Markov-switching GARCH models with application to insurance claims

Eivind Bjørnøy

June, 2020

Department of Mathematics
University of Bergen

Master of Science in Actuarial Science

Abstract

Attempting to model insurance claim data is usually done through fitting the data to a parametric distribution, irregardless of the time in which the claims occur. We attempt to view insurance claim data as a time series, and subsequently fit Markov-switching GARCH-models on the data. The methods we consider in this thesis are applied to a well-known insurance dataset through the use of the R-package `MSGARCH`(Ardia, et al., 2019)[7]. The possible model specifications, and the applicability of the models to the data are discussed. We compare the fitted models to some parametric distribution models suggested in a paper by Eling (2012)[21] on the same dataset. We also consider some tail-risk measures, and the practical evaluation of the risk measures Value-at-Risk and Expected Shortfall for the data at hand.

Acknowledgements

Firstly, I would like to reach out a thanks to my supervisor, prof. Antonello Maruotti, for fueling my inspiration, helping with choosing the topic of my thesis and bringing some valuable guidance to the process.
I would also like to thank my fantastic family and friends who have provided me with great support and encouragement, which has immensely helped me to be able to finish this thesis.
Lastly, a special thanks to the lovely Annelise Elde for being such a great comfort through the entire process of writing the thesis.

Contents

1 Introduction

The foremost objective of this thesis is to attempt to answer the question, "*Is it a good idea to view insurance claims data as a time series, and perform model estimation based on this?*" Usually, an actuary will group individual losses by size of loss and then fit a continuous positive distribution to the data of all the losses (Hewitt & Lefkowitz, 1979)[37]. So, when modelling insurance losses/claims, the preferred method it to simply fit the data to some distribution that captures the stylized effects of insurance claims best, as in Eling (2012) and Vernic (2006)[21][51]. Some of the more widespread distributions which usually perform model estimation well are the log-normal-, Gamma- and Weibull-distributions. In general, since insurance data usually has a very heavy right-tail, distributions that take this detail into consideration usually outperform the models which do not. We see the opportunity to explore model fitting which takes into consideration the time-aspect of some insurance data, as well as expanding from the single-distribution fits of general literature to a higher-order, several-state model.
As the distribution of the data in the tail is such an important feature of insurance data, this thesis has a particular focus on fitting models which are able to describe the behavior of the tail most accurately.
A method which lets us explore these opportunities is the Markov-switching GARCH-model specification by Haas (2004)[32], where we decide to treat insurance claims data as a time series that is allowed to exhibit autoregressive conditional heteroskedasticity (ARCH)-effects. This method also opens up the possibility of assuming that the data originates from several *regimes*, where the model is allowed to behave differently in each of the regimes.
We apply one- and two-regime Markov-switching GARCH models through the R-package `MSGARCH`(Ardia et. al, 2019)[7] on both the original and the log of a well-known dataset of Danish fire-reinsurance. We find that the models which were generated with a second regime describes the tail of the original insurance data quite well compared to the simple distribution fits of Eling (2012)[21] on the same data. However, we find that the general model fits of the Markov-switching models are competitive, but not better than the best simple distribution fit of Eling (2012). Another finding is that several of the MSGARCH-models actually outperform all of the benchmark models when applied to the log of the insurance data, both in general model fit and for describing the tail risk.

The thesis is constructed as follows: Section (2) contains a general discussion on time series and their properties. Section (2.2) introduces a variety of different well-known time series model and their applicability to different data. Section (2.3) contains a discussion on the differences between a general financial returns time series and an insurance time series, as well as the eligibility of applying some time

series models to an insurance loss dataset.
In Section (3), the autoregressive conditional heteroskedasticity(ARCH)-models are presented and discussed. Firstly, in section (3.1), the ARCH model is presented, and some of its properties are shown. Section (3.2) introduces the GARCH-model, and how it differs from, and compares to the ARCH-model. This section also presents the different methods of specifying the structure of the conditional volatility in a GARCH-model, as well as explaining the possibilities of letting the innovations follow some conditional distribution.
Section (4) finishes the groundwork for MSGARCH-models, as it contains discussion around Markov models. The section contains several Markov-models and their properties. Section (4.1) contains a discussion on the most simple Markov model, i.e. the Markov chain, often used as an underlying *hidden* force to drive some other process. Section (4.2) extends the discussion the the hidden Markov model (HMM), and section (4.3) is about the Markov-switching model.
As sections 2 - 4 work as building blocks, section (5) is the culmination of these building blocks into the Markov-switching GARCH-model. The section explains the possible specifications of the model, as well as builds the likelihood function, and explains the procedure behind performing ML-estimation. Section (5.1) contains some general attributes of the MSGARCH-model, and section (5.2) constructs the likelihood function and explains the numerical methods used to acquire the starting values of the ML-estimation. Section (5.3) describes the numerical process behind the actual ML-estimation. Section (5.4) briefly introduces the model comparison criterion AIC, and section (5.5) explains the h-step ahead prediction of a fitted model.
Section (6) contains a summary of different risk measures, their properties and their application. Firstly, section (6.1) and section (6.2) contains explanations of some desirable properties of a risk measure, i.e. coherency and elicitability. Section (6.3) introduces some univariate risk measures, and section (6.4) introduces some multivariate risk measures.
Section (7) marks the start of the empirical section of the thesis, and includes a presentation of the dataset used, the R-package used and the estimation-process. Section (7.1) presents the dataset, discusses the applicability of MSGARCH-models on said dataset, and makes some assumptions on how the models will perform. Section (7.2) is an introduction of the R-package `MSGARCH`, which is the main package used in the empirical analysis of the thesis. The functions of the package which are mainly used are also presented. Lastly, section (7.3) contains the estimation and prediction-process for each model, exemplified by four different specifications which yield different results. The section ends with a presentation of the results.

2 Time Series

Time series analysis has been a topic of much debate and vigorous research throughout many decades, with applications within areas such as economics, finance, medicine and insurance. The challenge with time series has mostly been to draw inferences from said time series, understand the way they evolve over time, and being able to recreate or replicate the movement in some kind of stochastic model. The main objective in time series analysis is, in simple terms, to set up a hypothetical model that is meant to represent the given data, estimate the parameters of said model, and hopefully use the fitted model to better understand the data. Another important objective in this field is prediction ahead of data. Researchers have been using time series analysis as an essential tool in attempting to forecast values such as market indices, financial asset prices, hospitalization numbers, monetary insurance losses from automobile crashes, or something as tangible as the weather. In short, a time series is a sequence of stochastic variables $\{Y_t\}_{t=1}^{T}$ which appear in chronological order from earliest to latest. Usually, a time series is evaluated at regular time intervals, but irregular time intervals are also possible, although it could alter the interpretation of the results.
In this thesis, we will attempt to model and forecast insurance losses through time series analysis instead of disregarding the time aspect and applying regular regression analysis. In general literature, time series analysis is seldom used in order to fit models that deal with insurance losses, so this thesis attempts to give a fairly unique perspective on insurance time series. To highlight this point, Figure (1) below shows examples of two time series that look and behave differently, the first (a) being losses in an insurance dataset, the second (b) being log-returns of a the Swiss market index (SMI), which is defined as $y_t = log\left(\frac{\pi_t}{\pi_{t-1}}\right)$, where π_t is the price of a financial asset.

The first thing we notice is that the plot of the SMI seem to be of a more symmetrical build, with unconditional mean close to zero, while the insurance data is absolutely positive, with some very high values scattered through seemingly randomly, and a positive unconditional mean. How this affects the applicability of some time series models on the insurance data will be a topic in this section. We will discuss some fundamental ideas of time series data, and also discern the difference in application of methodology between a financial time series and an insurance time series.

2.1 Fundamentals of Time Series

In order to develop models for time series, we summarize some of the fundamentals. Let $\{Y_t\}$ be a time series defined on $t = 1, 2, \ldots, T$, which has finite squared

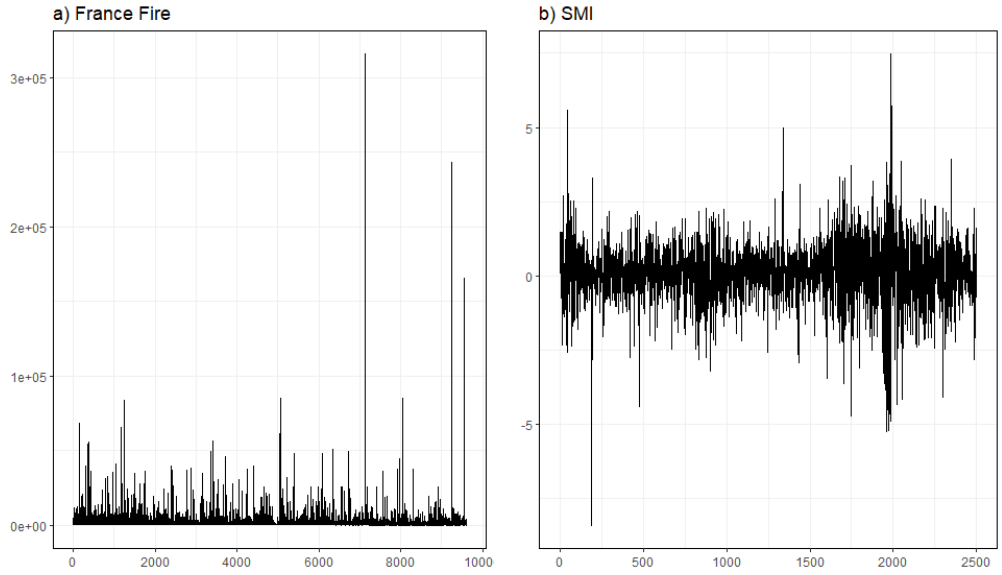


Figure 1: a) Commercial fire losses over the period 1982 to 1996 from the French Insurance Federation. b) Daily log-returns of the Swiss Market Index(SMI) over the period 1990 to 2000.

expectation $\mathbb{E}(Y_t^2) < \infty$. The moments of a time series is presented next.

2.1.1 Moments of a Time Series

The mean (2.1), variance (2.2), autocovariance(2.3) and autocorrelation (2.4)-functions of $\{Y_t\}$ at time t is given as follows (Brockwell & Davis, 2016, section 1.4) [13]:

$$\mu_t = \mathbb{E}(Y_t) \tag{2.1}$$

$$\sigma_t^2 = \mathbb{E}\left((Y_t - \mu_t)^2\right) \tag{2.2}$$

$$\text{cov}(Y_{t+k}Y_t) = \mathbb{E}\left((Y_{t+k} - \mu_{t+k})(Y_t - \mu_t)\right) \tag{2.3}$$

$$\text{corr}(Y_{t+k}Y_t) = \frac{\text{cov}(Y_{t+k}Y_t)}{\text{cov}(Y_tY_t)} \tag{2.4}$$

Autocovariance and autocorrelation are measures of serial correlation that explain how a time series observations depend on each of its previous values, called its *lagged* values. A significant ACF-value for a given lag k implies that the time series observation at time $t + k$ is dependent on its k-lagged value t.

2.1.2 Stationarity

An important assumption in the application of many statistical models is that the time series is *stationary*. In general, a time series is stationary if it has similar statistical properties at time t, Y_t as with the "time-shifted" series at time $t + h$, Y_{t+h} for all integers h. More formally, a time series is (weakly) stationary if μ_t is

independent of t and $\text{cov}(Y_{t+k}Y_t)$ is independent of t for each k. This type of weak stationarity that only considers independence from t in the first two moments is also called *covariance stationarity*, or *second-order stationarity*. It is not uncommon for a model to have stationarity *ranges* for its parameters, which means that the parameters need to be within a specific range in order for the model to ensure covariance stationary. *Strict stationarity* of a time series happens when $(Y_1, Y_2, \ldots, Y_n)$ and $(Y_{1+k}, Y_{2+k}, \ldots, Y_{n+k})$ have the same joint distribution for all integers k and $n > 0$ (Brockwell & Davis, 2016, section 1.4)[13]. As long as $\mathbb{E}(Y_t^2) < 0$, strictly stationary time series are also weakly stationary, however the opposite is not necessarily true. Non-stationarity in the second order is found in time series that seem to have a mean or covariance that is affected by the time of previous lags of the time series. A process would be non-stationary if there existed e.g. a trend or a constant periodic influence. As stationarity is a desirable trait in a time series, it is not uncommon to attempt to remove effects that disturb stationarity, e.g. by removing trend or seasonality. Figure (2) is an example of a non-stationary time series where a clear trend and what seems to be a seasonal effect appears.

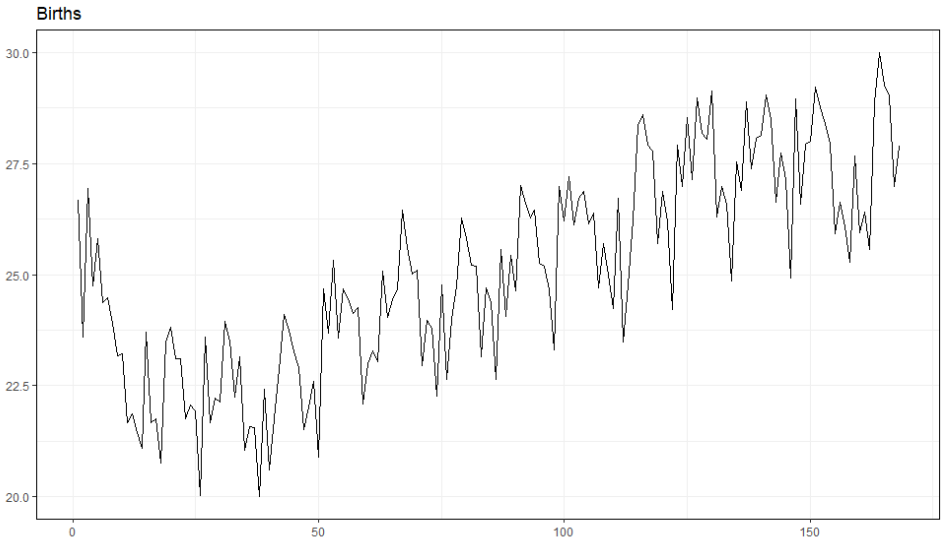


Figure 2: Example of a non-stationary time series. Amount of births in New York City from January 1946 to December 1959.

In this thesis, we will not be dealing with time series that exhibit any significant trend nor seasonality. Nonetheless, ensuring stationarity is an important requirement for accuracy in most time series models.

2.2 Time Series Models

In Brockwell & Davis (2016)[13], a time series model is defined as follows:

"A **time series model** for the observed data $\{x_t\}$ is a specification

of the joint distributions (or possibly only the means and covariances) of a sequence of random variables $\{X_t\}$ of which $\{x_t\}$ is postulated to be a realization."

2.2.1 Some simple models

Some time series models will be presented here, starting with the simplest one, *I.I.D noise* (independent, identically distributed noise). This model is simply a specification where there is no trend or seasonality, with the observations $\{Y_t\}$ being random variables that are independent and identically distributed with zero mean and no dependence between observations at all. This model is not as interesting in an of itself, but it plays an important role in more advanced models. I.I.D. noise is denoted here as

$$\{Y_t\} \sim \text{I.I.D.}(0, \sigma^2) \tag{2.5}$$

If the observations are uncorrelated variables with zero mean as defined above, but they are not independent, the time series is called *white noise*. Both white noise and I.I.D. noise are stationary time series models because they, by definition, keep a constant mean and exhibits no serial correlation.

The *random walk* is another simple time series model which is characterized with the process taking "steps" randomly towards the positive or negative direction. A random walk with zero mean is simply the cumulative sum of I.I.D. random variables with zero mean. Although a random walk has a constant zero mean, it is actually not weakly stationary because the covariance is time dependent. A one-step-ahead forecast of a zero mean random walk is therefore defined as $Y_t = Y_{t-1} + \eta_t$, where $\eta_t \sim \text{I.I.D.}(0, \sigma^2)$

Similarly to random walks, we also define *random walk with drift*, which is exactly the same as a random walk, but with a constant term a pushing the process in either direction for each observation. For $a > 0$ the process will have an upward trend, and for $a < 0$ the process will have a downward trend. For $a = 0$, it is simply a random walk. This process obviously does not have a zero mean, nor is it time-independent, therefore it is not stationary. A one-step-ahead forecast of a random walk with drift is defined as $Y_t = Y_{t-1} + a + \epsilon_t$. Fama (1965)[24] argues that the price of a financial asset follows this exact model. In this case, the price is given by

$$\pi_t = \pi_{t-1} + \mu_t + \epsilon_t. \tag{2.6}$$

Here, π_t is the price of the financial asset at time t, μ_t is the conditional mean of y_t, and also the drift term. ϵ_t is the random innovation-term with $\mathbb{E}(\epsilon_t) = 0$ and

$\mathbb{E}(\epsilon_t \epsilon_{t-\tau}) = 0$ for $\tau \neq 0$. Financial returns at time t are defined as the difference in the price of the financial asset from time t to time $t-1$,

$$x_t = \pi_t - \pi_{t-1}, \tag{2.7}$$

which in turn gives us another expression for the returns series x_t:

$$x_t = \mu_t + \epsilon_t. \tag{2.8}$$

Modeling log-returns plays a very important role in computational finance, as it allows for direct examination of the conditional variance. This is simply given by:

$$y_t := \log(x_t) = \log\left(\frac{\pi_t}{\pi_{t-1}}\right) \tag{2.9}$$

These processes will be much more important later on, as they have some interesting properties.

2.2.2 Conditional structure models

When we want to impose a specific conditional structure on the mean or the variance of a time series, we introduce *ARMA*-models and *(G)ARCH*-models. In general, ARMA-models are used when modelling realizations of a random process when one wishes to impose structure on the conditional **mean** of a process, while (G)ARCH-models are used when modelling realizations of a random process when one wishes to impose structure on the conditional **variance**. ARMA (Autoregressive Moving Average) are models used to provide a description of weakly stationary stochastic processes using two polynomials. These polynomials are the autoregressive(AR), and the moving-average(MA), both of whom are specified by the number of lags for which their observations are dependent on (Brockwell & David, 2016, chapter 2)[13]. In order to get an idea of which model to should be used, we look at sample ACF and sample PACF of the observed data.

We define the sample autocovariance function (sample ACVF, 2.10) and sample autocorrelation function (sample ACF, 2.11) as follows:

$$\widehat{\text{cov}}(Y_{t+k}Y_t) = n^{-1}\sum_{t=1}^{n-|k|}(y_{t+|k|} - \bar{y})(y_t - \bar{y}), \quad -n < k < n \tag{2.10}$$

$$\widehat{\text{corr}}(Y_{t+k}Y_t) = \frac{\widehat{\text{cov}}(Y_{t+k}Y_t)}{\widehat{\text{cov}}(Y_tY_t)} \tag{2.11}$$

The sample partial autocorrelation function (sample PACF, 2.12) is defined by the equations

$$\hat{\alpha}(0) = 1$$

$$\hat{\alpha}(k) = \hat{\phi}_{kk},\ k \geq 0 \tag{2.12}$$

where $\hat{\phi}_{kk}$ is the last component of

$$\hat{\boldsymbol{\phi}}_{\boldsymbol{k}} = \hat{\boldsymbol{\Gamma}}_k^{-1} \widehat{\mathbf{cov}}(\mathbf{Y_{t+k}} \mathbf{Y_t}) \tag{2.13}$$

where $\hat{\boldsymbol{\Gamma}}_k^{-1} = [\widehat{\text{cov}}(Y_{t+i-j} Y_t)]_{i,j=1}^{k}$ is the sample covariance matrix, and $\widehat{\mathbf{cov}}(\mathbf{Y_{t+k}} \mathbf{Y_t}) = [\widehat{\text{cov}}(Y_{t+1} Y_t), \widehat{\text{cov}}(Y_{t+2} Y_t), \ldots, \widehat{\text{cov}}(Y_{t+k} Y_t)]^\intercal$ (Brockwell & Davis, 2016, section 3.2) [13].

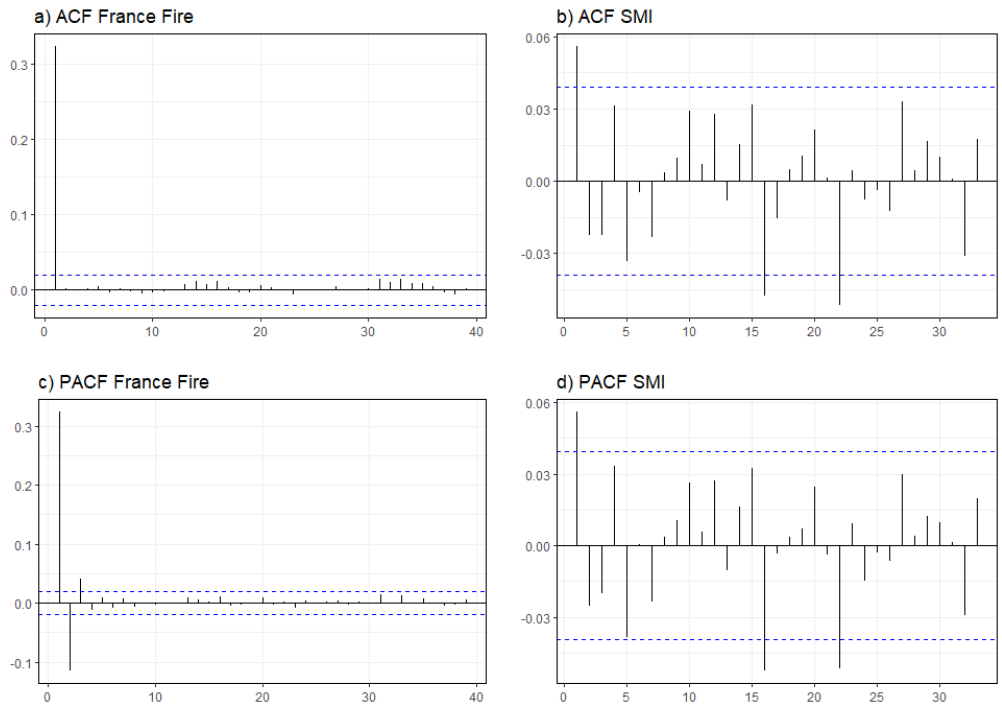


Figure 3: a, c) Sample ACF and PACF of commercial fire losses over the period 1982 to 1996 from the French Insurance Federation. b, d) Sample ACF and PACF of the daily log-returns of the Swiss Market Index(SMI). Inside the blue dotted lines indicates the area of non-significance.

Figure (3) shows the sample-ACF and sample-PACF of the examples from Figure (1). Here, we notice in a) and c) that the the insurance data has a significant ($\sim$0.3) autocorrelation with its lag-1-value, while it has significant ($\sim$0.3 and ~ -0.1) partial autocorrelation-values for both lag-1 and lag-2 that seems to quickly descent geometrically as the lags increase. This could be an indicator that the insurance dataset could be fitted with an MA(1)-model (Cryer & Chan, 2008, chapter 6)[19]. Further, we notice that the corresponding ACF and PACF-values for the SMI shows little to no significance, with the significant values only barely being so. This is expected for a financial log-returns time series, as they are notoriously difficult to predict. ACF and PACF-values similar to b) and d) can often be mistaken for a white noise process because the process seems entirely random when viewing just

these values. However, when viewing the plot of the SMI from Figure (1) we do not believe that the data can be explained by an I.I.D. noise process, since we notice that there appears to be periods where the volatility is higher than others. This phenomenon is called *volatility clustering*, and is indicative of serial correlation in the squared or absolute values of the time series. When squaring the time series of the SMI, we obtain the ability to only observe the severity of the returns, and to disregard the direction. In Figure (4), we plot the ACF and PACF of both our time series for the squared losses/returns. In b) and d), we notice that there is a quite significant serial correlation in the SMI, while the fire data a) and c) barely differs, except for the higher lags being even less significant. In fact, it is quite nonsensical to consider the squared fire losses because this data is absolute positive, as the severity of the claim is already the sole focus.

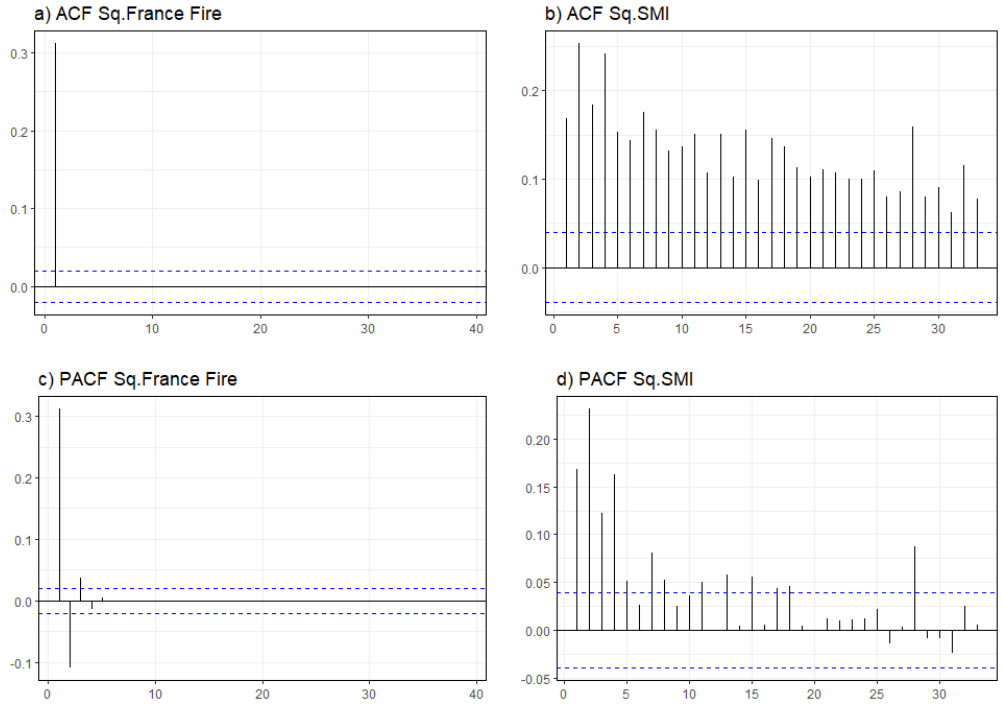


Figure 4: a, c) Sample ACF and PACF of squared commercial fire losses over the period 1982 to 1996 from the French Insurance Federation. b, d) Sample ACF and PACF of the squared daily log-returns of the Swiss Market Index(SMI). Inside the blue dotted lines indicates the area of non-significance.

Since the squared returns admits some significant autocorrelation, this autocorrelation gives strong evidence against the returns being independent and identically distributed. In fact, they show evidence for *autoregressive conditional heteroskedasticity* in the variance structure (Cryer & Chan, 2008, chapter 12)[19], which will be reviewed in section (3). Another useful test for the appearance of conditional heteroskedasticity is called the McLeod-Li test (McLeod & Li, 1983) [44]. The McLeod-Li test is a portmanteau test, which utilizes the characteristic that the sum

of squares of autocorrelations is approximated by the chi-squared distribution with K degrees of freedom if residuals are a realization of a I.I.D. or white-noise sequence (Shin, 2017)[48]. It is given by a statistic

$$Q^*_{ML}(K) = n(n+2)\sum_{k=1}^{K}\left\{\frac{\hat{\rho}^2_{yy}}{(n-t)}\right\}, \tag{2.14}$$

where $\hat{\rho}_{yy}$ is the sample autocorrelation function of the *squared* values of our observed value Y_t, K is a lagged value and n is the number of observations. We compare $Q^*_{ML}(K)$ with $\chi^2_{1-\alpha}(K)$, which is the $(1-\alpha)$-percentile of the chi-squared distribution with K degrees of freedom. If $Q^*_{ML}(K)$ exceeds $\chi^2_{1-\alpha}(K)$ for given lag K, we reject the hypothesis that the autocorrelation at given lag is equal to zero.

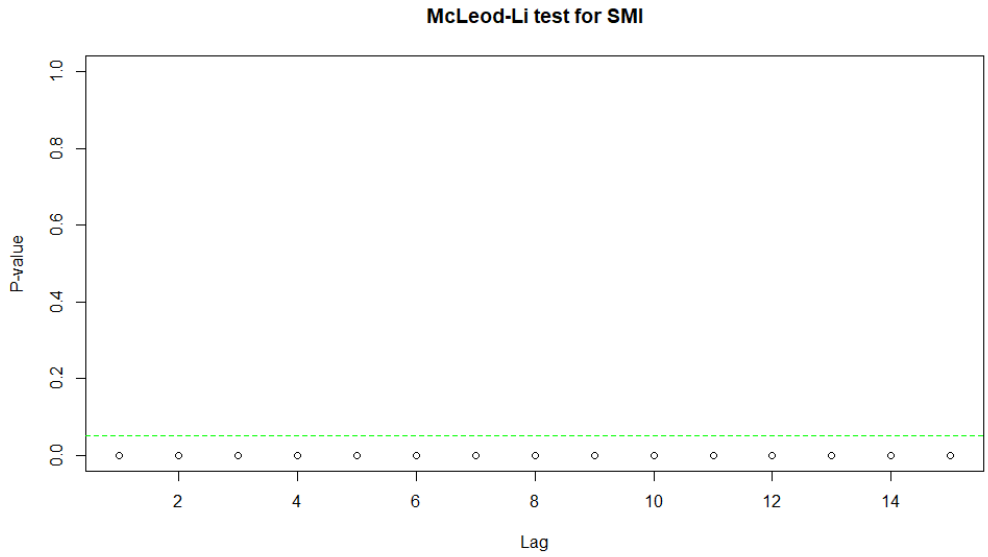


Figure 5: McLeod-Li test statistic for the Swiss Market Index (SMI). Above the red line indicates the area of hypothesis rejection

Figure (5) shows plotted SMI values of the McLeod-Li statistic for max lag = 15, with rejection area above the green dotted line. Here we obviously see that the McLeod-Li test cannot reject the alternative hypothesis, which is a good indicator that a model which takes conditional heteroskedasticity into account is suitable in this situation.

2.3 Insurance loss time series

The objective of this thesis is to model insurance loss data through time series analysis, and more specifically through Markov-switching GARCH models. This method is most commonly used on financial time series, so in this section we will discuss the applicability of this method to insurance data.
As we have discussed, the returns of a financial time series can be expressed as in

Equation 2.8. Harvey & Fernandes (1989)[36] find that insurance claims can be expressed in the same manner, that being a simple structural time series model that consists of the μ_t-term and the random disturbance term ϵ_t.

Previously, we have seen that the example fire dataset from France exhibits serial correlation in the first order, which is something that implies we should at the very least use an ARMA-model to model the time series. In addition to this, the application of (G)ARCH-methodology is most reasonable when applied to data that exhibits conditional variance-heteroskedasticity, and is simpler to implement for data that has conditional mean zero (Engle, 1982)[23]. In fact, much of the computational software that can model GARCH-effects assumes a unconditional mean that is equal to zero. The problem arises when this is not the case, as we see by the significant ACF and PACF-values from the fire insurance dataset in Figure (3). In this case, we would fit a ARMA-model to the data and use the characteristic that the residuals of this model should be without autocorrelation in the lags, and therefore have conditional mean zero. We then continue to deal with ARCH-properties on those residuals instead of on the observed time series (Haas, Mittnik & Paolella, 2004) [32]. If the time series seems to have a constant mean (i.e. the ACF and PACF-values are insignificant), testing and model fitting can be applied to the time series in excess of the sample mean. We define $y_t := x_t - \bar{x}$ as the de-meaned time series, where $\{X_t\}$ is the original time series which has sample mean $\bar{x}$. In this case, Equation (2.8) simply becomes

$$y_t = \epsilon_t. \tag{2.15}$$

When eliminating the conditional mean from the expression, we are actually left with only the random innovations term ϵ_t, and thus the observations are determined by how the random innovations behave. This is actually the case for the insurance time series that will be taken into question in section (7), and throughout the rest of this thesis we will assume that we are in the zero-mean case.

Another challenge when dealing with insurance loss data is that it is not uncommon for very large shocks to appear, as it is a known property of insurance data that they tend to have have a heavy tail. Because of this, significant previous lags before those shocks can end up being less significant than they should be. In order for the conditional heteroskedastic effects to be more pronounced and the differences in the data to be smaller, we can do model fitting on the log of the time series in addition to on the original data. Estimation will be done on both the original data as well as the log data in section (7).

Insurance time series can be considered special in the way they tend to have a

significant *skewness* and *kurtosis* in comparison to a normal distribution.

In order to quantify the skewness and kurtosis of our time series Y_t, which are defined simply as the third and fourth central moment of the time series, we use sample skewness (Equation 2.16) and sample kurtosis (Equation 2.17):

$$\widehat{\text{Skewness}}(Y) = \frac{1}{(1-N)\hat{\sigma}^3}\sum_{i=1}^{N}(y_i - \bar{y})^3, \tag{2.16}$$

$$\widehat{\text{Kurtosis}}(Y) = \frac{1}{(1-N)\hat{\sigma}^4}\sum_{i=1}^{N}(y_i - \bar{y})^4, \tag{2.17}$$

where N is the total amount of observations in the time series, and $\hat{\sigma}$ is the sample standard deviation derived from Equation (2.2).

The skewness, a measure of asymmetry, is very apparent in most insurance loss data because small-sized claims are usually way more frequent than the medium and large-sized claims (Lane, 2000) [42]. Financial time series show a lot less of this apparent asymmetry, and therefore are not as often modeled by skewed distributions (Brockwell & Davis, 2016, section 7.1) [13]

Market indices and financial return time series in general are leptokurtic (Bollerslev, 1986)[12], meaning they have a kurtosis in excess of the normal distribution (> 3). This implies that using the normal distribution as the conditional distribution for the returns might not yield the best model fit, as these data tend to have heavier tails than what we find in a normal distribution. Since the same property is known and present for insurance data, we can see that there might be some overlap in which conditional distributions has the best fit between financial time series and insurance time series.

In short, our absolute positive data with large shocks and constant nonzero mean may not be best suited for (G)ARCH-modelling. Therefore, we expect a better (G)ARCH model when fitting the de-meaned log of the total losses. Figure (6 a, c) shows the de-meaned versions of the original French fire data as well as their log. We have also included the same versions of the insurance data which will be discussed in section (7), a similar dataset of 2167 Danish fire insurance losses in the period 1980 to 1990 (Figure 6 b, d). We see that the two time series have very similar shapes, as one would expect with series of similar origin. The largest differences are that the amount of claims is higher for the French data (9613 observations compared to 2167 observations), the Danish data is scaled to be shown in million DKK instead of thousand 2007 EUR as in the French data, and that the French data seems to have more medium-sized losses compared to its small-sized losses than the Danish data. The main takeaway from this figure is that the de-meaned log versions of the insurance dataset bear a lot of the same characteristics as the financial time series

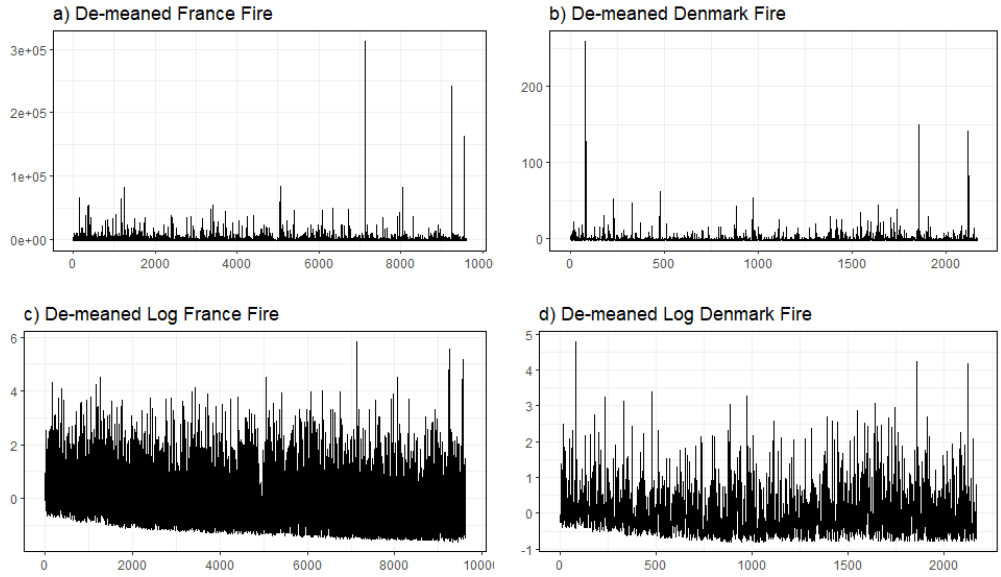


Figure 6: a & b: De-meaned time series plot of French fire losses and Danish fire losses, respectively. c & d: De-meaned log time series plot of French fire losses and Danish fire losses, respectively. Note that there are 9613 observations in the French dataset and only 2167 observations in the Danish dataset

of Figure 1, or at least more similar than the original data would be.

We also notice that there seems to be short periods in both insurance time series where we see larger claims, sometimes followed or preceded by a relatively large claim, and sometimes they seem to appear out of nowhere. This fact could imply that there are two distinct *regimes* in our time series where the structures of the volatility can differ for each of those regimes. We therefore wish to see if specifying a model where there are two distinct regimes will improve our fit, by reviewing *Markov-Switching Models* (Section 4).

Section (7) will go through the estimation and results of these models on the Danish Fire dataset. We fit those models for both de-meaned original data as well as de-meaned log data in order to be able to tell how much the transformation improves the model fit relative to using regular regression analysis.

3 ARCH models

In finance, insurance and other fields which involves financial risk, researchers have always been interested in modelling volatility in order to quantify the risk and price it. It is, in other words, quite important to attempt to model and predict volatility of returns, claims, assets, etc. Financial asset returns have been modelled as independent and identically distributed historically, but financial returns for high frequency data are actually not independent (Teräsvirta, 2009)[49]. Observations in such series may still be serially uncorrelated in the first degree, but there might exist higher-order dependence. This is where ARCH models excel, as the fitted model parameters are used in order to explain this dependence.
The theory on autoregressive conditional heteroskedasticity was introduced by Engle (1982)[23] to model time series data where the variance of the time series is in the main focus. With the original ARCH-model from Engle, it was possible to start modelling the variance of time series while assuming that it could vary over time, hereby the "Conditional" in the ARCH acronym. The first part, autoregressive, simply means that the time series model uses previous observations in order to explain current values. Lastly, heteroskedasticity can be present in the time series, which means that the conditional variance can vary over time.
ARCH-type models are great tools to use use in order to capture properties of a time series such as nonlinearities and asymmetries in the variance structure.
In this section, several ARCH models will be discussed, including the classic ARCH and the Generalized Autoregressive Conditional heteroskedastic (GARCH) model, introduced by Bollerslev (1986)[12]. Some more advanced specifications of the conditional variance will also be introduced, namely *tGARCH* and *gjrGARCH*. Lastly, we will relax the assumption of normality, and consider GARCH-models where the random innovations can follow more advanced conditional distributions.

3.1 ARCH

The original ARCH-model by Engle (1982)[23] for modelling a random variable y_t was postulated by first declaring that the random variable in question y_t is drawn from the conditional density function $f(y_t \mid y_{t-1})$, where the value of today is dependent upon the conditioning value y_{t-1}. The conditional expectation $\mathbb{E}(y_t \mid y_{t-1})$ has been a topic of much discussion, but the conditional variance, $\mathrm{Var}(y_t \mid y_{t-1})$, also implies that it depends on previous values, which is exactly what Engle's paper attempted to model with the introduction of the ARCH model. Engle suggested a model specification as follows for the zero mean-case:

$$y_t = \epsilon_t = \eta_t h_t^{\frac{1}{2}}, \tag{3.1}$$

$$h_t = \alpha_0 + \alpha_1 y_{t-1}^2. \tag{3.2}$$

This is the ARCH(1)-model, where $\{\eta_t\}$ is a sequence of I.I.D. (independent, identically distributed) random variables with zero mean and unit variance, and $h_t = \mathrm{Var}(y_t \mid y_{t-1}) = \mathbb{E}(y_t^2 \mid y_{t-1})$ is the conditional variance of y_t. α_j are constants for estimation. The sequence η_t is called the *standardized innovations*, since ϵ_t are the random innovations and $h_t^{\frac{1}{2}}$ is the conditional standard deviation. This model is called the ARCH(1)-model since it depends on the one-period lagged value of y_t. A more general ARCH(q)-model is specified as follows, by its conditional variance:

$$\mathrm{Var}(y_t \mid \mathcal{I}_{t-1}) = h_t = \alpha_0 + \sum_{j=1}^{q} \alpha_j y_{t-j}^2,$$

where $\alpha_0 > 0$ and $\alpha_j \geq 0$ for $j > 0$, and $\mathcal{I}_{t-1}$ is the information observed up to time $t-1$. Engle initially assumes normality of the sequence $\{\eta_t\}$, which makes y_t conditionally normal distributed with mean 0 and variance h_t,

$$y_t \mid \mathcal{I}_{t-1} \sim \mathcal{N}(0, h_t). \tag{3.3}$$

Since y_t in our case has conditional mean zero, y_t is a martingale difference sequence, which can be shown by the law of total expectation:

$$\mathbb{E}(y_t) = \mathbb{E}\left(\mathbb{E}(y_t \mid \mathcal{I}_{t-1})\right) = \mathbb{E}(0) = 0. \tag{3.4}$$

Furthermore, through again using the law of total expectation in addition to the already revealed properties of y_t, we can show that the observable process y_t is not only a martingale sequence, but shows no serial correlation:

$$\begin{aligned}
\mathrm{cov}(y_{t+k} y_t) &= \mathbb{E}(y_{t+k} y_t) - \mathbb{E}(y_{t+k})\mathbb{E}(y_t) && (3.5)\\
&= \mathbb{E}(y_{t+k} y_t) && (3.6)\\
&= \mathbb{E}\left(\mathbb{E}(y_{t+k} y_t \mid \mathcal{I}_{t+k-1})\right) && (3.7)\\
&= \mathbb{E}\left(y_t \mathbb{E}(y_{t+k} \mid \mathcal{I}_{t+k-1})\right) && (3.8)\\
&= 0. && (3.9)
\end{aligned}$$

As we recall, when a process has mean zero and doesn't show autocorrelation, it is called white noise, but not necessarily I.I.D. noise. This, as we have seen, and will continue to see, is an important assumption in applying the methods we will be using on our data, and is also a common property of financial returns-data, which is the area where ARCH/GARCH models are mostly used.

The greatest limitation of the ARCH-model lies in the fact that there are more parameters to estimate with every single additional considered lagged value of y_t. The ARCH(1)-model only uses one single lagged value in order to predict the conditional variance of the current value. In this case, if there were to incur a shock at time t, it would only have an effect on the outcome of the next time period. In many applications, especially financial time series, this is not in line with the empirical

applications. One would say that an ARCH(1)-model has very low persistence. Increasing the amount of considered lagged values , e.g. using an ARCH(q)-model with large enough q would result in low parsimony and inflexible variance structure for that model (Teräsvirta, 2008)[49].

3.2 GARCH

The introduction of the GARCH (Generalized autoregressive conditional heteroskeda -sticity)-model by Bollerslev (1986)[12] is in many ways a fix for the mentioned issues of the ARCH-model. The difference between the classic ARCH-model and the GARCH-model is that the model's conditional variance depends on lagged values of itself, in addition to lagged values of the observed process. Bollerslev defines the GARCH(1,1)-model as follows:

$$y_t \mid \mathcal{I}_{t-1} \sim \mathcal{N}(0, h_t), \tag{3.10}$$

$$h_t = \alpha_0 + \alpha_1 y_{t-1}^2 + \beta_1 h_{t-1}, \tag{3.11}$$

$$\eta_t = \frac{y_t}{h_t^{\frac{1}{2}}} \sim \mathcal{N}(0, 1). \tag{3.12}$$

Here, $\beta_1 \geq 0$ is the parameter that scales the lagged value of the conditional variance, and η_t are the standardized innovations. The GARCH(1,1) is actually a special case of the ARCH-model, i.e. the ARCH(∞)-model, which helps to understand that the GARCH-model is a more flexible version of ARCH, but with few parameters. This equivalence is shown, from Equation (3.11);

$$h_t = \alpha_0 + \alpha_1 y_{t-1}^2 + \beta_1 h_{t-1}.$$

We replace $h_{t-1} = \alpha_0 + \alpha_1 y_{t-2}^2 + \beta_1 h_{t-2}$,

$$\begin{aligned} h_t &= \alpha_0 + \alpha_1 y_{t-1}^2 + \beta_1 \left(\alpha_0 + \alpha_1 y_{t-2}^2 + \beta_1 h_{t-2}\right) \\ &= \alpha_0(1 + \beta_1) + \alpha_1 y_{t-1}^2 + \beta_1 \alpha_1 y_{t-2}^2 + \beta_1^2 h_{t-2}. \end{aligned}$$

We replace h_{t-2} and get

$$h_t = \alpha_0(1 + \beta_1 + \beta_1^2) + \alpha_1 y_{t-1}^2 + \beta_1 \alpha_1 y_{t-2}^2 + \beta_1^2 \alpha_1 y_{t-3}^2 + \beta_1^3 h_{t-3}.$$

Repeating the above steps for h_{t-j}, $j > 2$, we get

$$h_t = \alpha_0(1 + \beta_1 + \beta_1^2 + \beta_1^3 + \dots) + \alpha_1 \sum_{j=1}^{\infty} (\beta_1^{j-1} y_{t-j}^2). \tag{3.13}$$

This shows that a GARCH(1,1)-model can be written as a combination of parameters α_0, α_1 and β_1 in conjunction with infinite lagged values of our observed process y_t, which is a variation of an ARCH(∞)-model. Obviously, one would prefer the

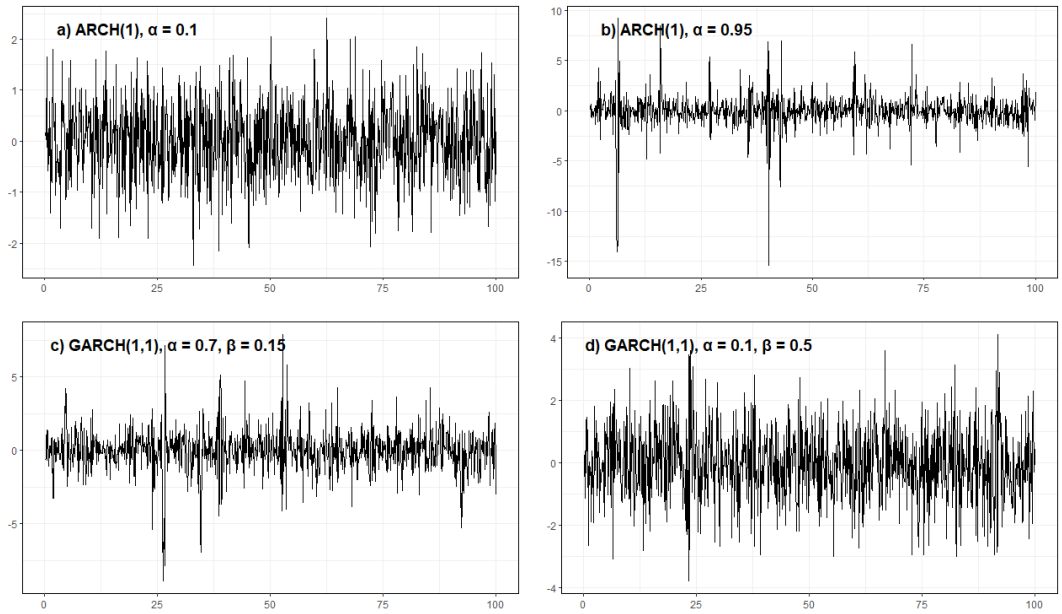


Figure 7: Simulated ARCH(1) and GARCH(1,1) time series. $\alpha_0 = 0.5$ for all processes, and the other parameters α_1, β_1 are differing. The simulation was done in R.

three-parameter GARCH(1,1)-model over directly using an ARCH(∞)-model because of its improved parsimony.
In Figure 7, we have shown some simulated time series of ARCH(1) and GARCH(1,1)-specification in order to better visualize the difference in these two models.

In Figure 7 we notice how a) differs from b) in that α_1 is much higher in the latter, making its current conditional variance depend more on the value of the previous value. We notice that there is very little persistence in b) in the way that the shocks last a very short time. In a), the α_1 is so low that it can be difficult to differentiate it from a regular white-noise process. c) and d) differ in that d) has a higher β_1 (0.5) and lower α_1 (0.1) than c) (0.15 & 0.7, respectively). Due to small α_1 in d), it responds weakly to the last period's return, but the relatively high β_1 ensures that the shocks decay slowly, making it seem like high volatility-periods appear very often. There is still a discernible difference between b) and c), since there is a weak β_1-effect in play in c). We can observe this effect by the fact that the shocks in c) seem to last a bit longer than in b), and one can see some more pronounced clustering. One last thing to note is that a) and d) may look quite similar, although this is not because they are equivalent. a) and d) may look the same because of the high β_1 of d), so the duration of the shocks last so long that they seem to overlap each other and mimic a white-noise process.
Extending the GARCH(1,1)-process to the more general GARCH(p, q), the condi-

tional variance becomes

$$h_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i y_{t-i}^2 + \sum_{j=1}^{q} \beta_j h_{t-j}, \tag{3.14}$$

where $\alpha_0 > 0, \alpha_i, \beta_j \geq 0$ to ensure that the conditional variance is strictly positive. The GARCH(p,q) turns into the classic ARCH(p) if $q = 0$. We will not be concerned with the GARCH(p, q)-model where $p, q > 1$ in this thesis because the GARCH(1,1)-model is the most popular application of GARCH-models, and a GARCH(1,1) model does quite well by itself. [49]. The GARCH(p,q)-model is *weakly* stationary if the following condition is met: (Bollerslev, 1986) [12]

$$\sum_{i=1}^{p} \alpha_i + \sum_{j=1}^{q} \beta_j < 1 \tag{3.15}$$

3.2.1 Conditional variance dynamics in GARCH

Since the introduction of the GARCH-model, several other specifications of the conditional variance has come to the surface, with the intent to create models that catch more of the conditional variance dynamics in time series data. In addition to the ARCH and GARCH-models already presented, we also consider the gjrGARCH (Glosten, Jagannathan & Runkle, 1993)[28] and tGARCH (Zakoian, 1994)[54] specifications, which are four of the specifications available in the R package `MSGARCH` (Ardia, Bluteau, Boudt, Catania & Trottier, 2019)[7]. We denote $\boldsymbol{\theta}$ as the vector that contains the parameters which are to be estimated for each model specification. As we recall, the classic ARCH(1)-model has the following specification for its conditional variance:

$$h_t = \alpha_0 + \alpha_1 y_{t-1}^2. \tag{3.16}$$

Here we have $\boldsymbol{\theta} = \{\alpha_0, \alpha_1\}^\intercal$. To ensure positivity and covariance stationarity, it is required that $\alpha_0 > 0$ and $0 \leq \alpha_1 < 1$.

We also reacall the GARCH(1, 1)-models conditional variance, defined as:

$$h_t = \alpha_0 + \alpha_1 y_{t-1}^2 + \beta_1 h_{t-1} \tag{3.17}$$

Here we have $\boldsymbol{\theta} = \{\alpha_0, \alpha_1\, \beta_1\}^\intercal$. To ensure positivity, it is required that $\alpha_0 > 0, \alpha_1 \geq 0$ and $\beta_1 \geq 0$, and covariance stationarity is ensured by taking $p = 1, q = 1$ of Equation 3.15, i.e. $\alpha_1 + \beta_1 < 0$.

The *gjrGARCH*-model presented by Glosten, Jagnnathan & Runkle (1993)[28] was introduced when the authors realized that the standard GARCH-model may not

be rich enough to encompass asymmetry in the conditional volatility. The model they presented as an adjustment to this is the following:

$$h_t = \alpha_0 + (\alpha_1 + \alpha_2 \mathbb{I}\{y_{t-1} < 0\}) y_{t-1}^2 + \beta_1 h_{t-1} \tag{3.18}$$

Here, $\mathbb{I}(\cdot)$ is the indicator function which takes the value one when the condition inside is held, and zero if the condition is not held. In effect, this means that the conditional variance includes an extra parameter α_2 that depends on the previous squared return y_{t-1}^2 if, and only if, the previous return y_{t-1} is negative. This means that α_2 is the parameter that controls the degree of asymmetry in the conditional variance. In this model specification, we have $\boldsymbol{\theta} = \{\alpha_0, \alpha_1, \alpha_2, \beta_1\}^\intercal$. To ensure positivity, it is required that $\alpha_0 > 0, \alpha_1 \geq 0, \alpha_2 \geq 0$ and $\beta_1 \geq 0$. Covariance stationarity is ensured by requiring $\alpha_1 + \alpha_2 \mathbb{E}\left(\eta_t^2 \mathbb{I}\{\eta_t < 0\}\right) + \beta_1 < 0$.

The *tGARCH* (threshold GARCH)-model, introduced by Zakoian (1994)[54], differs from the other specifications because it attempts to model the conditional standard deviation instead of the conditional variance. The conditional standard deviation is defined in the tGARCH-model as follows:

$$h_t^{\frac{1}{2}} = \alpha_0 + (\alpha_1 \mathbb{I}\{y_{t-1} \geq 0\} - \alpha_2 \mathbb{I}\{y_{t-1} < 0\}) y_{t-1} + \beta_1 h_{t-1}^{\frac{1}{2}}. \tag{3.19}$$

The tGARCH-model is quite similar to the gjrGARCH, with both models attempting to model asymmetry in the conditional variance. The difference, however, is that tGARCH includes one parameter to be estimated for negative past values and one parameter for positive past values. Zakoian argues that this difference allows the conditional standard deviations to have different reactions to different signs of the lagged values of the time series. In addition to this, the author has changed the focus from conditional variance and squared lagged values, to conditional standard deviation and absolute lagged values. The reason for this is motivated by the fact that David and Carroll (1987) [20] have found that "absolute residuals yield more efficient variance estimates than squared residuals." Here, we have $\boldsymbol{\theta} = \{\alpha_0, \alpha_1, \alpha_2, \beta_1\}^\intercal$. Positivity conditions are $\alpha_0 > 0, \alpha_1 \geq 0, \alpha_2 \geq 0$ and $\beta_1 \geq 0$, and stationarity is ensured by $\alpha_1^2 + \beta_1^2 - 2\beta_1^2(\alpha_1 + \alpha_2)\mathbb{E}\left(\eta_t^2 \mathbb{I}\{\eta_t < 0\}\right) - (\alpha_1^2 - \alpha_2^2)\mathbb{E}\left(\eta_t^2 \mathbb{I}\{\eta_t < 0\}\right) < 1$ (Franq & Zakoian, 2019)[27].

3.2.2 Conditional Distributions

In the original ARCH-paper by Engle (1982)[23], the author initially assumed that the innovations ϵ_t followed the conditional normal distribution, i.e. $\epsilon_t \mid \mathcal{I}_{t-1} \sim \mathcal{N}(0, h_t)$. However, this specification is not required, and not always satisfactory when the structure of the observations exhibits non-normal tendencies, e.g. excess skewness or kurtosis. In this section we will present some conditional distributions of

the standardized innovations η_t that are available in the R package `MSGARCH`. We will be presenting these conditional distributions: *normal*(`norm`), *Student-t*(`std`), *generalized error distribution*(`ged`) as well as the same distributions with the ability to capture skewness: *skew-normal*(`snorm`), *skew-Student-t*(`sstd`) and *skew-generalized error distribution*(`sged`). All distributions are standardized to have mean zero and unit variance, because they are the distributions of the standardized innovations η_t, and not the observations themselves. We define $\boldsymbol{\zeta}$ as the vector that contains the shape parameters in the conditional distribution.

The standardized normal distribution (`norm`) has probability density function (PDF)

$$f_N(\eta) \equiv \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\eta^2}, \ \ \eta \in \mathbb{R} \tag{3.20}$$

There are no additional parameters in a standardized normal distribution, therefore $\boldsymbol{\zeta} = [\ ]^\intercal$

The standardized Student-t (`std`) has PDF

$$f_T(\eta;\nu) \equiv \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{(\nu-2)\pi}\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{\eta^2}{(\nu-2)}\right)^{-\frac{\nu+1}{2}}, \ \ \eta \in \mathbb{R}, \nu > 0, \tag{3.21}$$

where $\Gamma(\cdot)$ is the Gamma function. $\nu > 2$ is required in order to ensure that the second order moment exists. Student-t distributions bear a lot of similarities to the normal distribution, but it has a higher capacity to capture kurtosis. The lower ν, the higher the kurtosis. If $\nu = \infty$, the Student-t distribution is equivalent to the normal distribution. Here, $\boldsymbol{\zeta} = [\nu]^\intercal$

The standardized generalized error distribution (`ged`) has PDF

$$f_{GED}(\eta;\nu) \equiv \frac{\nu e^{-\frac{1}{2}|\frac{\eta}{\lambda}|^\nu}}{\lambda 2^{\left(1+\frac{1}{\nu}\right)}\Gamma\left(\frac{1}{\nu}\right)}, \ \ \eta \in \mathbb{R}, \nu > 0, \tag{3.22}$$

where λ is defined as follows:

$$\lambda \equiv \left(\frac{\Gamma\left(\frac{1}{\nu}\right)}{4^{\frac{1}{\nu}}\Gamma\left(\frac{3}{\nu}\right)}\right)^{\frac{1}{2}} \ \ \nu > 0,$$

where ν is the shape parameter. This distribution becomes a normal distribution if $\nu = 2$, a Laplace distribution if $\nu = 1$ and a uniform distribution if $\nu \to \infty$. The GED, like the Student-t, has a large capacity for capturing kurtosis. The difference between the two lies in the fact that the GED acquires a cusp at its origin when approximating data with heavy tails, while the Student-t has a smooth peak when approximating data with heavy tails. Here, $\boldsymbol{\zeta} = [\nu]^\intercal$.

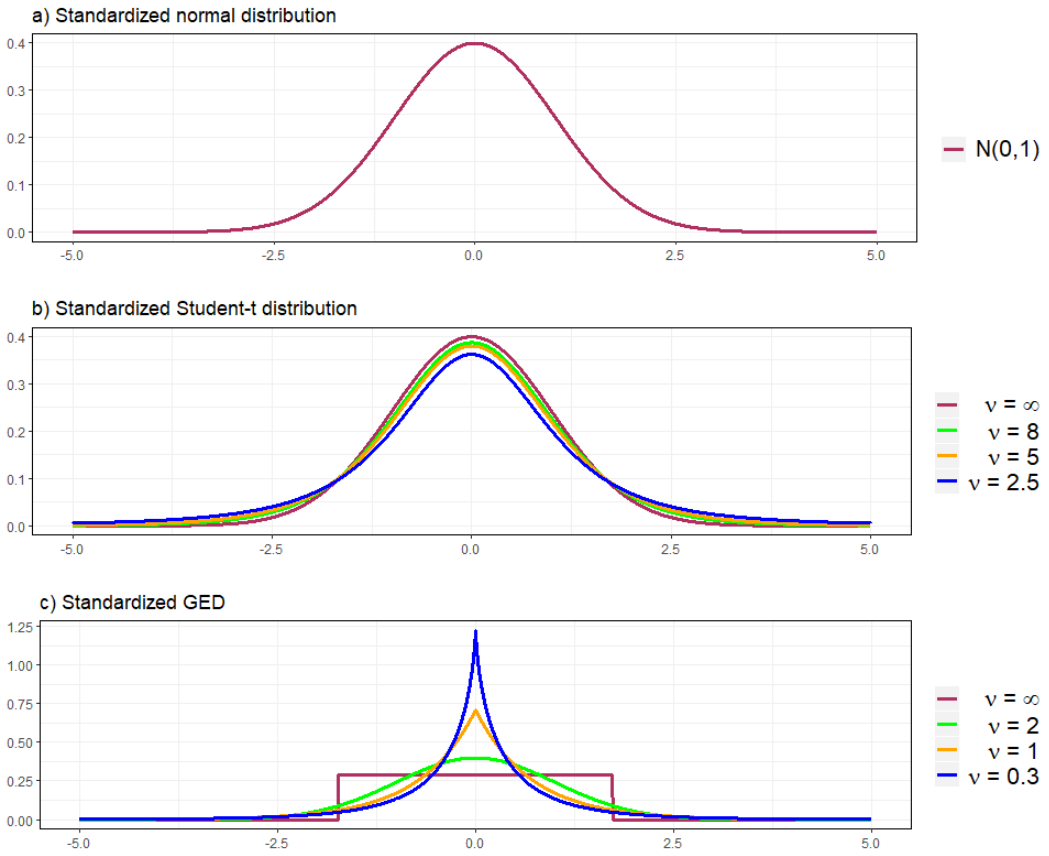


Figure 8: Example PDFs of standardized a) normal, b) Student-t and c) generalized error distribution for differing ν

The difference between the standardized distributions is shown graphically in Figure (8) for different values of ν.

In order to account for possible skewness in these three distributions, a transformation is necessary. Fernández & Steel (1998) [25] proposed a transformation that introduces skewness into any distribution that is unimodal and symmetric around 0. The density that accounts for skewness, and that has been standardized to have zero mean and unit variance (Lambert & Laurent, 2001)[40], can be written as follows:

$$f_\xi(\eta) = \frac{2\sigma_\xi}{\xi + \frac{1}{\xi}} f_1(\eta_\xi), \tag{3.23}$$

where η_ξ is given by

$$\eta_\xi \equiv \begin{cases} \frac{1}{\xi}(\sigma_\xi \eta + \mu_\xi) & \text{if } \eta \geq -\frac{\mu_\xi}{\sigma_\xi} \\ \xi(\sigma_\xi \eta + \mu_\xi) & \text{if } \eta < -\frac{\mu_\xi}{\sigma_\xi} \end{cases}.$$

Here, $\mu_\xi \equiv \mathrm{M}_1\left(\xi - \frac{1}{\xi}\right)$, $\sigma_\xi^2 \equiv (1 - \mathrm{M}_1^2)\left(\xi^2 + \frac{1}{\xi^2}\right) + 2\mathrm{M}_1^2 - 1$ and $\mathrm{M}_1 \equiv 2\int_0^\infty u f_1(u) du$. Also, $\xi > 0$ is the skewness parameter which controls the amount of skewness in the data. If $\xi = 1$, the distribution is symmetric. For $\xi > 1$, the distribution is said to be "right-skewed", and for $\xi < 1$, the distribution is said to be "left-skewed". Therefore, $f_1(\cdot)$ is the standardized distribution with no skewness. Skew-normal, skew-Student-t and skew-GED distributions have respectively, $\boldsymbol{\zeta} = [\xi]^\intercal$, $\boldsymbol{\zeta} = [\nu, \xi]^\intercal$ and $\boldsymbol{\zeta} = [\nu, \xi]^\intercal$. Figure (9) shows some example PDFs of skewed distributions.

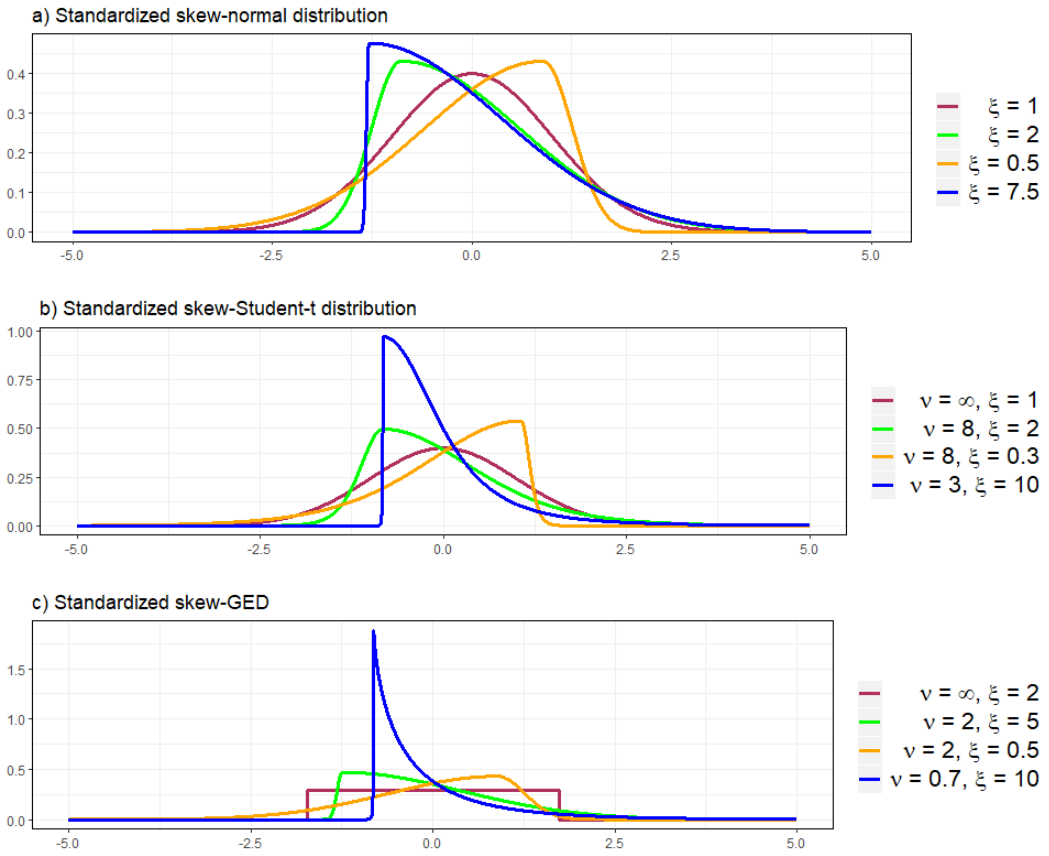


Figure 9: Example PDFs of standardized a) skew-normal, b) skew-Student-t and c) skew-generalized error distribution for differing ν & ξ

4 Markov models

We can have a preconception that a system has a set of available *states* it can find itself in, like a store being open or closed, or a machine being on, off or broken. If we assume that the realization of one such future state only depend on the current state, and not on what happened before, we turn to *Markov models*. Markov models are stochastic models used to model systems that change, with the property that the model is memoryless. These models have been extremely important in the process of modelling such systems, and can be applied to almost every field of research.
Markov models are usually built on the assumption that our observed variable can behave differently according to which specific state it is connected to. These states may or may not be observable, and this generalization provides a lot of flexibility while modeling. In the case of this thesis, Markov models are applied to a time series of insurance claims losses. The reasoning behind this is that insurance loss data are prone to sudden shocks of large losses, and the values of the losses in these short periods could be assumed to originate from a different unobserved state which is governed by a Markov chain.
This section provides an introduction to some important Markov models, with special attention guided towards the *Markov-switching model*.

4.1 Markov chain

A *Markov chain* $\{S_t, t = 1, 2, \dots T\}$ is a Markov model, which is characterized by the property that the probabilities of the current event are determined based on the state $\{1, 2, \dots, K\}$ of the previous event. The process S_t fulfills the Markov property, formalized as

$$\mathbb{P}(S_t = s_t \mid S_{1:t-1} = s_{1:t-1}) = \mathbb{P}(S_t = s_t \mid S_{t-1} = s_{t-1})$$

for times $t = \{1, \dots, T\}$, where $S_{1:t-1} = (S_1, S_2, \dots S_{t-1})$, and the realizations of those events $s_{1:t-1} = (s_1, s_2, \dots, s_{t-1})$. In Figure 10, the dependence structure of a simple Markov chain is displayed graphically for two time periods.

Here, $p_{i,j} = \mathbb{P}(S_t = j \mid S_{t-1} = i), \forall i, j \in \{1, 2, \dots K\}$ are the transition probabilities, or the probability that we are in state j at time t given that we were in state i in the previous time $t - 1$. We denote $\mathbf{P}$ as the transition probability matrix for the matrix process S_t:

$$\mathbf{P} = \begin{bmatrix} p_{1,1} & \dots & p_{1,K} \\ \vdots & \ddots & \vdots \\ p_{K,1} & \dots & p_{K,K} \end{bmatrix}, \; 0 < p_{i,j} < 1 \, \forall \, i, j \; \in \; \{1, \dots, K\}$$

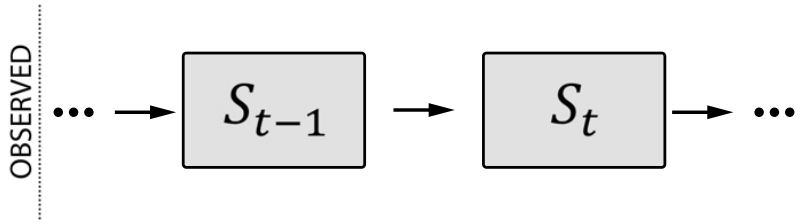


Figure 10: Dependence structure of a Markov chain. Here, $\{S_t\}$ is the observable chain which is shown over two time periods S_{t-1}, and the later time S_t which is dependent on the previous.

The i'th row of $\mathbf{P}$ is the probability density of S_t given that $S_{t-1} = i$. Furthermore, it also holds that $\sum_{j=1}^{K} p_{i,j} = 1 \; \forall i \in \{1, 2, \ldots, K\}$ which tells us that there occurs transitions also when the state remains the same. This probability, $p_{i,i}$, is called the *staying probability*. When the $p_{i,i}$ are large in comparison to the other transition probabilities, the model has highly persistent regimes, which means that the model will seldom switch regimes, and few transitions occur. In this thesis we will mainly be focusing on the two-state Markov chain, i.e. $K = 2$. The transition probability matrix becomes

$$\mathbf{P} = \begin{bmatrix} p_{1,1} & p_{1,2} \\ p_{2,1} & p_{2,2} \end{bmatrix} = \begin{bmatrix} p_{1,1} & 1 - p_{1,1} \\ 1 - p_{2,2} & p_{2,2} \end{bmatrix},$$

where we have used the property that $p_{1,1} + p_{1,2} = 1 \rightarrow p_{1,2} = 1 - p_{1,1}$.

Markov models are useful in a large range of applications where one is interested in analyzing the behavior of the variables in question for different, categorical states. This could mean a given market being in different states, like a financial crisis which is, in general, reflected in the market volatility, or smaller fluctuations could also be of interest.

4.2 Hidden Markov models

A *hidden Markov model* (HMM) is a Markov chain observed in noise (Cappé, Moulines & Rydén, 2005, section 1.1)[17]. The Markov chain $\{S_t\}$ is now hidden, which means we can not observe the states $\{S_t\}$ we find ourselves in. However, these hidden states govern the distribution of another stochastic process $\{Y_t\}$, which

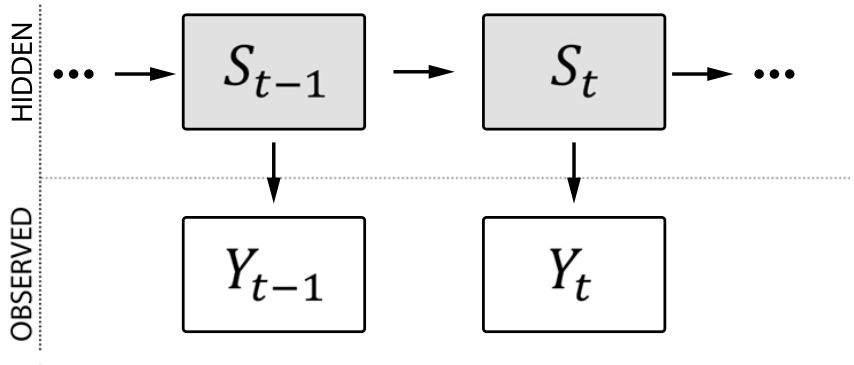


Figure 11: Dependence structure of a hidden Markov model. Here, $\{S_t\}$ is hidden, and $\{Y_t\}$ is the observable process, which are both shown over two time periods S_{t-1}, and the later time S_t which is dependent on the previous. In this case, $\{Y_t\}$ is dependent only on the hidden, "underlying" process

we do observe. Actually, one of the key assumptions of a HMM is that the observed values Y_t are conditionally independent of all other variables given their state variable $S_t, t = 1, 2, \ldots, K$. $\{Y_t\}$ can follow a distribution, and the parameters of said distribution will then be decided by the hidden process. In other words, the states which are hidden to us solely decide how our observable process behaves. Figure 11 is the graphical representation of the dependence structure of a simple hidden Markov model, with an attempt to highlight the fact that the observed process depends on the outcome of the underlying hidden process. Hidden Markov models are a very useful tool within fields such as econometrics, time series analysis and computational science because of the many applications in day-to-day life where unseen states influence the behavior of the values at interest.

One of the simplest specifications of hidden Markov models is the *normal hidden Markov model*, which is a model where the conditional distribution of Y_t given S_t follows a Gaussian distribution. The state-dependent distribution of Y_t then becomes

$$Y_t|(S_t = s_t) \sim \mathcal{N}\left(\mu_{s_t}, \sigma^2_{s_t}\right),$$

where μ_{s_t} and $\sigma^2_{s_t}$ are the parameters of the normal distribution that corresponds specifically to the state $S_t = s_t$. The variable in question thus follows a normal distribution at every time point t, but with the capability of having a changing mean and variance when the process enters a different state. Y_t can also follow different conditional distributions, depending on the specification.

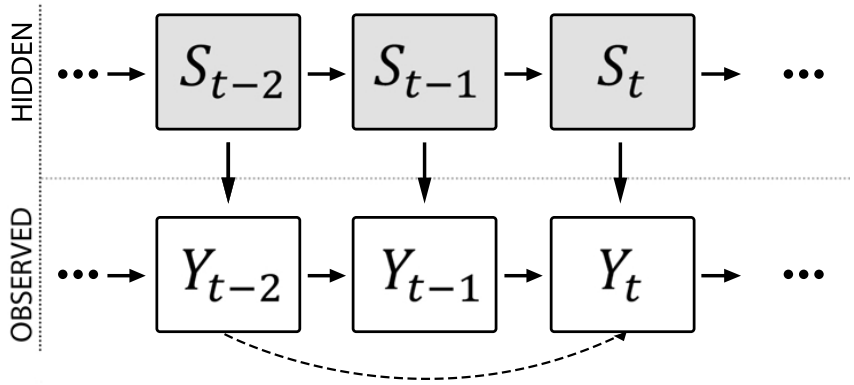


Figure 12: Dependence structure of a Markov-switching model. Here, $\{S_t\}$ is hidden, and $\{Y_t\}$ is the observable process, which are shown over three time periods. Here, the Markov chain behaves as normal, while the observed process now is dependent on its previous values, in addition to being dependent on the state variable

4.3 Markov-switching models

Markov-switching models(MSM) are a generalization of HMMs (Cappé et al., 2005, section 1.2)[17], and it was first introduced by Hamilton (1989)[33]. The difference between HMMs and MSMs lie in that the conditional distribution of the observed process Y_t does not only depend on the state S_t, but also depends on Y_{t-1}, or additional of the lagged values of $\{Y_t\}$. The conditional distribution of the observed process in a Markov-switching model, given the hidden process, could in theory depend on all previous values of itself, but the dependence on the Markov chain still upholds the Markov property, i.e.

$$f(Y_t = y_t \mid S_{1:t} = s_{1:t}, Y_{1:t-1} = y_{1:t-1}) = f(Y_t = y_t \mid S_t = s_t, Y_{1:t-1} = y_{1:t-1}) \quad (4.1)$$

where $f(\cdot)$ would be a generic pdf e.g. the normal distribution, and $Y_{1:t-1} = (Y_1, Y_2, \ldots, Y_{t-1})$.
We now see that the observed values Y_t are no longer conditionally independent of the other variables given the state variable. The dependence structure for MSMs are shown in Figure 12

The general Markov-switching model specification can be expressed as

$$Y_t \mid (S_t = s_t, \mathcal{I}_{t-1}) \sim \mathcal{D}\left(\mu_{s_t}(Y_{1:t-1}), \sigma^2_{s_t}(Y_{1:t-1}), \boldsymbol{\zeta}_{s_t}\right), \quad (4.2)$$

where $\mathcal{I}_{t-1} \equiv \{Y_t - i\}, i > 0$, is the information observed up to time $t-1$.
$\mathcal{D}\left(\mu_{s_t}(Y_{1:t-1}), \sigma^2_{s_t}(Y_{1:t-1}), \boldsymbol{\zeta}_{s_t}\right)$ is a continuous distribution with location and scale parameters μ_{s_t} and $\sigma^2_{s_t}$ that are dependent on both the realized state $S_t = s_t$, and possibly several lagged values of Y_t, up to time $t-1$. $\boldsymbol{\zeta}_{s_t}$ is the vector of additional

state-dependent shape parameters from section 3.2.2 that may be used, depending on the distribution in question.

Markov-switching models are used extensively to model financial asset-returns. Fama (1965)[24] argues that the price of a financial asset follows a random walk with drift, as discussed in section (2.3), where an expression for the returns time series was presented as $y_t = \mu_t + \epsilon_t$. We choose to use this model as an example in order to understand how regime-switching influences a time series.

After introducing the different regimes from the Markov-switching framework, y_t depends not only on time, but also on the regime variable S_t. y_t then becomes, for the $L = 2$ case:

$$y_t = \left\{ \begin{array}{ll} \mu_{t,1} + \epsilon_{t,1} & \text{if } S_t = 1 \\ \mu_{t,2} + \epsilon_{t,2} & \text{if } S_t = 2 \end{array} \right\}.$$

Here, μ_{t,s_t} is the time and regime-dependent conditional mean, and ϵ_{t,s_t} is the conditional random innovations-term. In the case where the conditional mean is constant over time for the different regimes, we have argued that we can switch our focus to a de-meaned version of the time series, and therefore only be concerned with ϵ_{t,s_t}. Adapting Equation (2.15) to the two-regime case gives the expression:

$$y_t = \left\{ \begin{array}{ll} \epsilon_{t,1} & \text{if } S_t = 1 \\ \epsilon_{t,2} & \text{if } S_t = 2 \end{array} \right\},$$

and consequently, $\mathbb{E}(y_t) = 0$ and $\mathbb{E}(y_t y_{t-\tau}) = 0$ for $\tau \neq 0$. This is, in other words, the application of a Markov-switching model to a zero-mean random innovations term with no serial correlation for the $K = 2$ case.

In the case where we regard the conditional mean as constantly zero, the Markov-switching model specification from Equation (4.2) can be simplified to

$$Y_t \mid (S_t = s_t, \mathcal{I}_{t-1}) \sim \mathcal{D}\left(0, \sigma^2_{s_t}(Y_{1:t-1}), \boldsymbol{\zeta}_{s_t}\right). \tag{4.3}$$

We know from section (3) that when we are interested in modeling a time series where we impose a specific structure on the conditional variance, letting this conditional variance h_t follow a (G)ARCH model is the usual method of approach. In the next section, we will see how we can include GARCH-models in the Markov-switching framework.

5 Markov-switching GARCH

In the previous sections we have discussed (G)ARCH-models and Markov-switching models separately. The motivation between combining the two comes from characteristic that the GARCH model may overestimate the persistence in variance because of possible existence of deterministic structural shifts in a model, which may be neglected (Lamoureux & Lastrapes, 1990)[41]. Consequently, GARCH forecasts may return too large values in high-volatility periods because a regular GARCH-model may struggle to adapt to the high persistence of those shocks. Klaassen (2001)[39] has shown this exact phenomenon in twenty years of daily data on USD exchange rates vs. GBP, DM and JPY, and therefore suggests combining the Markov-switching model, originally presented by Hamilton (1989)[33] with the GARCH-model. We consider the possibility that structural regimes that carry different parameters in each regime may improve a model where we observe sudden shocks that are followed by short periods of high volatility in the data. We only see the need for one additional regime in this case, partly because of parsimony and partly because we have a preconceived notion that data switches between a low-volatility regime and a high-volatility regime. To show the implementation of a Markov-switching GARCH model we refer to Figure (13), which illustrates the log-returns from the Swiss Market Index (left) and the Danish fire insurance losses (right), with the most likely state-path from some fitted Markov-switching GARCH-model superimposed on top. The plots were created with the `MSGARCH`-package in R.

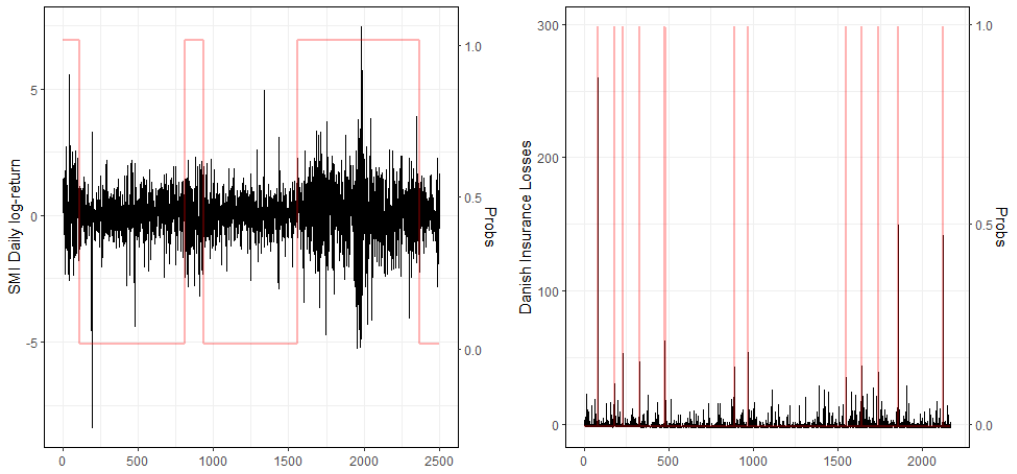


Figure 13: Swiss Market Index with MSGARCH-states estimated (left). Danish fire insurance losses with MSGARCH-states estimated(right). The red line is the most likely state-path which is equal to "0" in state one (low volatility) and "1" in state two (high volatility).

We see that the MSGARCH-model is quite good at distinguishing between volatility regimes in the log-return data, but it seems to also manage to capture the sudden shocks in the insurance loss dataset quite well without over-estimating the persistence of said shocks.
Another upside of using Markov-switching GARCH compared to regular GARCH methodology is that the different regimes are not required to follow the same conditional distribution, or the same conditional variance dynamics. An example of this would be a time series where observations from regime 1 tend to have more skewed observations with heavier tails and asymmetry in the conditional volatility, and the observations in regime 2 come from a less skewed distribution with more symmetry in the conditional volatility. If this were the case, we can see that it could be beneficial to fit a Markov-switching model where regime 1 follows a skew-Student-t distributions and has conditional volatility on the form of *gjrGARCH*, and regime 2 follows a normal distribution with a regular *GARCH* specification of the conditional volatility. The one-regime GARCH-framework would not allow for such flexibility, and this flexibility is especially apparent when attempting to model the tails of the predictive distribution, as being able to increase specificity in the high-volatility areas of a time series could give more accurate predictions in the tail.
In this section, we piece together the fundamental theory from earlier sections in order to build the Markov-switching GARCH-model and its different possible specifications. We then construct the likelihood function for estimation purposes, as well as presenting the Akaike Information Criterion (Akaike, 1998)[6], which will be used for model comparison. Lastly, we consider h-step-ahead prediction of the conditional volatility and draws.

5.1 Constructing the MSGARCH

As we have just discussed the reasons for using Markov-switching GARCH in lieu of the the regular single-regime GARCH, this section aims to put together the pieces in order for us to have a complete model specification. In this thesis, we will be implementing MSGARCH in the same manner as in Haas et al. (2004)[32]. This version has improved capabilities over the original MSGARCH-paper from Gray (1996)[31], namely that estimation is easier, it has higher analytical tractability, and the dynamic properties of the conditional variance has higher interpretability. In general, the MSGARCH needs three components that have been mentioned earlier. These components are the regime-switching probabilities which are defined in section 4.1, the conditional variance and the conditional distribution of the standardized GARCH innovations.

The Haas (2004)[32]-specification of the Markov-switching model lets the zero-

mean time series in question y_t satisfy

$$y_t = \epsilon_t = \eta_{s_t,t} h_{s_t,t}^{\frac{1}{2}} \tag{5.1}$$

where $\eta_{s_t,t}$ is the state and time-dependent standardized innovations, which follow a conditional distribution on the following form:

$$\eta_{s_t,t} \mid (S_t = s_t, \mathcal{I}_{t-1}) = \frac{y_t}{h_{s_t,t}^{\frac{1}{2}}} \mid (S_t = s_t, \mathcal{I}_{t-1}) \overset{I.I.D.}{\sim} \mathcal{D}(0, 1, \boldsymbol{\zeta_{s_t}}). \tag{5.2}$$

Consequently, the innovations themselves follow a similar conditional distribution with variance $h_{s_t,t}$:

$$y_t \mid (S_t = s_t, \mathcal{I}_{t-1}) \sim \mathcal{D}(0, h_{s_t,t}, \boldsymbol{\zeta_{s_t}}). \tag{5.3}$$

$h_{s_t,t}$ is the conditional variance of the model, $\mathcal{I}_{t-1}$ is the observed information in the time series up until time $t-1$ and $\mathcal{D}(\cdot)$ denotes a continuous distribution of choice with mean 0 and additional regime-dependent shape-parameter vector $\boldsymbol{\zeta_{s_t}}$, which takes on values depending on which distribution $\mathcal{D}(\cdot)$ adapts.
We notice that the specification of the model in Equation (5.1) is very similar to the single-regime (G)ARCH model specification from Equation (3.1), with the only difference being the that conditional variance $h_{s_t,t}$ and the standardized innovations $\eta_{s_t,t}$ are allowed to depend on which regime the time series is in. We also notice that the expression in Equation (5.2) is also very similar to the distribution specification from Equation (4.3), with the only difference being that the variance $\sigma_{s_t}^2$ of the former is replaced with the structural conditional variance h_{s_t} from (G)ARCH-models.

5.1.1 Conditional variance

The conditional variances of the regular Markov-switching ARCH(1) (5.4) and Markov-switching GARCH(1,1) (5.5) models take the following forms:

$$h_{s_t,t} = \alpha_{0,s_t} + \alpha_{1,s_t} y_{t-1}^2, \tag{5.4}$$

$$h_{s_t,t} = \alpha_{0,s_t} + \alpha_{1,s_t} y_{t-1}^2 + \beta_{1,s_t} h_{s_t,t-1} \tag{5.5}$$

The regime of the MS-(G)ARCH model determines the parameters $\alpha_{0,s_t}, \alpha_{i,s_t}$ and β_{1,s_t} for $i \in \mathbb{Z}^+$. As an example, for a MS-GARCH model with two regimes that both follow the standard ARCH(1) specification of the conditional variance, there are only 4 conditional variance parameters to be estimated $(\alpha_{0,1}, \alpha_{0,2}, \alpha_{1,1}, \alpha_{1,1})$.

A problem of intractability appears when fitting a Markov-switching GARCH model, and that problem is observed in the $h_{s_t,t-1}$-term of Equation (5.5). If we solve this equation recursively, we notice that the conditional variance $h_{s_t,t}$ depends on the full history of the time series, i.e. $\{y_{t-1}, y_{t-2}, \ldots, y_0, s_t, s_{t-1}, \ldots, s_1\}$. Estimation

becomes a problem because the number of possible paths of the process will grow exponentially as t grows. This proves to be a large hindrance in effectively fitting such a model, as much so that resarch papers by Hamilton & Susmel (1994)[35] and Cai (1994)[16] confined themselves to ARCH dependencies instead. Gray (1996)[31] proposed a way to circumvent this problem by claiming that the conditional variance should, instead of being generated by Equation (5.5), be generated by

$$h_{s_t,t} = \alpha_{0,s_t} + \alpha_{1,s_t} y_{t-1}^2 + \beta_{1,s_t} \tilde{h}_{t-1}, \tag{5.6}$$

where $\tilde{h}_{t-1}$ is given by

$$\sum_{i=1}^{K} \hat{\zeta}_{i,t-1|t-2} \left(\alpha_{0,i} + \alpha_{1,i} y_{t-2}^2 + \beta_{1,i} \tilde{h}_{t-2} \right). \tag{5.7}$$

Here, $\hat{\boldsymbol{\zeta}}_{t-1|t-2}$ is a probability vector whose i'th element corresponds to $\mathbb{P}(S_{t-1} = i \mid \boldsymbol{\Theta}; \Omega_{t-2})$. K is the amount of regimes in our model, Ω_t is the information gathered from only the observations up to time t, and $\boldsymbol{\Theta}$ is a vector of parameters in the model. The important thing to notice here, is how $h_{s_t,t}$ only depends on the history of the observations Ω_{t-1}, and not on the history of the states. This results in a much more tractable method of estimating a MSGARCH model.
Haas et al. (2004)[32] argued that this method of circumventing the problem was insufficient, mostly because of the lack of reasonable interpretation of the GARCH-parameters. The workaround was to hypothesize K separate GARCH-processes, whose value of conditional variance $h_{i,t}$ all exist as latent variables at time t, i.e.

$$h_{i,t} = \alpha_{0,i} + \alpha_{1,i} y_{t-1}^2 + \beta_{1,i} h_{i,t-1}, \tag{5.8}$$

We show that this only depends on the history of the previous observations $_{t-1}$ by inverting the expression of the whole vector $\boldsymbol{h_t}$:

$$\boldsymbol{h_t} = \boldsymbol{\alpha_0} + \boldsymbol{\alpha_1} y_{t-1}^2 + \boldsymbol{\beta_1} \boldsymbol{h_{t-1}} \tag{5.9}$$

$$= (I - \boldsymbol{\beta_1})^{-1} \boldsymbol{\alpha_0} + \sum_{i=1}^{\infty} \boldsymbol{\beta_1}^{i-1} \boldsymbol{\alpha_1} y_{t-i}^2 \tag{5.10}$$

Here, $\boldsymbol{\beta_1} = \text{diag}(\beta_{1,1}, \beta_{1,2}, \ldots, \beta_{1,K})$, $I \overset{1\times K}{=} diag(1, 1, \ldots, 1)^\intercal$,
$\boldsymbol{\alpha_i} = [\alpha_{i,1}, \alpha_{i,2}, \ldots, \alpha_{i,K}]^\intercal$ and $\boldsymbol{h_t} = [h_{1,t}, h_{2,t}, \ldots, h_{K,t}]^\intercal$
Diagonality of $\boldsymbol{\beta_1}$ implies, for the j'th element:

$$h_{j,t} = \alpha_{0,j}(1 - \beta_{1,j})^{-1} + \alpha_{1,j} \sum_{i=1}^{\infty} \beta_{1,j}^{i-1} y_{t-i}^2. \tag{5.11}$$

We observe that this equation only depends on the history of the observed values Ω_{t-1}, and not on the regime-history. The inversion of the preceding expression is actually very similar to what was done in section (3.2) when showing that a

GARCH(1,1)-model can be shown to follow a specific type of ARCH(∞)-model (Equation 3.13). By utilizing this exact property, Haas remarks that the interpretation of the GARCH-parameters α_1 and β_1 becomes much easier. More precisely, α_1 reflects the magnitude of a shock's immediate impact on the variance of the next period, and β_1 is a parameter of inertia, which explains the memory in the variance. The R package that will be mostly used in this thesis, `MSGARCH`, uses this version of approximating the conditional variance, which greatly helps in the understanding of the parameters that are fitted.

We also allow for other specifications of the conditional variance in the Markov-switching GARCH, as we did in section 3.2.1. There is very little difference between applying a specific GARCH-structure to a multi-regime GARCH model, than it is to apply it to a single-regime GARCH model. In fact, the single-regime GARCH is just a special case of the MSGARCH, where $K = 1$, and therefore the GARCH-parameters remain constant through the entire data. The advantage of introducing regimes, however, is that the different regimes are allowed to follow different specifications of the conditional variance. Table (1) contains an overview of the already mentioned specifications of the conditional variance that will be estimated in this thesis.

Conditional volatility models	
Model	Equation
ARCH	$\mathrm{h}_{s_t,t} = \alpha_{0,s_t} + \alpha_{1,s_t} y_{t-1}^2$
GARCH	$\mathrm{h}_{s_t,t} = \alpha_{0,s_t} + \alpha_{1,s_t} y_{t-1}^2 + \beta_{1,s_t} h_{s_t,t-1}$
gjrGARCH	$\mathrm{h}_{s_t,t} = \alpha_{0,s_t} + (\alpha_{1,s_t} + \alpha_{2,s_t}\mathbb{I}\{y_{t-1} < 0\}) y_{t-1}^2 + \beta_{1,s_t} h_{s_t,t-1}$
tGARCH	$\mathrm{h}_{s_t,t}^{\frac{1}{2}} = \alpha_{0,s_t} + (\alpha_{1,s_t}\mathbb{I}\{y_{t-1} \geq 0\} - \alpha_{2,s_t}\mathbb{I}\{y_{t-1} < 0\}) y_{t-1}$ $+ \beta_{1,s_t} h_{s_t,t-1}^{\frac{1}{2}}$

Table 1: Overview of specifications of the conditional variance. These specifications are all available in the R package `MSGARCH`

5.1.2 Conditional distribution

As we have mentioned in section (3.2.2), we will be presenting distributions for the *standardized* innovations, meaning the distributions all have mean zero and unit variance. The innovations y_t, however, are distributed with zero mean and variance $h_{s_t,t}$. For the time being, we consider the case where the standardized innovations for all regimes follow a normal distribution, i.e. $\eta_{s_t,t} \overset{\text{I.I.D.}}{\sim} \mathcal{N}(0,1)$. Here, y_t as mentioned follows a normal distribution with zero mean and variance $h_{s_t,t}$, which can be expressed as follows, in the two-regime case:

$$y_t \mid (S_t, \mathcal{I}_{t-1}) \sim \begin{cases} \mathcal{N}(0, h_{1,t}), \text{ with probability } P(S_t = 1 \mid \mathcal{I}_{t-1}) \\ \mathcal{N}(0, h_{2,t}), \text{ with probability } P(S_t = 2 \mid \mathcal{I}_{t-1}), \end{cases} \tag{5.12}$$

where $P(S_t = j \mid \mathcal{I}_{t-1})$ is the probability of being in state j given the information up to time $t-1$.
Having one distribution specification for each distinct regime can further extend the flexibility of our model compared to a single regime GARCH-model. The distributions from section (3.2.2) are reiterated for the MSGARCH case in Table (2)

Conditional PDFs	
Model	PDF
Normal	$\mathrm{f}_{N,s_t}(\eta) \equiv \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\eta^2}, \ \eta \in \mathbb{R}$
Student-t	$\mathrm{f}_{T,s_t}(\eta;\nu_{s_t}) \equiv \frac{\Gamma\left(\frac{\nu_{s_t}+1}{2}\right)}{\sqrt{(\nu_{s_t}-2)\pi}\Gamma\left(\frac{\nu_{s_t}}{2}\right)} \left(1+\frac{\eta^2}{(\nu_{s_t}-2)}\right)^{-\frac{\nu_{s_t}+1}{2}}, \ \eta \in \mathbb{R}, \nu_{s_t} > 0$
GED	$\mathrm{f}_{GED,s_t}(\eta;\nu_{s_t}) \equiv \frac{\nu_{s_t} e^{-\frac{1}{2}\left\lvert\frac{\eta}{\lambda}\right\rvert^{\nu_{s_t}}}}{\lambda 2^{\left(1+\frac{1}{\nu_{s_t}}\right)}\Gamma\left(\frac{1}{\nu_{s_t}}\right)}, \ \eta \in \mathbb{R}, \nu_{s_t} > 0$
Skew-normal	Normal PDF used in Eq. (3.23)
Skew-Student-t	Student-t PDF used in Eq. (3.23)
Skew-GED	GED PDF used in Eq. (3.23) height

Table 2: Overview of specifications of the conditional distributions. These specifications are all available in the R package `MSGARCH`

5.2 Maximum likelihood estimation

We have the fundamental building blocks needed to construct the likelihood function. Estimation will be done through maximum likelihood estimation (MLE). The model parameters which we want to estimate are contained in the vector $\boldsymbol{\Theta} = [\boldsymbol{\theta_1}, \boldsymbol{\zeta_1}, \ldots, \boldsymbol{\theta_K}, \boldsymbol{\zeta_K}, \mathbf{P}]$, where $\boldsymbol{\theta_j}$ is the vector of GARCH-parameters in regime j, $\boldsymbol{\zeta_j}$ is the vector of distribution parameters in regime j and $\mathbf{P}$ is the transition probability matrix. The likelihood function is given by:

$$\mathcal{L}(\boldsymbol{\Theta}) = \prod_{t=1}^{T} f(y_t \mid \boldsymbol{\Theta}, \Omega_{t-1}), \tag{5.13}$$

where $f(y_t \mid \boldsymbol{\Theta}, \Omega_{t-1})$ is the distribution of the observations given the history of observations until time $t-1$, Ω_{t-1}, and the model parameters $\boldsymbol{\Theta}$. For evaluating $f(y_t \mid \boldsymbol{\Theta}, \Omega_{t-1})$ we consider first the $K = 2$ case. To obtain the complete conditional PDF of y_t for a MSGARCH, we need to sum together the densities of y_t conditional on being in state j. For current state $j = 1$, where $K = 2$ there are two possible "paths" to consider; one where the previous state was $j = 1$ and the current state is $j = 1$, and one where the previous state was $j = 2$ and the current state is $j = 1$. We therefore need to multiply the probability of being in state $j = 1$ at time $t-1$ by the probability of going from state 1 to 1 in order to get the probability of the first "path". The second path is acquired by multiplying the probability of being in

state $j = 2$ at time $t - 1$ by the probability of going from state 2 to 1. We do this for both current states $j = 1, 2$. We end up with four parts in this $K = 2$ case:

$$\begin{aligned} f(y_t \mid \boldsymbol{\Theta}; \Omega_{t-1}) &= p_{1,1} z_{1,t-1|t-1} f(y_t \mid s_t = 1, \boldsymbol{\Theta}; \Omega_{t-1}) + \\ &= p_{1,2} z_{1,t-1|t-1} f(y_t \mid s_t = 2, \boldsymbol{\Theta}; \Omega_{t-1}) + \\ &= p_{2,1} z_{2,t-1|t-1} f(y_t \mid s_t = 1, \boldsymbol{\Theta}; \Omega_{t-1}) + \\ &= p_{2,2} z_{2,t-1|t-1} f(y_t \mid s_t = 2, \boldsymbol{\Theta}; \Omega_{t-1}) \\ &= \sum_{i=1}^{2} \sum_{j=1}^{2} p_{i,j} z_{i,t-1} f(y_t \mid s_t = j, \boldsymbol{\Theta}; \Omega_{t-1}) \end{aligned}$$

This can be generalized for the K case: (Ardia et al. 2019a)[7]

$$f(y_t \mid \boldsymbol{\Theta}; \Omega_{t-1}) = \sum_{i=1}^{K} \sum_{j=1}^{K} p_{i,j} z_{i,t-1|t-1} f(y_t \mid s_t = j, \boldsymbol{\Theta}; \Omega_{t-1}) \tag{5.14}$$

In both these cases, $z_{i,t|t}$ is the probability of being in state i at time t, conditioned on the parameter vector and the history of the time series until time t, i.e. $\mathbb{P}(s_t = i \mid \boldsymbol{\Theta}; \Omega_t)$. These probabilities are not known, as the states are hidden and we therefore can not know which regime the process was in at every time point. In reality, these probabilities are 1 when the process is in state i, and 0 when the process is not in state i. Since we cannot determine these exact probabilities, we use Hamilton's filtered probabilities (Hamilton, 1994, chapter 22)[34], which are derived from a probabilistic inference that is a generalization of $\mathbb{P}(s_t = i \mid \boldsymbol{\Theta}; \Omega_t)$. We denote this inference as $\hat{z}_{i,t|t}$.

In order to derive this filter, which is the conditional probability of s_t, we first define $\boldsymbol{v}_t$ as the $(K \times 1)$ vector whose j'th element is the density of y_t conditioned on s_t being in state j at time t. We also define $\hat{\boldsymbol{z}}_{\boldsymbol{t|t}}$ as the $(K \times 1)$ vector which contains our inference of conditional probabilities of being in state j at time t based on the information up to time t, and $\mathbf{1}$ is a $(K \times 1)$ vector of ones. These three vector are defined as follows:

$$\boldsymbol{v}_t = \begin{bmatrix} f(y_t \mid s_t = 1, \boldsymbol{\Theta}; \Omega_{t-1}) \\ f(y_t \mid s_t = 2, \boldsymbol{\Theta}; \Omega_{t-1}) \\ \vdots \\ f(y_t \mid s_t = K, \boldsymbol{\Theta}; \Omega_{t-1}) \end{bmatrix}, \quad \hat{\boldsymbol{z}}_{\boldsymbol{t|t}} = \begin{bmatrix} \mathbb{P}(s_t = 1 \mid \boldsymbol{\Theta}; \Omega_t) \\ \mathbb{P}(s_t = 2 \mid \boldsymbol{\Theta}; \Omega_t) \\ \vdots \\ \mathbb{P}(s_t = K \mid \boldsymbol{\Theta}; \Omega_t)) \end{bmatrix}, \quad \mathbf{1} \overset{K \times 1}{=} \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

Next, the element-by-element multiplication between $\hat{\boldsymbol{z}}_{\boldsymbol{t|t-1}}$ and $\boldsymbol{v}_t$ is denoted by the following $(K \times 1)$-vector:

$$(\hat{\boldsymbol{z}}_{t|t-1} \odot \boldsymbol{v}_t) = \begin{bmatrix} \mathbb{P}(s_t = 1 \mid \boldsymbol{\Theta}; \Omega_{t-1}) \times f(y_t \mid s_t = 1, \boldsymbol{\Theta}; \Omega_{t-1}) \\ \mathbb{P}(s_t = 2 \mid \boldsymbol{\Theta}; \Omega_{t-1}) \times f(y_t \mid s_t = 2, \boldsymbol{\Theta}; \Omega_{t-1}) \\ \vdots \\ \mathbb{P}(s_t = K \mid \boldsymbol{\Theta}; \Omega_{t-1}) \times f(y_t \mid s_t = K, \boldsymbol{\Theta}; \Omega_{t-1}) \end{bmatrix}, \tag{5.15}$$

where $\odot$ symbolizes element-by-element multiplication. This is actually a vector where the j'th element can be interpreted as the conditional joint density distribution of y_t and s_t, because of the definition of conditional probability (5.17). The marginal density of the observations y_t is therefore given by summing the K values of the joint probability density (Equation 5.15), i.e.

$$f(y_t \mid \boldsymbol{\Theta}; \Omega_{t-1}) = \mathbf{1}^\intercal (\hat{\boldsymbol{z}}_{t|t-1} \odot \boldsymbol{v}_t) \tag{5.16}$$

In order to derive an expression for the conditional probability of s_t that is conditioned on times up to time t, i.e. $\hat{\boldsymbol{z}}_{t|t}$, we use the definition of conditional probability:

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \tag{5.17}$$

Here, $\mathbb{P}(A \cap B)$ is given by the j'th element of the conditional joint density distributions of y_t and s_t from Equation (5.15), i.e.

$$\mathbb{P}(s_t = j \mid \boldsymbol{\Theta}; \Omega_{t-1}) \times f(y_t \mid s_t = j, \boldsymbol{\Theta}; \Omega_{t-1}) \tag{5.18}$$

$$= p(y_t, s_t = j \mid \boldsymbol{\Theta}; \Omega_{t-1}) \tag{5.19}$$

and $\mathbb{P}(B)$ is expression from Equation (5.16), i.e. the marginal distribution of y_t conditioned on the past observations. $\mathbb{P}(A \mid B)$ is thus the conditional probability of s_t, given the parameters $\boldsymbol{\Theta}$, the past observations until time $t-1$, y_{t-1} and the value of y_t. As we can see, this is actually the conditional probability of s_t given the parameters and the information of the observed values until time t:

$$\mathbb{P}(A \mid B) = \mathbb{P}(s_t = j \mid \boldsymbol{\Theta}; y_t, \Omega_{t-1}) = \mathbb{P}(s_t = j \mid \boldsymbol{\Theta}; \Omega_t) \tag{5.20}$$

Equation (5.17) now becomes:

$$\mathbb{P}(s_t = j \mid \boldsymbol{\Theta}; \Omega_t) = \frac{p(y_t, s_t = j \mid \boldsymbol{\Theta}; \Omega_{t-1})}{\mathbf{1}^\intercal (\hat{\boldsymbol{z}}_{t|t-1} \odot \boldsymbol{v}_t)}. \tag{5.21}$$

As we recall, the left side of Equation (5.21) is the j'th element of $\hat{\boldsymbol{z}}_{\boldsymbol{t|t}}$, and the right side of the equation is the j'th element of $(\hat{\boldsymbol{z}}_{t|t-1} \odot \boldsymbol{v}_t)$. We can thus recollect the K elements of the vectors in the above equation, which produces a $(K \times 1)$ vector, which we present:

$$\hat{\boldsymbol{z}}_{\boldsymbol{t|t}} = \frac{(\hat{\boldsymbol{z}}_{t|t-1} \odot \boldsymbol{v}_t)}{\mathbf{1}^\intercal (\hat{\boldsymbol{z}}_{t|t-1} \odot \boldsymbol{v}_t)}. \tag{5.22}$$

These filtered probabilities are thus a first-order recursive process, which almost finishes the definition of our likelihood function (5.13). The ML estimator $\hat{\boldsymbol{\Theta}}$ is obtained by maximizing the logarithm of Equation (5.13), but first we need to choose some starting values for the maximum likelihood estimation.

5.2.1 Choosing starting values

The starting values of the maximum likelihood algorithm will be decided by the following process, as it has been implemented in `MSGARCH`:

1. (**P**): Choosing the transition probabilities by estimating a static version of the model ($h_{s_t,t} = \bar{h}_{s_t}$) by the Baum-Welch algorithm.

2. (Decoding): Assigning each of the observations to a regime through the Viterbi algorithm (Viterbi, 1967)[52].

3. $(\alpha_{0,j}, \alpha_{1,j}, \alpha_{2,j}, \beta_{1,j}, \nu_j, \xi_j)$: Estimating the the remaining parameters through Maximum Likelihood estimation, where we fit the model once for each regime j, independent of the others.

Firstly, the **static estimation** will be done by the *Baum-Welch algorithm*, which is a generalization of the EM-algorithm. The BW-algorithm was introduced through some papers by Leonard E. Baum and Lloyd R. Welch during the 1960's and 1970's. Cappé et al. (2005, chapter 5)[17] provides an insight in how the algorithm works. Its objective is to calculate the probability of a given observation sequence y_t, which returns a maximum-likelihood estimate of the the transition probability matrix $\mathbb{P}$, an initial state probability vector $\boldsymbol{\pi}_i$ and the emissions matrix $\mathbf{C}$ of size $(K \times T)$ whose element $c_j(y_i) = \mathbb{P}(Y_t = y_i \mid S_t = j)$ is the probability of observing $Y_t = y_i$ from state $S_t = j$. We are also able to calculate an estimate for the unconditional variances $\sigma^2_{s_t}$, which will be used as their starting value.
We set $\theta = (\mathbf{P}, \mathbf{C}, \boldsymbol{\pi})$ as the parameter vector which we want to estimate. Initial values for these parameters could be chosen at random. In our case they are chosen as follows:

P is chosen by letting the values on the diagonal be equal to 0.9, and the remaining values are chosen to be equal to each other, based on the condition that the row-sums are equal to 1. E.g. for a $K = 2$ case: $p_{1,1} = p_{2,2} = 0.9, p_{1,2} = p_{2,1} = 0.1$.

C is chosen by fitting K normal distributions to the observations y_t. Here, the variances are chosen by multiplying the sample variance of the observations by a scaling-value which is different for each of the K states. E.g. for a $K = 2$ case: $\sigma^2_1 = 0.8\bar{\sigma}^2, \sigma^2_2 = 1.2\bar{\sigma}^2$. The probabilities of those normal distributions are used as emission probabilities.

$\boldsymbol{\pi}$ is chosen by the steady-state vector of the transition matrix $\mathbf{P}$. The steady-state vector is the vector $\boldsymbol{\pi}$ who satisfies the equation $\mathbf{P}\boldsymbol{\pi} = \boldsymbol{\pi}$. Another way of explaining it is that it is an eigenvector of $\mathbf{P}$ who is associated to the eigenvalue 1. This vector is calculated as follows:

$$\boldsymbol{\pi} = ((I - \mathbf{P} + \mathbb{1})^{\intercal})^{-1}\mathbf{1} \tag{5.23}$$

The Baum-Welch algorithm takes use of *forward filtering backward sampling*, which is a smoothing algorithm that computes the posterior marginal distributions of the states for each of K. The algorithm goes through two methods, one going forward in time, and one going backwards in time, hence the name.
The forward procedure computes $\alpha_i(t) = \mathbb{P}(Y_1 = y_1, Y_2 = y_2, \ldots, Y_t = y_t, S_t = i \mid \theta)$, the joint probability of the observations up to time t and being in state i. Define the $(K \times T)$ matrix $\mathbf{A} = [\boldsymbol{\alpha_1}, \boldsymbol{\alpha_2}, \ldots, \boldsymbol{\alpha_K}]$. The values of $\boldsymbol{\alpha}$ are found recursively by these equations:

$$\alpha_i(1) = \pi_i c_i(y_1), \tag{5.24}$$

$$\alpha_i(t+1) = c_i(y_{t+1}) \sum_{j=1}^{K} \alpha_j(t) p_{j,i} \tag{5.25}$$

The backward procedure computes $\beta_i(t) = \mathbb{P}(Y_{t+1} = y_{t+1}, Y_{t+2} = y_{t+1}, \ldots, Y_T = y_t \mid S_t = i, \theta)$, the probability of the remaining y_t until the final value T, given that the process was in state i at time t. Define the $(K \times T)$ matrix $\mathbf{B} = [\boldsymbol{\beta_1}, \boldsymbol{\beta_2}, \ldots, \boldsymbol{\beta_K}]$. The values of $\boldsymbol{\beta}$ are found recursively by these equations:

$$\beta_i(T) = 1, \tag{5.26}$$

$$\beta_i(t) = \sum_{j=1}^{K} \beta_j(t+1) p_{i,j} c_j(y_{t+1}), \tag{5.27}$$

We can obtain the joint probability of being in state i at time t as well as the probabilities of the entire history of y_t defined as $Y = (Y_1 = y_1, Y_2 = y_2, \ldots, Y_T = y_T)$ by simply multiplying the two vectors for state i and time t:

$$\mathbb{P}(S_t = i, Y_1 = y_1, Y_2 = y_2, \ldots, Y_T = y_T \mid \theta) = \alpha_i(t)\beta_i(t) \tag{5.28}$$

The marginal probability of the observations Y_t over the entire history $[1, T]$ is gained by summing the joint probability over all states, i.e. $\sum_{j=1}^{K} \alpha_j(t)\beta_j(t)$. To obtain the smoothed probability of being in state i at time t given the full history of the observations, we utilize Bayes' rule:

$$\gamma_i(t) = \mathbb{P}(S_t = i \mid Y, \theta) = \frac{\mathbb{P}(S_t = i, Y \mid \theta)}{\mathbb{P}(Y \mid \theta)} = \frac{\alpha_i(t)\beta_i(t)}{\sum_{j=1}^{K} \alpha_j(t)\beta_j(t)}, \tag{5.29}$$

We also want to define the probability of being in state i at time t, and then transitioning to state j at time $t+1$, given the entirety of the observations Y. Similar to $\gamma_i(t)$, we define, by using Bayes' rule:

$$\delta_{i,j}(t) = \mathbb{P}(S_t = i, S_{t+1} = j \mid Y, \theta) = \frac{\mathbb{P}(S_t = i, S_{t+1} = j, Y \mid \theta)}{\mathbb{P}(Y \mid \theta)} \tag{5.30}$$

$$= \frac{\alpha_i(t)p_{i,j}\beta_{t+1}c_j(y_{t+1})}{\sum_{j=1}^{K}\sum_{l=1}^{K} \alpha_j(t)p_{j,l}\beta_l(t+1)c_l(y_{t+1})} \tag{5.31}$$

Now, we can update the equation for the transition probability matrix. If we sum $\delta_{i,j}$ over t values 1 through $T-1$, we get the total expected amount of transitions between i and j. If we compare these values with the sum of γ_i from time 1 to time $T-1$, which is the total amount of transitions that start in state i, we naturally obtain the rates of transitioning between state i and j to the total transitions. This is the definition of the updated transition probabilities, whose matrix $\mathbf{P}^*$ now has entries which will be written as follows:

$$p^*_{i,j} = \frac{\sum_{t=1}^{T-1} \delta_{i,j}(t)}{\sum_{t=1}^{T-1} \gamma_i(t)}, \tag{5.32}$$

We also define expressions for the sample unconditional mean and variance of the time series as follows:

$$\mu^*_j = \frac{\sum_{t=1}^{T} (\gamma_j(t)y_t)}{\sum_{t=1}^{T} \gamma_j(t)} \tag{5.33}$$

$$\sigma^{2*}_j = \frac{\sum_{t=1}^{T} \left(\gamma_j(t)(y_t - \mu^*_j)^2\right)}{\sum_{t=1}^{T} \gamma_j(t)} \tag{5.34}$$

When all these values are calculated, the Baum-Welch algorithm has completed one iteration. For the next iteration, we recalculate the emission probabilities $\mathbf{C}$ by the same method as when choosing the starting values for the algorithm, i.e. by fitting K normal distributions to the observations y_t vector, although we now use the updated σ^{2*}_j in the distributions. $\boldsymbol{\pi}$ is also recalculated in the same way as when choosing starting values, i.e. by computing the steady-state vector for the updated transition matrix $\mathbf{P}^*$. The algorithm is repeated iteratively until some tolerance condition is met.

Next up in the estimation process is **decoding** through using the Viterbi algorithm. This algorithm is used to find the most likely regime-path our time series will follow. The Viterbi algorithm returns a $(T \times 1)$ vector $\boldsymbol{V} = [v_1, v_2, \ldots, v_T]$ of

integers that can take values $v \in K$. The input for the algorithm, other than the amount of states K and the observations Y, is the initial state vector $\boldsymbol{\pi}^*$, the transition probability matrix $\mathbf{P}^*$ and the emission probability matrix $\mathbf{C}^*$, which were all estimated by the BW-algorithm.

To start the algorithm, two $(K \times T)$ vectors $T_1[i,j]$ & $T_2[i,j]$ are defined. $T_1[i,j]$ contains the probabilities of the most probable state sequence $\mathbb{P}(\hat{V}_1, \hat{V}_2, \ldots, \hat{V}_j, Y_1, Y_2, \ldots, Y_j)$ responsible for the first j observations, and where the sequence is in state i at time j. $T_2[i,j]$ contains $\hat{V}_{j-1}$ extracted from the most likely path until the j'th observation. We define these values formally as:

$$T_1[i,j] = \max_k(T_1[k,j-1]p_{k,i}c_{i,j}) \tag{5.35}$$

$$T_2[i,j] = \operatorname*{argmax}_k(T_1[k,j-1]p_{k,i}) \tag{5.36}$$

$T_2[i,j]$ is called a back-pointer as it extracts which state was used to obtain the value of $T_1[i,j]$. The starting value for $T_1[i,1]$ is the initial state vector times the emission probabilities from observation 1, i.e. $T_1[i,1] = \pi_i c_{i,1}$.
After recursively determining values of $T_1[i,j]$ and $T_2[i,j]$ for all values of states $i \in [1:K]$ and observations $j \in [1:T]$, we simply define the Viterbi path as the states belonging to the most likely path determined by $T_2[i,j]$.

We now have a vector of the most likely path that our time series follows. The next part of choosing the starting values involves creating K vectors of the observations corresponding to each of the K regimes, based on the Viterbi path. These K vectors of decoded observations are then used to fit the remaining parameters, i.e. the shape parameters and the (G)ARCH parameters. This ML-estimation is carried out in the same way as for the ML-estimation of the likelihood function 5.13 in section 5.3, i.e. by the BFGS-algorithm.

The starting values of our parameter-vector $\boldsymbol{\Theta}$ is now completely decided, and we are able to carry out the maximum likelihood estimation. We define this vector as

$$_0\boldsymbol{\Theta} = [_0\boldsymbol{\theta_1}, {}_0\boldsymbol{\zeta_1}, \ldots, {}_0\boldsymbol{\theta_K}, {}_0\boldsymbol{\zeta_K}, {}_0\mathbf{P}], \tag{5.37}$$

where $_0\mathbf{P}$ only contains $(K \times K) - K$ values for $p_{i,j}$, as each row in the matrix sums to 1, and thus each row's K'th value can be found by subtracting the row's previous $K-1$ values from 1.

5.3 The optimization process

We recall that the likelihood function of our model is given by Equation (5.13):

$$\mathcal{L}(\boldsymbol{\Theta}) = \prod_{t=1}^{T} f(y_t \mid \boldsymbol{\Theta}, \Omega_{t-1}).$$

Direct maximization of this function is infeasible in most cases as it usually involves high-dimensional integration and is therefore sufficiently complex enough so that we cannot use simple optimization techniques (Cappé et al., 2005, section 10.1.1)[17]. We therefore choose an iterative optimization method, as these are generally easier and less computationally demanding to use. The method we are using for our optimization is called the BFGS(Broyden 1970, Fletcher 1970, Goldfarb 1970, Shanno 1970)[15][26][30][47]-algorithm, which is named after the four people who first discovered the algorithm. This algorithm belongs to a class of optimization techniques called *quasi-Newton*. A regular Newton-method is an optimization technique which evaluates the Hessian for each iteration. The Hessian $\mathbf{H}$ is defined as the matrix containing the second-order partial derivative of a scalar-valued function. Quasi-Newton-methods are more computationally cheap and faster to compute than Newton-methods, as these methods does not require the Hessian to be evaluated for each iteration, but instead use an approximation of the Hessian based on gradient evaluations. The basis for the following text is Nocedal & Wright (2006, section 6.1)[45], as it gives a good explanation of the BFGS method.
The function we want to maximize is $\log\mathcal{L}(\boldsymbol{\Theta})$. As BFGS is mainly used for minimization problems, we adjust our function to being negative, i.e. $\mathcal{M}(\boldsymbol{\Theta}) := -\log\mathcal{L}(\boldsymbol{\Theta})$
The quasi-Newton methods of optimization is of an iterated scheme:

$$_{i+1}\boldsymbol{\Theta} = {}_{i}\boldsymbol{\Theta} - \alpha_i \mathbf{H}_i^* \boldsymbol{\nabla}\mathcal{M}({}_{i}\boldsymbol{\Theta}), \tag{5.38}$$

where $\mathbf{H}_i^*$ is the approximation of the inverse Hessian $\mathbf{H}_i = [\boldsymbol{\nabla}^2\mathcal{M}({}_{i}\boldsymbol{\Theta})]^{-1}$ at iteration i, $\boldsymbol{\nabla}\mathcal{M}({}_{i}\boldsymbol{\Theta})$ is the gradient of our function and α_i is a step-size of the iterations. We define $\mathbf{h}_k = -\mathbf{H}_i^*\boldsymbol{\nabla}\mathcal{M}({}_{i}\boldsymbol{\Theta})$ as the *search direction* of the algorithm, which is a vector that moves the method closer to the optimized value.
Before we can define the approximation to the Hessian used in the BFGS algorithm, we define two vectors

$$\boldsymbol{s}_i = {}_{i+1}\boldsymbol{\Theta} - {}_{i}\boldsymbol{\Theta}, \qquad \boldsymbol{y}_i = \boldsymbol{\nabla}\mathcal{M}({}_{i+1}\boldsymbol{\Theta}) - \boldsymbol{\nabla}\mathcal{M}({}_{i}\boldsymbol{\Theta}).$$

We also note that quasi-Newton methods requires the next approximation matrix $\mathbf{H}_{i+1}^*$ to satisfy the secant equation:

$$\mathbf{H}_{i+1}^* \boldsymbol{y}_i = \boldsymbol{s}_i \tag{5.39}$$

In order to uniquely determine $\mathbf{H}_{i+1}^*$ for the updates, we also impose another condition, which is that among all symmetric matrices which satisfies the secant equation (Equation 5.39), $\mathbf{H}_{i+1}^*$, is closest to the previous-step matrix $\mathbf{H}_i^*$, i.e. we solve the following:

$$\min_{\mathbf{H}^*} \|\mathbf{H}^* - \mathbf{H}_i^*\| \tag{5.40}$$

subject to $\mathbf{H}^* = (\mathbf{H}^*)^\intercal$. $\|\cdot\|$ is the norm of a function. In addition, we require the approximate inverse Hessian to be positive definite in order for the method to be conclusive. We can ensure this by imposing another condition, given by

$$\boldsymbol{s}_i^\intercal \boldsymbol{y}_i > 0. \tag{5.41}$$

As we have these conditions in place, Equation (5.40) has an unique solution given by: (Nocedal Wright, 2006)[45]

$$\mathbf{H}_{i+1}^* = (I - \rho_i \boldsymbol{s}_i \boldsymbol{y}_i^\intercal)\mathbf{H}_i^*(I - \rho_i \boldsymbol{y}_i \boldsymbol{s}_i^\intercal) + \rho_i \boldsymbol{s}_i \boldsymbol{s}_i^\intercal \tag{5.42}$$

where $\rho_i = (\boldsymbol{y}_i^\intercal \boldsymbol{s}_i)^{-1}$.
To pass the positive definite property over from $\mathbf{H}_i^*$ to $\mathbf{H}_{i+1}^*$, Quasi-newton methods utilizes Wolfe line-search that needs to satisfy the *Wolfe conditions* in order to determine the step-size α_i, which ensures a positive curvature at each iteration, i.e. Equation (5.41)(Wolfe, 1969)[53]. In Wolfe line-search, the idea is to minimize $\mathcal{M}(\cdot)$ by solving the sub-problem

$$\min_{\alpha_i} \mathcal{M}({}_i\boldsymbol{\Theta} + \alpha_i \mathbf{h}_i). \tag{5.43}$$

We recall that $\mathbf{h}_i$ is the search direction at iteration i. So this is the optimization of ${}_{i+1}\boldsymbol{\Theta}$ as defined in Equation (5.38).The Wolfe conditions, which need to be met in order to pass on the positive definiteness, are given as follows:

$$\mathcal{M}({}_i\boldsymbol{\Theta} + \alpha_i \mathbf{h}_i) \leq \mathcal{M}({}_i\boldsymbol{\Theta}) + c_1 \alpha_i \mathbf{h}_i^\intercal \boldsymbol{\nabla} \mathcal{M}({}_i\boldsymbol{\Theta}) \tag{5.44}$$

$$-\mathbf{h}_i^\intercal \boldsymbol{\nabla} \mathcal{M}({}_i\boldsymbol{\Theta} + \alpha_i \mathbf{h}_i) \leq -c_2 \mathbf{h}_i^\intercal \boldsymbol{\nabla} \mathcal{M}({}_i\boldsymbol{\Theta}, \tag{5.45}$$

for some values $0 < c_1 < c_2 < 1$. Based on these conditions, we can find values for α_i used in the estimation process.

The starting value of our Hessian, $\mathbf{H}_0^*$, can be chosen in several ways, e.g. as the identity matrix, or calculating an approximate Hessian by finite differences at ${}_0\boldsymbol{\Theta}$. We are now ready to define the BFGS-algorithm's iterative scheme:

Algorithm 1: BFGS

Initial values: starting parameter vector ${}_0\boldsymbol{\Theta}$, convergence tolerance $\epsilon > 0$,
starting inverse Hessian approx. $\mathbf{H}_0^*$
$i \leftarrow 0$;
while $\|\mathcal{M}({}_i\boldsymbol{\Theta})\| > \epsilon$ **do**
Compute search direction
$\boldsymbol{h}_i = -\mathbf{H}_i^* \boldsymbol{\nabla} \mathcal{M}({}_i\boldsymbol{\Theta})$
Decide α_i *from a line-search satisfying the Wolfe conditions (Eq. 5.44)*
Update ${}_{i+1}\boldsymbol{\Theta} = {}_i\boldsymbol{\Theta} + \alpha_i \mathbf{h}_i$
Define $\boldsymbol{s}_i = {}_{i+1}\boldsymbol{\Theta} - {}_i\boldsymbol{\Theta}$
Define $\boldsymbol{y}_i = \boldsymbol{\nabla} \mathcal{M}({}_{i+1}\boldsymbol{\Theta}) - \boldsymbol{\nabla} \mathcal{M}({}_i\boldsymbol{\Theta})$.
Compute $\mathbf{H}_{i+1}^*$ *by 5.42*
i++
end

Applying this algorithm to our negative log-likelihood yields a fairly fast and computationally effective estimate for the parameter vector ${}_i\boldsymbol{\Theta}$, and thus we conclude the optimization process.

5.4 Model comparison

The biggest downside of introducing more regimes in our model is, of course, that there are more parameters to estimate. A single regime GARCH model that assumes a normal distribution only has *three* parameters to be estimated (α_0, α_1 & β_0). A two-regime gjrGARCH model which follows a skew-Student-t distribution, however, has 14 parameters to be estimated ($\alpha_{0,1}, \alpha_{0,2}, \alpha_{1,1}, \alpha_{1,2}, \alpha_{2,1}, \alpha_{2,2}, \beta_{0,1}, \beta_{0,2}, \nu_1, \nu_2, \xi_1$, $\xi_2, p_{1,1}$ and $p_{2,1}$). This disparity in number of parameters could be affect our results. Specifically, more parameters could lead to overfitting, and therefore affecting the ability of the model to make accurate predictions. In general, the more parameters in a model, the better the model fits the data it has been fitted on, but the probability of higher prediction error increases. We need a way to compare our models, and the Akaike Information Criterion will be our main source of model selection criteria. AIC is an estimator that admits the sample log-likelihood, as well as the number of parameters in the model as a penalty. Subsequently, it returns a value which can be used in comparison with other models. Lower AIC implies a better model fit, however it must be noted that a better AIC value does not necessarily mean that the model is better, as there are several other caveats that one needs to consider. In this thesis, however, AIC will be used as the main basis for comparing models. AIC is defined as follows, for our case:

$$\text{AIC} = 2z - 2log\mathcal{L}(\boldsymbol{\Theta}), \tag{5.46}$$

where z is the number of parameters in the model.

In addition to the goodness-of-fit testing that AIC yields, we also want to estimate some tail risk measures for the data, as these values compared to their empirical counterparts provide some valuable insight on the data. Section (6) proves an overview of some risk measures and their properties.

5.5 Prediction

An important aspect of this thesis is to be able to simulate an h-step-ahead prediction from the models, as these predictions will be the basis for computing the Value-at-Risk and Expected Shortfall risk measures. We recall from section 5.2 that we used the Hamilton filter in order to get the filtered probabilities of being in state i at time t conditioned on the history of the time series until time t, i.e. $z_{i,t|t} = \mathbb{P}(s_t = i \mid \boldsymbol{\Theta}; \Omega_t)$. In order to be able to compute h-step ahead

predictions, we need to define an expression for the prediction probabilities, i.e. $z_{i,t+1|t} = \mathbb{P}(s_{t+1} = i \mid \boldsymbol{\Theta}; \Omega_t)$, which is the probability of being in state i at time $t+1$, given the information of the observations until time t, i.e. the *one-step-ahead* prediction probabilities. Hamilton (1994, section 22.4)[34] formulates the expression for this as follows:

$$\hat{\boldsymbol{z}}_{\boldsymbol{t+1|t}} = \mathbf{P}\hat{\boldsymbol{z}}_{\boldsymbol{t|t}}, \tag{5.47}$$

where $\hat{\boldsymbol{z}}_{\boldsymbol{t|t}}$ is the Hamilton filter-vector from Equation (5.22) and $\mathbf{P}$ is the transition probability matrix.

In order to obtain a prediction for the model at time $T+1$ we use the prediction probabilities in conjunction with random samples of both the states and the innovations. Sampling of the state at time $T+1$ is simply the process of extracting a random state s^*_{T+1}, which is determined based on the prediction probabilities from Equation (5.47).

Sampling innovations takes a random value from the standardized innovations function $\eta_{s_t,t}$, which has its shape parameters determined by the sample state s^*_{T+1}, i.e. drawing deviations from specification:

$$\eta_{s_{T+1},T+1} \mid (S_t = s^*_{T+1}, \mathcal{I}_T) \sim \mathcal{D}\left(0, 1, \boldsymbol{\zeta}_{s^*_{T+1}}\right). \tag{5.48}$$

Finally, the prediction of the variable in question y_t is determined by Equation (5.2), so the one-step ahead prediction of the model is defined by:

$$y_t \mid (S_t = s^*_{T+1}, \mathcal{I}_T) \sim \mathcal{D}\left(0, h^*_{s_{T+1},T+1}, \boldsymbol{\zeta}_{s^*_{T+1}}\right), \tag{5.49}$$

where $h^*_{s_{T+1},T+1}$ is random samples of the conditional variance equation, where the parameters are decided by the sampled state s^*_{T+1}.

This entire process can be repeated as many times as desired, and the resulting vector of one-step ahead predictions helps us in making valuable observations, and more specifically allows us to see the effect of conditional variance and how the state-dependence affects the model. In our case, the resulting vector is used in order to find the one-step-ahead Value-at-Risk and Expected Shortfall, which we will be discussing in the next section.

6 Risk Measures

The theory of measuring risk is a field that has attracted much interest in the fields of economics, finance, insurance, banking and mathematics. More recent discussion in these fields has examined how to obtain a more realistic measure of risk when one takes into consideration the effect of the system which each institution is a part of, called *systemic* risk. Risk measures are helpful in many applications, as they are used for better understanding of data and can assist in financial decision-making. Some measures are also used in order to quantify a reserve requirement, e.g. the Basel Committee's minimum capital requirement function for banks' credit risk is based on a risk measure called the Expected Shortfall (ES).
A risk measure ρ has certain advantages when it satisfies some conditions in order to become *coherent* or *elicitable*. Sections (6.1 & 6.2) covers these topics. In section (6.3), we introduce the univariate risk measures Value-at-Risk and Expected Shortfall, as well as Entropic VaR (EVar). Section (6.4) covers some multivariate risk measures, and section (6.5) contains a discussion the applicability of these methods to insurance data.

6.1 Coherency

A coherent risk measure is a risk measure that satisfies a set of four axioms which reflect realistic properties of a risk measure. It is desirable that risk measures behave in accordance to the properties that corresponds to how risk typically arises. Artzner et al. (1999)[8] suggested that introducing coherency axioms to a risk measure upholds this desirable behavior. These properties are *monotonicity, subadditivity, positive homogeneity*, and *translational invariance*. The following is a presentation of these axioms.
Consider a set $\mathscr{G}$ of real-valued functions defined on an appropriate probability space. Let $\rho : \mathscr{G} \to \mathbb{R}$ be a functional that is said to be a coherent risk measure if the stated axioms are satisfied.
Translational invariance ensures that adding (subtracting) a sure amount β to our initial portfolio X, decreases (increases) the risk measure by β:

$$\rho(X + \beta) = \rho(X) - \beta$$

Subadditivity is a natural requirement that, when upheld, states that the combined risk of two portfolios can not exceed the risk of the two portfolios on their own. This is closely connected with the principle of diversification. For all X_1 and $X_2 \in \mathscr{G}$:

$$\rho(X_1 + X_2) \leq \rho(X_1) + \rho(X_2)$$

Positive homogeneity for a risk measure refers to the risk of a portfolio being proportional to the size of the portfolio. For all $\lambda \geq 0$ and all $X \in \mathscr{G}$:

$$\rho(\lambda X) = \lambda \rho(X)$$

Monotonicity ensures that portfolios with greater future returns are less risky, or at least not more risky than portfolios with lesser future returns. For all $X_1, X_2 \in \mathscr{G}$ with $X_1 \leq X_2$:

$$\rho(X_2) \leq \rho(X_1)$$

6.2 Elicitability

The desirable and realistic behavior of a risk measure was the topic of coherency as a framework. Being able to score and compare different methods is also of interest. Elicitability is a mathematical property that finds its roots in decision theory, e.g. in Savage (1971)[46], however formalized by Gneiting (2011)[29]. A risk measure ρ is defined as elicitable if it can be defined by an expected scoring function. Elicitabiliy is a desirable property for risk measures because it is closely connected with the ability for backtesting, which is the process of periodically comparing forecast risk measures with realized values of the variable of interest in order to try and assess the accuracy of the forecasting. A risk measure ρ is called elicitable if there exists a scoring function S, such that for any $F \in \mathcal{F}$, the expected value $\mathbb{E}[S(x, Y)]$, where Y is a random variable that follows distribution F, takes its unique minimum at $x = \rho(F)$:

$$\rho = \arg\min_x \mathbb{E}[S(x, Y)]$$

Gneiting (2011)[29] finds that the ES (expected shortfall) is actually not elicitable, while VaR (Value at risk) is. This can prove to become an issue when it comes to attempts at model selection, estimation, forecast comparison and forecast ranking. Backtesting becomes a more challenging task when the risk measure is non-elicitable, and strict backtesting is impossible. However it has been shown for different measures like ES (Acerbi & Szekely, 2014)[1] that approximate backtesting is possible in many cases.

6.3 Univariate risk measures

A risk measure being univariate means that the risk measure only depends on one institution's capital/loss/profit/etc., and is not affected by other exogenous factors. The **VaR** (Value at Risk) is a risk measure formalized by Jorion (2000)[38] that, given a confidence level, returns the minimum loss that can occur in a "worst case scenario", which is defined by a confidence level τ, $\tau \in [0, 1]$. If τ is the confidence level, VaR corresponds to the $1 - \tau$ lower tail level of the distribution of gains/losses

over a given time horizon. When Y is the random variable for the profit/loss function, VaR can be calculated by the inverse distribution function:

$$VaR_\tau = F_Y^{-1}(1-\tau) = \hat{y}^\tau \tag{6.1}$$

In the case of this thesis, we are not considering a profit/loss function, but simply a loss function. In this case, positive values indicate a loss, whereas in the profit/loss case, a positive value indicates profit. So, for the loss function Y, we define $\alpha = 1-\tau$ as the confidence level for a pure loss function. The definition of VaR then becomes:

$$VaR_\alpha = F_Y^{-1}(\alpha) = \hat{y}^\alpha, \tag{6.2}$$

and the focus is now shifted to the right tail of the realized observations. Figure (14) shows the $\alpha = 0.90$ VaR level (5.2847) on a loss function created by a normal distribution with mean 4 and unit variance. For the remainder of section (6), we will be discussing the VaR for the profit/loss function. VaR has been widely used to express risk because it is easy to understand - there is only one number to consider, and it takes into consideration that returns aren't always distributed normally. However, VaR does have its limitations. Namely that yields little information about what we can expect when the losses exceed the VaR, and the fact that it is not a coherent risk measure, as it is not subadditive. So, VaR does not take into consideration that diversifying our portfolio usually means that we lower our risk. VaR has been widely accepted and the most used risk measure for regulatory purposes. The Basel Accords, which are recommendations on banking regulations issued by the Basel Committee on Banking Supervision (BCBS), have in their first and second issue (I and II) used VaR as their preferred risk measure for regulation of banks. This choice went under much discussion, e.g. Acharya et al. (2017)[4] argued that VaR was never meant to be used as a regulatory measure in that it doesn't take into consideration how the returns of financial institutions in the market react to total systemic changes.

ES (Expected Shortfall), also called Conditional Value at Risk (CVaR) or Tail Conditional Expectation (TCE) is a risk measure that is based on the VaR. Contrary to VaR, ES has the desirable attribute of being coherent, see Acerbi and Tasche (2002)[2]. ES is, given confidence level τ, the expected value of the profit/loss function Y truncated below the $VaR_\tau = \hat{y}^\tau$. Define the risk measure

$$ES_\tau = \mathbb{E}[Y|Y \leq \hat{y}^\tau] \tag{6.3}$$

Equivalently, for the loss function with confidence level α, the ES is defined as follows:

$$ES_\alpha = \mathbb{E}[Y|Y \geq \hat{y}^\alpha] \tag{6.4}$$

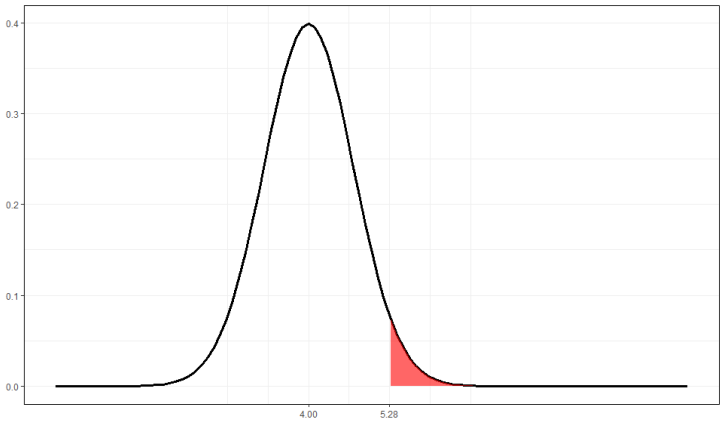


Figure 14: VaR for a loss distribution at confidence level $\alpha = 0.90$. The expectation over the values of x within the red area is the $ES_{\alpha=0.90}$

This measure effectively answers the question "What is the expected return in the $(1-\tau)100\%$ of worst cases?". The expectation over the values of x within the red area is the $ES_{\alpha=0.90}$
In the latest issue of the Basel Accords, Basel III [10], expected shortfall has taken the place of VaR as the preferred regulatory risk measure.

EVaR (Entropic Value at Risk) is a measure introduced by Ahmadi-Javid (2012)[5] that is defined as an upper bound of the VaR and the ES, obtained by means of using the Chernoff inequality. Ahmadi-Javid (2012)[5] discusses that there is a need for EVaR because of the lack of coherency (subadditivity) of VaR, as well as the difficulty to efficiently compute VaR and ES. The chernoff inequality, by Chernoff (1952)[18], for a constant a and a random variable Y for which the moment-generating function $M_Y(z) = \mathbb{E}(e^{zY})$ exists for all $z \in \mathbb{R}$ is defined by:

$$\mathbb{P}(Y \geq a) = \mathbb{P}(e^{zY} \geq e^{za}) \leq \frac{M_Y(z)}{e^{za}}, \forall z > 0,$$

Define $\tau \equiv \frac{M_X(z)}{e^{za}}$ and solve for a. with a confidence level $\tau \in [0, 1]$:

$$a_Y(\tau, z) \equiv z^{-1} \ln\left(\frac{M_Y(z)}{\tau}\right)$$

Here, for each $z > 0, a_Y(\tau, z)$ is an upper bound for $\text{VaR}_{1-\tau}(Y)$. Entropic value at risk of Y for confidence level τ can now be defined as

$$\text{EVaR}_\tau(Y) \equiv \inf_{z>0}\{a_Y(\tau, z)\} = \inf_{z>0}\left\{z^{-1} \ln\left(\frac{M_Y(z)}{\tau}\right)\right\},$$

which is the "tightest possible upper bound that can be obtained from the Chernoff inequality" (Ahmadi-Javid, 2012[5]). Furthermore, he shows that EVaR is an upper bound for both the VaR and the CVaR, making it a more risk-averse measure than

the other two, which may make it less desirable to use for institutions that do not wish to allocate more funds than necessary. However, the fact that EVaR is computationally tractable in more cases than CVaR, and because it has the desirable property of being coherent could make it a good contestant to VaR and ES.

6.4 Multivariate Risk Measures

Multivariate risk measures are a widely discussed topic, as they can help better calculate risks, since they take into consideration more than the institution in question. This can effectively mean that the risk measure takes into consideration J other institutions, or it might take into consideration how the system(e.g. the market) itself affects the risk of the institution in question. In this thesis we do not include the implementation of these multivariate risk measures to data, as this is beyond the scope of the thesis, and the implementation requires some additional data which we do not have access to for the insurance loss dataset we are considering. Nevertheless, some multivariate risk measures are still presented here, as it could, in theory, be applied to insurance data. Section (6.5) contains a short discussion on this choice.

CoVaR (Conditional Value at Risk) was introduced by Adrian and Brunnermeier (2016)[50], and measures a single financial institution's contribution to systemic risk. $\Delta CoVaR$ is the difference between the CoVaR conditional on the institution being in distress and the CoVaR condtional of the institution's median state. For two confidence levels $\tau_j \in (0,1)$ for $j = 1, 2$, CoVaR can be defined implicitly by

$$\mathbb{P}(Y_i \leq CoVaR_{i|j}^{\tau_1|\tau_2} \mid Y_j = VaR_j^{\tau_1}) = \tau_2$$

where Y_i and Y_j are the random variables corresponding to the profit/loss of institutions i and j, and $VaR_j^{\tau_1}$ is the "univariate" Value at Risk for institution j. In some cases, the institution i is instead denoted as a comprehensive index that corresponds to the whole financial system. We say that there is a $100\tau_2\%$ chance that the institution i (or the financial system) is less than the $CoVaR_{i|j}^{\tau_1|\tau_2}$ within a given time frame, given that the returns of institution j is at its $100\tau_1\%$ VaR-level. CoVaR gives us an important edge over VaR because it takes into consideration the fact that the situations an institution find themselves in often has an effect on other institutions, and vice versa. This difference allows us to analyze events like financial crises, where tail-dependency between an institution and its financial system as a whole is a measure of great interest for historical and predictive analysis. VaR, on the other hand, focuses on the risk of an individual institution in isolation, and largely neglects its connection to systemic risk.

Since we have the definition of CoVaR, it is natural to also look at the **CoES** (Conditional Expected Shortfall). $CoES_{i|j}^{\tau_1|\tau_2}$ is defined as the expected returns for institution i, conditional on said returns being less than $CoVaR_{i|j}^{\tau_1|\tau_2}$:

$$CoES_{i|j}^{\tau_1|\tau_2} = \mathbb{E}(Y_i \mid Y_i \leq CoVaR_{i|j}^{\tau_1|\tau_2})$$

or as defined in Bernardi, Maruotti & Petrella (2017)[11], the expected return of institution i conditional on said return being less than its τ_2-level VaR, as well as the return of institution j equalling its "univariate" expected shortfall $ES_j^{\tau_1}$. The last condition corresponds to institution j being in distress.

$$CoES_{i|j}^{\tau_1|\tau_2} = \mathbb{E}(Y_i \mid Y_i \leq VaR_{i,\tau_2}, Y_j = ES_j^{\tau_1})$$

In order for CoVaR and CoES to measure how much of the risk institution j contributes to institution i (or the financial system) when institution j is in distress, Adrian and Brunnermeier (2016)[50] introduces $\Delta CoVaR$ and $\Delta CoES$. These measures are defined by the difference between the CoVaR (CoES) conditional on institution j being in distress and the CoVaR (CoES) of institution i conditional on the institution j being at its median state. Defined by

$$\Delta CoVaR_{i|j}^{\tau_1|\tau_2} = CoVaR_{i|j}^{\tau_1|\tau_2} - CoVaR_{i|j}^{50|\tau_2}$$

$$\Delta CoES_{i|j}^{\tau_1|\tau_2} = CoES_{i|j}^{\tau_1|\tau_2} - CoES_{i|j}^{50|\tau_2}$$

$\Delta CoVaR$ and $\Delta CoES$ are directional, which means that swapping institutions does not necessarily return the same value, e.g. switching from $\Delta CoVaR_{i|j}^{\tau_1|\tau_2}$ to $\Delta CoVaR_{j|i}^{\tau_2|\tau_1}$ where i is the whole financial system. In this case, the interpretation of $\Delta CoVaR$ is completely different in that the question changes from "How does the fact that institution j is in economic distress affect the risk of the whole system?" to "How does it affect institution j's returns that the whole financial system is in distress?"

MES (Marginal Expected Shortfall) (Acharya et al. 2017 [4]) was developed as a way to analyze systemic risk of an economy. By looking at the whole economy's returns as a weighted sum of the individual institution's returns, where each institution i's weight ω_i holds information about the influence it has over the economy. The returns of the system Y gets decomposed into the sum of each of the institutions' returns y_i, so that $Y = \sum_i \omega_i y_i$. From the definition and properties of expected shortfall we get the ES for the whole sum

$$ES_\tau = \sum_i \omega_i \mathbb{E}[y_i \mid Y \leq \hat{y}^\tau]$$

and taking the change in overall ES with respect to the weight of each institution i

$$MES_\tau^i \equiv \frac{\partial ES_\tau}{\partial \omega_i} = \mathbb{E}[y_i \mid Y \leq \hat{y}^\tau]$$

is called the marginal expected shortfall (MES) of institution i. This measures the marginal contribution of the institution i to systemic risk. Acharya et al. (2017)[4] explain that they saw the need for a risk measure that took into consideration effects on systemic risk, particularly because the regulatory precedent at the time set by the Basel I and II frameworks did not.

Long-run MES (**LRMES**) was introduced by Acharya, Engle & Richardson (2012)[3] as the mean of the returns of the institution over a given time (typically six months), restricted to only the cases where the system index drops below a certain crisis-threshold C. $\text{LRMES}_{i,t}$ is the long-run MES of institution i over a given time interval h, defined as

$$\text{LRMES}_{i,t} = \mathbb{E}_t(Y_{i,t+1:t+h} \mid Y_{m,t+1:t+h} < C)$$

where $Y_{m,t+1:t+h}$ is the return of the market between times $t+1$ and $t+h$, and $Y_{i,t+1:t+h}$ is the return of institution i over the same time interval.

SRISK was discussed by Brownlees and Engle (2016)[14], and is defined as the expected shortfall on the capital of a financial institution conditioned on a systemic event, which is mostly a prolonged market decline.
Capital shortfall is here, for institution i of N total financial institutions in the system, defined as the capital reserves the institution needs to hold depending on its regulation minus the institution's equity. For time t, capital shortfall is defined by Brownlees and Engle (2016)[14] as

$$\text{CS}_{i,t} = k\text{A}_{i,t} - \text{W}_{i,t} = k(\text{D}_{i,t} + \text{W}_{i,t}) - \text{W}_{i,t}$$

where, for institution i at time t, $\text{W}_{i,t}$ is the market value of equity, $\text{D}_{i,t}$ is the book value of debt, $\text{A}_{i,t}$ is the value of quasi-assets and k is the capital fraction of the assets that the institution can use, depending on regulatory or prudential decisions. When this value is positive (there is capital shortfall), the institution is experiencing financial distress. Finding the expectation of this value given a systemic event is the aim of SRISK. A systemic event is defined as a total market decline below a given threshold C over a given time period h, for this risk measure. SRISK is then defined, first symbolically by Acharya et al. (2012)[3] as

$$\text{SRISK}_{i,t} = \mathbb{E}_{t-1}(\text{Capital Shortfall}_i \mid \text{Crisis})$$

and formally as

$$\begin{aligned}\text{SRISK}_{i,t} &= \mathbb{E}_t(\text{CS}_{i,t+h} \mid Y_{m,t+1:t+h} < C) \\ &= k\mathbb{E}_t(D_{i,t+h} \mid Y_{m,t+1:t+h} < C) - (1-k)\mathbb{E}_t(W_{i,t+h} \mid Y_{m,t+1:t+h} < C)\end{aligned}$$

We assume that the debt remains constant over the given period of incurred crisis, so $\mathbb{E}_t(D_{i,t+h} \mid Y_{m,t+1:t+h} < C) = D_{i,t}$. It is normal during crises that one cannot renegotiate debt, and the amount of short-term loans decreases. From this assumption, it follows that

$$\begin{aligned}
\text{SRISK}_{i,t} &= kD_{i,t} - (1-k)\mathbb{E}_t(W_{i,t+h} \mid Y_{m,t+1:t+h} < C) \\
&= kD_{i,t} - (1-k)W_{i,t}\mathbb{E}_t\left(\frac{W_{i,t+h}}{W_{i,t}} \mid Y_{m,t+1:t+h} < C\right) \\
&= kD_{i,t} - (1-k)W_{i,t}(1 + \text{LRMES}_{i,t}) \\
&= W_{i,t}(k\text{L}_{i,t} - (1-k)\text{LRMES}_{i,t} - 1)
\end{aligned}$$

where $L_{i,t}$ is the quasi-leverage ratio $\frac{D_{i,t}+W_{i,t}}{W_{i,t}}$of institution i at time t.
SRISK differs from other risk measures because it does not only depend on the volatility, correlation or other moments of equity return for an institution, but it also depends on the size and degree of leverage the institution holds. It also depends on the long-run expected shortfall over a given time period conditional on a market decline. The measure can provide a prediction of the degree of capital shortfall an institution would experience in the case of a systemic event.

In order to use SRISK as a way to measure total systemic risk for the entire financial system, we sum the positive values of all individual institutions in the system's SRISKs

$$\text{SRISK}_t = \sum_{i=1}^{N} (\text{SRISK}_{i,t})_+.$$

Sometimes, it is beneficial to look at the share each institutions systemic risk has on the financial system, which can be defined as the percentage SRISK:

$$\text{SRISK}\%_{i,t} = \frac{\text{SRISK}_{i,t}}{\text{SRISK}_t}, \text{ for } \text{SRISK}_{i,t} > 0$$

Brownlees and Engle (2012)[14] argue that SRISK is a superior risk measure to the MES-framework developed by Acharya et al. (2017)[4] because SRISK can be estimated without structural assumptions and it does not require the observation of a realization of a systemic crisis, while MES does. Thus, SRISK improves on MES in that it can be used for ex-ante analysis.
Another way SRISK differs from many other risk measures is that it does not only depend on moments of the returns of the institutions in the market, but it also explicitly takes into consideration exogenous variables about the institutions. Namely, it depends on the size and the degree of leverage the institutions have on the financial system.

MCoVaR Multiple-CoVaR (Bernardi et al. 2017[11]) is a generalization of CoVaR in that it takes into account interconnections among all market participants, whereas CoVaR is a pairwise measure that only explains the risk of an institution/the market conditional on one other institution/the market being in distress. During financial crisis, it is likely that more than one institution is experiencing instances of distress. Bernardi et al. (2017)[11] thus saw the need for MCoVaR and MCoES which takes into consideration that more than one institution can be in distress, and more than one institution can be at the median state. Here, we have a set of p total institutions $\mathcal{S} = \{1, 2, \ldots, p\}$, and the set of d institutions in distress is $\mathcal{J}_D = \{j_1, j_2, \ldots, j_d\} \subset {}_d\mathbb{C}_{p-1}, d \leq p-1$, where ${}_d\mathbb{C}_{p-1}$ is the set denoting all possible combinations of $p-1$ elements from d. $\mathcal{J}_N = \overline{\mathcal{J}_D}$ is the set of all institutions being at their median state. We then define the $\text{MCoVaR}^{\tau_1|\tau_2}_{i|\mathcal{J}_D}$ as the multiple CoVaR, where $\tau_j \in (0, 1)$ are confidence levels for $j = 1, 2$. This can be interpreted as the value at risk of institution $i \in \mathcal{S}$ at confidence level τ_1, conditional on the set of distress-institutions $\mathcal{J}_D$ being at their individual VaR_{τ_2}-level, and the set of median-level-institutions $\mathcal{J}_N$ being at their median $VaR_{0.5}$-level. The vector of returns for the p total institutions is $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_p)$. The VaR-values corresponding to institutions being in distress are given by $\hat{\mathbf{y}}^{\tau_2}_{\mathcal{J}_D} = (\hat{y}^{\tau_2}_{j_1}, \hat{y}^{\tau_2}_{j_2}, \ldots, \hat{y}^{\tau_2}_{j_d})$, and the VaR-values corresponding to institutions being in their median state are given by $\hat{\mathbf{y}}^{0.5}_{\mathcal{J}_N} = (\hat{y}^{0.5}_{d+1}, \hat{y}^{0.5}_{d+2}, \ldots, \hat{y}^{0.5}_{p-1})$. MCoVaR is then defined implicitly by

$$\mathbb{P}(Y_i \leq \text{MCoVaR}^{\tau_1|\tau_2}_{i|\mathcal{J}_D} \mid \mathbf{Y}_{\mathcal{J}_D} = \hat{\mathbf{y}}^{\tau_2}_{\mathcal{J}_D}, \mathbf{Y}_{\mathcal{J}_N} = \hat{\mathbf{y}}^{0.5}_{\mathcal{J}_N}) = \tau_1,$$

for $i = 1, 2, \ldots, p$

Due to the lack of coherency(subadditivity) in VaR and CoVaR, it is natural to extend the scope of MCoVaR to **MCoES**. MCoES is defined very similarly, and can be viewed as a generalization of CoES (Adrian and Brunnermeier, 2016) that takes the pairwise measure and extends it to include several, or all, institutions in the market. We wish to define the multivariate expected shortfall that is conditional on the distress-institutions $\mathcal{J}_D$ being at their marginal distress-ES_{τ_2}-level $\hat{\psi}_{\mathbf{y}\mathcal{J}_D}(\hat{\mathbf{y}}^{\tau_2}_{\mathcal{J}_D}) = \left(\hat{\psi}_{yj_1}(\hat{y}^{\tau_2}_{j_1}), \hat{\psi}_{yj_2}(\hat{y}^{\tau_2}_{j_2}), \ldots, \hat{\psi}_{yj_d}(\hat{y}^{\tau_2}_{j_d})\right)$, and the median-institutions being at their marginal median-$ES_{0.5}$-level $\hat{\psi}_{\mathbf{y}\mathcal{J}_N}(\hat{\mathbf{y}}^{0.5}_{\mathcal{J}_N}) = \left(\hat{\psi}_{yj_{d+1}}(\hat{y}^{0.5}_{j_{d+1}}), \ldots, \hat{\psi}_{yj_{p-1}}(\hat{y}^{0.5}_{j_{p-1}})\right)$. The MCoES for institution $i = 1, 2, \ldots p$ given confidence level τ_1 and returns $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_p)$ is then given by

$$\text{MCoES}^{\tau_1|\tau_2}_{i|\mathcal{J}_D} = CoES_{\tau_1}\left(Y_i \mid \mathbf{Y}_{\mathcal{J}_D} = \hat{\psi}_{\mathbf{y}\mathcal{J}_D}(\hat{\mathbf{y}}^{\tau_2}_{\mathcal{J}_D}), \mathbf{Y}_{\mathcal{J}_N} = \hat{\psi}_{\mathbf{y}\mathcal{J}_N}(\hat{\mathbf{y}}^{0.5}_{\mathcal{J}_N})\right).$$

The introduction of MCoVaR and MCoES-measures extend the perspective of systemic risk measures entirely. The risk measures do not only consider how one institution reacts to a systemic event, or conversely how one institution's leverage

affects the degree of a systemic event in the market. Bernardi et al. (2017)[11] finds that, when combining these measures for all p institutions into an overall systemic risk indicator that attributes the risk to the individual institutions, the resulting measure does explains movement during financial crises well when applied to market participants of the Standard & Poor's 500 index (S&P500). MCoVaR and MCoES do not, however, take into consideration other exogenous information like how SRISK utilizes the size and leverage of the institutions.

6.5 Application to insurance data

In the section covering empirical analysis of data (7), we only consider and compute the Value at Risk and Expected Shortfall risk measures. There are several reasons behind this choice. Firstly, we are only considering a single univariate time series, and the multivariate risk measures can only be computed when there are more similar institutions, or some measure of the system the institutions belong to, that exhibits similar data in the same time period as our institution. We still choose to include the presentation of the different multivariate risk measures, as we can imagine a case where it would apply to insurance data. For example, imagine a scenario where we had access to several time series from different insurance companies that contain losses associated with fire damages, where all data is from the same country over the same time period. In the case of some large-scale wildfire, we imagine that the increase of insurance claims paid out by institution i, who is closer to the fire, could affect the future risk associated with the claims of institution j, who is farther away. Another reason for not calculating the multivariate risk measures is that the VaR and ES are the most widely used risk measures, with the simplest interpretation and generally has the highest applicability. VaR and ES-values are also computed in the paper of Eling (2012), which gives grounds for comparison of these risk measures. Lastly, several of the multivariate risk measures are beyond the scope of this thesis to calculate, as there are no readily available programming-methods available to the public.

7 Data & Results

In the theoretical analysis of this thesis, time series models have been introduced, and the (G)ARCH model has been extended to the more advanced Markov-switching GARCH model, which is expected to be able to capture more intricate behaviors in our time series as it is a more flexible method. In particular, we expect the MSGARCH-extension to be able to capture the behavior of the shocks in the data better than the single-regime (G)ARCH-model. Introducing the ability to alter the structure of the conditional volatility, as well as different specifications of the conditional distribution should further allow us to improve our model fit. In this section, these claims will be put to the test, as the several possible models will be applied to a dataset, and then these models will be compared in performance. We will also be comparing our results with the results of another paper, "*Fitting insurance claims to skewed distributions: Are the skew-normal and skew-student good models?*" by Eling (2012)[21]. Eling's paper attempts to fit skew-normal and skew-Student-t distributions to the same dataset which is used in this thesis, and concludes that the two models are reasonably good models compared to 18 other benchmark models. Comparison to this paper can give us some valuable insight into whether or not using Markov-switching GARCH models on insurance datasets can yield better model fit than simply fitting the data to a distribution. The most important distinction between the methods applied here and in Eling's paper is that Eling's paper does not take the aspect of time into consideration for fitting the models, while this thesis views the insurance losses as a time series. Section 7.1 contains a presentation and discussion of our dataset, section 7.2 discusses the R-package `MSGARCH` and section 7.3 contains the construction and comparison of the models.

7.1 Data

The data used in this thesis is comprised of Danish fire reinsurance losses, which was first presented in Embrechts, Klüppelberg & Mikosch (1997)[22]. In section (2.3), this dataset was also used in the discussion of insurance time series, and in this section we will formally introduce it. The dataset consists of several variables that comprise different parts of the total individual losses. We are only applying our models to a time series of the total losses. The total losses are 2,167 observations of individual losses above 1 million Danish kroners (DKK), in the time frame January 3rd 1980 to December 31st 1990. The data is adjusted for inflation in order to portray 1985 values. The dataset is available through a great deal of R-packages, including in `CASdatasets`, where the univariate version of the dataset is named `danishuni`. We take a look at the first 6 values of the dataset:

```
> head(danishuni)
        Date     Loss
1 1980-01-03 1.683748
2 1980-01-04 2.093704
3 1980-01-05 1.732581
4 1980-01-07 1.779754
5 1980-01-07 4.612006
6 1980-01-10 8.725274
```

The first thing we notice is that each entry is one individual claim, and therefore the `Date`-variable does not have constant increments between each observation, and there may even be more than one observation on the same date. This, however, is not necessarily a problem, as none of the methods presented are severely affected by non-constant time increments. Figure (15) displays the time series of the data. When discussing the applicability of conditional structure models in section

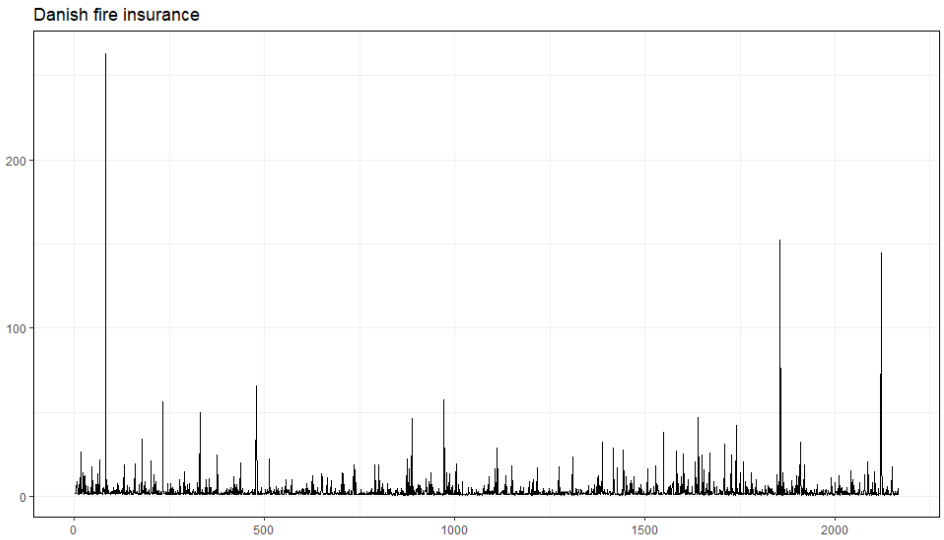


Figure 15: This figure shows the original Danish fire insurance losses data from January 3rd 1980 to December 31st 1990

(2.2.2), we recall that there are two major models to be considered, i.e. ARMA and (G)ARCH. A good pointer for whether we should use ARMA, GARCH or both is to examine serial correlation for the original data, as well as the log-data. The sample autocorrelation and sample partial autocorrelation of the original and log data is shown in Figure (16) This figure shows little evidence, if any, to prove that there exists serial correlation in the non-squared dataset. We therefore do not believe that fitting an ARMA-model to our data will significantly improve our model fit. Recall that the example French insurance dataset from section 2 did, in fact, show significant serial correlation on the non-squared data. We can therefore not claim anything about whether or not the general insurance loss time series exerts serial

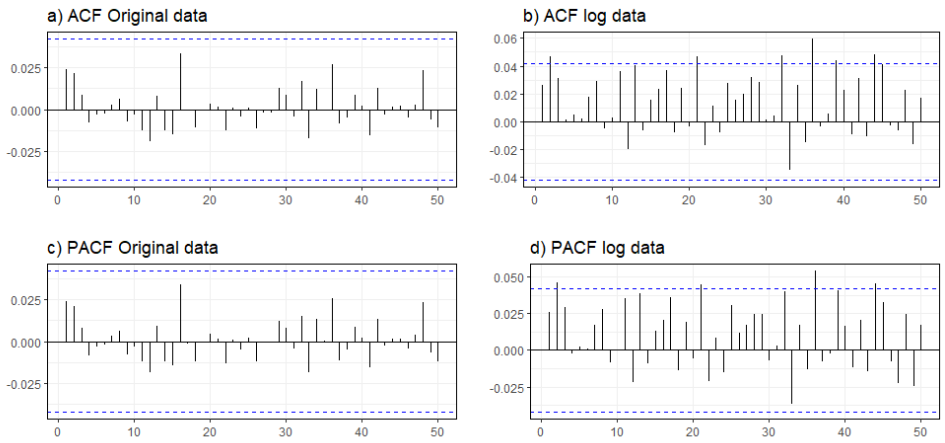


Figure 16: a) & b) is the sample ACF and PACF of the original data. c) & d) is the sample ACF and PACF of the log data. All plots are shown up to 50 lags.

correlation or not. Because of this result for the Danish data, we decide to continue the discussion on the de-meaned data instead of the original data, as we have only defined the (G)ARCH models for the zero-mean case for simplicity. Checking for serial autocorrelation on the squared de-meaned data helps us to see if there exists conditional heteroskedasticity-effects in the data, and therefore allows us to make a decision on if (G)ARCH-models are applicable to our situation. We recall that the McLeod-Li test (Equation 2.14) can give us an idea about the existence of any ARCH effects in the data. The McLeod-Li test on the danish fire data is displayed in Figure (17) below.

This figure actually diminishes the probability of the original data having any heteroskedasticity in the variance, as none of the lags are remotely close to being significant. The log of the de-meaned data, however, does show that the first two lags are significant in the test. Therefore, the original data seems to not show much serial dependence at all, and we do not expect the GARCH-model to give a significant improvement over standard regression used in Eling's paper. The log data might improve on Eling's results, though, as some of the ARCH-effects should be caught by the models. This is the motivation behind using (G)ARCH models for this dataset, although the motivation is not as strong as we see in most financial time series, e.g. the SMI-example from Figure (5).

We now shift our focus to the shape of the distribution of the insurance losses, which is thoroughly discussed in Eling's paper, and the basis for choosing skewed models to model the data. Figure (18 c and d) visualises the histograms of the de-meaned original data, as well as the de-meaned log data.

We see what is usually the case for insurance loss datasets, i.e. a large amount of small claims, with some large shocks that result in a heavy right tail and negative.

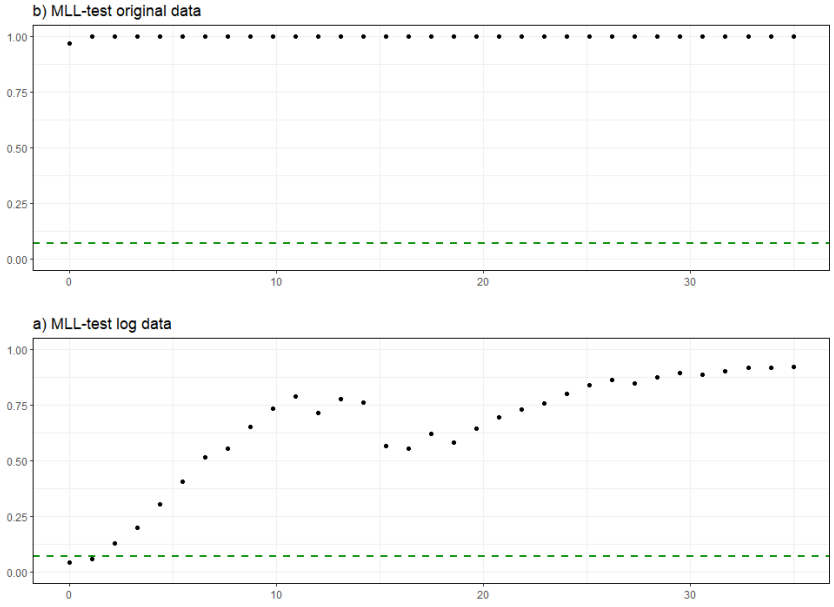


Figure 17: In this figure, the McLeod-Li test for the original data (a), and the log data (b) is displayed. The green dotted line indicates the area of significance.

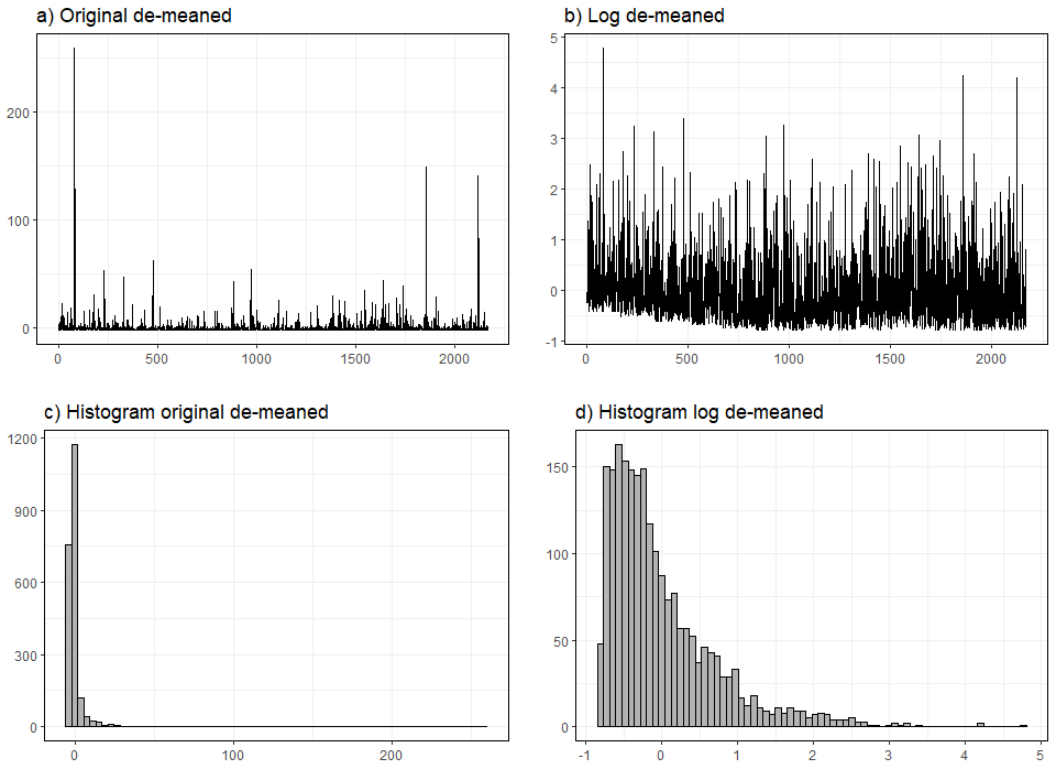


Figure 18: This figure contains a time series representation of the de-meaned Danish fire reinsurance dataset (original a, log b), and the histogram of the same data (original c, log d)

Since the original data is absolutely positive (≥ 1), we obviously observe some positive skewness in the data. The sample skewness and kurtosis of the data is checked:

```
> c(skewness(Total.dm), skewness(log.Total.dm),
    kurtosis(Total.dm), kurtosis(log.Total.dm))
[1]  18.736849   1.761092 482.197970   4.179029
```

Both the original data as well as the log data shows positive skewness, especially for the original data. We also have that kurtosis is larger than for a normal distribution (>3) for both, making the distribution leptokurtic. This is again especially the case for the original data case. These findings, in addition to viewing the example PDFs of normal, skew-normal, Student-t, skew-Student-T, GED and skew-GED from Figure (8 & 9), we start to get a sense that non-skewed models are not suitable for fitting our data. We also do not believe that the skew-normal distribution is able to catch enough of the kurtosis that the insurance dataset displays. The skew-Student-t and the skew-GED distributions, however, are expected to reproduce the data much better.

So far, the applicability of (G)ARCH-models and the conditional distribution herein has been discussed. The conditional variance structure models that was presented in section 3.2.1 will all be utilized when fitting our models, and we expect the *gjrGARCH* and the *tGARCH* models to perform the best, simply because they are both more complex specifications that can account for asymmetry in the conditional variance. We assume that the time series consists of two distinct regimes, one which catches the majority of the small claims, and one which catches the shocks. In addition to this, we will be fitting single-regime GARCH models to examine if extending to the two-regime ($K = 2$) case will improve our model fit. This all culminates into a few questions which we are attempting to answer in the upcoming sections:

- Are we improving our model by taking into consideration the possible ARCH-effects of the data?
- What conditional distribution best fits the data?
- What conditional variance structure best fits the data?
- Does model fitting increase results when assuming two distinct regimes?
- Which model can predict the most realistic tail behavior?

7.2 MSGARCH

The R package `MSGARCH`[1], developed by Ardia, Bluetau, Boudt, Catania & Trottier (2019)[7] will be used to carry out our model specification, model fitting and prediction. The package is quite comprehensive, as it allows for fitting of Markov-switching GARCH(1,1) models, as described by Haas (2004)[32], with a large variety of specifications of the model. The package has no equation for the mean, so having a de-meaned time series as observation variable is a requirement for a competitive model fit. Following is an explanation of the main functions which are used:

`CreateSpec`

```
CreateSpec(
  variance.spec = list(model = c("sGARCH", "sGARCH")),
  distribution.spec = list(distribution = c("norm", "norm")),
  switch.spec = list(do.mix = FALSE, K = NULL),
  constraint.spec = list(fixed = list(), regime.const = NULL),
  prior = list(mean = list(), sd = list())
)
```

This function is used to create the specification which we wish to fit to our model data. `variance.spec` contains a vector of K different specifications of the conditional volatility, which can take values `"sARCH", "sGARCH", "gjrGARCH, "tGARCH"` and `"eGARCH"`, which we do not use. `distribution.spec` contains a vector of K different conditional distribution specifications, which can take values `"norm", "std", "ged", "snorm", "sstd"` and `"sged"`. The i'th entry of this vector corresponds to the i'th entry of the variance specification vector. `do.mix = FALSE` indicates that we are fitting a Markov-switching model instead of a mixture GARCH model. `constraint.spec` is used when we wish to fix, or keep constant, some of the parameters.

`FitML`

```
FitML(spec, data, ctr = list())
```

This function uses a specification created by `CreateSpec` in combination with a vector of the data in order to fit a ML-estimation of the parameter vector, as explained in sections 5.2 & 5.3. The function returns optimized values of the parameters, the transition probability matrix, the stable probabilities of being in a given state, the value of the maximized log-likelihood value and the AIC of the model.

`predict`

```
predict(
  object,
  newdata = NULL,
  nahead = 1L,
```

[1] https://CRAN.R-project.org/package=MSGARCH

```
  do.return.draw = FALSE,
  do.cumulative = FALSE,
  ctr = list(),
  ...
)
```

This is a function that performs h-step-ahead predictions as in section 5.5. `object` is the fitted object from `FitML`, `newdata = NULL` indicates that we are doing predictions based on the original data and `nahead` is the amount of steps-ahead predictions we wish to compute. `do.return.draw = FALSE` indicates that the function does not return a value for y_{T+h}, but only returns the conditional volatility $h_{s_{T+1},T+1}$. In our case, we change this to `TRUE`, as we are interested in predicting one-step-ahead VaR and ES, which is based on the draws. In the `ctr` argument we can define the number n of simulations to be carried out. `nsim` defaults to 10000.

`simulate`

```
simulate(object,
  nsim = 1L,
  seed = NULL,
  nahead = 1L,
  nburn = 500L, ...)
```

This function is only used here for visualization purposes. The function creates n amounts of h-step ahead simulated paths.

7.3 Model fitting & results

In this section, all the mentioned models are fitted for both original data and log data. Four of these are used for visualization purposes, while the key values of all models are compared later in the section. The four example models are:

- the one-regime ARCH with normal distribution
- the two-regime GARCH with Student-t distribution
- the two-regime GARCH with skew-Student-t distribution
- the two-regime tGARCH with skew-GED

Appendix A contains the R code that is used for the two-regime tGARCH model with skew-GED. Note that the same methods are applied to every other considered model.
Tables (3 & 4) contains the values of the transition probabilities, the fitted parameters and the AIC of these four models as they have been fitted through ML-estimation. We do not display the parameters estimations of all the models, as there are a too large amount of parameter estimates to feasibly show them.

ORIGINAL DATA - Comparison of parameters, transition probabilities and AIC				
	MS1 ARCH NORM	MS2 GARCH T	MS2 GARCH S-T	MS2 tGARCH S-GED
P	-	$\begin{bmatrix} 0.9298 & 0.0702 \\ 0.9010 & 0.0990 \end{bmatrix}$	$\begin{bmatrix} 0.9954 & 0.0046 \\ 0.7966 & 0.2034 \end{bmatrix}$	$\begin{bmatrix} 0.9462 & 0.0538 \\ 0.9131 & 0.0869 \end{bmatrix}$
$\alpha_{0,1}$	69.7	3.3674	8.9404	0.1074
$\alpha_{1,1}$	0.0	0.0	0.0	0.0
$\alpha_{2,1}$	-	-	-	0.0001
$\beta_{1,1}$	-	0.0093	0.7943	0.9627
ν_1	-	99.9778	0.0001	0.7083
ξ_1	-	-	6.4530	38.2273
$\alpha_{0,2}$	-	137.5088	2211.8726	29.8576
$\alpha_{1,2}$	-	0.0	0.0025	0.0006
$\alpha_{2,2}$	-	-	-	0.0216
$\beta_{1,2}$	-	0.9167	0.9474	0.0766
ν_2	-	2.1552	2.1001	0.7033
ξ_2	-	-	0.1808	0.7118
AIC	15425.36	10218.26	7212.48	7172.13

Table 3: Values for the probability matrix, the model parameters and AIC corresponding to the four example models which are fitted to the original data.

LOG DATA - Comparison of parameters, transition probabilities and AIC				
	MS1 ARCH NORM	MS2 GARCH T	MS2 GARCH S-T	MS2 tGARCH S-GED
P	-	$\begin{bmatrix} 0.8675 & 0.1325 \\ 0.8815 & 0.1185 \end{bmatrix}$	$\begin{bmatrix} 0.9985 & 0.0015 \\ 0.7907 & 0.2093 \end{bmatrix}$	$\begin{bmatrix} 0.9576 & 0.0424 \\ 0.0076 & 0.9924 \end{bmatrix}$
$\alpha_{0,1}$	0.4994	0.0001	0.0068	0.0075
$\alpha_{1,1}$	0.0267	0.0007	0.0001	0.0001
$\alpha_{2,1}$	-	-	-	0.0580
$\beta_{1,1}$	-	0.9988	0.9817	0.9651
ν_1	-	99.9633	24.9297	0.9427
ξ_1	-	-	91.2613	21.9491
$\alpha_{0,2}$	-	1.8601	108.1350	0.0042
$\alpha_{1,2}$	-	0.9999	0.0009	0.0019
$\alpha_{2,2}$	-	-	-	0.0065
$\beta_{1,2}$	-	0.0	0.4792	0.9914
ν_2	-	99.6136	2.1001	1.0416
ξ_2	-	-	88.1383	12.7597
AIC	4701.98	4377.38	3382.93	3140.56

Table 4: Values for the probability matrix, the model parameters and AIC corresponding to the four example models which are fitted to the log data.

There are many things to examine from these tables. Firstly, the transition probability matrix is actually quite similar in almost all of the models, i.e. very high probability of staying in state 1, and a high probability of returning to state 1 when in state 2. This translates well to what we already assumed about the data, i.e. that the longer periods of low volatility and small claims corresponds to state 1. State 1 is much more abundant than state 2 which should contain the shocks. The only deviation is the two-regime tGARCH skew-GED model for the log data, which has high staying-probability for both states, and especially state 2. The reason behind this is that the volatility regimes are much less apparent in the log data than they are in the original data, so the models have a more difficult time catching the regimes. The effect of the shocks are more or less diminished by taking the log of the data.

For the original data, the conditional volatility intercept $\alpha_{0,2}$ appears, as expected, quite high for all the two-state models, since the sudden shocks are disproportionately large compared to the claims in state 1. $\alpha_{1,j}$, the direct effect of the previous claim on the conditional volatility, is almost zero for most models. The models which have smaller $\alpha_{1,j}$ typically have larger $\beta_{1,j}$, i.e. the persistence of these very small effects is high. For the model with high $\alpha_{1,2}$ (two-regime GARCH T), the corresponding $\beta_{1,2}$ is low, i.e. the persistence of these large effects is low. Both these kinds of combinations of the parameters $\alpha_{1,j}$ and $\beta_{1,j}$ can have similar effects in on end results, i.e. the effect the previous observation has on the conditional volatility is either quite small, or has quite small persistence. Thus, the applicability of (G)ARCH models, especially for the original data, will only slightly improve model fit, as the effects of this addition are small for most models.

The shape parameter ν_j for these models generally show leptokurtic tendencies. For the two-regime GARCH with skew-T distribution, ν_1 and ν_2 are sufficiently low to not behave like the normal distribution. For the two-regime tGARCH with skew-GED, ν_1 and ν_2 are clearly lower than 2 which corresponds to the normal distribution. ν_1 for the two-regime MS2 GARCH student-T distribution on the original and log data is quite high with values around 99, so these models are fairly close to normal.

The skewness parameters ξ_j in the two models with skewed distributions are all quite high for both regimes, with the exception of ξ_2 for both models when estimating the original data. In these two cases, ν_2 is < 1, and therefore exerts left skewness.

The AIC-values reflect mostly what we already assumed; the more complex models outperform the others, especially the models that admit skewness. There might

be a slight advantage for the tGARCH skew-GED compared to the GARCH skew-T. This might be because the insurance data shows evidence of having the "cusp" at the most frequent values, especially for the original data.

To better understand how well these four different models fit the original data and the log of the original data, we simulate 500-steps ahead for each model and combine these 500 values with the last 1000 values of the dataset. Figure (19) displays these two time series together, which allows us to see if the model has any big flaws in the specification.

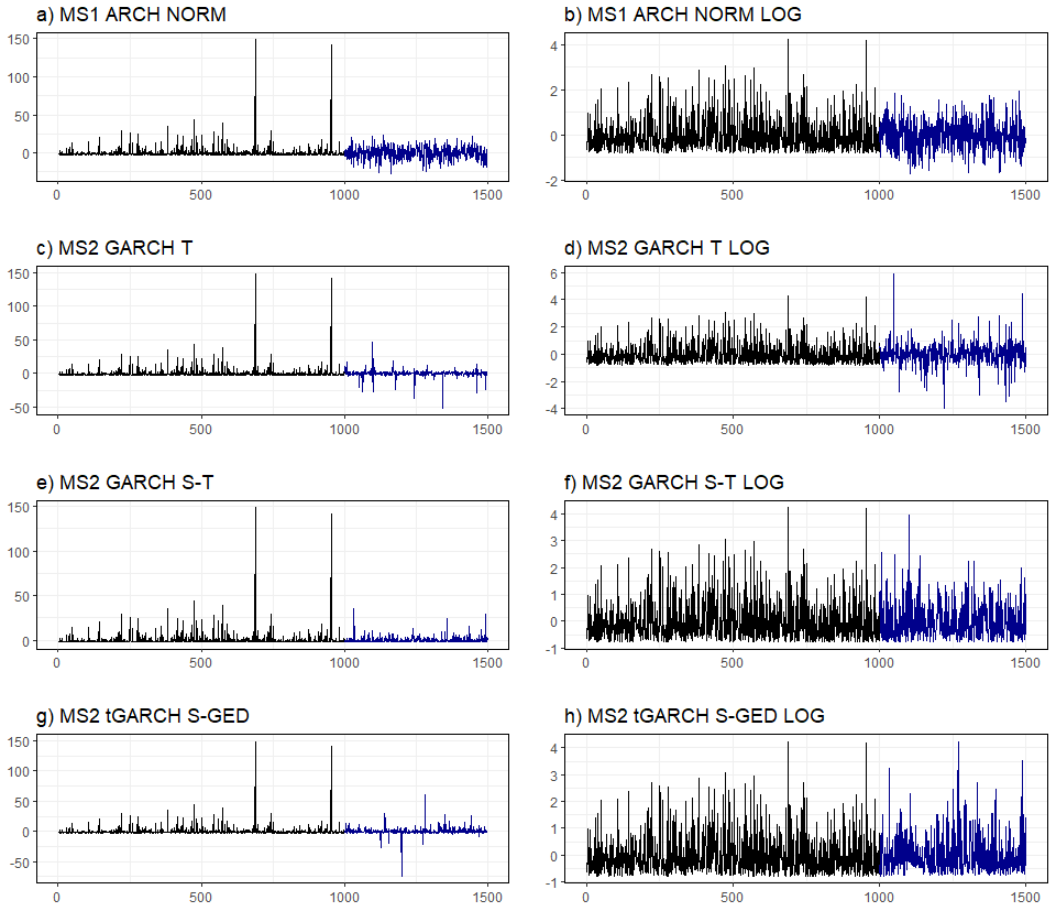


Figure 19: Plot of the 500-step simulation of the eight models (blue) in combination with the last 1000 observations of the dataset (black)

We can obviously see in the simulation of the original data the importance of adding skewness-distributions, as the first two model specifications (a - d) does not capture the skewness at all, and therefore predicts values that are much too low to fit the data, which originally were all larger than or equal to 1. It seems that the two-regime GARCH with Student-t distribution outperforms the single-regime normal distribution quite a lot, as expected. As for the two-regime skewed models, it may seem that the skew-GED model is better at capturing the size of the shocks, while it sometimes wrongly predicts the direction of the skewness.

As another method of graphically comparing the models, we create one-step ahead predictions and simulate this prediction process 100,000 times, and then visualize the histograms in conjunction with the histograms of the original data. This better shows the shape of the distributions in use. Figure (20) contains these histograms for our four example models, where the red histogram corresponds to the original data.

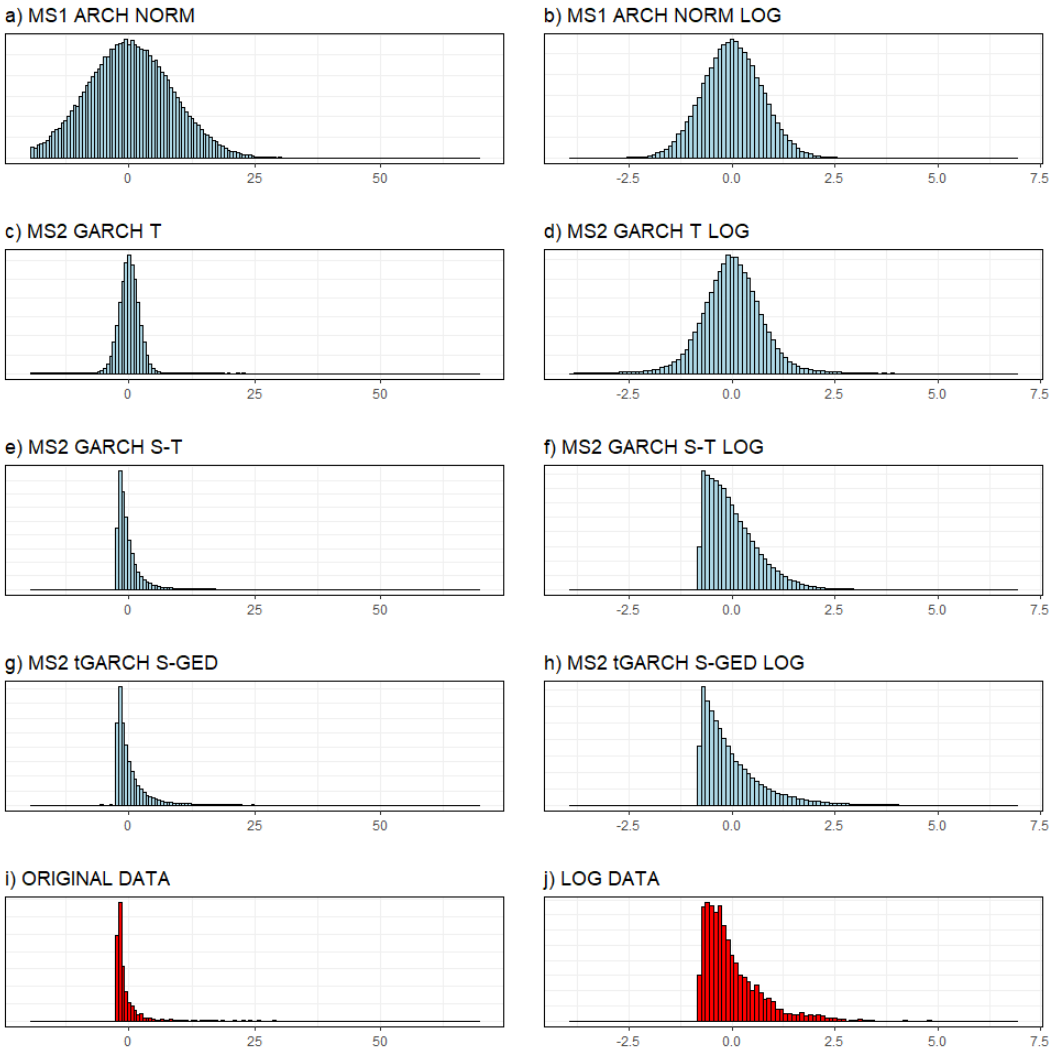


Figure 20: Histograms of the considered models' one-step ahead predictions, for both original and log data. i) and j) in red are the histograms of the original data. We have set the limits to better view the shape of the predicted values, and thus some of the values are out of bounds.

Viewing the figure yet again confirms the notion that skewed distributions are better for our candidate models. The figure leaves out values on the x axis (< -20 & > 70) for the original data and (< -4 & > 7) for the log data, so that it is easier to scope the shape of the distributions. These histograms, however, does not give a

	$VaR_{.99}$ Original	$ES_{.99}$ Original	$VaR_{.99}$ Log	$ES_{.99}$ Log
MS1 ARCH NORM	19.82	22.65	1.67	1.90
MS2 GARCH T	15.45	33.03	2.17	2.81
MS2 GARCH S-T	15.95	35.37	1.88	2.49
MS2 tGARCH S-GED	20.68	33.90	2.80	3.59
Empirical	22.66	55.20	2.47	3.03

Table 5: 99% Value-at-Risk and Expected Shortfall-values for the original and log-fitted models.

good depiction on the effect of having several regimes in the model, as these effects mostly come into play for the fewer, larger values (i.e. the tail). Instead, taking these same one-step-ahead predictions and calculating the VaR and ES from these gives us a better idea on how the tails of the models behave.

The empirical 99% Value-at-Risk for the original and log de-meaned data are 22.66 and 2.47 respectively. The 99% ES-values are 55.20 and 3.03. Naturally, if the VaR and ES from a predicted model is close to the empirical values, one can assume that the model has a fairly good tail fit. This is especially the case for VaR and ES-values with confidence levels close to 1, as these values tend to have much more inaccuracy, as there are fewer, more volatile values in the end of the tail. Table (5) contains the relevant values for the four example models.

From what we observe in the table, it seems that none of the models give entirely bad values for the 99% VaR and ES-approximations. We notice that the single regime normal distribution has a much too small jump from its VaR-value to its ES-value. This is expected, since the normal distribution not only doesn't take heavier tails into consideration, but it also does not have a second regime that can catch the more extreme values. The other three models, however, are much better at catching the effect of the few extreme claims. All models seem to underestimate the empirical values of the original data, especially for the expected shortfall. Out of these four models, it seems that the two-regime GARCH skew-T model and the two-regime tGARCH skew-GED model performs the best, which reflects all the previous observations we have made so far in this process. This table is extended to include all candidate models in the next section.
Figure (21) displays all example models' plotted VaR and ES values for both original and log data. The x-axis is the confidence levels between 90% and 99.95% for VaR and ES. The black line represents the empirical values of the VaR and ES for the varying confidence levels.

On the original data (a and c), we notice that the relative differences between the example models and the empirical values is not actually that large for all the models except the single-regime normal-ARCH, which generally underestimates the

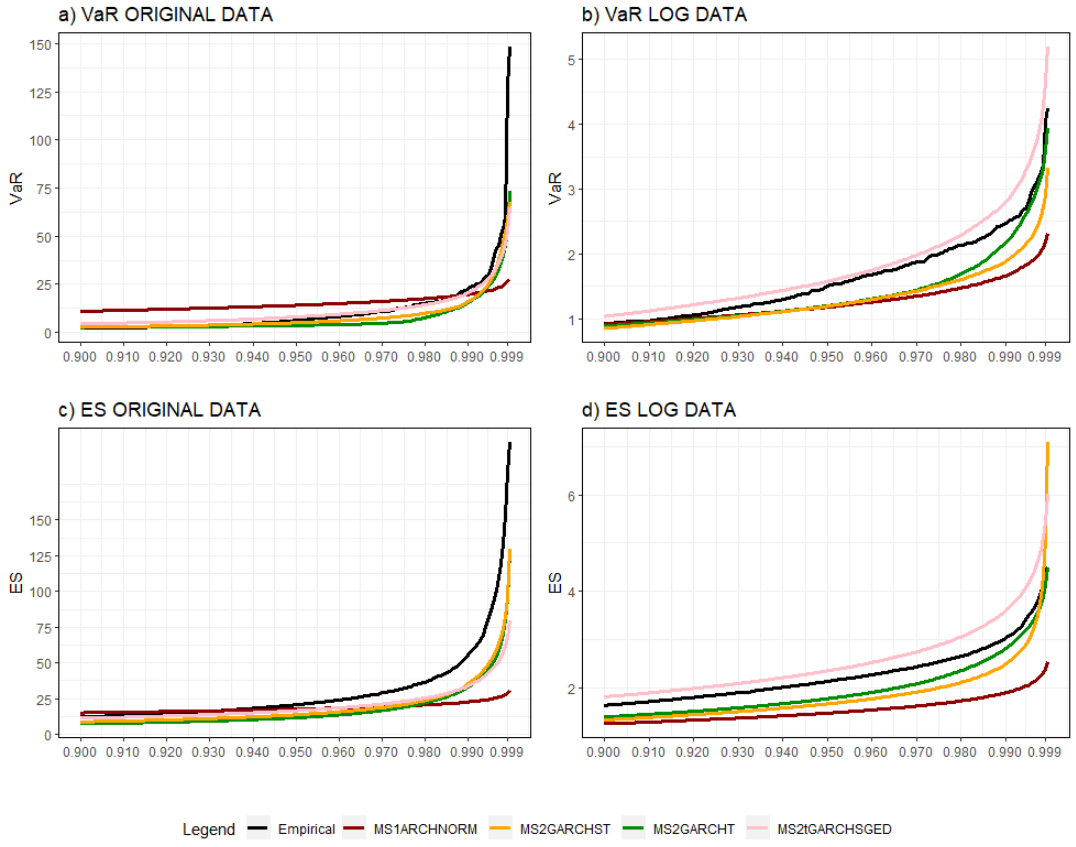


Figure 21: Value at Risk and Expected Shortfall for various confidence levels.

values at higher confidence levels. Most of the models are quite adequate at being able to follow the sudden increase in tail risk as the confidence level grows. Another thing to note is how, for the Expected Shortfall values (c, d), there appears to be a more significant estimation error than for the VaR values. This is attributed to the fact that there is much less information in the outer tails of the data, resulting in less accurate estimation. Appendix B contains these plots for all contending models. This concludes the process of estimation, prediction and calculation of risk measures for these four example models. We take a look at the results of all the models below.

7.3.1 Results

Eling (2012)[21] fitted skew-normal and skew-Student-t models on the original data and the log-data, while this thesis has used the de-meaned versions of these. AIC-values of Eling's regression models do not change when shifting the mean of the functions, and therefore those models have here been re-fitted on the de-meaned data, in order to ease the comparison-process. In addition to the skewed models, Eling also compared his findings with 18 benchmark models using different distributions. We will include some of them here, i.e. the exponential(exp)-, chi-squared , cauchy-, gamma-, normal-, logistic-, log-normal-, Student-t, Weibull-, symmetric/asymmet-

ric hyperbolic(S-Hyp & A-Hyp)-, symmetric/asymmetric generalized hyberbolic(S-Ghyp & A-Ghyp)-, symmetric/asymmetric NIG(S-NIG & A-NIG)- and symmetric/asymmetric variance gamma(S-VG & A-VG)- distributions. These different models have been estimated through the R packages `sn` (Azzalini, 2020)[9], `ghyp` (Luethi & Breymann, 2016)[43] and `MASS`(base R). As for the models considered in this thesis, we have fitted one-regime models for each combination of conditional distributions and conditional volatility specification, and we have fitted all combinations of two-regime models where both regimes are specified the same way, as not to get hundreds of models. Tables (6, 7, 8 & 9) include all log-likelihood values and AIC-values for the considered models.

For the original data, our models do *not* beat the skew-Student-T model of Eling in model fit. If we recall the discussion of the data in section (7.1), we could not bring forth any evidence of ARCH-effects in the data, so a simpler skewed model with fewer parameters will expectedly outperform our models, which are all fitting some kind of (G)ARCH-effects *in addition* to the conditional distribution. Adding the second regime to the model has not done enough to improve the model fit in general for the original data. Nevertheless, the two-regime models on the original data generally outperform their single-regime counterparts, if only just slightly. This result is also reflected in the estimated conditional volatility parameters for the original data models, as $\alpha_{0,i}, \alpha_{1,j}, \alpha_{2,j}$ and $\beta_{1,j}$ generally yield insignificant effects, similarly to how the conditional volatility parameters in the four-model example were quite insignificant.

For the log of the original data, however, several of the fitted MS-models improve on the best model of Eling. These models are the one-regime tGARCH with skew-GED conditional distribution, and several of the two-regime models with skewed distributions. Notably, the two-regime tGARCH with skew-GED conditional distribution appeared to have the best model fit. In general, the most advanced models with more parameters did better for the log case, which is not so strange. With a medium-large dataset and still not a tremendous amount of parameters, the punishment for choosing a more complex model is not that severe.

We also wish to say something about the shape parameters ν_j and ξ_j for the different models. For both the original and the log data, all the skewed-models actually provided values larger than 1 for ξ_1, which is a good indication that the model has been able to capture the effect of right-skewness in the main regime. For the models with two regimes, the values of ξ_2 greatly differ between left-skewness and right-skewness. This is probably because of the small amount of observations which are allocated to the second regime.

The kurtosis-parameter ν_j in the specification of the skew-Student-t distributions for the original data appears to get quite close to 2 for most models and irregardless of regime-choice, which signifies that these methods are estimating some quite

extreme tails in both regimes. Even on the log data, most of the models have quite low ν_j-values for the skew-Student-t distributions, as most models appear to have ν_j-values around 4, which still signifies quite a heavy tail. As for the skew-GED, the ν_j parameter tells a similar story, as the ν_j-values of the models on the original data stay around 0.5, and on the log data stay around 1.3, both of which signifies quite heavy tails, as seen in Figures (8) and (9).

We wish to examine the tail behavior next. Table (10) contains the 99% VaR and ES-values for all of the contending models, and some select models from Eling (2012), including the skew-normal and skew-Student-T distributions. Notably, all of the one-regime MSGARCH-models underestimate the empirical VaR and ES for the original data, while their two-regime counterparts have, in general, a better fit at the 99% confidence level. This effect is seen especially for the ES of the original data, where almost all of the two-regime MSGARCH models outperform any of the one-regime MSGARCH models, as well any of the models from Eling (2012). Here we see the clear advantage that Markov-switching models yields when describing tail behavior for this insurance data, since the additional regime seems to be able to capture the large shocks in the tail better than through fitting the data to a skewed model. As we recall, the log data has much less pronounced shocks, and therefore the effect of adding a second regime to the model seems to barely improve the VaR and ES estimates compared to the empirical value. Still, the values are quite good for most of the MSGARCH-models compared to the benchmark models, the skew-normal and the skew-Student-t. The MSGARCH-models that use skew-Student-T and skew-GED as their conditional distribution does particularly well.
Visualizing how the models of this thesis and the models of Eling (2012) behave for 90% to 99.95% VaR and ES is done in Appendix B. We can make some remarks about what is seen in this compilation of plots:

- On the original data, almost all of the benchmark models, as well as almost all of the one-regime MSGARCH-models, severely underestimate the VaR and ES at the very end of the tail. The main exception is the Skew-T distribution, which quite largely overestimates the VaR and ES in the end of the tail. None of these mentioned models fit the data particularly well. Some of the MS1 skewed models, and most of the MS2 models, however, fits the data much better, and are a lot better at catching the effect of the large shocks at the very end of the tail. In particular, the one-regime MSGARCH models with skew-Student-t distribution, as well as almost all of the two-regime MSGARCH models with skewed distributions does significantly better than the skew-T and the rest of the models for the original data.

- On the log data, most of all the models perform much better than on the original data, since there are more similar observations in the tail. It seems

that the skew-T, the Weibull, the exponential and the Gamma distribution-models do particularly well out of the regular model fits. For the MSGARCH-models, there is a large divide between the models that perform poorly, i.e. the models fitted with conditional distributions `norm, std, ged` or `snorm`, and the models which perform well, i.e. the models fitted with conditional distributions `sstd` or `sged`. There seems to be not much difference between the one-regime models and their two-regime counterparts for the log data, but the two-regime models do seem to perform a little bit better. Some MSGARCH models perform outstandingly well for the log data, e.g. the one-regime ARCH model following skew-GED and its two-regime counterpart, and the big winner: The two-regime GARCH model following the skew-GED.

- It appears that the choice of conditional variance structure does not seem to be much of a deciding factor for most models. This supports what we discussed earlier about adding the (G)ARCH-specification only having a slight effect on the model fit, especially for the original data.

This concludes the presentation of the results.

LOG-LIKELIHOOD ORIGINAL DATA										
K=1	norm	std	ged	snorm	sstd	sged	Skew-norm	-6301.18	Weibull	-4803.62
ARCH	-7710.68	-5278.63	-5655.60	-6897.39	-3629.90	-3625.66	Skew-t	**-3337.51**	S-Hyp	-5213.58
GARCH	-7710.67	-5278.29	-5659.15	-6897.39	-3587.55	-3622.88	Exp	-4809.40	S-NIG	-4115.81
gjr	-7676.23	-5277.72	-5658.67	-6820.62	-3586.44	-4136.94	Chi-Squared	-4802.55	S-VG	-4541.88
t	-7590.21	-5279.27	-5652.10	-6700.94	-3784.59	-3625.17	Cauchy	-4118.09	S-Ghyp	-4108.33
K=2	norm	std	ged	snorm	sstd	sged	Gamma	-4767.10	A-Hyp	-4050.63
ARCH	-7772.05	5099.12	-5565.27	-5124.58	-3788.79	-3777.12	Normal	-7713.76	A-NIG	-3399.40
GARCH	NA	-5099.13	-5565.27	NA	-3594.24	-3611.21	Logistic	-5737.85	A-VG	-3699.88
gjr	NA	-5095.71	-5566.04	-6988.92	-3601.81	-3503.08	Log-normal	-4057.90	A-Ghyp	-3382.93
t	NA	-5097.59	-5566.90	-6264.78	-3764.91	-3572.07	Student-T	-4115.93		

Table 6: Log-likelihood values of all fitted values for the original data + the models of Eling (2012) w/ benchmark models. Some models did not converge, and are therefore "NA".

LOG-LIKELIHOOD LOG DATA										
K=1	norm	std	ged	snorm	sstd	sged	Skew-norm	-1717.75	Weibull	-1645.13
ARCH	-2348.99	-2217.68	-2272.36	-1721.27	-1634.17	-1634.68	Skew-t	-1633.11	S-Hyp	-2178.95
GARCH	-2348.65	-2217.60	-2270.70	-1721.03	-1636.41	-1682.50	Exp	-1647.81	S-NIG	-2148.96
gjr	-2346.98	-2208.69	-2268.93	-1700.71	-1707.61	-1665.59	Chi-Squared	-1909.12	S-VG	-2184.59
t	-2347.54	-2209.05	-2269.11	-1699.46	-1712.62	-1579.59	Cauchy	-2292.69	S-Ghyp	-2145.10
K=2	norm	std	ged	snorm	sstd	sged	Gamma	-1647.81	A-Hyp	-1649.92
ARCH	-2188.86	-2187.57	-2118.36	-1780.86	-1707.40	-1565.68	Normal	-2352.58	A-NIG	-1685.07
GARCH	-2176.65	-2178.69	-2134.00	-1700.01	-1679.46	-1637.50	Logistic	-2208.59	A-VG	-1636.53
gjr	-2169.31	-2171.53	-2130.78	-1650.40	-1593.08	-1587.77	Log-normal	-2750.31	A-Ghyp	-1677.23
t	-2169.10	-2172.01	-2111.16	-1651.05	-1586.50	**-1556.28**	Student-T	-2146.95		

Table 7: Log-likelihood values of all fitted values for the log data + the models of Eling (2012) w/ benchmark models.

AIC ORIGINAL DATA										
K=1	norm	std	ged	snorm	sstd	sged	Skew-norm	12608.36	Weibull	9611.24
ARCH	15425.36	10563.26	11317.20	13800.78	7267.79	7259.33	Skew-t	**6681.02**	S-Hyp	10433.17
GARCH	15427.35	10564.57	11326.30	13802.77	7185.10	7255.77	Exp	9620.79	S-NIG	8237.61
gjr	15360.47	10565.43	11327.33	13651.24	7184.87	8285.89	Chi-Squared	9607.09	S-VG	9089.76
t	15188.42	10568.54	11314.20	13411.87	7581.18	7262.34	Cauchy	8240.17	S-Ghyp	8224.65
K=2	norm	std	ged	snorm	sstd	sged	Gamma	9538.19	A-Hyp	8109.27
ARCH	15556.1	10214.25	11146.53	10265.17	7597.58	7574.24	Normal	15431.52	A-NIG	6806.79
GARCH	NA	10218.26	11150.53	NA	7212.48	7246.41	Logistic	11479.71	A-VG	7407.76
gjr	NA	10215.42	11156.08	14001.85	7231.63	7034.17	Log-normal	8119.79	A-Ghyp	6775.85
t	NA	10219.18	11157.80	12553.55	7557.81	7172.13	Student-T	8237.85		

Table 8: AIC values of all fitted values for the original data + the models of Eling (2012) w/ benchmark models. Some models did not converge, and are therefore "NA".

AIC LOG DATA										
K=1	norm	std	ged	snorm	sstd	sged	Skew-norm	3441.49	Weibull	3294.27
ARCH	4701.98	4441.36	4550.71	3448.54	3276.34	3277.36	Skew-t	3272.21	S-Hyp	4363.90
GARCH	4703.31	4443.20	4549.40	3450.05	3282.82	3375.00	Exp	3297.61	S-NIG	4303.93
gjr	4701.95	4427.39	4547.87	3411.42	3427.220	3343.18	Chi-Squared	3820.24	S-VG	4375.17
t	4703.09	4428.09	4548.21	3408.92	3437.24	3171.18	Cauchy	4589.38	S-Ghyp	4298.21
K=2	norm	std	ged	snorm	sstd	sged	Gamma	3299.61	A-Hyp	3307.83
ARCH	4389.73	4391.15	4252.71	3577.71	3434.80	3151.35	Normal	4709.15	A-NIG	3378.14
GARCH	4369.30	4377.38	4288.01	3420.02	3382.93	3299.00	Logistic	4421.17	A-VG	3281.06
gjr	4358.61	4367.05	4285.55	3324.80	3214.16	3203.55	Log-normal	5504.62	A-Ghyp	3364.47
t	4358.19	4368.02	4246.33	3326.09	3201.00	**3140.56**	Student-T	4299.90		

Table 9: AIC values of all fitted values for the log data + the models of Eling (2012) w/ benchmark models.

$VaR_{0.99}$ **ORIGINAL DATA**								
K=1	norm	std	ged	snorm	sstd	sged	Skew-norm	20.39
ARCH	19.88	9.21	12.03	15.16	12.56	11.35	Skew-t	4.75
GARCH	19.68	9.07	11.96	15.02	13.59	11.52	Chi-squared	11.52
gjr	18.55	8.94	10.61	12.68	13.59	12.00	Gamma	12.00
t	15.34	9.09	10.70	9.93	14.67	11.51	Normal	19.74
K=2	norm	std	ged	snorm	sstd	sged	Exp	12.27
ARCH	19.47	15.54	12.85	39.31	19.59	13.45	Log-normal	8.21
GARCH	NA	15.53	12.85	NA	15.85	11.41	Logistic	6.28
gjr	NA	15.58	12.56	14.25	13.35	12.38	Weibull	12.80
t	NA	15.54	12.93	40.10	14.30	20.32	Empirical	22.66
$VaR_{0.99}$ **LOG DATA**								
K=1	norm	std	ged	snorm	sstd	sged	Skew-norm	2.11
ARCH	1.67	1.83	1.81	1.76	2.67	2.47	Skew-t	2.66
GARCH	1.66	1.85	1.80	1.75	2.74	3.10	Chi-squared	6.46
gjr	1.66	1.98	1.86	1.80	3.12	2.95	Gamma	2.83
t	1.67	2.04	1.89	1.79	2.40	2.73	Normal	1.67
K=2	norm	std	ged	snorm	sstd	sged	Exp	2.83
ARCH	2.18	2.10	2.42	3.02	1.78	2.59	Log-normal	40.02
GARCH	NA	2.20	1.99	NA	1.89	2.47	Logistic	1.57
gjr	NA	2.12	2.47	2.20	2.65	2.66	Weibull	2.68
t	NA	2.09	2.01	2.03	2.87	2.80	Empirical	2.47
$ES_{0.99}$ **ORIGINAL DATA**								
K=1	norm	std	ged	snorm	sstd	sged	Skew-norm	23.23
ARCH	22.81	15.13	16.30	17.66	22.47	15.37	Skew-t	7.39
GARCH	22.40	14.30	16.56	17.81	27.27	15.40	Chi-squared	10.34
gjr	21.21	14.94	14.36	14.91	27.62	16.35	Gamma	13.02
t	17.72	14.46	14.27	11.73	30.74	15.55	Normal	22.59
K=2	norm	std	ged	snorm	sstd	sged	Exp	15.63
ARCH	22.25	36.26	37.50	84.98	30.94	35.49	Log-normal	11.72
GARCH	NA	36.55	37.50	NA	34.76	291853.30	Logistic	7.87
gjr	NA	35.21	34.16	16.73	27.44	28.15	Weibull	16.43
t	NA	36.36	32.41	48.88	28.45	35.40	Empirical	55.20
$ES_{0.99}$ **LOG DATA**								
K=1	norm	std	ged	snorm	sstd	sged	Skew-norm	2.47
ARCH	1.92	2.47	2.17	2.07	3.91	3.08	Skew-t	3.76
GARCH	1.90	2.47	2.16	2.07	3.89	4.03	Chi-squared	8.32
gjr	1.91	2.61	2.23	2.13	4.81	3.74	Gamma	3.61
t	1.92	2.74	2.26	2.11	3.16	3.41	Normal	1.91
K=2	norm	std	ged	snorm	sstd	sged	Exp	3.62
ARCH	2.9	2.80	2.96	4.31	2.10	3.24	Log-normal	102.53
GARCH	NA	2.88	2.60	NA	2.42	3.05	Logistic	1.94
gjr	NA	2.77	2.95	2.76	3.68	3.34	Weibull	3.40
t	NA	2.73	2.41	2.55	4.09	3.55	Empirical	3.03

Table 10: 99% Value-At-Risk and Expected Shortfall values for all models, including select models from Eling (2012). Some models did not converge, and are therefore "NA".

8 Conclusion

This thesis' main purpose was to examine how a Markov-switching GARCH-framework would suit a time series of insurance losses, and if this implementation could yield models which accomplishes two things; beats models which are estimated through probability distribution fitting, and better describes the right-tail of the loss distribution.
We explained that (G)ARCH-models are not usually applied to insurance time series, and the fact that insurance losses are not traditionally viewed as a time series at all. Since the data in consideration is absolute positive, we argued that we could perform model fitting on the de-meaned version of the data, as there seemed to be no first-degree serial correlation in the original- or log of the original-data.
When examining the insurance loss data, we found that the sample skewness and sample kurtosis of the data implied that there was a significant appearance of both skewness and kurtosis in the original and log data, so we assumed that using a skewed version of a distribution which fits a shape parameter that alters kurtosis would result in a better model fit than the converse. The original data seemed like it had such a heavy tail that somehow adding another regime for capturing the higher-valued claims could be a good idea.
There was little evidence to back up the hypothesis that there were any conditional heteroskedasticity in the original data at all. As a consequence of this, changing the structure of the conditional volatility should not yield much of a difference in how well the model fit becomes for the original data. We therefore did not expect the GARCH-component of the MSGARCH-models on the original data to be the reason for a better model fit. However, the Markov-switching-component of the MSGARCH-model was shown to be able to capture the effects of large shocks in the data better, and not overestimate the persistence of said shocks in the original data. We therefore expected the MSGARCH-models to better describe the right tail of the data.
For the log of the original data, we performed the McLeod-Li test where we found that it did exhibit a slight bit of conditional heteroskedasticity. As a result of this, the GARCH-component of the MSGARCH-model could be the reason for a slight improvement in the model fit. Since the presence of the large shocks in the original data was greatly diminished when taking the log of the data, the improvement-effect of adding more regimes to the models should be reduced for the log data.

The MSGARCH-framework allows for changing of the structure of the conditional volatility, and changing of the conditional distribution of the observed variables. In total, we considered six conditional distributions and four conditional volatility structures, for a total of 24 possible model combinations, irregardless of the amount of regimes in the model. We fitted 24 models for the one-regime case

and 24 models for the two-regime case through ML-estimation. The BFGS-algorithm was used for the optimization process. We also re-fitted the models of Eling (2012) for the de-meaned data.

After executing the estimation, we found that models which used skewed distributions with an additional shape parameter improved the model fit significantly over their non-skewed counterparts. We also found that, for the original data, all candidate models performed worse than the skew-Student-t model of Eling (2012), when simply looking at the AIC. This was not unexpected, as there were no evidence of any conditional heteroskedasticity in the original data. As a consequence of this, the effect of the different structures of the conditional volatility was very small for the original data. In addition to this, the MSGARCH-models sometimes wrongly predicted the "direction" of the skew, and the skew-Student-t of Eling (2012) contained fewer parameters, which improves AIC slightly.
For the log data, several of the MSGARCH-models with skewed distributions improved on the models of Eling (2012), when just looking at the AIC. We attribute this to the fact that there were some slight conditional heteroskedasticity in the log de-meaned data. The structure of the conditional volatility seemed to have more of an effect on the log data, and the *tGARCH* generally outperformed the other structural specifications, possibly because of some slight asymmetry in the effect of previous observations on the conditional volatility.

As for the tail risk measures, the models of Eling (2012) consequently underestimated the tail-risk-values at the very far right tail of the original data. The MSGARCH-models with two regimes severely improved on the tail-risk-values for confidence levels between 90% and 99.95%. This is as expected, as the second regime is specifically added to capture the effects of the large shocks, which end up in the far right tail. For the log-data, the improvement is not as clear, since some of the models of Eling(2012) perform very well on this data. However, there are still several of the MSGARCH-models for both the one-regime- and two-regime case which outperform the best models of Eling (2012). The VaR and ES-values for the one-regime models for the log data are actually not very different from the two-regime models for the log data, and one type does not consequently outperform the other. This confirms our hypothesis that adding a second regime for the log-data is not as significant for the log-data, since the effect of the large shocks in the data is diminished when taking the logarithm.

All in all, we note that applying MSGARCH to the de-meaned version of this insurance time series can, in fact, yield some results that outperform regular probability distribution fitting. In particular, the application of this method should be

considered for insurance time series if one wishes to attain high-accuracy estimation of tail risk measures, since the addition of additional regimes could increase the focus on the tail. It is difficult to say if it is a good idea to use Markov-switching GARCH models on the general insurance loss time series if the aim is to get the best general model fit, as one cannot be sure that there will exists any (G)ARCH-effects in an insurance time series. Usually, insurance loss data are not viewed as a time series, which would suggest that most insurance loss data does not exhibit any kind of serial correlation.

References

[1] C. Acerbi and B. Szekely. Back-testing expected shortfall. *Risk*, 27(11):76–81, 2014.

[2] C. Acerbi and D. Tasche. On the coherence of expected shortfall. *Journal of Banking & Finance*, 26(7):1487–1503, 2002.

[3] V. Acharya, R. Engle, and M. Richardson. Capital shortfall: A new approach to ranking and regulating systemic risks. *American Economic Review*, 102(3):59–64, 2012.

[4] V. V. Acharya, L. H. Pedersen, T. Philippon, and M. Richardson. Measuring systemic risk. *The Review of Financial Studies*, 30(1):2–47, 2017.

[5] A. Ahmadi-Javid. Entropic value-at-risk: A new coherent risk measure. *Journal of Optimization Theory and Applications*, 155(3):1105–1123, 2012.

[6] H. Akaike. Information theory and an extension of the maximum likelihood principle. In *Selected papers of hirotugu akaike*, pages 199–213. Springer, 1998.

[7] D. Ardia, K. Bluteau, K. Boudt, L. Catania, and D.-A. Trottier. Markov-switching GARCH models in R: The MSGARCH package. *Journal of Statistical Software*, 91(4):1–38, 2019.

[8] P. Artzner, F. Delbaen, J.-M. Eber, and D. Heath. Coherent measures of risk. *Mathematical finance*, 9(3):203–228, 1999.

[9] A. Azzalini. *The R package* `sn`*: The Skew-Normal and Related Distributions such as the Skew-t (version 1.6-2).* Università di Padova, Italia, 2020.

[10] Bank for International Settlements. Minimum capital requirements for market risk, 2019, Accessed 2019-12-13. https://www.bis.org/bcbs/publ/d457.pdf.

[11] M. Bernardi, A. Maruotti, and L. Petrella. Multiple risk measures for multivariate dynamic heavy–tailed models. *Journal of Empirical Finance*, 43:1–32, 2017.

[12] T. Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of econometrics*, 31(3):307–327, 1986.

[13] P. J. Brockwell and R. A. Davis. *Introduction to time series and forecasting*. springer, 2016.

[14] C. Brownlees and R. F. Engle. Srisk: A conditional capital shortfall measure of systemic risk. *The Review of Financial Studies*, 30(1):48–79, 2016.

[15] C. G. Broyden. The convergence of a class of double-rank minimization algorithms 1. general considerations. *IMA Journal of Applied Mathematics*, 6(1):76–90, 1970.

[16] J. Cai. A markov model of switching-regime arch. *Journal of Business & Economic Statistics*, 12(3):309–316, 1994.

[17] O. Cappé, E. Moulines, and T. Rydén. *Inference in hidden Markov models*. Springer Science & Business Media, 2005.

[18] H. Chernoff et al. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, 23(4):493–507, 1952.

[19] J. D. Cryer and K.-S. Chan. *Time series analysis: With applications in R*. Springer, 2008.

[20] M. Davidian and R. J. Carroll. Variance function estimation. *Journal of the American Statistical Association*, 82(400):1079–1091, 1987.

[21] M. Eling. Fitting insurance claims to skewed distributions: Are the skew-normal and skew-student good models? *Insurance: Mathematics and Economics*, 51(2):239–248, 2012.

[22] P. Embrechts, C. Klüppelberg, and T. Mikosch. *Modelling extremal events for Insurance and Finance*. Springer, 1997.

[23] R. F. Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica: Journal of the Econometric Society*, pages 987–1007, 1982.

[24] E. F. Fama. The behavior of stock-market prices. *The journal of Business*, 38(1):34–105, 1965.

[25] C. Fernández and M. F. Steel. On bayesian modeling of fat tails and skewness. *Journal of the American Statistical Association*, 93(441):359–371, 1998.

[26] R. Fletcher. A new approach to variable metric algorithms. *The computer journal*, 13(3):317–322, 1970.

[27] C. Francq and J.-M. Zakoian. *GARCH models: structure, statistical inference and financial applications*. John Wiley & Sons, 2019.

[28] L. R. Glosten, R. Jagannathan, and D. E. Runkle. On the relation between the expected value and the volatility of the nominal excess return on stocks. *The journal of finance*, 48(5):1779–1801, 1993.

[29] T. Gneiting. Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494):746–762, 2011.

[30] D. Goldfarb. A family of variable-metric methods derived by variational means. *Mathematics of Computation*, 24(109):23–26, 1970.

[31] S. F. Gray. Modeling the conditional distribution of interest rates as a regime-switching process. *Journal of Financial Economics*, 42(1):27–62, 1996.

[32] M. Haas, S. Mittnik, and M. S. Paolella. A new approach to markov-switching garch models. *Journal of financial econometrics*, 2(4):493–530, 2004.

[33] J. D. Hamilton. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica: Journal of the Econometric Society*, pages 357–384, 1989.

[34] J. D. Hamilton. *Time series analysis*, volume 2. Princeton New Jersey, 1994.

[35] J. D. Hamilton and R. Susmel. Autoregressive conditional heteroskedasticity and changes in regime. *Journal of econometrics*, 64(1-2):307–333, 1994.

[36] A. Harvey and C. Fernandes. Time series models for insurance claims. *Journal of the Institute of Actuaries*, 116(3):513–528, 1989.

[37] C. C. Hewitt and B. Lefkowitz. Methods for fitting distributions to insurance loss data. In *Proceedings of the casualty actuarial society*, volume 66, pages 139–160, 1979.

[38] P. Jorion. *Value at risk*. McGraw-Hill Professional Publishing, 2000.

[39] F. Klaassen. Improving garch volatility forecasts with regime-switching garch. In *Advances in Markov-switching models*, pages 223–254. Springer, 2002.

[40] P. Lambert and S. Laurent. Modelling financial time series using garch-type models with a skewed student distribution for the innovations. Technical report, 2001.

[41] C. G. Lamoureux and W. D. Lastrapes. Persistence in variance, structural change, and the garch model. *Journal of Business & Economic Statistics*, 8(2):225–234, 1990.

[42] M. N. Lane. Pricing risk transfer transactions 1. *ASTIN Bulletin: The Journal of the IAA*, 30(2):259–293, 2000.

[43] D. Luethi and W. Breymann. *ghyp: A Package on Generalized Hyperbolic Distribution and Its Special Cases*, 2016. R package version 1.5.7.

[44] A. I. McLeod and W. K. Li. Diagnostic checking arma time series models using squared-residual autocorrelations. *Journal of time series analysis*, 4(4):269–273, 1983.

[45] J. Nocedal and S. Wright. *Numerical optimization*. Springer Science & Business Media, 2006.

[46] L. J. Savage. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66(336):783–801, 1971.

[47] D. F. Shanno. Conditioning of quasi-newton methods for function minimization. *Mathematics of computation*, 24(111):647–656, 1970.

[48] Y. Shin. *Time Series Analysis in the Social Sciences: The Fundamentals*. Univ of California Press, 2017.

[49] T. Teräsvirta. An introduction to univariate garch models. In *Handbook of Financial time series*, pages 17–42. Springer, 2009.

[50] A. Tobias and M. K. Brunnermeier. Covar. *The American Economic Review*, 106(7):1705, 2016.

[51] R. Vernic. Multivariate skew-normal distributions with applications in insurance. *Insurance: Mathematics and economics*, 38(2):413–426, 2006.

[52] A. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE transactions on Information Theory*, 13(2):260–269, 1967.

[53] P. Wolfe. Convergence conditions for ascent methods. *SIAM Review*, 11(2):226–235, 1969.

[54] J.-M. Zakoian. Threshold heteroskedastic models. *Journal of Economic Dynamics and control*, 18(5):931–955, 1994.

Appendices

A R Code

```
#Creating specifications for the four example models
ms2.t.sg <- CreateSpec(list(model = c("tGARCH", "tGARCH")),
                       list(distribution = c("sged", "sged")))

#Fitting the models to the de-meaned original and log-data
fitms2.t.sg <- FitML(ms2.t.sg, data = Total.dm)
fitms2.t.sg.log <- FitML(ms2.t.sg, data = log.Total.dm)

#Viewing the models: (some of the information has been removed)
> fitms2.t.sg
Specification type: Markov-switching
Specification name: tGARCH_sged tGARCH_sged
-------------------------------------------
Fitted parameters:
         Estimate Std. Error
alpha0_1   0.1074     0.0027
alpha1_1   0.0000     0.0000
alpha2_1   0.0001     0.0000
beta_1     0.9627     0.0009
nu_1       0.7083     0.0001
xi_1      38.2273     0.2683
alpha0_2  29.8576     0.2712
alpha1_2   0.0006     0.0000
alpha2_2   0.0216     0.0002
beta_2     0.0766     0.0008
nu_2       0.7003     0.0000
xi_2       0.7118     0.0056
P_1_1      0.9462     0.0009
P_2_1      0.9131     0.0006
-------------------------------------------
Transition matrix:
      t+1|k=1 t+1|k=2
t|k=1  0.9462  0.0538
t|k=2  0.9131  0.0869
-------------------------------------------
LL: -3572.0655
AIC: 7172.131
BIC: 7251.6664
-------------------------------------------

> fitms2.t.sg.log
Specification type: Markov-switching
Specification name: tGARCH_sged tGARCH_sged
```

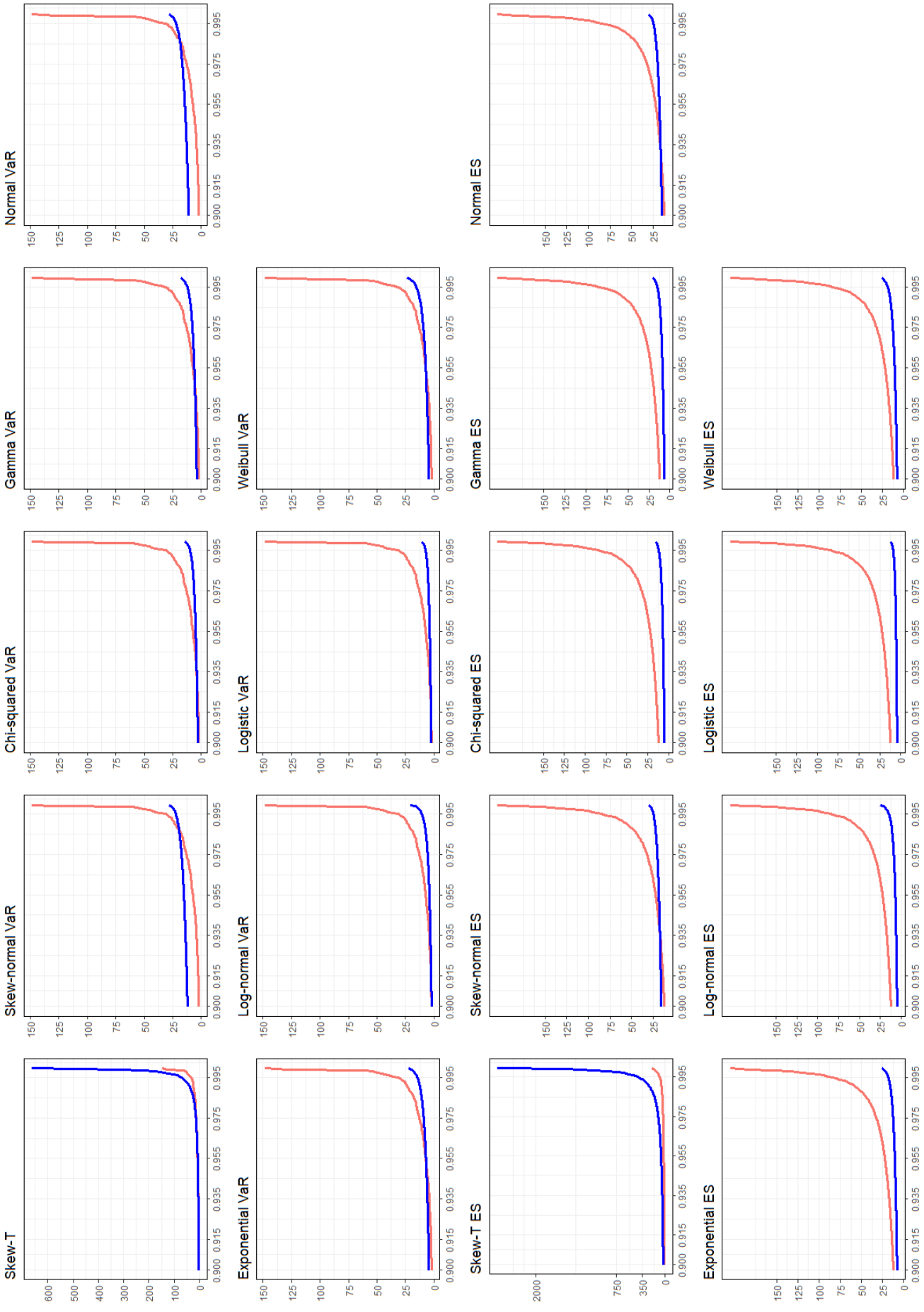
```
-----------------------------------------
Fitted parameters:
         Estimate Std. Error
alpha0_1   0.0075     0.0001
alpha1_1   0.0001     0.0000
alpha2_1   0.0580     0.0004
beta_1     0.9651     0.0001
nu_1       0.9427     0.0017
xi_1      21.9491     0.6358
alpha0_2   0.0042     0.0002
alpha1_2   0.0019     0.0001
alpha2_2   0.0065     0.0004
beta_2     0.9914     0.00026
nu_2       1.0416     0.0271
xi_2      12.7597     0.3966
P_1_1      0.9576     0.0003
P_2_1      0.0076     0.0015
-----------------------------------------
Transition matrix:
      t+1|k=1 t+1|k=2
t|k=1  0.9576  0.0424
t|k=2  0.0076  0.9924
-----------------------------------------
LL: -1556.2814
AIC: 3140.5629
BIC: 3220.0982
-----------------------------------------


#Predicting one-step ahead draws
y.ms2.t.sg <- predict(fitms2.t.sg, nahead = 1,
                      do.return.draw = TRUE, ctr = list(nsim = 100000))
y.ms2.t.sg.log <- predict(fitms2.t.sg.log, nahead = 1,
                      do.return.draw = TRUE, ctr = list(nsim = 100000))

#Simulating a 500-step ahead path
y.ms2.t.sg.sim <- simulate(fitms2.t.sg, nahead = 500, nsim = 1)
y.ms2.t.sg.log.sim <- simulate(fitms2.t.sg.log, nahead = 500, nsim = 1)

#Calculating 99% VaR
VaR <- quantile(y.ms2.t.sg$draw[1, ], 0.99)
> VaR
     99%
20.69505
VaR.log <- quantile(y.ms2.t.sg.log$draw[1, ], 0.99)
> VaR.log
     99%
```

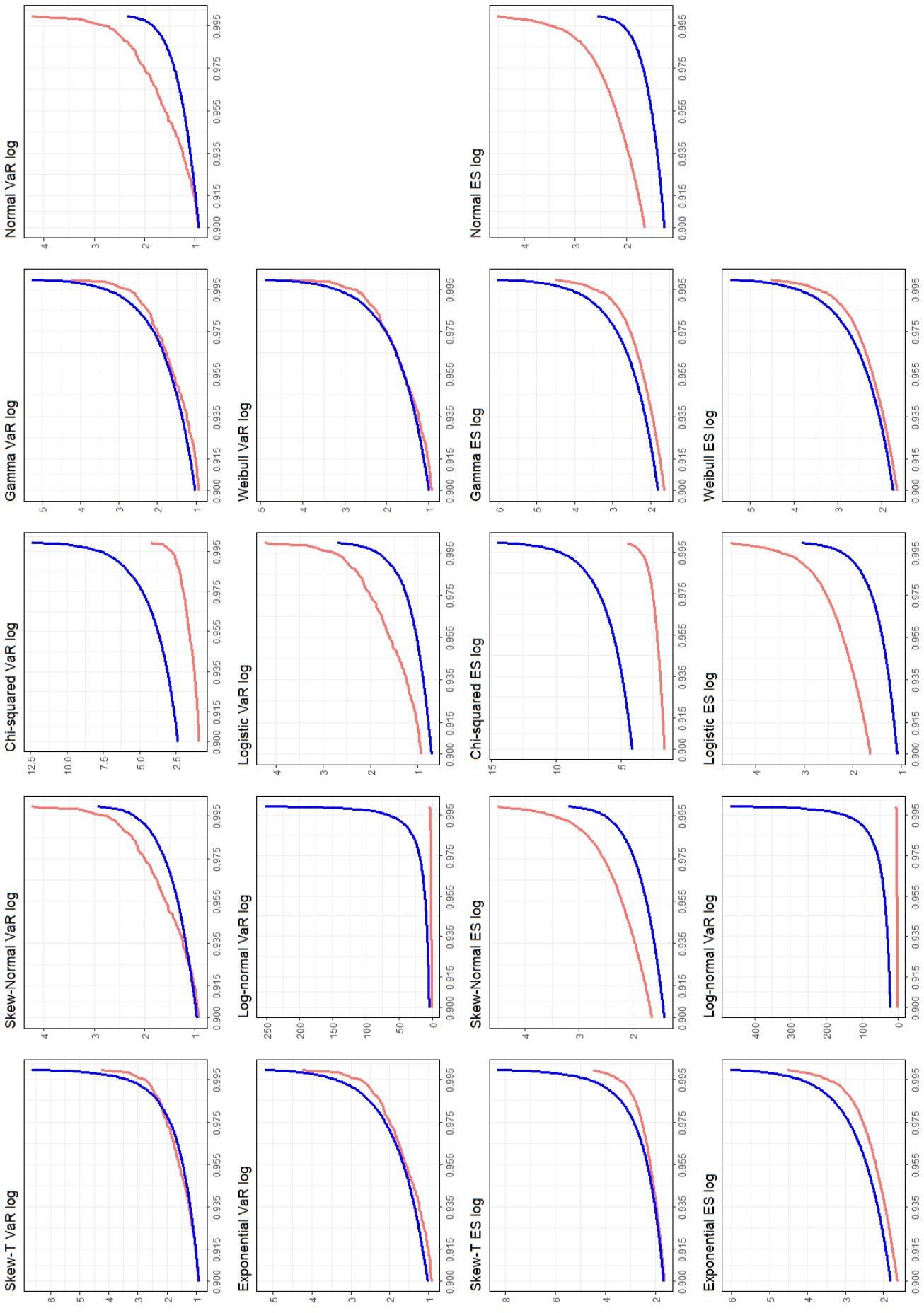
```
2.799733

#Calculating the 99% ES
ES <- mean(y.ms2.t.sg$draw[1, ][y.ms2.t.sg$draw[1, ] > VaR])
> ES
[1] 34.05283
34.05283
ES.log <- mean(y.ms2.t.sg.log$draw[1, ][y.ms2.t.sg.log$draw[1, ] > VaR.log])
> ES.log
[1] 3.604707
```

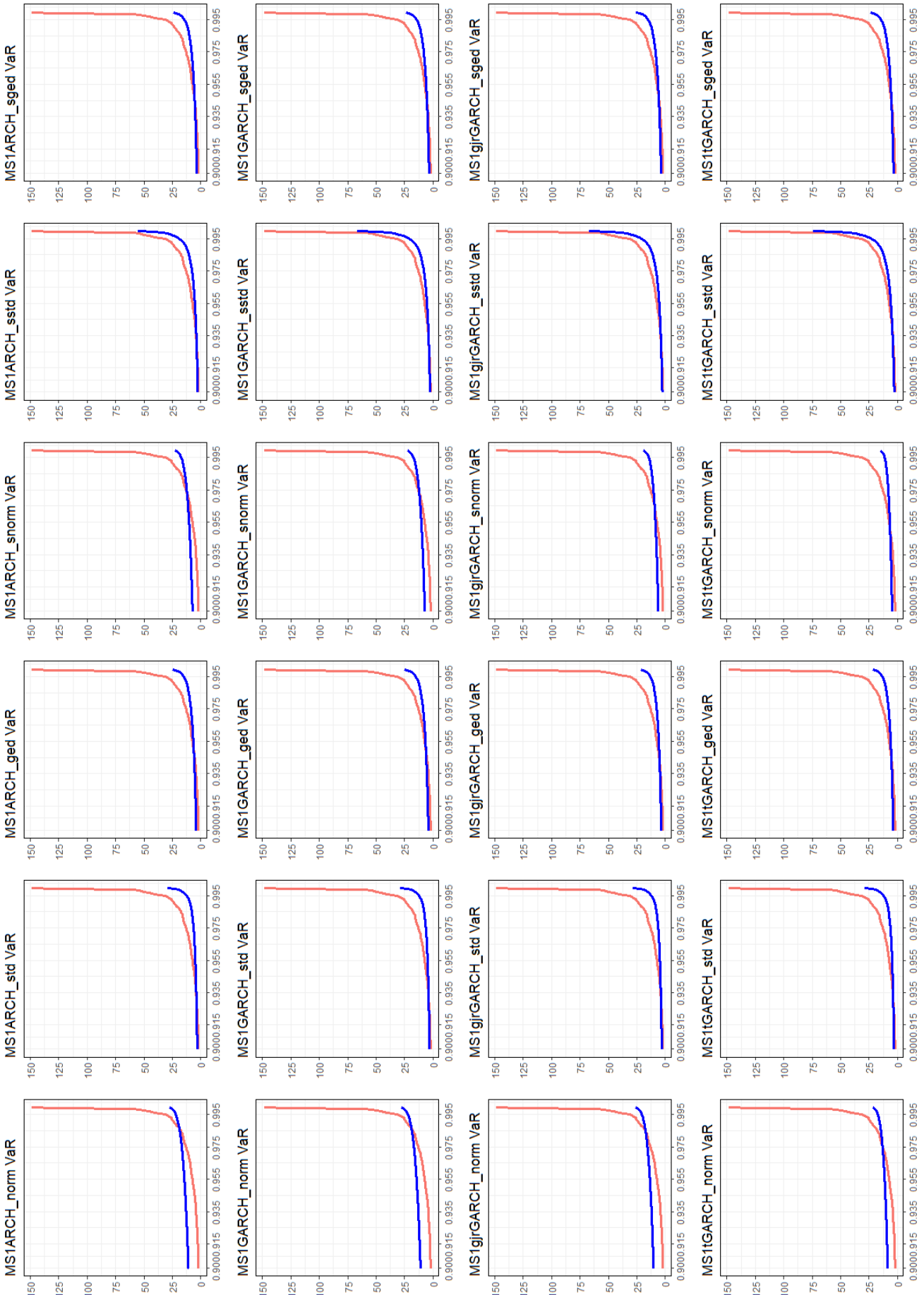
B VaR & ES: Plots

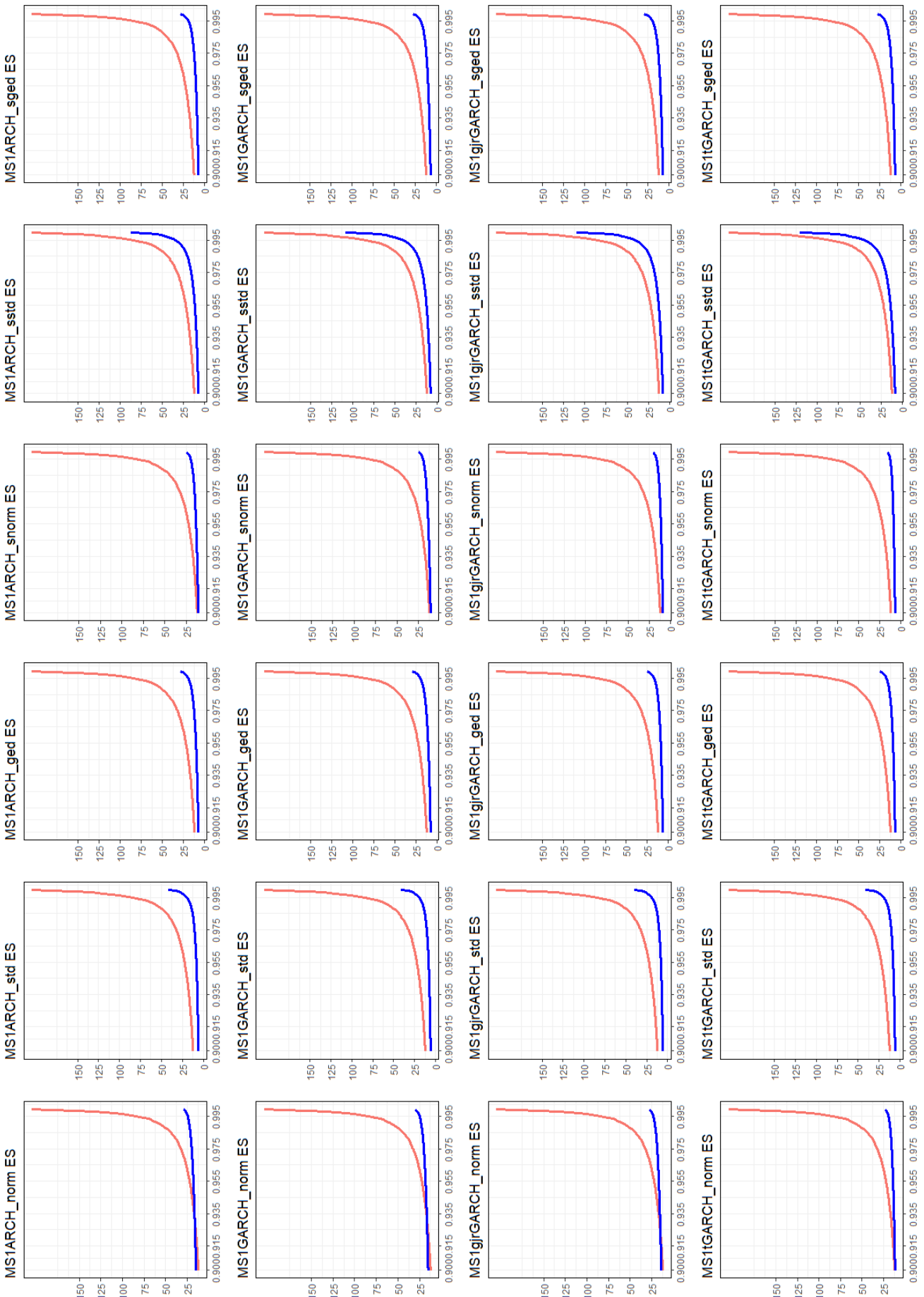
In the following pages are plots of the VaR and ES-values for confidence levels between 90% and 99.95%. Some models did not converge. These areas are intentionally left blank, in order to preserve the matrix form of the plots. For every plot, the red line is the VaR or ES corresponding to the empirical value from the original or log de-meaned data. The next next pages are structured as follows:

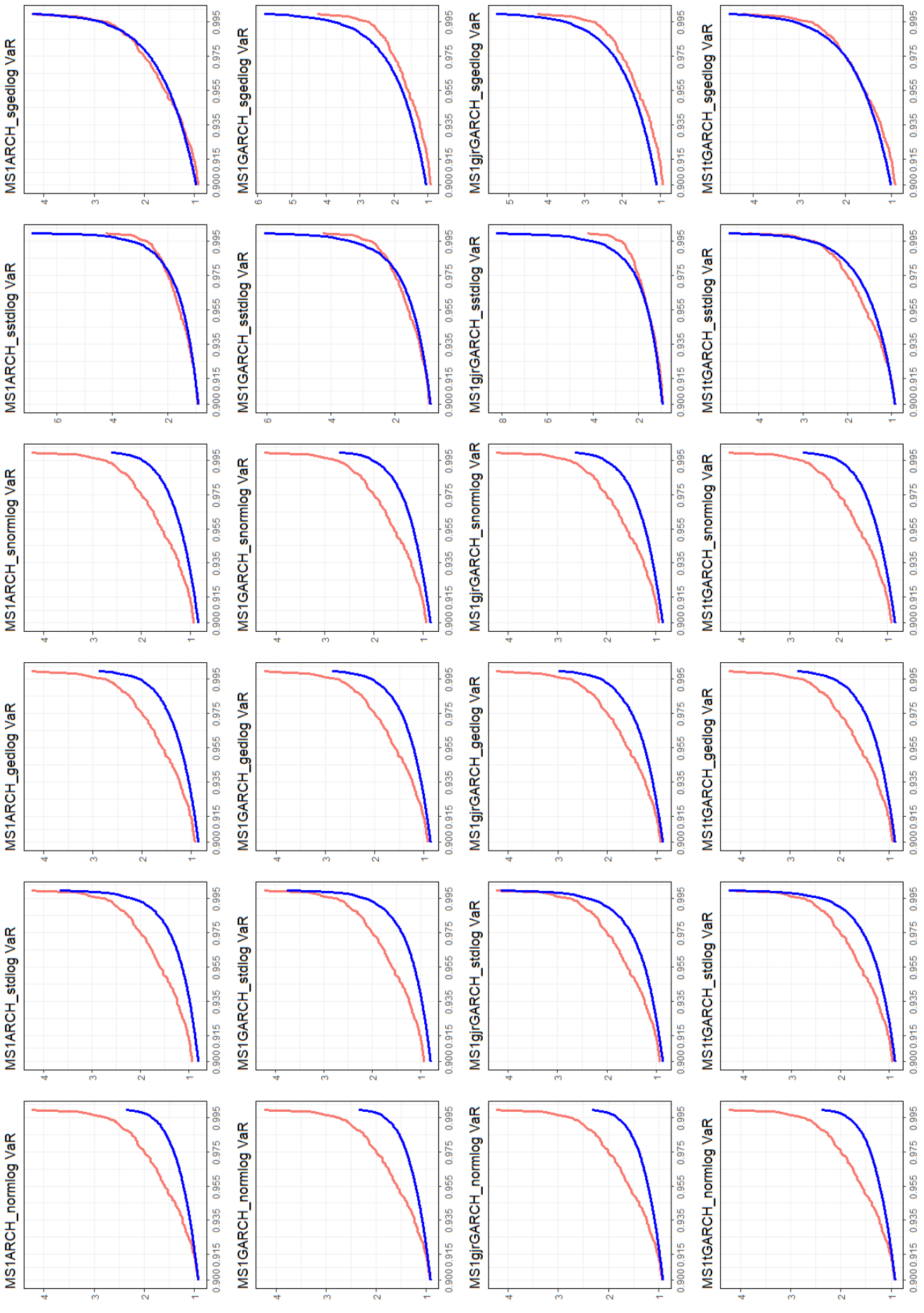
- **Page 89 - 90:** Each of the selected benchmark models and the Skew-normal and Skew-T from Eling (2012) on the original and log data
- **Page 91 - 94:** Each of the 24 fitted one-regime models on the original and log data.
- **Page 95 - 98:** Each of the 24 fitted two-regime models on the original and log data

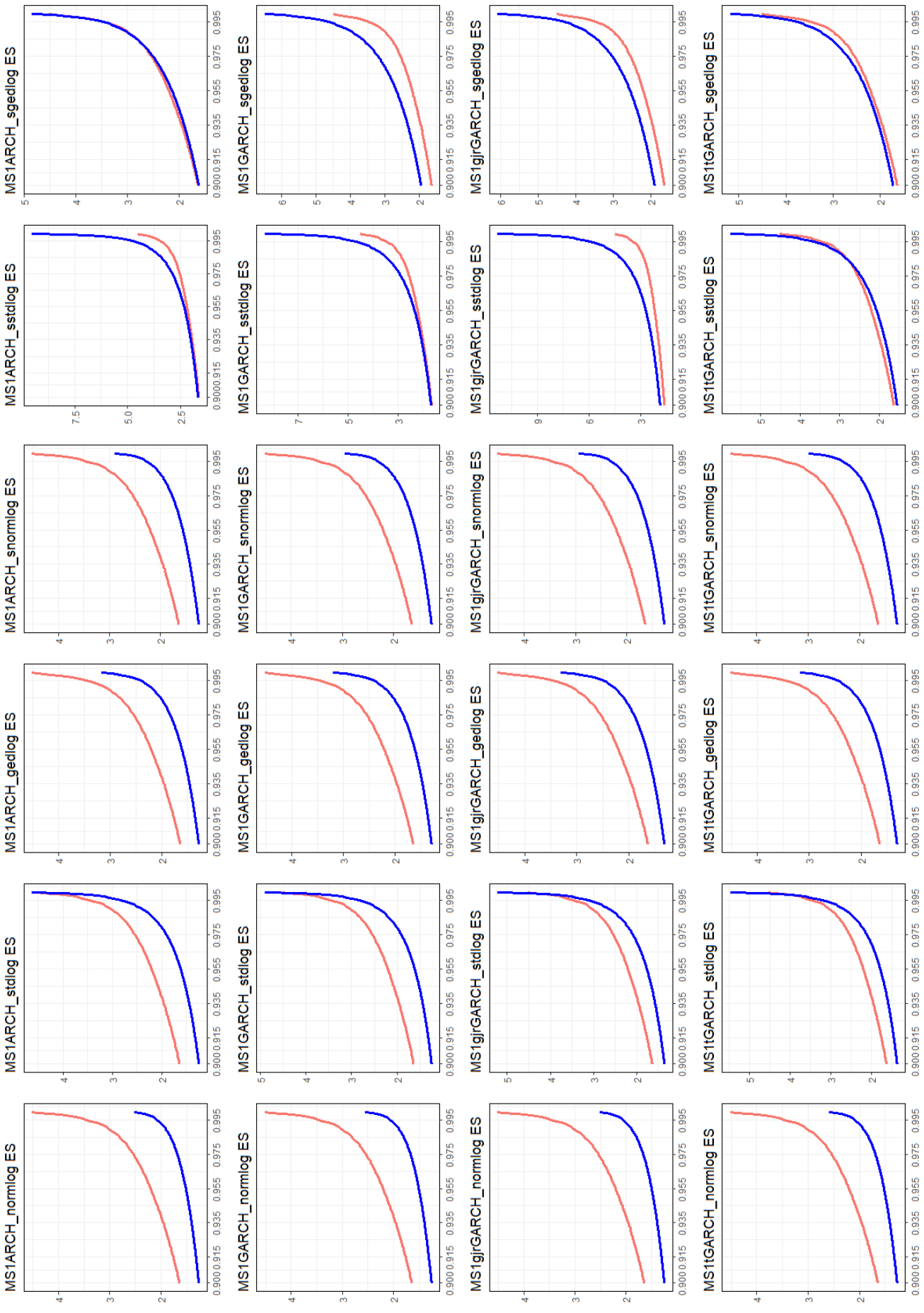

Normal VaR
Normal ES
Gamma VaR
Gamma ES
Weibull VaR
Weibull ES
Chi-squared VaR
Chi-squared ES
Logistic VaR
Logistic ES
Skew-normal VaR
Skew-normal ES
Log-normal VaR
Log-normal ES
Skew-T
Skew-T ES
Exponential VaR
Exponential ES

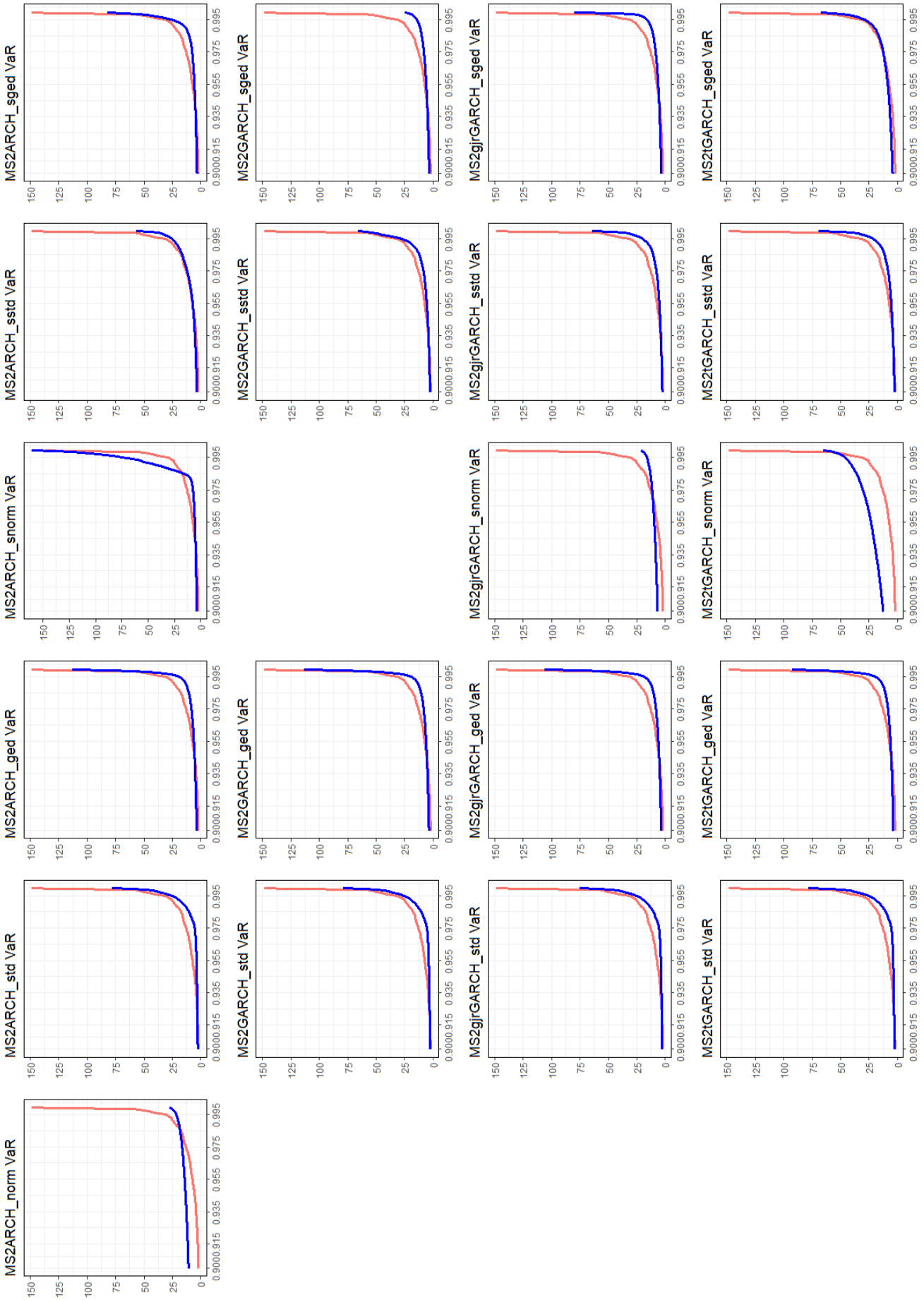

Normal VaR log
Gamma VaR log
Chi-squared VaR log
Skew-Normal VaR log
Logistic VaR log
Weibull VaR log
Log-normal VaR log
Skew-T VaR log
Exponential VaR log
Normal ES log
Gamma ES log
Chi-squared ES log
Skew-Normal ES log
Logistic ES log
Weibull ES log
Log-normal VaR log
Skew-T ES log
Exponential ES log

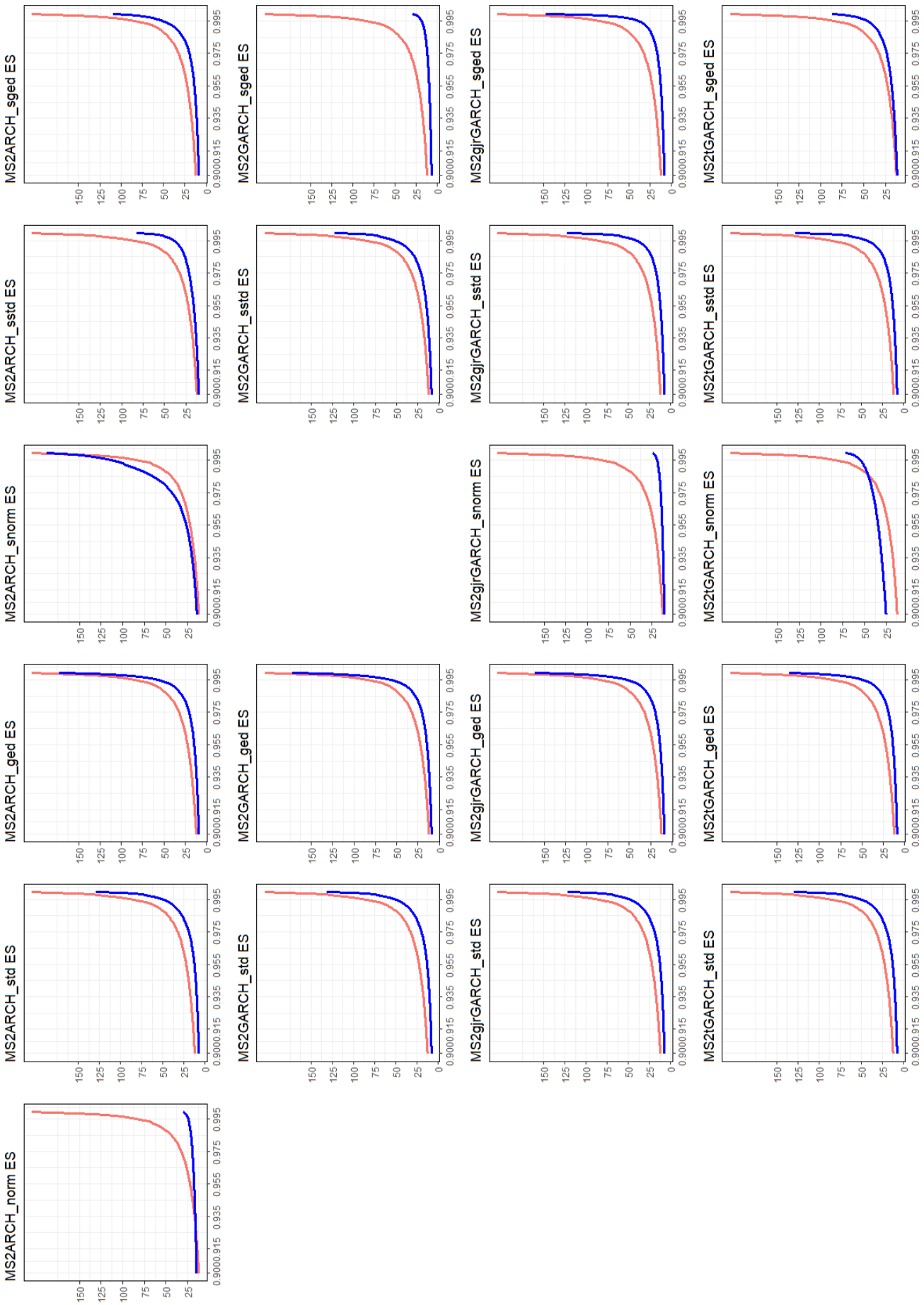

MS1ARCH_norm VaR
MS1ARCH_std VaR
MS1ARCH_ged VaR
MS1ARCH_snorm VaR
MS1ARCH_sstd VaR
MS1ARCH_sged VaR
MS1GARCH_norm VaR
MS1GARCH_std VaR
MS1GARCH_ged VaR
MS1GARCH_snorm VaR
MS1GARCH_sstd VaR
MS1GARCH_sged VaR
MS1gjrGARCH_norm VaR
MS1gjrGARCH_std VaR
MS1gjrGARCH_ged VaR
MS1gjrGARCH_snorm VaR
MS1gjrGARCH_sstd VaR
MS1gjrGARCH_sged VaR
MS1tGARCH_norm VaR
MS1tGARCH_std VaR
MS1tGARCH_ged VaR
MS1tGARCH_snorm VaR
MS1tGARCH_sstd VaR
MS1tGARCH_sged VaR
0 25 50 75 100 125 150
0.9000 0.915 0.935 0.955 0.975 0.995

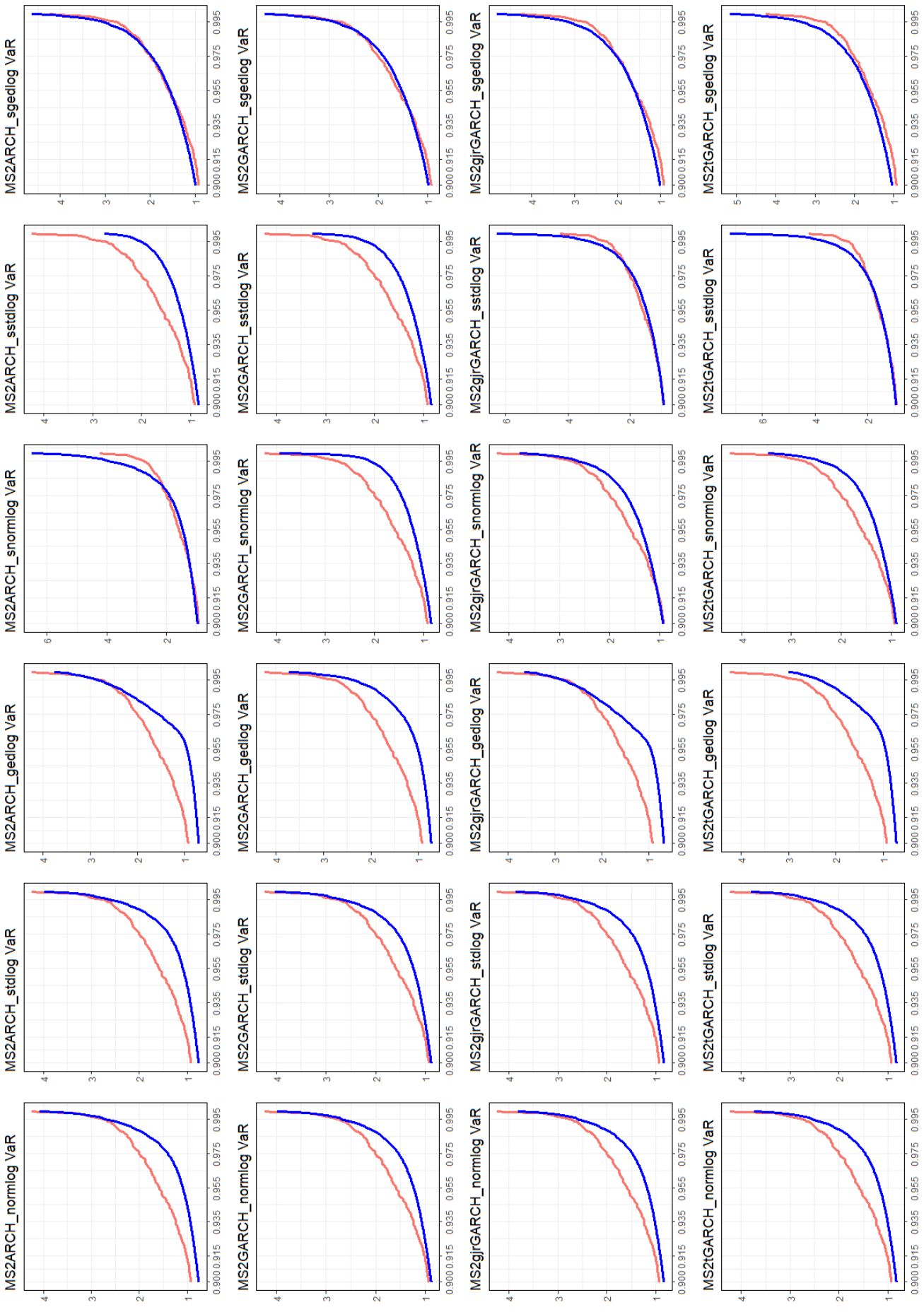

MS1ARCH_norm ES
MS1ARCH_std ES
MS1ARCH_ged ES
MS1ARCH_snorm ES
MS1ARCH_sstd ES
MS1ARCH_sged ES
MS1GARCH_norm ES
MS1GARCH_std ES
MS1GARCH_ged ES
MS1GARCH_snorm ES
MS1GARCH_sstd ES
MS1GARCH_sged ES
MS1gjrGARCH_norm ES
MS1gjrGARCH_std ES
MS1gjrGARCH_ged ES
MS1gjrGARCH_snorm ES
MS1gjrGARCH_sstd ES
MS1gjrGARCH_sged ES
MS1tGARCH_norm ES
MS1tGARCH_std ES
MS1tGARCH_ged ES
MS1tGARCH_snorm ES
MS1tGARCH_sstd ES
MS1tGARCH_sged ES
0
25
50
75
100
125
150
0.9000.915 0.935 0.955 0.975 0.995

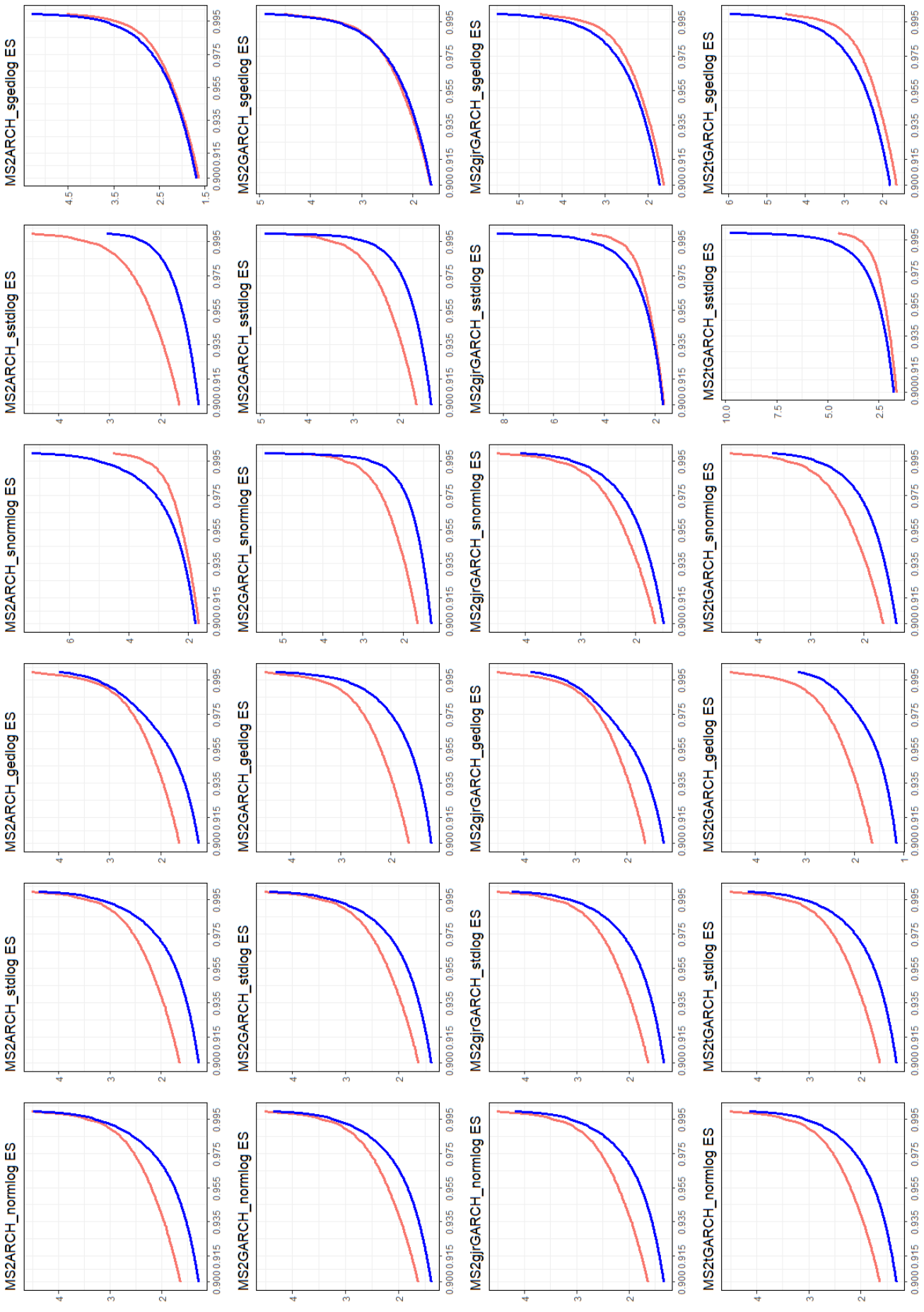

MS1ARCH_normlog VaR
MS1ARCH_stdlog VaR
MS1ARCH_gedlog VaR
MS1ARCH_snormlog VaR
MS1ARCH_sstdlog VaR
MS1ARCH_sgedlog VaR
MS1GARCH_normlog VaR
MS1GARCH_stdlog VaR
MS1GARCH_gedlog VaR
MS1GARCH_snormlog VaR
MS1GARCH_sstdlog VaR
MS1GARCH_sgedlog VaR
MS1gjrGARCH_normlog VaR
MS1gjrGARCH_stdlog VaR
MS1gjrGARCH_gedlog VaR
MS1gjrGARCH_snormlog VaR
MS1gjrGARCH_sstdlog VaR
MS1gjrGARCH_sgedlog VaR
MS1tGARCH_normlog VaR
MS1tGARCH_stdlog VaR
MS1tGARCH_gedlog VaR
MS1tGARCH_snormlog VaR
MS1tGARCH_sstdlog VaR
MS1tGARCH_sgedlog VaR
0.900 0.915 0.935 0.955 0.975 0.995


MS1ARCH_normlog ES
MS1ARCH_stdlog ES
MS1ARCH_gedlog ES
MS1ARCH_snormlog ES
MS1ARCH_sstdlog ES
MS1ARCH_sgedlog ES
MS1GARCH_normlog ES
MS1GARCH_stdlog ES
MS1GARCH_gedlog ES
MS1GARCH_snormlog ES
MS1GARCH_sstdlog ES
MS1GARCH_sgedlog ES
MS1gjrGARCH_normlog ES
MS1gjrGARCH_stdlog ES
MS1gjrGARCH_gedlog ES
MS1gjrGARCH_snormlog ES
MS1gjrGARCH_sstdlog ES
MS1gjrGARCH_sgedlog ES
MS1tGARCH_normlog ES
MS1tGARCH_stdlog ES
MS1tGARCH_gedlog ES
MS1tGARCH_snormlog ES
MS1tGARCH_sstdlog ES
MS1tGARCH_sgedlog ES


MS2ARCH_norm VaR
MS2ARCH_std VaR
MS2ARCH_ged VaR
MS2ARCH_snorm VaR
MS2ARCH_sstd VaR
MS2ARCH_sged VaR
MS2GARCH_std VaR
MS2GARCH_ged VaR
MS2GARCH_sstd VaR
MS2GARCH_sged VaR
MS2gjrGARCH_std VaR
MS2gjrGARCH_ged VaR
MS2gjrGARCH_snorm VaR
MS2gjrGARCH_sstd VaR
MS2gjrGARCH_sged VaR
MS2tGARCH_std VaR
MS2tGARCH_ged VaR
MS2tGARCH_snorm VaR
MS2tGARCH_sstd VaR
MS2tGARCH_sged VaR
0
25
50
75
100
125
150
0.900
0.915
0.935
0.955
0.975
0.995


MS2ARCH_sged ES
MS2GARCH_sged ES
MS2gjrGARCH_sged ES
MS2tGARCH_sged ES
MS2ARCH_sstd ES
MS2GARCH_sstd ES
MS2gjrGARCH_sstd ES
MS2tGARCH_sstd ES
MS2ARCH_snorm ES
MS2gjrGARCH_snorm ES
MS2tGARCH_snorm ES
MS2ARCH_ged ES
MS2GARCH_ged ES
MS2gjrGARCH_ged ES
MS2tGARCH_ged ES
MS2ARCH_std ES
MS2GARCH_std ES
MS2gjrGARCH_std ES
MS2tGARCH_std ES
MS2ARCH_norm ES
0
25
50
75
100
125
150
0.9000.915.935.955.975.995


MS2ARCH_normlog VaR
MS2ARCH_stdlog VaR
MS2ARCH_gedlog VaR
MS2ARCH_snormlog VaR
MS2ARCH_sstdlog VaR
MS2ARCH_sgedlog VaR
MS2GARCH_normlog VaR
MS2GARCH_stdlog VaR
MS2GARCH_gedlog VaR
MS2GARCH_snormlog VaR
MS2GARCH_sstdlog VaR
MS2GARCH_sgedlog VaR
MS2gjrGARCH_normlog VaR
MS2gjrGARCH_stdlog VaR
MS2gjrGARCH_gedlog VaR
MS2gjrGARCH_snormlog VaR
MS2gjrGARCH_sstdlog VaR
MS2gjrGARCH_sgedlog VaR
MS2tGARCH_normlog VaR
MS2tGARCH_stdlog VaR
MS2tGARCH_gedlog VaR
MS2tGARCH_snormlog VaR
MS2tGARCH_sstdlog VaR
MS2tGARCH_sgedlog VaR


MS2ARCH_normlog ES
MS2ARCH_stdlog ES
MS2ARCH_gedlog ES
MS2ARCH_snormlog ES
MS2ARCH_sstdlog ES
MS2ARCH_sgedlog ES
MS2GARCH_normlog ES
MS2GARCH_stdlog ES
MS2GARCH_gedlog ES
MS2GARCH_snormlog ES
MS2GARCH_sstdlog ES
MS2GARCH_sgedlog ES
MS2gjrGARCH_normlog ES
MS2gjrGARCH_stdlog ES
MS2gjrGARCH_gedlog ES
MS2gjrGARCH_snormlog ES
MS2gjrGARCH_sstdlog ES
MS2gjrGARCH_sgedlog ES
MS2tGARCH_normlog ES
MS2tGARCH_stdlog ES
MS2tGARCH_gedlog ES
MS2tGARCH_snormlog ES
MS2tGARCH_sstdlog ES
MS2tGARCH_sgedlog ES