converges to the estimated model. One explanation as to why the likelihood of the esti-

mated model is higher than the likelihood of real model, could be that because the model is simulated, it will only be approximately equal to the real model, and because only 2.000 observations were simulated, this approximation could be somewhat inaccurate.

	Estimated model	Real model
log-likelihood	-1706.6	-1747.692

Table 8: Log-likelihood calculated using the real model parameters and states, and the log-likelihood calculated the estimated parameters and states.

Using the knowledge from this section, we will attempt to estimate the MS-GARCH model on a financial time series. When performing this estimation we will further test if the model can be used to estimate the volatility differently when the markets are in a financial crisis, compared to when they are not.

9 Estimating MS-GARCH on S&P 500

In this section we will implement a model which accounts for the structural changes in the volatility, caused by financial crises, by using the MS-GARCH model. This will help determine whether the MS-GARCH model can be used in practice when modelling empirical data. It will further allow the possibility of comparing how well the model works compared to other models such as the GARCH model.

9.1 Data

The underlying data used for this analysis will be the *S&P 500 index* (S&P 500), which is an index measuring the performance of 500 companies traded at stock exchanges in the United States.[1]

This index has been chosen since it covers a wide spectrum of branches, providing a good representations of the American stock market. For simplicity we will be using the daily closing prices of index, and the daily return will be calculated as

$$r_t = \frac{p_t - p_{t-1}}{p_{t-1}}.$$

Since one the objective of this thesis is to model risk by accounting for the structural changes in the volatility, and because of the computational load associated with using large data sets in EM algorithms, we will not be using the entire index for the estimation. The index will instead be limited from 01-Jan-1999 to 01-Jan-2011. This interval encapsulates most of the Dot-com bubble in the late 90's, as well as the financial crisis in 2008, while also having a less volatile interval in between these two crises. This interval also makes the length of the time series roughly 3000 observations, which limits the calculation time of the estimation.

Figure 30 shows the S&P500 index, as well as the interval that will be used in this thesis for the estimation. In the returns it is clear to see that the beginning and end of the interval behaves much more extremely than the middle.

[1]The historical quotes for this index have been found on Yahoo finance https://finance.yahoo.com/

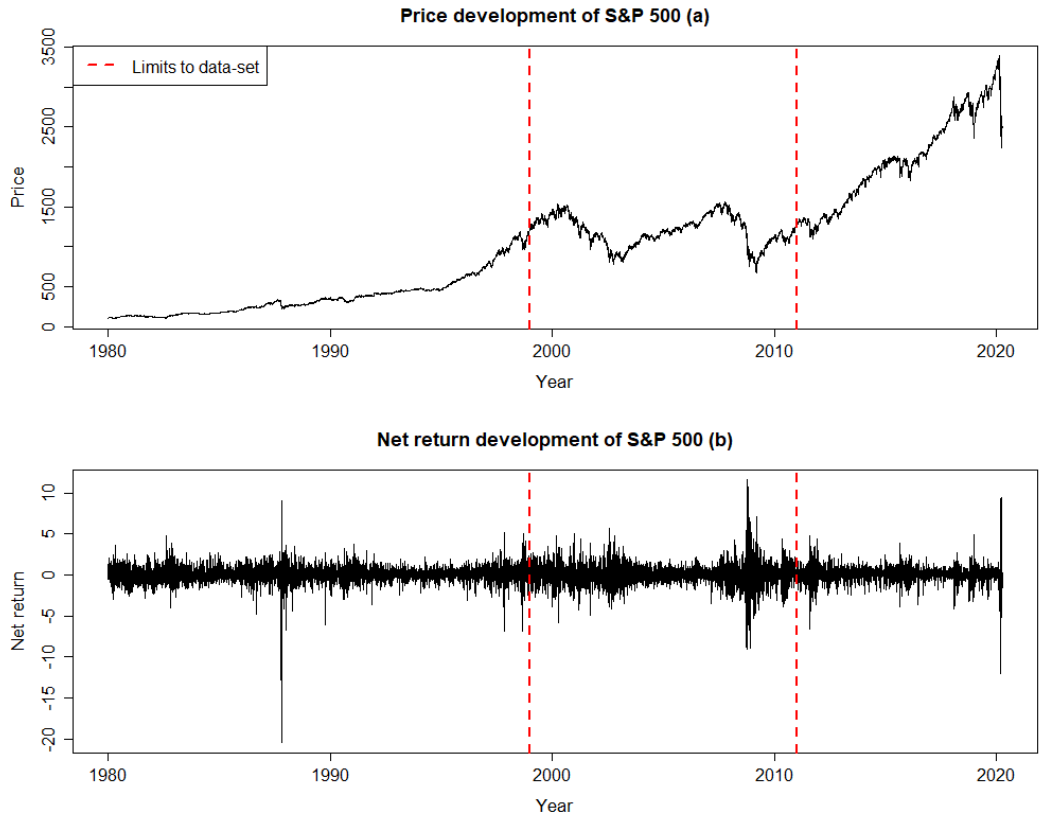


Figure 30: S&P 500 index from 05-01-1980 to 05-04-2020. Figure (a) shows the actual price development, while Figure (b) shows the net return.

A histogram of the returns can be seen in Figure 31 where it is clear to see that the returns in this time series are very heavy tailed. This also makes sense since two financial crises are included.

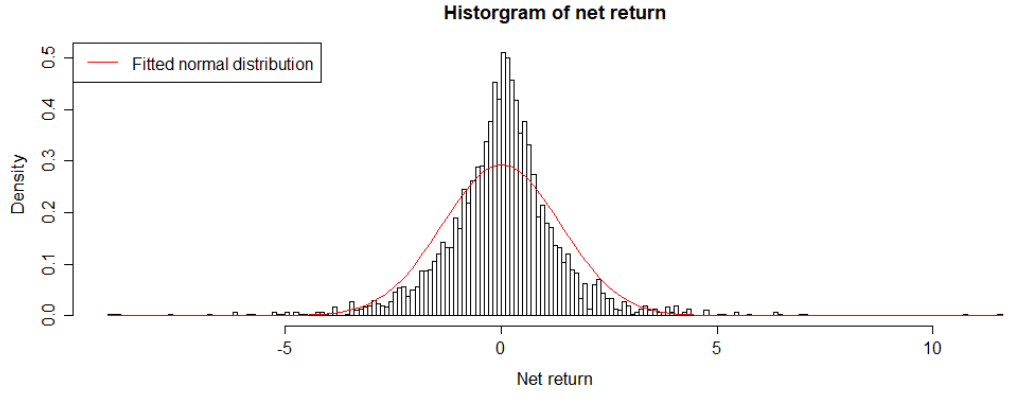


Figure 31: Histogram of the S&P500 index between 01-Jan-1999 and 01-Jan-2011. The normal distribution is fitted on the data where $\mu = 0.009$ and $\sigma = 1.3$

The central moments of the returns can also be seen in Table 9, where we also observe a

very high excess kurtosis.

Mean	Variance	Skewness	Excess kurtosis
0.009	1.359	0.082	7.474

Table 9: Central moments of the S&P500 index between 01-Jan-1999 and 01-Jan-2011

In Figure 32 the autocorrelations of the returns and the squared returns can be seen. Here we see a strong autocorrelation in the squared returns, however there does not seem to be any autocorrelation in the returns.
The heteroskedasticity indicated by the autocorrelation plots, as well as the high kurtosis and low skewness, seen in Table 9, indicates that a GARCH or MS-GARCH model could be a proper choice to model the volatility.

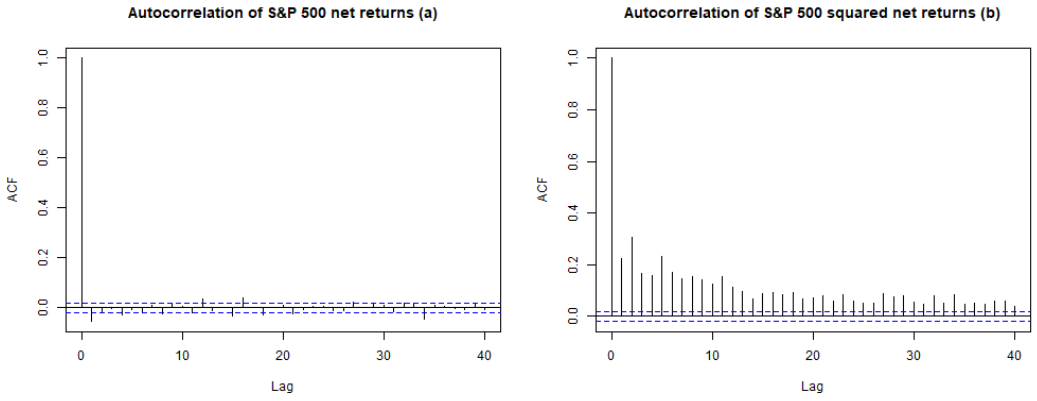


Figure 32: Autocorrelations of the S&P500 index between 01-Jan-1999 and 01-Jan-2011. Figure (a) shows the auto correlation of the returns, and Figure (b) shows the autocorrelation of the squared returns.

9.2 Model estimation

As in section 8, some starting parameters have to be defined before the estimation algorithm can be run. The process of choosing these will follow the same procedure as in the previous section, where regime 1 will be a less volatile and non-explosive regime, and regime two will be more volatile and explosive.

9.2.1 Initial values of the estimation

The starting point of the regime states will be the same as in the previous section, where each observation was be defined as regime 1.

The chosen starting point of the model parameters can be seen in Table 10. These parameters have been chosen since they are very different volatility processes, and the low probability of changing between regimes seem suitable since we are expecting the model to only have two regime changes (after the Dot-com bubble, and before the financial crisis in 2008).
The prior distributions have all been chosen as uniform distributions, and the boundaries of the prior distributions can be seen in Table 10.
These boundaries mainly differentiate the regimes by β and μ, which hopefully should be sufficient in order to keep the regimes from switching labels.

	Starting parameters	Limits	
		Min	Max
ω_1	0.10	0.01	0.60
α_1	0.05	0.01	0.10
β_1	0.40	0.20	0.60
μ_1	0.00	0.01	0.30
ω_2	0.40	0.01	0.30
α_2	0.20	0.01	0.15
β_2	0.80	0.70	0.95
μ_2	0.00	-0.40	0.10
$\eta_{1,1}$	0.99	0.00	1.00
$\eta_{2,2}$	0.99	0.00	1.00

Table 10: Starting parameters and limits of the prior distributions used in the estimation algorithm.

9.2.2 Parameter estimation

Figure 33 shows a trace-plot of 50.000 iterations of the estimation algorithm, using the starting parameters and prior distributions defined in Table 10. In this figure we see that the model converges rather quickly, and that the boundaries are wide enough to capture the probability mass for most of the parameters.

Looking at the kernel regressions of the trace-plot, a burn-in period of 10.000, like the one used in the previous section, seem sufficient. However the running average converges rather slowly for the α_1 and β_1 parameters, and therefore a longer burn-in period of 20.000 has been chosen in order to ensure the convergence of the process.

The posterior distributions calculated after removing the burn-in period can be seen in Figure 34. Here we see that posterior distributions of the parameters in regime two are very smooth, and lie well within the boundaries. Looking at the posterior distributions for the parameters in regime 1, show that the α_1 parameter converges towards zero, and the β_1 parameter is almost uniformly distributed. This will be investigated further in this section.

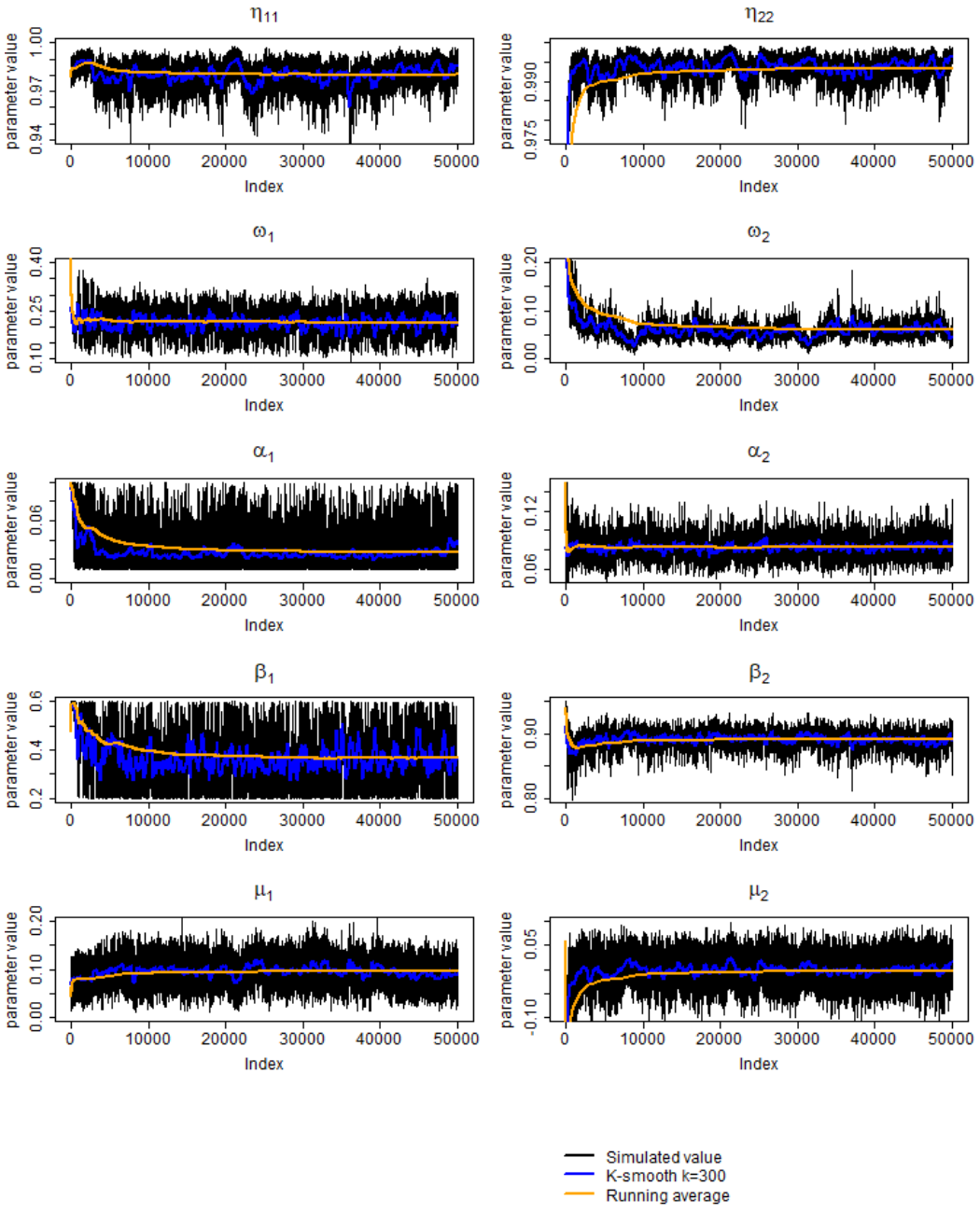


Figure 33: A trace plot showing the parameter estimates for each iteration of the Gibbs-sampler, for the S&P500 index.

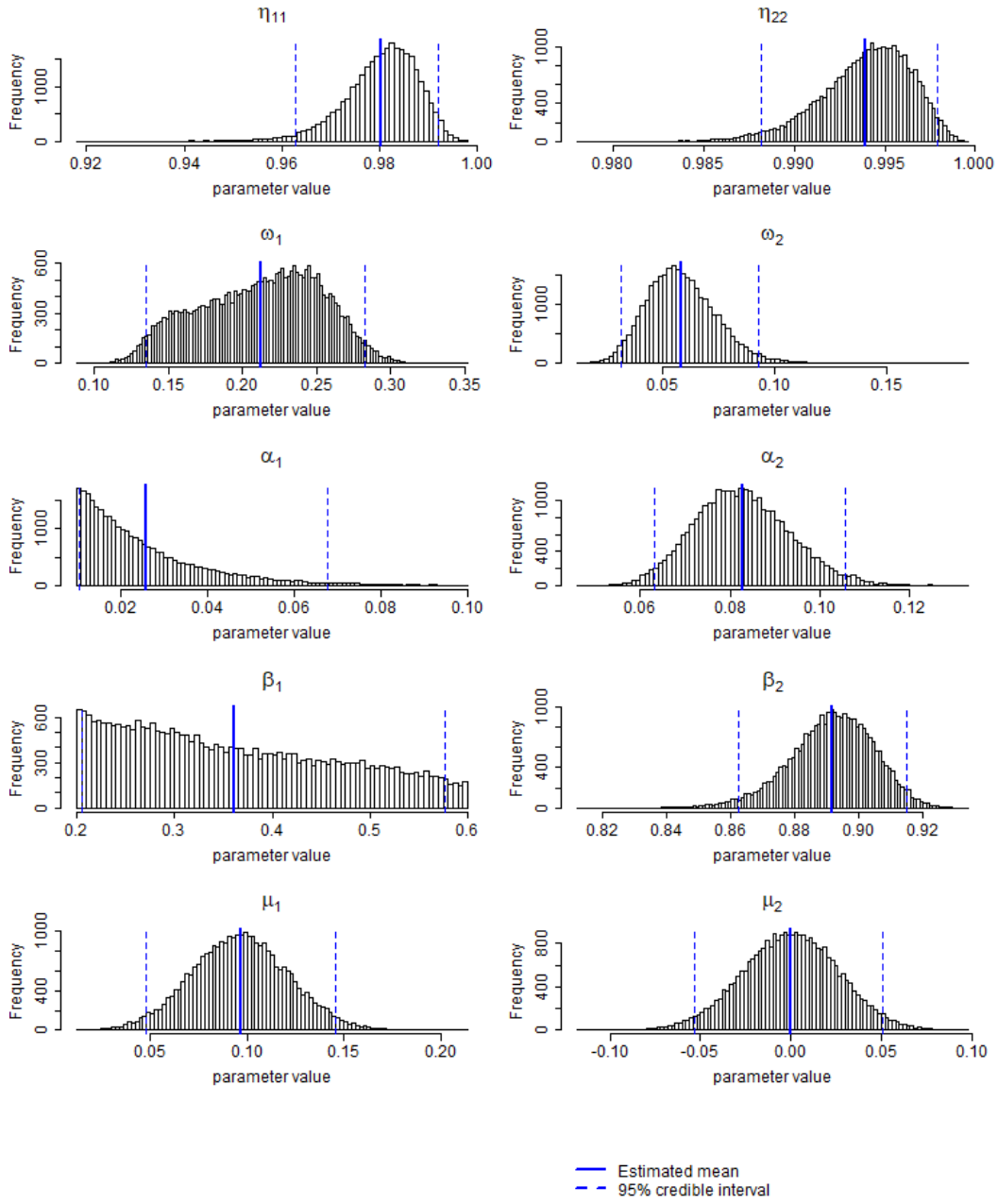


Figure 34: Histograms showing the posterior distributions of the parameters, after removing the burn-in period of 20.000 observations.

In previous section it was determined that the means of the posterior distributions produced good estimates for the model parameters. The model parameters calculated using the means of the posterior distributions, can be seen in Figure 11.
Looking at the transition probabilities $\eta_{1,1}$ and $\eta_{2,2}$ we see that the expected length of the

regimes will be

$$\begin{aligned} regime_1 &= \frac{1}{1-0.9804} \approx 51 \\ regime_2 &= \frac{1}{1-0.9939} \approx 163 \end{aligned} \tag{25}$$

which are quite small compared to the length of the time series. This means that estimated high volatility regime is most likely not confined to the two financial crises.

	Estimation	
	Mean	Std.
ω_1	0.2118	0.0412
α_1	0.0255	0.0149
β_1	0.3619	0.1118
μ_1	0.0959	0.0245
ω_2	0.0593	0.0162
α_2	0.0829	0.0109
β_2	0.8909	0.013
μ_2	-0.0002	0.0264
$\eta_{1,1}$	0.9804	0.0072
$\eta_{2,2}$	0.9939	0.0024

Table 11: Parameter estimates, and the standard deviation of the posterior distributions, for the MS-GARCH parameters estimated by the estimation algorithm for the S&P500 index.

In Figure 35 the estimated regime states are shown. These regime states are estimated using the method explained in Section 8, where an additional 10.000 iterations of the algorithm are used to generate the regimes states. In Figure 35 it is clearly seen that regime 2 is not limited to the financial crises, as hoped. The regimes do however capture most to the volatility periods within the time series.

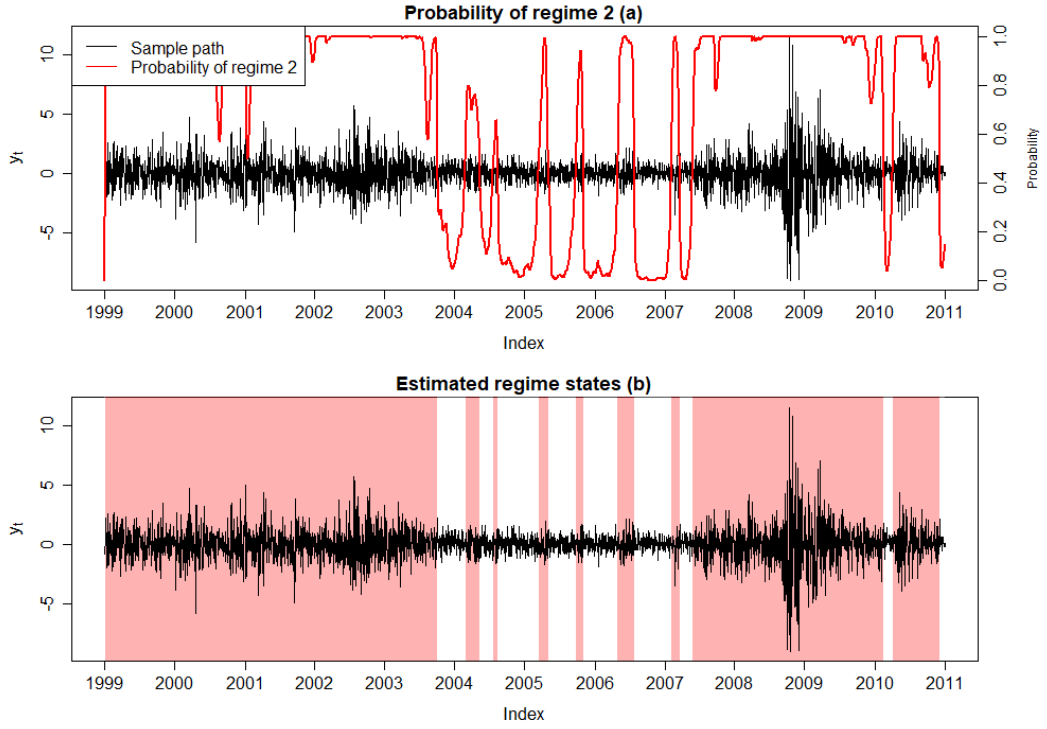


Figure 35: Estimated regimes and regime probabilities for the S&P500 index. Figure (a) shows the probability of being in regime 2 for each observation in the sample-path. Figure (b) shows the regimes are estimated by classifying regimes which have a probability higher than 50% of being in regime 2, as regime 2. The white background indicated regime 1 and the red background indicated regime 2.

9.2.3 Evaluation of the α and β parameters of regime 1

Going back to the posterior distributions of α_1 and β_1 in Figure 34, a contour plot showing the two parameters can be seen in Figure 36 (a). Here it is seen that when α_1 is low, the chance of β_1 being low will also increase.

One explanation for this correlation, is that, as α converges to zero, the GARCH process converges to a white noise process, as explained in section 4.1. When the GARCH model becomes a white noise process, i.e. when $\alpha \to 0$, β will only effect the variance level of the model.

If regime 1 has converged to a white noise process, it would be expected that the unconditional variance of the underlying GARCH model, which is defined in Equation 5, would converge as well. This would be expected since the only relevant attribute of a normally distributed white noise process, is the variance, and assuming the process has converged, the unconditional variance should therefore have converged as well.

Given that the unconditional variance has converged, a relationship between ω and β would be expected, since these parameters both effect the unconditional variance of the

underlying GARCH model. In Figure 36 (b), a contour plot showing samples of β_1 and α_1 can be seen, which show a clear correlation between the two parameters. This again indicates that regime one has converged to a white noise process.

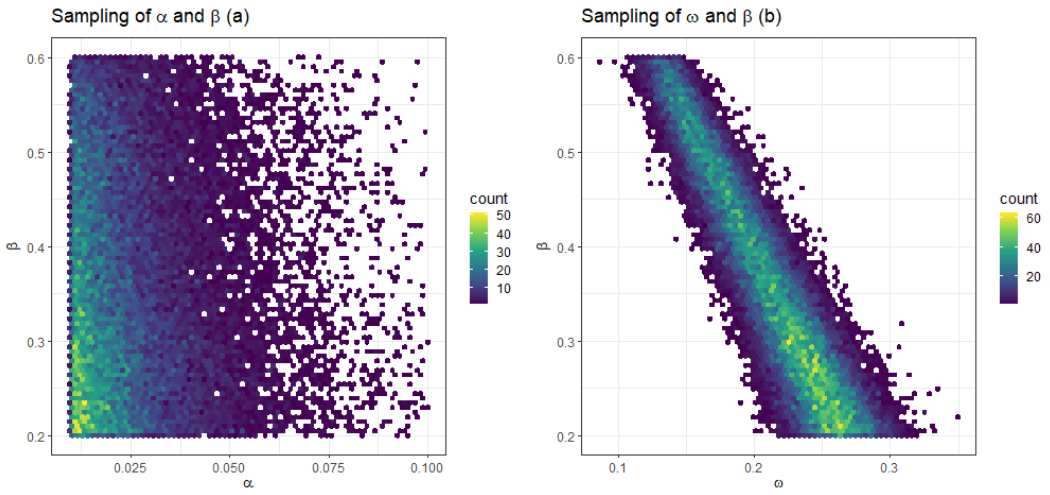


Figure 36: contour plot showing correlations between the GARCH parameters in regime 1. Figure (a) show the correlation between α_1 and β_1, and Figure (b) show the correlation between ω_1 and β_1

Figure 37 shows the unconditional variance of regime one. Here we see that the unconditional variance has converged to a fixed level, which also indicates that the process has converged to a white noise process.
This convergence helps explain why the observations of β_1 are almost uniformly distributed, since the only significance of the β_1 is to ensure that the unconditional variance stays relatively constant.

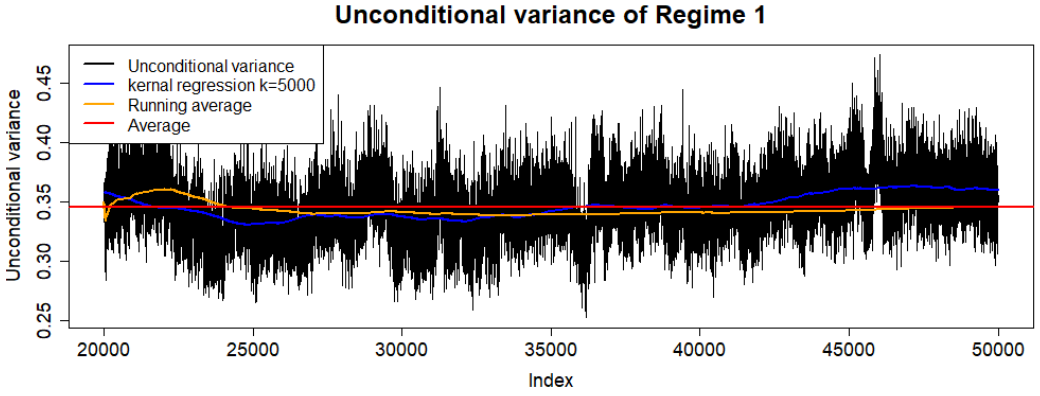


Figure 37: A trace plot showing the unconditional variance of the GARCH model in regime 1 calculated using Equation 5 on page 16

9.3 Modelling regime 1 with constant variance

Since regime 1 converges to a white noise process, it would be interesting to test how the estimation performs if α_1 and β_1 are fixed to zero. This would be equivalent to defining regime 1 as a white noise process with ω_1 variance.
Before doing this it has to be considered whether fixing α_1 and β_1 to zero, is allowed given the maximum-likelihood expression. To do this, consider the maximum-likelihood expression of the MS-GARCH model, explained in section 5

$$p(Y, S \mid \mu, \theta, \Gamma) = \prod_{t=1}^{t=T} \frac{1}{\sqrt{2\pi}\sigma_t} \exp\left[-\frac{(y_t - \mu_{s_t})^2}{2\sigma_t^2}\right] \eta_{s_{t-1}, s_t}.$$

In this expression neither α nor β appear directly, but instead they appear through σ. Since every maximum-likelihood expression used for the estimation algorithm is based directly on this expression, fixing α and β to zero should not effect the expressions.
Another point worth noting is that the stationarity requirement of the MS-GARCH model assumes $\alpha > 0$ and $\beta > 0$, which is obviously broken if α and β are fixed to zero. However since a white noise process is stationary, the MS-GARCH process will also be stationary is regime 2 is stationary. If regime 2 is not stationary, we will investigate the stationarity of the process further.

9.3.1 Parameter estimation

This estimation where regime 1 has been fixed to a white noise process has also been conducted using 50.000 iterations, and using the same starting parameters as the previous estimation. The prior distributions used are also the same, except for α_1 and β_1, of which the prior distributions are negated, given that these are constant. The trace plots and histograms for the posterior distributions of the parameters can be seen in appendix A.1 and A.2. These are not included here, as they look almost identical to the ones from the previous estimation.
Similarly to the previous estimation, it seems to have converged after 20.000 iterations, so this burn-in period will be kept. The estimated model parameter found after removing the burn-in period can be seen in Table 12.

	Estimation Regime 1 = GARCH		Estimation Regime 1 = white noise	
	Mean	Std.	Mean	Std.
ω_1	0.2118	0.0412	0.3437	0.0273
α_1	0.0255	0.0149	-	-
β_1	0.3619	0.1118	-	-
μ_1	0.0959	0.0245	0.0964	0.0246
ω_2	0.0593	0.0162	0.0594	0.0153
α_2	0.0829	0.0109	0.0825	0.0110
β_2	0.8909	0.0130	0.8915	0.0138
μ_2	-0.0002	0.0264	0.0000	0.0263
$\eta_{1,1}$	0.9804	0.0072	0.9799	0.0074
$\eta_{2,2}$	0.9939	0.0024	0.9939	0.0023

Table 12: Comparison between the parameter estimates when regime 1 is assumed to be a GARCH model, and when regime 1 is assumed to be a white noise process.

Here we see that $\alpha_2 + \beta_2 < 1$, which means that regime 2 is stationary, and the entire process will therefore also be stationary. We also see that the parameters in regime 2 are almost identical to what they were in the previous estimation.
Comparing the parameters of regime 1 is not as straight forward, as the underlying models of the estimations are different. However the μ_1 parameters of the two estimations are equivalent.
One way of comparing regime 1 in the two estimations, is to compare the unconditional variance, from the first estimation, with ω_1 from the model where regime 1 has been defined as a white noise process. This comparison can be seen in Figure 38, where the empirical densities of the two estimations are shown. Here we see that the unconditional variance, and the variance of the white noise process are almost identical.

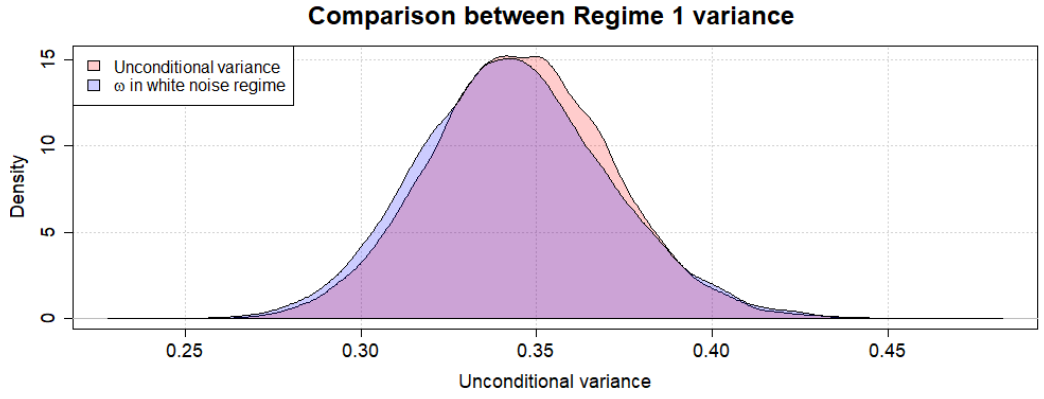


Figure 38: Comparison between the posterior distribution of the uncnditional variance of regime 1, when assumed to be a GARHC model, and ω of regime 1, when assumed to be a white noise process. The purple area indicated the overlap of the blue and red areas.

In Figure 39 the estimated regime states, and regime probabilities are shown. Here we see that the regimes are also almost identical to the ones found in the previous estimation. This was also to be expected, since the rest of the model has been almost identical between the two estimations.

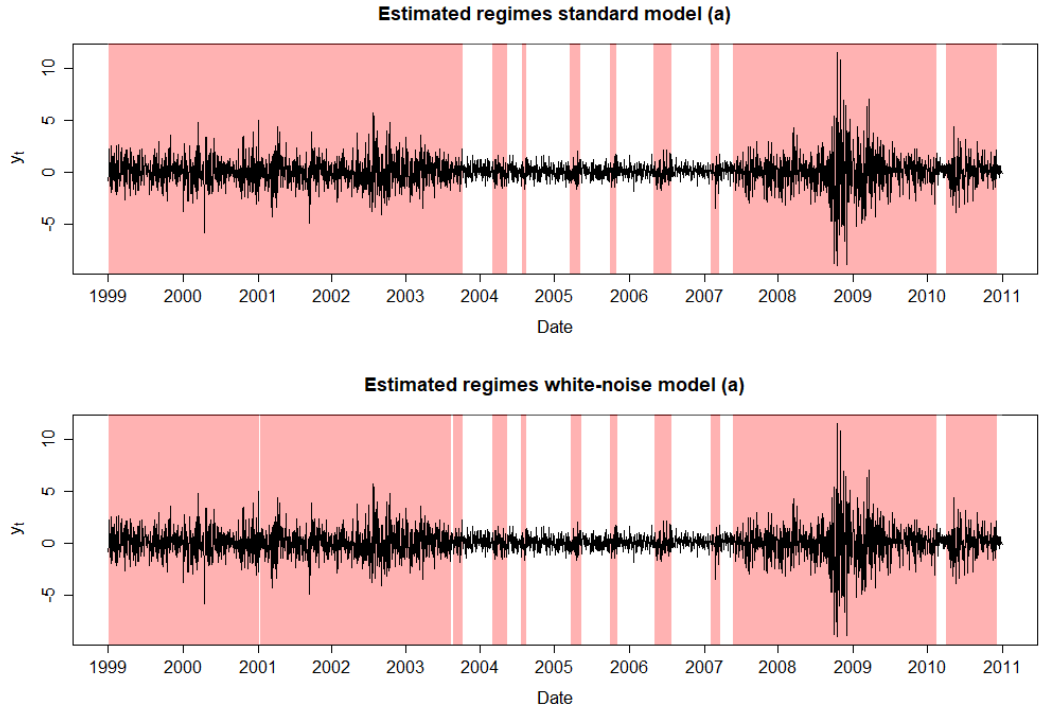


Figure 39: Estimated regimes using the standard MS-GARCH model, and the estimated regimes when fixing regime 1 to a white noise process. The white background indicated regime 1 and the red background indicated regime 2.

The fact that the MS-GARCH model estimates regime one as a white noise process is quite interesting. Since risk is usually modelled using either very simple models, such as a white noise process, or more complex models such as the GARCH model, it therefore seems fitting that the MS-GARCH model suggest that both models work in different scenarios. Since modelling regime 1 as a GARCH model only provides a more complex model, this model will no longer be considered. Instead regime 1 will be fixed to a white noise process moving forward.

9.4 Separation of regimes

As shown in Figure 39, regime 2 is not only based on the financial crises. Therefore we will consider alternatives which may split up the regimes such that regime 2 is solely based on the financial crises.

One solution which may help further spilt up the regimes is smoothing the state probabilities. Doing this could generate the desired states, which could then be fixed, after which another estimation can be run using the fixed states. For smoothing the regime state probabilities, the smoothing expression shown in Equation 8 on Page 21 will be used.

After smoothing the regime state probabilities the regime states are as shown in Figure 40. Here it is seen that some of the observations which are classified by both of the regimes in 50% of the iterations, will be pushed into one of the two regimes. This gives 'cleaner' regimes, which could be quite useful, however it does not resolve the issue, as there are still areas between the two financial crises which are defined by regime 2.

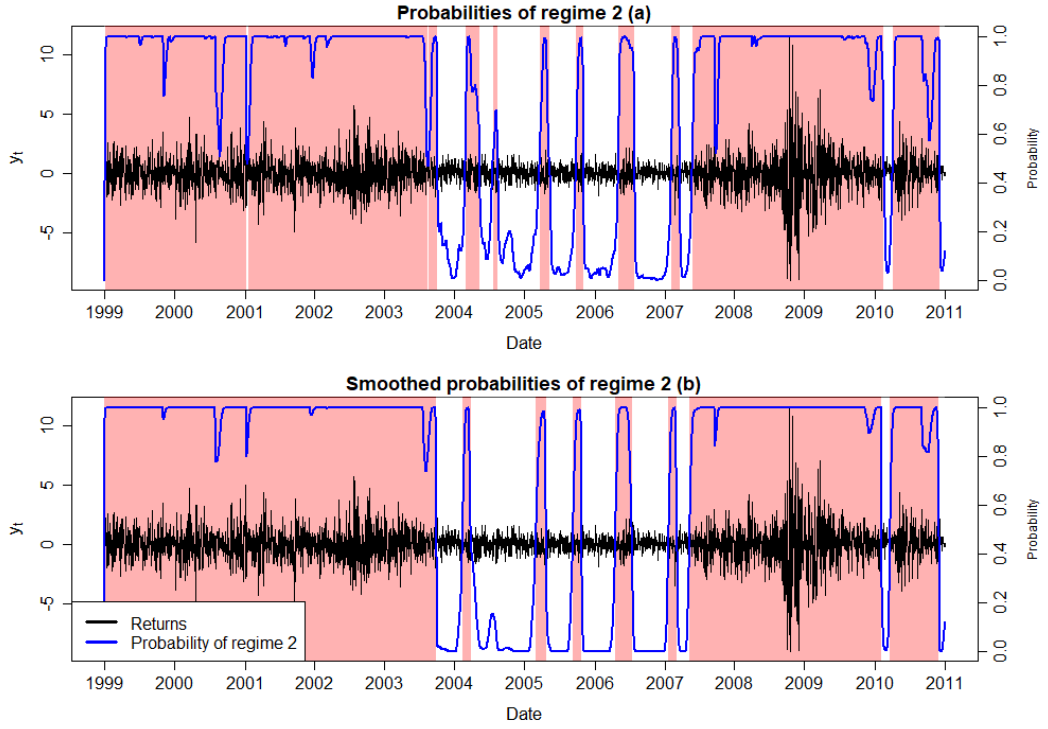


Figure 40: Smoothed probabilities of being in regime 2 calculated using Equation 8.

Another way of splitting up the two regimes as desired, is by using another prior distribution for the transition probabilities. This might help the process converge to another local maximum, which only includes regime 2 in times of financial crises.

The new prior distribution will be chosen such that it increases the probability of sampling high values of $\eta_{1,1}$ and $\eta_{2,2}$. To do this two different options will be considered. The first option will be using a truncated distribution with a limit in 0.99. The truncated distribution will remain uniform, but hopefully the small sampling interval will help the estimation algorithm converge at a higher level for $\eta_{1,1}$ and $\eta_{2,2}$.

The second option will be to apply a skewed beta prior distribution which has a very high probability of sampling close to 1, but still not predetermine the posterior distribution. For doing this we have chosen a skewed beta distribution with $shape_1 = 200$ and $shape_2 = 1.1$. This distribution can be seen in Figure 41.

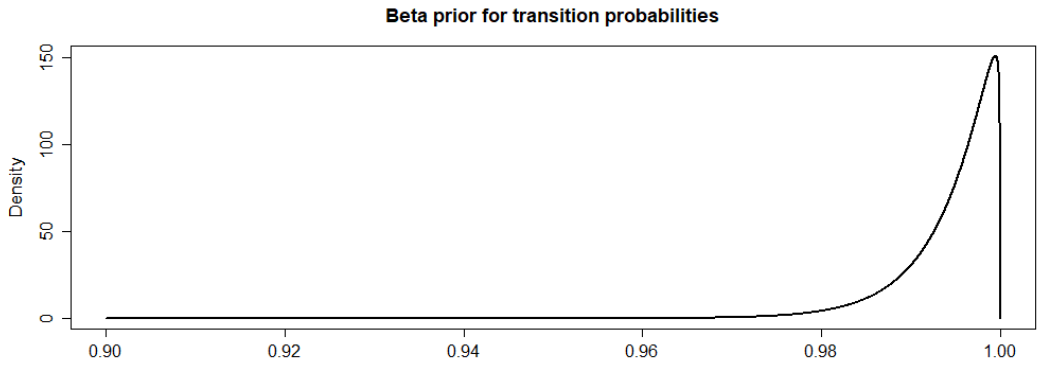


Figure 41: The skewed beta distribution used as a prior distribution for the transition probabilities in the estimation algorithm.

The trace plots from running the estimation algorithm with the new prior distributions for the transition probabilities can be seen in appendix B.1 and C.1. In the trace plots it is seen that the estimation algorithm converges after 20.000 iterations as with the rest of the estimations. Therefore this burn-in period will be used for these estimations as well. The state probabilities generated using the new prior distributions can be seen in Figure 42. Here we see that the skewed beta distribution achieves the desired result where the regimes are split into financial crises, and non-financial crises.
The truncated distribution however failed to achieve this result. The reason for this is most likely that the boundaries of the truncation were too wide, which caused the estimation algorithm to converge to the previous regime states.

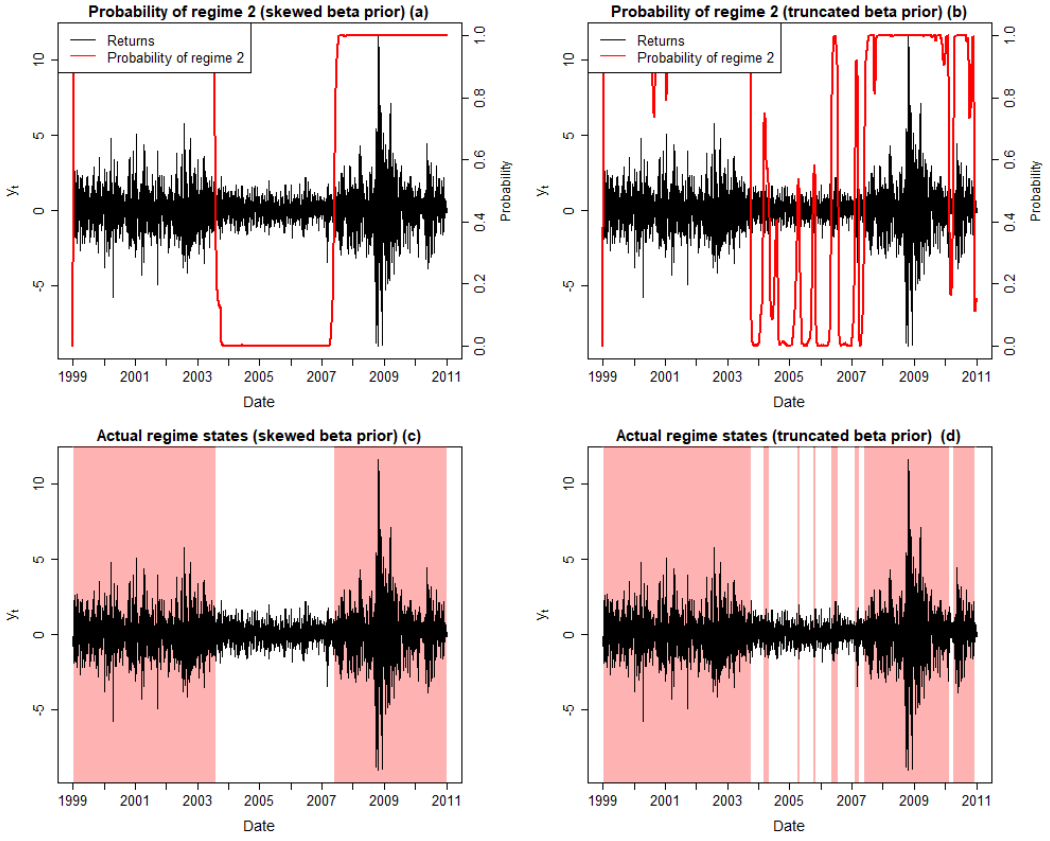


Figure 42: Estimated regimes and regime probabilities using the skewed and truncated beta distributions as priors. Figure (a) and (c) show the regimes found using the skewed beta distribution. Figure (b) and (d) show the regimes found using the truncated distribution. The white background indicated regime 1 and the red background indicated regime 2.

This explanation is also supported by Figure 43, which shows the trace plot of the $\eta_{1,1}$ and $\eta_{2,2}$ parameters, in the estimation where a truncated prior distribution is used.
In this figure it is clear to see that $\eta_{1,1}$ is sampled in the boundary of the truncation. This means that the truncated prior distribution could most likely produce the desired regimes if it was truncated at a higher level. However since the skewed beta distribution already achieves this, we will stick to using this prior distribution.

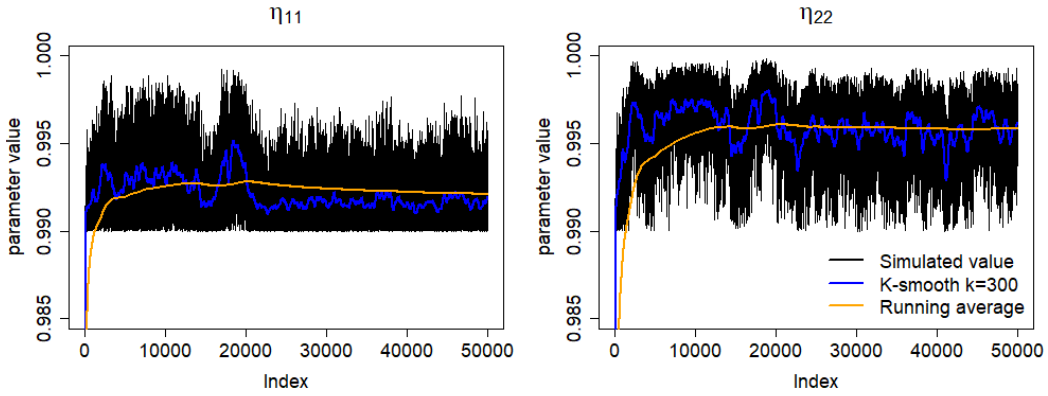


Figure 43: trace plots showing the transition probabilities from the estimation algorithm where a truncated prior distribution is used for $\eta_{1,1}$ and $\eta_{2,2}$.

We now have two different estimated models,

- The *white-noise model*, where a white noise process was used to estimate regime 1
- The *beta-prior model*, where a white noise process was used to estimate regime 1, and a skewed beta prior distribution was used for the transition probabilities.

The posterior distributions of the *beta-prior model* are shown in Figure 44. We see that the posterior distributions lie within the boundaries, and seem to have converged to smooth distributions. It is only the posterior distribution for $\eta_{2,2}$ which look unusual, however this is simply because the probability of switching away from regime 2 is very low.
We also see that the posterior distributions are distinctly different, even though the same prior distribution has been used for both $\eta_{1,1}$ and $\eta_{2,2}$. This means that the prior distribution has not predetermined the shape of the posterior distributions, but instead it has led to a new local optima.

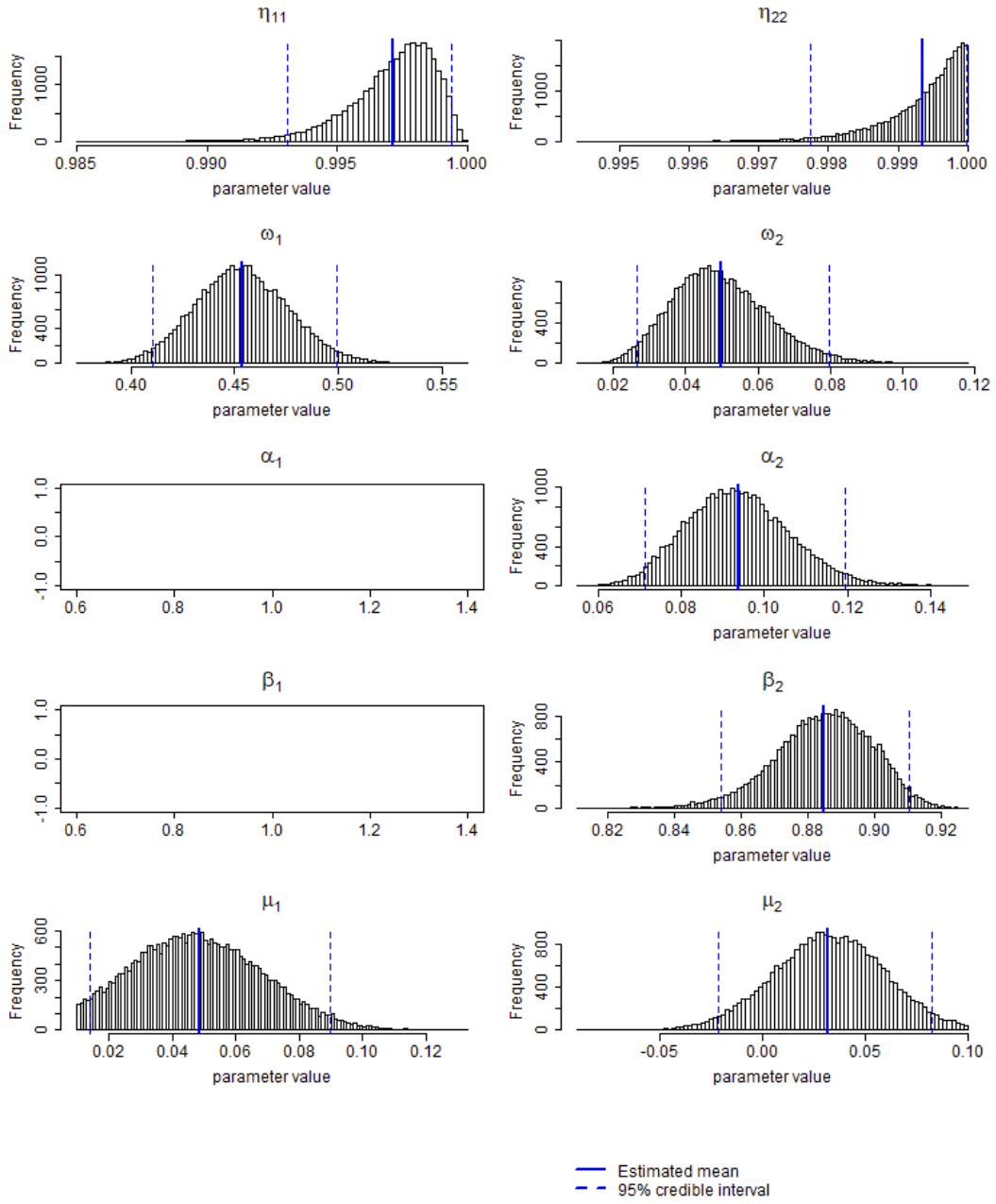


Figure 44: Histograms showing the posterior distributions of the parameters, after removing the burn-in period of 20.000 observations, and defining regime 1 as a white noise process as well as using a skewed beta prior for the η parameters.

The estimated model parameters of the *beta-prior model* can be seen in Table 13. Here we see that regime 2 is not much different than it was before changing the prior distribution. The only noticeable differences is that the drift becomes positive, whereas it was zero before changing the prior distribution. The process has also become slightly more explosive

since α_2 has increased.
In regime 1 the variance has increased a lot. This is most likely due to the fact that the periods which were previously classified as regime 2, has now been classified as regime 1. Since these periods are more volatile than the rest of regime 1, they will increase the volatility of regime 1.

	Estimation *white-noise model*		Estimation *beta-prior model*	
	Mean	Std.	Mean	Std.
ω_1	0.3437	0.0273	0.4536	0.0226
μ_1	0.0964	0.0246	0.0487	0.0198
ω_2	0.0594	0.0153	0.0497	0.0135
α_2	0.0825	0.0110	0.0937	0.0124
β_2	0.8915	0.0138	0.8849	0.0144
μ_2	0.0000	0.0263	0.0322	0.0266
$\eta_{1,1}$	0.9799	0.0074	0.9971	0.0016
$\eta_{2,2}$	0.9939	0.0023	0.9994	0.0006

Table 13: Comparison between the parameter estimates of the *white-noise model* and the *beta-prior model*.

9.5 Model comparison and evaluation

Since a prior distribution had to be used in order to get the *beta-prior model*, it is worth investigating if this assumption, has made the model worse when modelling the returns than the *white-noise model*. In order to evaluate this we will look at the residuals and see if one of the models perform significantly worse than the other.
The residuals are calculated equivalently to how they were calculated in Section 8, where both the residuals as well as the squared residuals are calculated. Figure 45 shows qq-plots of the residuals of both models. Here we see that none of the models are able to explain the most extreme returns, as the residuals have heavy tails. We also see that the *beta-prior model* is slightly worse at modelling the variance when looking at the qq-plots for the squared residuals. However this is not a massive difference, and the two models should therefore perform quite similarly.

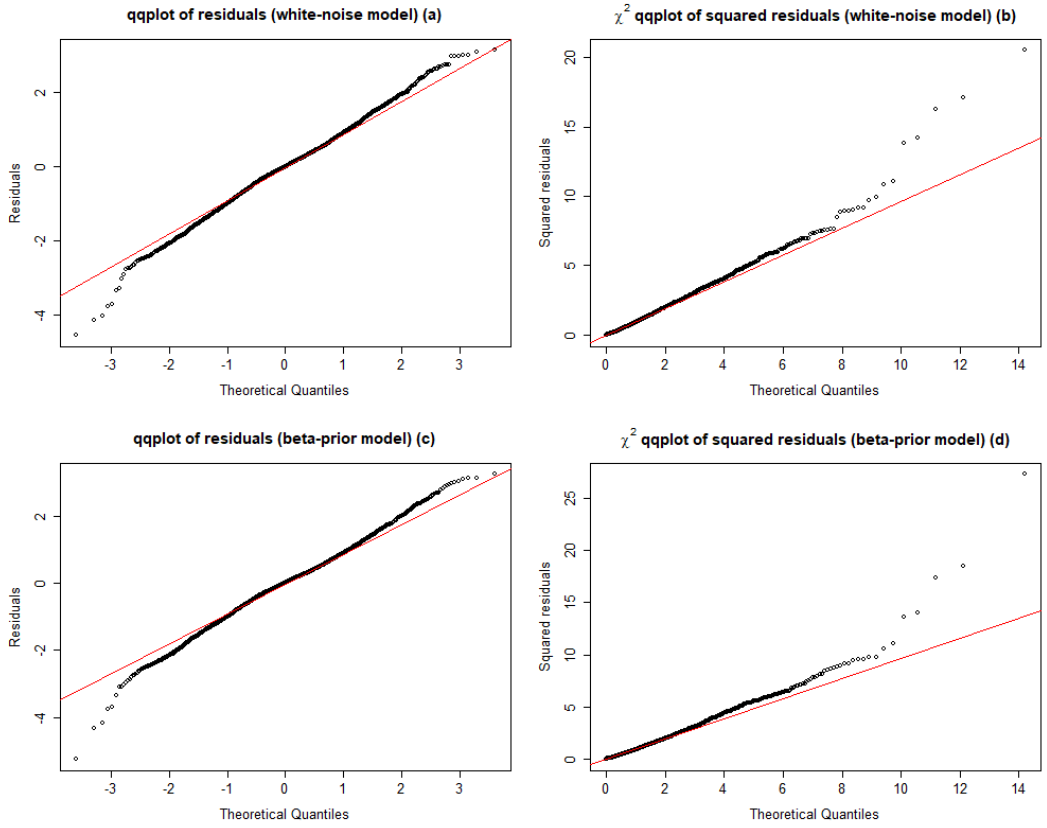


Figure 45: qq-plots for evaluating the residuals of the models achieved with and without a skewed beta prior for the transition probabilities.

The autocorrelations of the residuals and squared residuals are also shown in Figure 46. Here we see that none of the residual show any significant autocorrelation. Therefore the models main has difficulty, is modelling the extreme returns which are seen in financial time series from time to time.

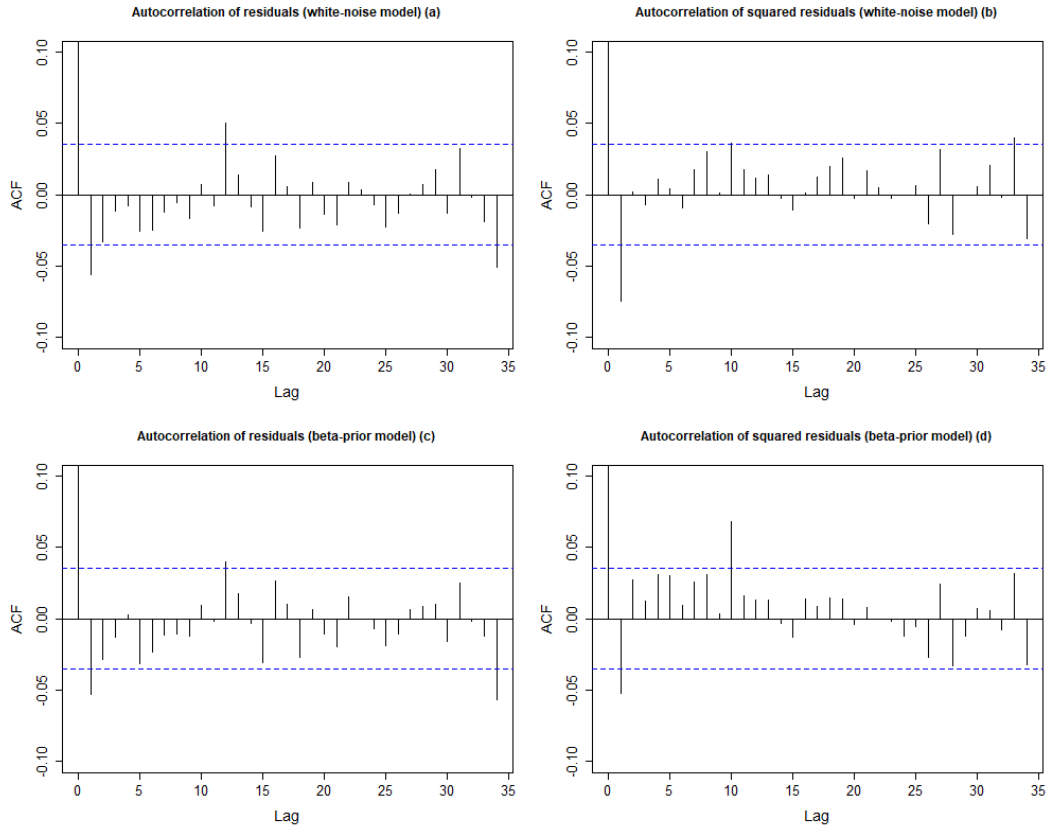


Figure 46: Autocorrelation plots for evaluating the residuals of the models achieved with and without a skewed beta prior for the transition probabilities.

Through estimation of the MS-GARCH model on the S&P 500 index from 01-Jan-1999 to 01-Jan-2011, we hoped to be able to split up the period, such that the structural changes in the volatility process could be accounted for. By doing so it was found that periods of high volatility were described well by a GARCH process, and volatility was constant in periods of low volatility.
It was further shown that, if a skewed beta distribution was used as a prior for the transitions probabilities, the estimated periods could be split up such that regime 2 was solely based on the financial crises.

In the next section we will apply the *beta-prior model* to another part of the S&P 500 index, in order to evaluate its capabilities as a risk model.

10 Evaluation of MS-GARCH

The purpose of the empirical study in the previous section, was to explore the possibilities of implementing an MS-GARCH model which is able to model the structural changes in the volatility caused by financial crises.
We proposed the *beta-prior model* which was found using a skewed beta distribution as prior distribution for the transitions probabilities, and modelled regime 1 as white noise. this model gave a clear separation of highly volatile and low volatile periods.
In this section we wish to examine if, and how, this model can be used for modelling the daily risk of financial assets.

10.1 Comparison of MS-GARCH and GARCH

The parameter estimates from the previous section will be used, to estimate the risk on the remaining part of the S&P 500 index, i.e. the period after 01-Jan-2011. This series is also shown in Figure 47.

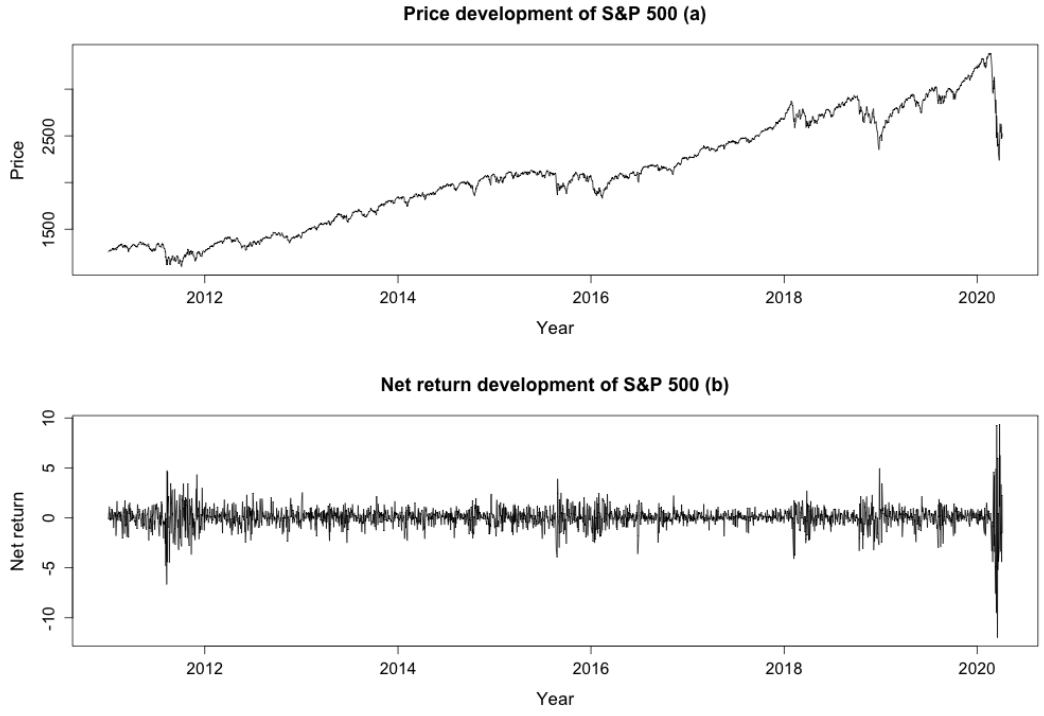


Figure 47: S&P 500 index from 01-Jan-2011 to 05-Apr-2020. Figure (a) shows the actual price development, while Figure (b) shoes net return.

When estimating the risk we will use the value-at-risk metric described in section 3, where

the parametric method will be used with a normal distribution, where the variance is defined as the conditional variance of the MS-GARCH model.
In order to evaluate how well the MS-GARCH model estimates the risk, it will be compared with a standard GARCH(1,1) model, which was described in Section 9. The parameters of the GARCH model will found similarly to how the parameters in the MS-GARCH model were found, using the S&P 500 index from 01-Jan-1999 to 01-Jan-2011. The estimation of the parameters in the GARCH model is done using the *rugarch* package in R.
This model will then be used to estimate the risk in the S&P 500 index after 01-Jan-2011, in order to compare it to the risk estimated by the MS-GARCH model.
For this comparison a higher lagged GARCH(p,q) model could also be used, however since the MS-GARCH model is based on the GARCH(1,1) model, it makes sense to compare these.

Table 14 shows estimated parameters of the MS-GARCH found in Section 9.2 and the GARCH model parameters found using the *rugarch* package in R.

	Estimation *beta-prior model*		Garch(1,1)	
	Mean	Std.	Estimate	Std. error
ω_1	0.4536	0.0226	0.012	0.003
α_1	-	-	0.077	0.009
β_1	-	-	0.917	0.009
μ_1	0.0487	0.0198	0.044	0.017
ω_2	0.0497	0.0135	-	-
α_2	0.0937	0.0124	-	-
β_2	0.8849	0.0144	-	-
μ_2	0.0322	0.0266	-	-
$\eta_{1,1}$	0.9971	0.0016	-	-
$\eta_{2,2}$	0.9994	0.0006	-	-

Table 14: Parameter comparison between MS-GARCH model and GARCH(1,1) with $u_t \sim \mathcal{N}(0,1)$ estimated using the S&P 500 index between 01-Jan-1999 to 01-Jan-2011.

As shown in Table 14 the parameters of the standard GARCH model does not deviate

much from the second regime in the MS-GARCH model.
This is quite interesting as the standard GARCH model has to account for both volatile and non-volatile periods whereas regime 2 of the MS-GARCH model only has to account for the volatile periods.
Nonetheless it is still interesting to compare the two models, and thereby examine if low volatile periods can be explained by a white noise process.

When applying the MS-GARCH model to a new series, the underlying regimes have to be found, in order to produce risk estimates for the MS-GARCH model.
The first issue with estimating the regime states with fixed parameters, is whether the transitions probabilities should be fixed or re-estimated.
The estimated transition probabilities, $\eta_{1,2}$ and $\eta_{2,1}$, are roughly equal to the number of regime shifts in the estimation interval, divided by the total length of the estimation interval. Because of this the transition probabilities will depend on where the limits of the estimation interval are, as a single regime shift can have a large effect on the estimated transition probability.
However because the rest of the model parameters have been estimated using those transition probabilities they will be kept constant. This might affect the precision of the estimated regime states, but this is an issues that arise, when a model is fitted on a different interval than it is applied to.
The rest of the parameters of the MS-GARCH model, namely the θ and μ will be kept constant as well.

Because the regime states found in section 9 are estimated directly using the full estimation algorithm, these will also be kept constant. This allows for some simplifying conditions, as it gives a starting regime s_0, and starting variance σ_0^2 for estimating the new regimes.
Keeping all the parameters constant removes a large part of the estimation procedure described in Section 7, and the only remaining part will be the regime state estimation described in Section 7.3.

The regimes estimated using the modified estimation algorithm, where the model parameters have been fixed can be seen in Figure 48. Here we see that the model is able to split the new series very well, as it is only highly volatile periods that are classified as regime two.

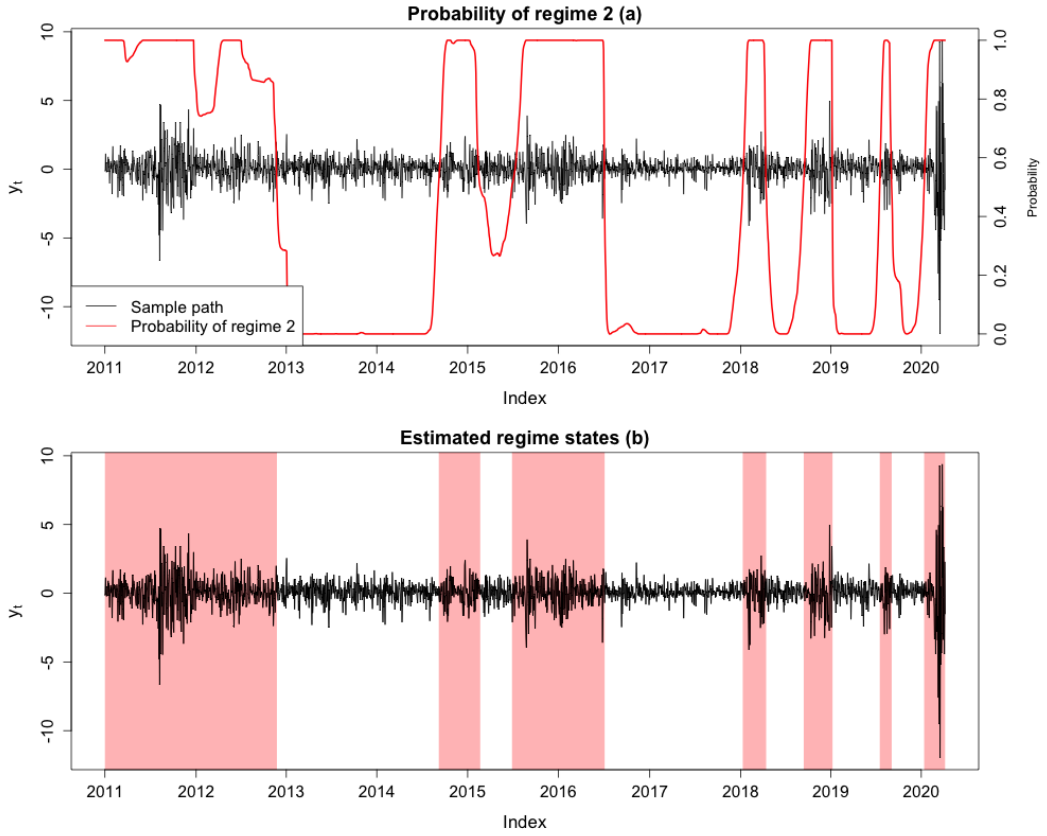


Figure 48: Estimated regimes and regime probabilities for the S&P500 index after 01-Jan-2011. Figure (a) shows the probability of being in regime 2 for each observation in the sample-path. Figure (b) shows the regimes are estimated by classifying regimes which have a probability higher than 50% of being in regime 2, as regime 2. The white background indicated regime 1 and the red background indicated regime 2.

Using these regimes, it is possible to calculate the estimated variance process using the MS-GARCH model. The variance process of the MS-GARCH model compared with the equivalent variance process calculated using the standard GARCH model with the parameters from Table 14, can be seen in Figure 49. Here we see that the variance of the two models are similar when the MS-GARCH model is in regime 2, however when it is in regime 1, they are quite difference.

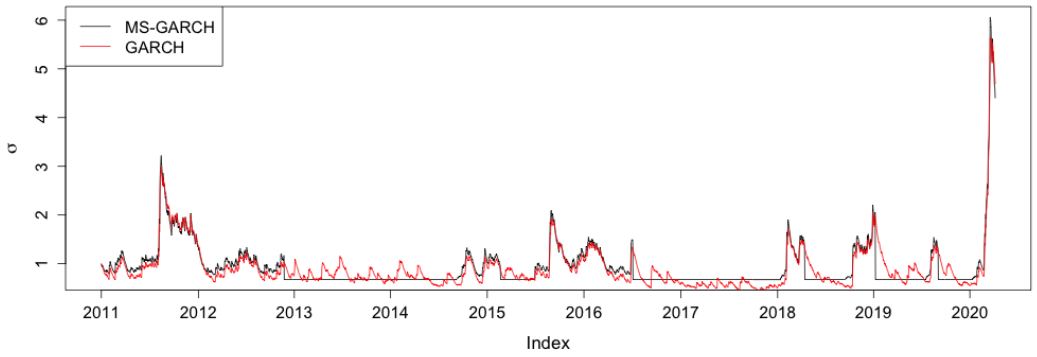


Figure 49: Comparison of the estimated variance using the MS-GARCH model and a standard GARCH model.

Using the conditional variance processes shown in Figure 49, it is also possible to calculated the VaR_α estimates. This has been done for a confidence level of $\alpha = 99\%$ in Figure 50, where the two models seem to perform similarly.
One interesting thing regarding the VaR estimates in Figure 50, is that Regime 1 acts almost as a lower bound on the VaR estimate. This lower bound on the conditional variance seem quite fitting, as no matter how limited the movement of a financial asset is, there will always be an underlying risk, which has to be accounted for.

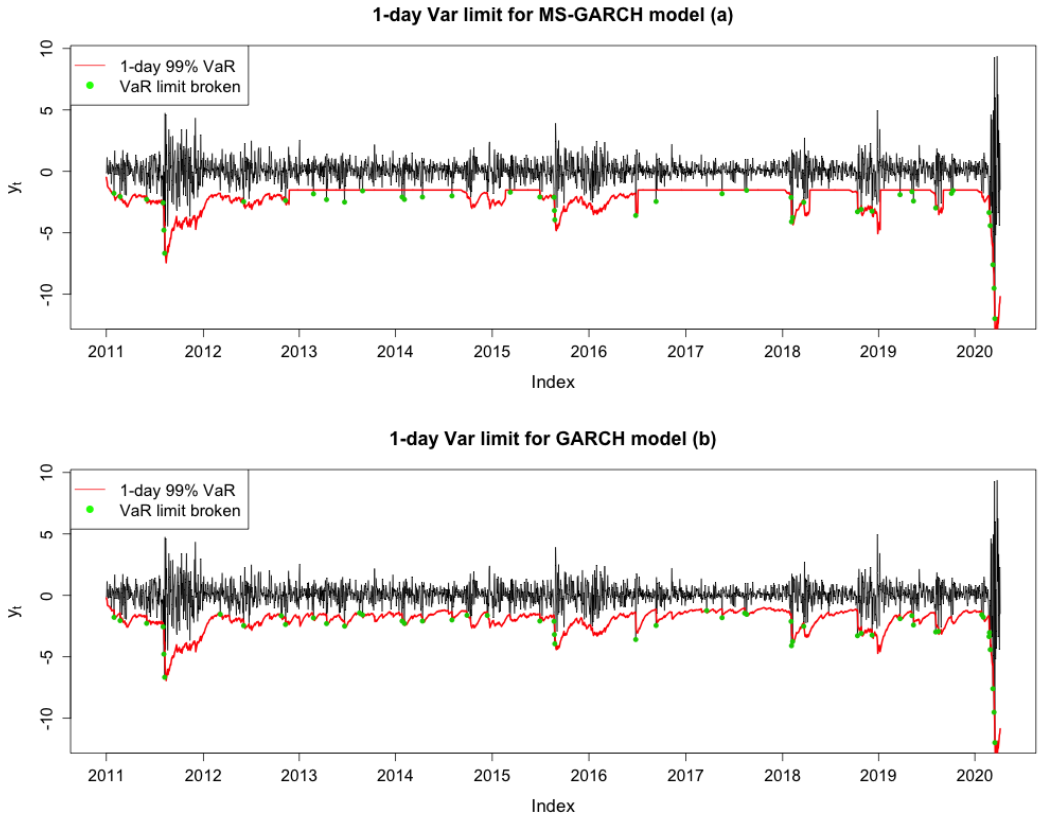


Figure 50: $VaR_{99\%}$ estimated calculated using the conditional variance from the MS-GARCH model and the GARCH model.

Using the two series of VaR estimates, we can also calculate how often the returns break the VaR_α limit, and compare this with what would be expected at the given α level.
As there are 2331 returns in the used series, and since we calculate the VaR on a $\alpha = 99\%$ level, we would expect the VaR_α limit to be broken roughly 23 times. In Table 15 the number of times the VaR_α limit was broken using the MS-GARCH and the GARCH model are shown.
Here we see that the MS-GARCH model is closer than the GARCH model to the expected number of broken VaR_α limits. While this is not the only important aspect of the risk modelling, it still indicates that the MS-GARCH model is slightly better at modelling the risk.

	Expected	MS-GARCH	GARCH
Number of broken VaR_α limits	23	45	51

Table 15: Number of times the estimated $VaR_{99\%}$ estimates were broken for the MS-GARCH and GARCH model.

10.2 Daily value at risk estimation

Since the likelihood expression of S depends on all future values of Y, as described in section 7.3, every observation, except for the last one, is conditioned on unavailable information, in the context of a risk model. This means that the estimated state of a date will be conditioned on every observed return after that date, and since this information will not be available at the time of the risk estimate, these risk estimate will be unattainable.

Instead of evaluating the model as a historical estimate, it would be interesting to evaluate it as a true risk model, which only condition on known information.
When doing so, the $S_t = \{s_1, ..., s_t\}$ process will have to be recalculated for each day, while only conditioning on $Y_t = \{y_1, ..., y_t\}$ in order to calculate σ_t. Doing this presents another problem, as this allows for the regimes before t to change, as time passes, and the risk on a given day can change as more information becomes available.
In order to show this effect, Figure 51 shows how the regime states are estimated using the first 1500 observation of the time-series, and how they were estimated using the full time-series. Here we see that the area in the beginning of 2012 and 2016, are estimated differently by the two models, and given the path dependency of the variance, this can also have an effect on the risk estimated in the future.

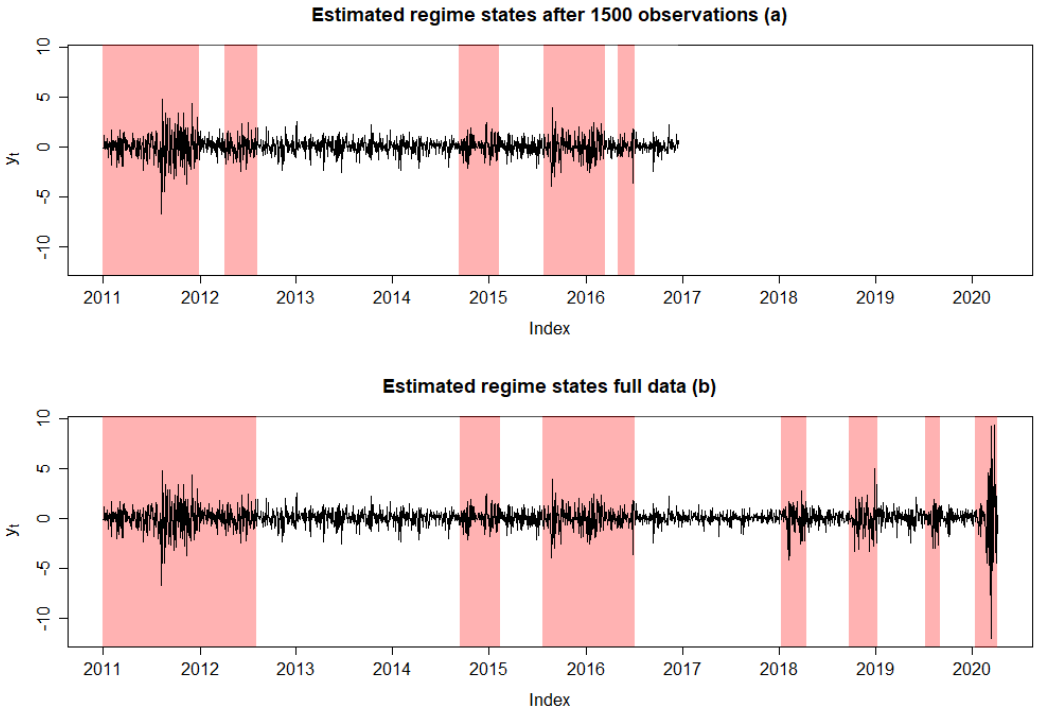


Figure 51: Estimated regimes using the first 1500 observation, and using the full time-series from 01-Jan-2011 to 05-Apr-2020.

A way of avoiding this, is to fix the estimated states every day. This could also be convenient as the regime states could then be used as an indicator of how the next period of time is expected to behave, i.e. the model will tell when we are in a crisis scenario.
The issue with this approach is that the estimation procedure will no longer take the path dependence into account. Effectively this means that the estimated regime state will be based directly on the last term of the likelihood expression, as well as the transition probability from the previous state. The state probabilities will therefore be possible to calculate using

$$\begin{aligned} q(s_t|S_{\neq t},\mu,\theta,\Gamma,y) &\propto \log\left(\eta_{s_{t-1},1}^{2-s_t}\eta_{s_{t-1},2}^{s_t-1}\eta_{s_t,1}^{2-s_{t+1}}\eta_{s_t,2}^{s_{t+1}-1}\right)+\sum_{j=t}^{T}-\log(\sigma_j)-\frac{(y_j-\mu_{s_j})^2}{2\sigma_j^2} \\ &\propto \log\left(\eta_{s_{t-1},1}^{2-s_t}\eta_{s_{t-1},2}^{s_t-1}\right)-\log(\sigma_t)-\frac{(y_t-\mu_{s_t})^2}{2\sigma_t^2}, \end{aligned} \tag{26}$$

which can be calculated for $s_t = 1$ and $s_t = 2$ in order to calculate the proportional probability, as done previously.
Calculating the regime states using this method, yields the probabilities shown in Figure 52. Here it is important to note that the second axis has been scaled, and that every observation will effectively be classified as regime 2.
This method can therefore not be used, which also makes sense, given that the path-dependency was not accounted for.

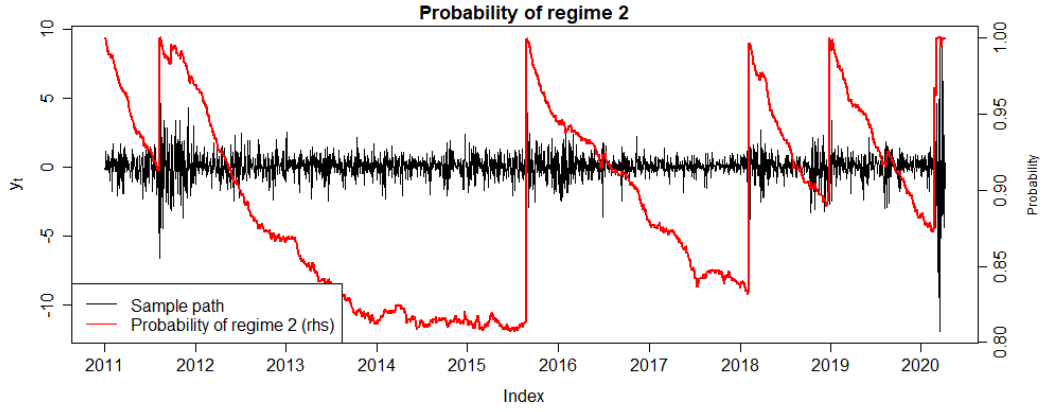


Figure 52: Transition probabilities found by keeping the previous regimes constant.

Because of this, the variance of each observation has to be based on a individually estimated series of regime states, calculated using every previous observation. The variances calculated when only conditioning on the known information is shown in Figure 53. Here we see that the two models work similarly in the high volatility periods, however in the low volatility periods, the model which is only based on the known information is less stable.

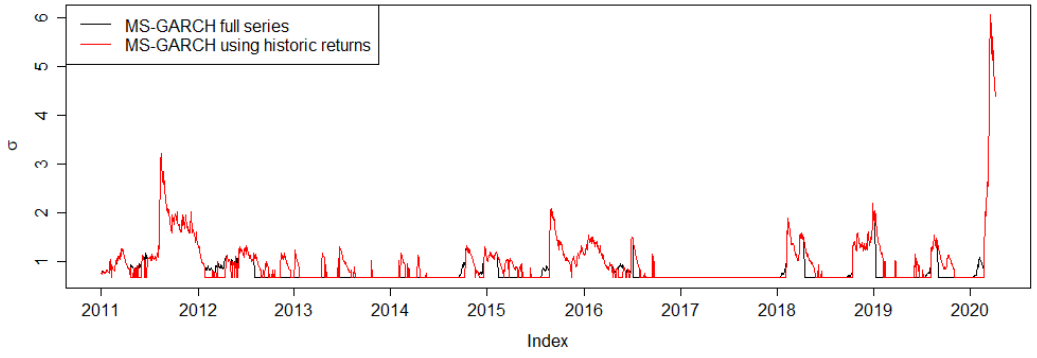


Figure 53: The conditional variance generated by the MS-GARCH model, assuming knowledge of the full time-series from 01-Jan-2011 to 05-Apr-2020 and assuming available knowledge are included.

These conditional variance estimates have also been used to estimate the VaR, which can be seen in Figure 54. It is difficult to tell if the estimates are better or worse than the previous ones, however they are definitely not as clean as the risk estimates generated using the full series.

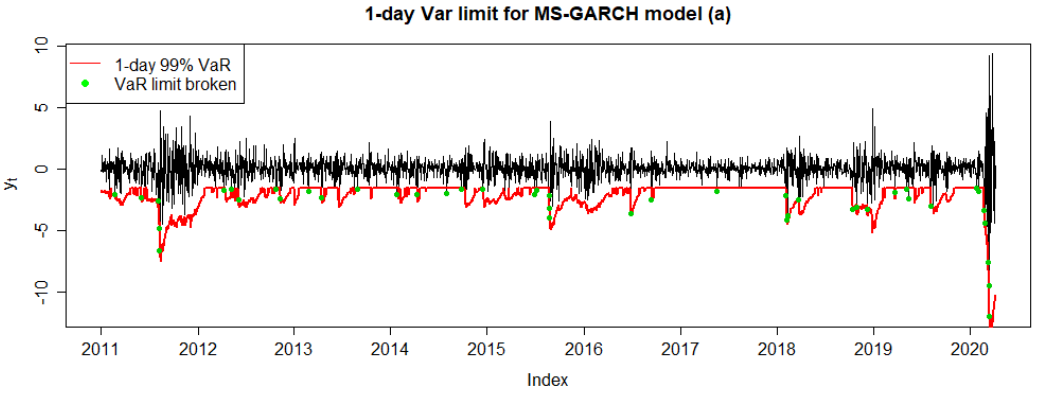


Figure 54: $VaR_{99\%}$ estimated calculated using the conditional variance from the MS-GARCH model assuming available knowledge.

In table 16 the number of times the $VaR_{99\%}$ estimates have been broken for each of the estimated models is shown.
Here the performance of the MS-GARCH model which is only conditioned on known information performs equivalently to the MS-GARCH model found using the full time series. However the VaR estimates were not as 'clean' in the low volatile periods.

	Expected	GARCH	MS-GARCH	
			Full	Limited
Number of broken VaR_{α} limits	23	51	45	45

Table 16: Number of times the estimated $VaR_{99\%}$ estimates were broken for the MS-GARCH and GARCH model. For the MS-GARCH model both the estimates assuming knowledge of the full time-series from 01-Jan-2011 to 05-Apr-2020 and only assuming available knowledge are included.

Using the *beta-prior model* found in section 9 it is possible to estimate the regimes such that the regimes are split up well. The value at risk estimates generated from this model is also quite good compared to a standard GARCH model, and the constant volatility in regime 1 does not seem to negatively affect the risk estimates.
However both models underestimate the risk, since the $VaR_{99\%}$ limit is broken more than would be expected.

The risk estimated by only conditioning on the known information, yielded a result, which did not split up the regimes very well, since the estimated risk process often switched to the high volatility regime after seeing a single large return. However the MS-GARCH model still does not perform worse that the standard GARCH model, and is therefore not specifically bad in risk modelling.

11 Conclussion

The main objective with this thesis was to implement a model, that can describe the volatility, of financial asset returns, by accounting for periods in which, the variance behave structurally different. We found that the GARCH model is able to describe for the heteroscedasticity, as well as some of the high kurtosis, seen in financial asset returns. We further discovered that regime shifting models, make it possible to account for periods with structurally different behaviour.
The combination of these models resulted in the MS-GARCH model, that include the desired attributes of both of the models.
The regime shifting capabilities of the MS-GARCH model were especially useful, as they allowed for the variance to be estimated differently in periods of high volatility, and periods of low volatility.

The estimation of the parameters in the MS-GARCH model, can not be done using Maximum likelihood, due to the path dependence, in the conditional variance process. Instead a Bayesian approach was chosen, and specifically a Gibbs sampling algorithm was used, because of the high dimensionality of the estimation procedure.
A few modifications had to be implemented in order for the estimation procedure to work accurately. These modification included a log transformation, of the conditional posterior distributions, which kept the computations numerically stable. Another modification was the implementation of an alternative numerical integration procedure, which made the integration accurate under any circumstance, with only a small computational demand.

Using the estimation procedure, the parameter set of the MS-GARCH model was estimated based on the asset returns of the S&P 500 index. The result of the estimation procedure showed that the variance of the financial asset returns, in financial crises, are described well with a GARCH model. However when the financial markets were not in a state of crises, the variance could be considered constant.
An issue encountered in the estimation was that the volatile, and stable periods were not completely separated, as a few low volatile periods were estimated as high volatile. In order to deter the estimation procedure from estimating periods of low volatility as high volatility regimes, a prior distributions was applied in the estimation of the transitions probabilities. It was found that a skewed beta distribution, would allow the model to converge, such that the high and low volatile regimes were completely separated.

The estimated model was then applied to a different period of the S&P 500 index, in order to test how well it could estimate the risk. Here we saw that the risk estimates generated by the MS-GARCH model, were slightly more accurate than the ones generated by a standard GARCH model. An issue was, that the estimated regime states for the MS-GARCH model, were based on the entire information set. The model was therefore estimated for each day individually, where the regime states were only conditioned on the known information to that day.
The risk estimated by only conditioning on the known information, yielded a result, which did not split up the regimes as well, since the estimated risk process often switched to the high volatility regime after seeing a single large return.

The MS-GARCH model is therefore difficult to use as a risk model, since it is not stable when modelling the variance of the present time.
However, it does work rather at modelling historical volatility, and the fact that it allows for low volatility periods to be modelled with a constant variance, is very interesting.

11.1 Discussion

In order to get a model which clearly separates regimes of high and low volatility, some assumptions were made regarding the prior distributions for the transition probabilities. While these prior assumptions are allowed when using a Bayesian approach for the estimation, it could be argued that they were too extreme and the impact on the posterior distribution was therefore too large.
An alternative approach could have been to include a third regime, which could possibly help explain periods that were too extreme to be classified as regime one, but not extreme enough to be classified as regime 2. However since we decided to only focus on models with two regime, the prior assumptions was made in order clearly separate the regimes.

The model estimated in Section 9 was based on a limited section of the S&P 500 index. If another part of the index would have been chosen, the estimated model might have been different. Especially the transition probabilities could have been affected much by the choice of interval, given that the number of regime changes were very low.
An alternative approach could have been to fit the model on the full data-set for each estimation. The issue with this approach is that the computational demand is very high, and limiting the interval was therefore necessary.

The regime states found when only conditioning on the known information, available at that time, were often estimated to be in regime 2, as shown in section 10. The process was therefore modelled mainly as a standard GARCH model.
The reason why this becomes an issue, is due to the fact the the GARCH model does an exceptional job at modelling the volatility of financial asset returns. Since the GARCH model describes both periods of high and low volatility well, regime 1 in the MS-GARCH model becomes less important.
Because of this it could be questioned whether the MS-GARCH model introduce a high level of complexity, without much gain. Since the volatility modelling closely resembles that of a standard GARCH model, it could be argued that the regime shifting capabilities of the MS-GARCH model are unnecessary in a risk modelling framework.
However the MS-GARCH model still does not perform worse that the standard GARCH model, and is therefore not specifically bad in risk modelling. Furthermore the MS-GARCH model allows for periods of low volatility to be modelled using a constant variance, which could be interesting to consider when modelling risk.

11.2 Future work

Throughout the thesis, the noise term of the MS-GARCH model was defined as draws from a standard normal distribution. However the residuals generated with the estimated MS-GARCH model, on the empirical data, showed signs of heavy tails. A suggestion can therefore be to use a more heavy tailed distribution, Such as the *t-distribution*, as the noise terms of the MS-GARCH model. Doing so might remove some of the heavy tails seen in the residuals, which could possibly make the MS-GARCH model even more accurate in capturing the more extreme returns. Another aspect of the analysis which

was kept constant throughout the thesis, was the number of regimes in the MS-GARCH model. Since only two regimes were investigated, the analysis was limited to focus on regime states of financial crisis, and regime states of financial stability. For future work it might be interesting to consider more periods for the volatility process. By including more regimes in the MS-GARCH model it could perhaps be useful explaining different business cycles. Lastly it could be interesting to consider a combination of a threshold

model and a MS-GARCH model. When exploring performance of the MS-GARCH model in a risk framework, we experienced a difficulty in predicting regimes, when no information regarding the future is available. A combination of a threshold model and a MS-GARCH model might be useful in detecting regimes changes more accurately.

References

Luc Bauwens, Arie Preminger, and Jeroen V. K. Rombouts. Theory and inference for a markov switching garch model. *Econometrics Journal*, 13:218–244, 2010.

Tim Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, pages 307–327, 1986.

Chris Brooks. A double-threshold garch model for the french francdeutschmark exchange rate. *Journal of Forecasting*, 20:135–143, 2001.

Jun Cai. Markov model of unconditional variance in arch. *Journal of Business and Economics Statistics*, 12:307–333, 1994.

Robert F. Engel. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica*, 50:987–1007, 1982.

Yahoo Finance. https://finance.yahoo.com/. Accessed: 04-06-2020.

Christian Francq and Jean-Michel Zakoïan. *GARCH Models Structure, Statistical Inference and Financial Applications*, volume 2. John Wiley & Sons Inc, 2010.

Dani Gamerman and Hedibert F. Lopes. *Markov Chain Monte Carlo*. Chapman & HallCRC, 2006.

Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian data analysis*. CRC Press, 2014.

Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Xplore*, 1984.

Stephen M. Goldfeld and Richard E. Quandt. A markov model for switching regressions. *Journal of Econometrics*, pages 3–16, 1973.

Stephen F. Gray. Modeling the conditional distribution of interest rates as a regime-switching process. *Journal of Financial Economic*, pages 27–62, 1996.

Ute Hahn. Markov chain monte carlo methods, 2013-2014.

James D. Hamilton. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57:357–384, 1989.

James D. Hamilton. *Time series analysis*. Princeton University Press, 1994.

James D. Hamilton and Raul Susmel. Autoregressive conditional heteroskedasticity and changes in regime. *Journal of Econometrics*, 64:307–333, 1994.

Wilfred Keith Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 1970.

Daniel F. Waggoner James D. Hamilton and Tao Zha. Normalization in econometrics. *Econometric reviews*, 26:221–252, 2007.

Gregory F. Lawler. *Introduction to Stochastic Processes*, volume 2. Chapman & HallCRC, 2006.

Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 1953.

Petra Posedel. Properties and estimation of garch(1,1) model. *Metodoloski zvezki*, 2: 243–257, 2005.

Christian Ritter and Martin A. Tanner. Facilitating the gibbs sampler: The gibbs stopper and the griddy-gibbs sampler. *Taylor & Francis*, page 861–868, 1992.

David Ruppert and David S. Matteson. *Statistics and Data Analysis for Financial Engineering*, volume 2. Springer, 2015.

Jan R. M. Röman. *Analytical finance*, volume 1. Palgrave macmillan, 2017.

Howell Tong. On a threshold model. in pattern recognition and signal processing. *Sijthoff & Noordhoff*, page 101–141, 1978.

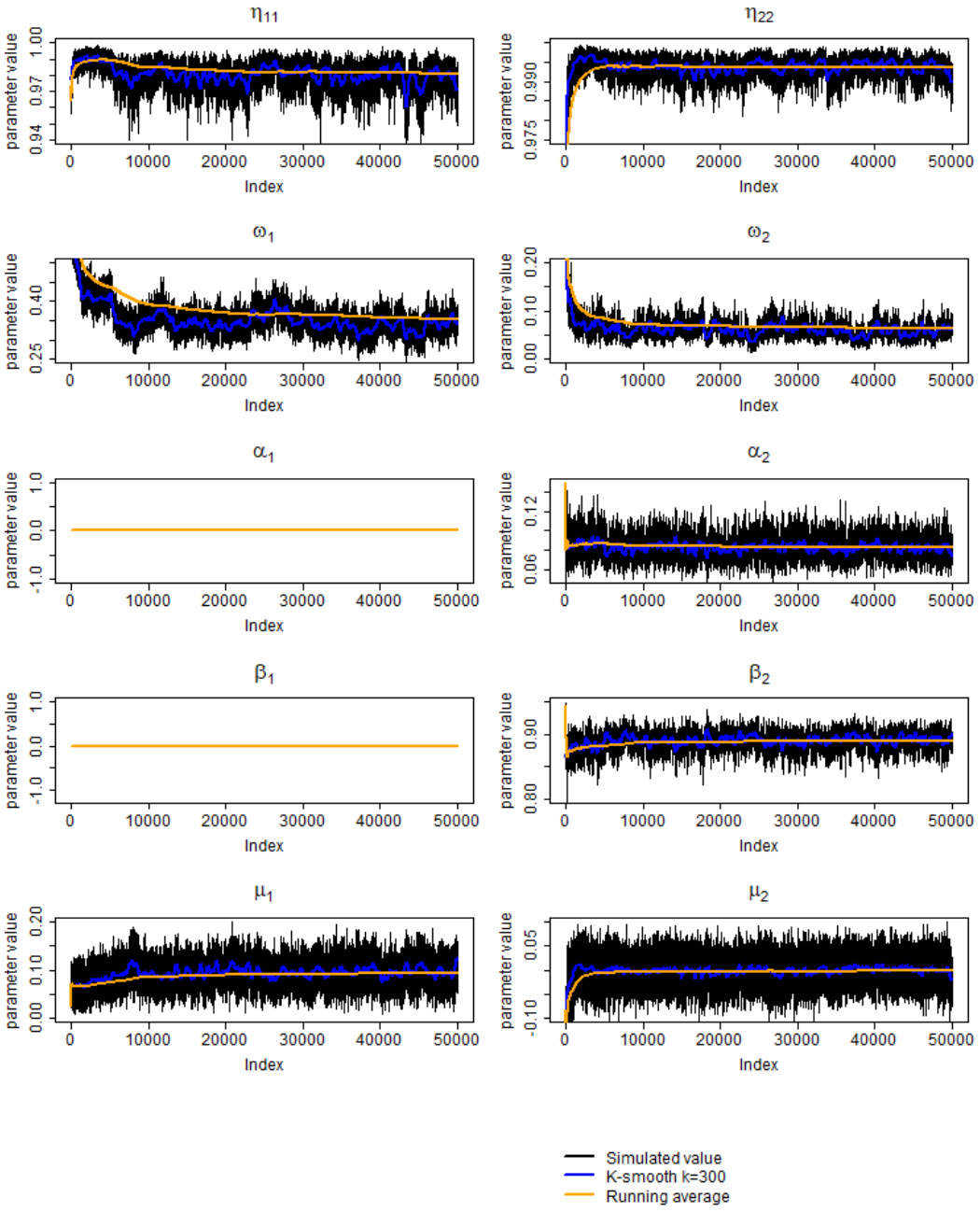
Reuy S. Tsay. *Analysis of financial time series*. Wiley Interscience, 2005.

Zheng Rong Yang. *Machine Learning Approaches To Bioinformatics*. World Scientific Publishing Co. Pte. Ltd, 2010.

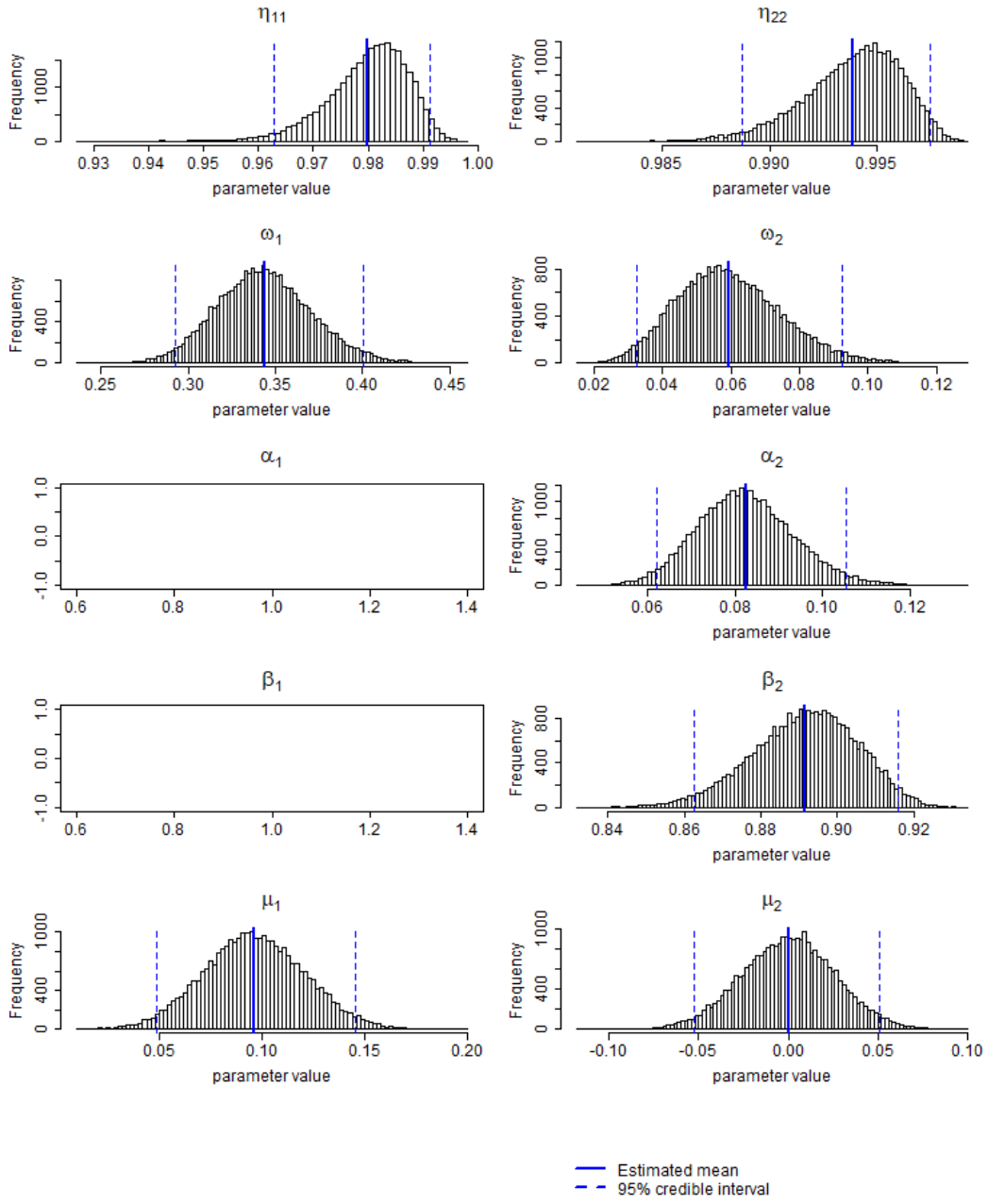
Appendices

A Estimation: regime 1 with constant volatility

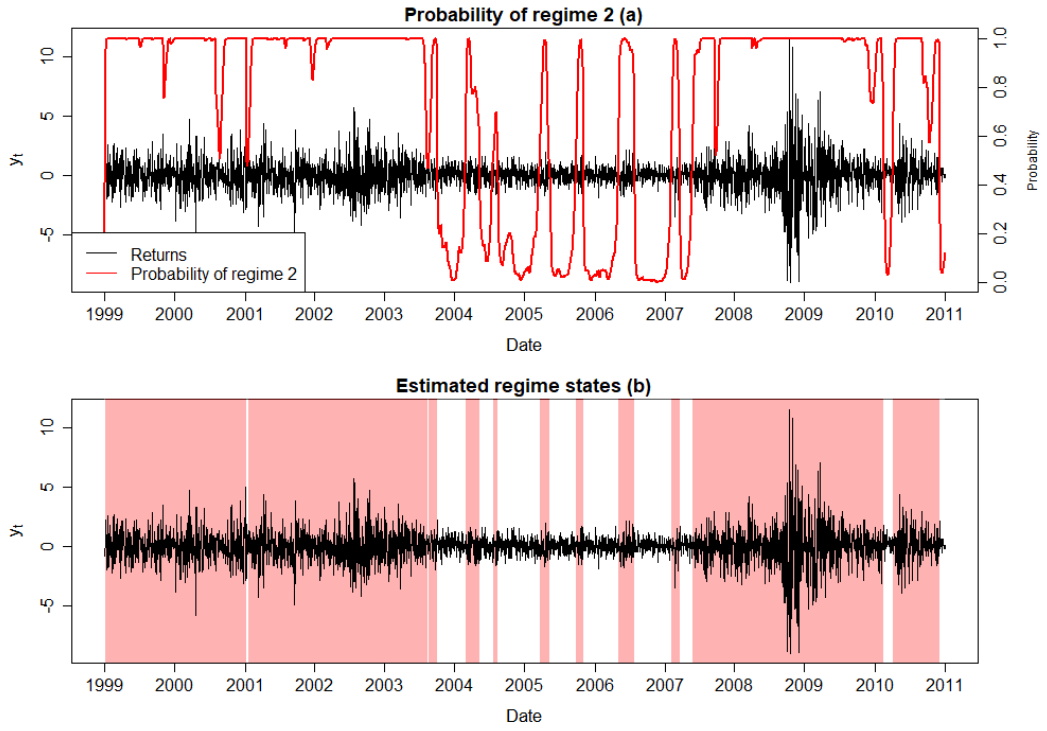
A.1 Trace-plot



A.2 Histogram of posterior distributions with a burn-in of 20.000

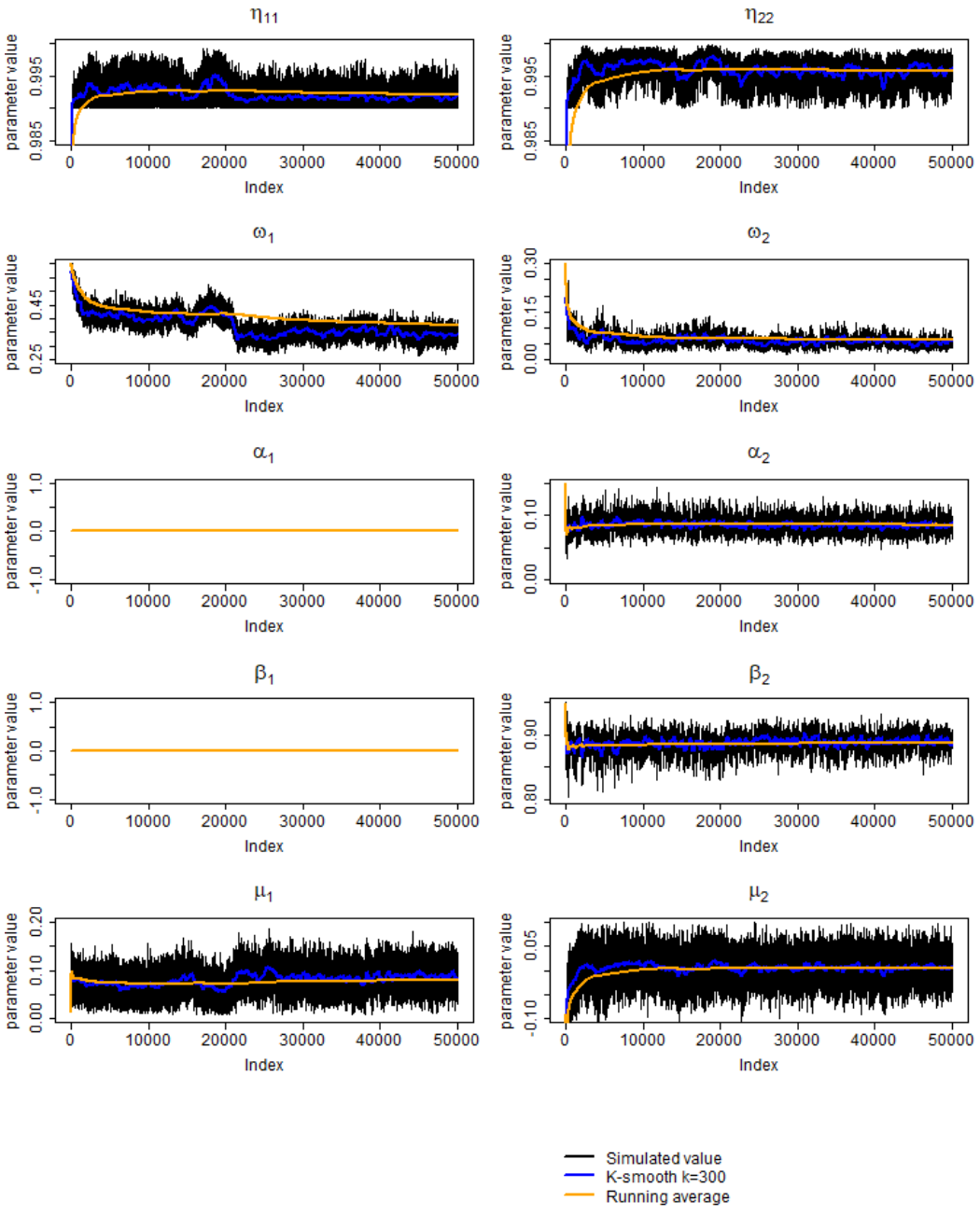


A.3 State probabilities and estimated states

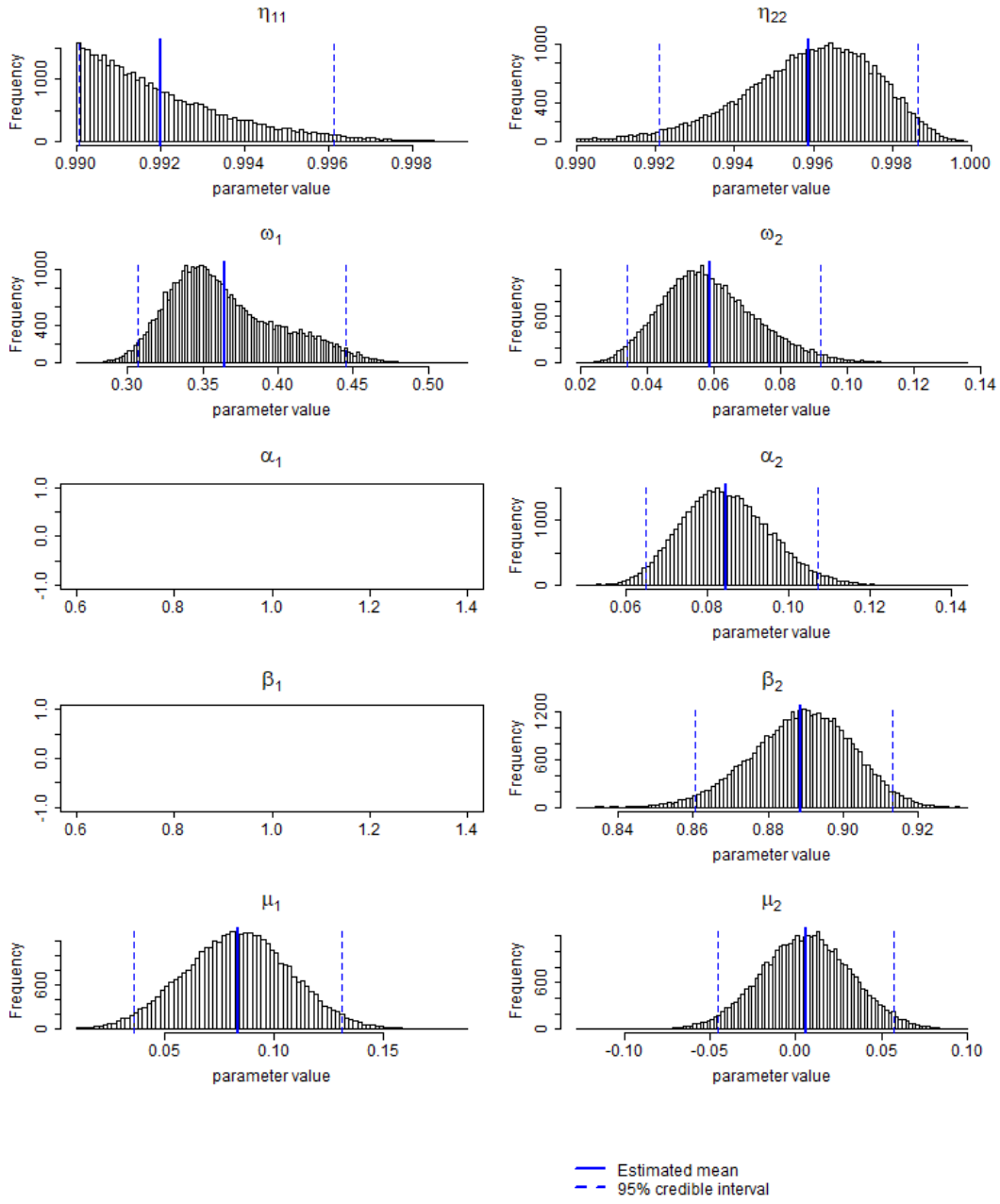


B Estimation: Truncated distribution for transition probabilities

B.1 Trace-plot

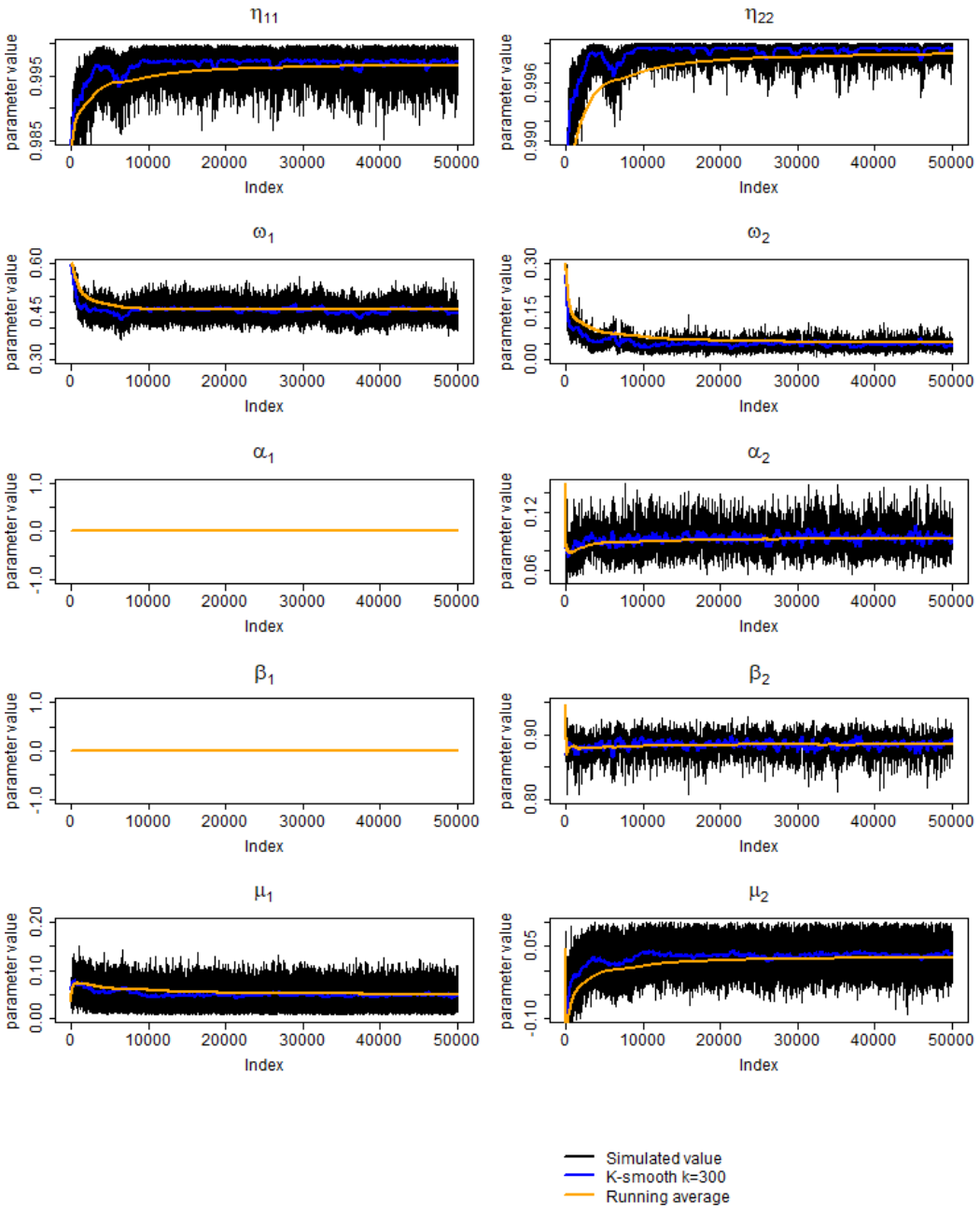


B.2 Histogram of posterior distributions with a burn-in of 20.000



C Estimation: Skewed beta distribution for transition probabilities

C.1 Trace-plot



C.2 Histogram of posterior distributions with a burn-in of 20.000