INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Autogarch: An R package to automatically estimate and select GARCH models

Ricardo Filipe Leal Correia

Master in Finance

Supervisor:

PhD José Joaquim Dias Curto, Associate Professor with Habilitation, ISCTE-IUL

October, 2021

BUSINESS
SCHOOL

Department of Finance

Autogarch: An R package to automatically estimate and select GARCH models

Ricardo Filipe Leal Correia

Master in Finance

Supervisor:

PhD José Joaquim Dias Curto, Associate Professor with Habilitation, ISCTE-IUL

October, 2021

Acknowledgements

A word of appreciation for my family, for their words of wisdom. A big thank you to my girlfriend, for the friendship, and unconditional support. A special thanks to Prof. José Dias Curto, for helping me complete this work and leaving a mark in my academic journey.

Resumo

A inovação impulsionada pela evolução tecnológica nas finanças aumenta a procura por soluções de software. O desenvolvimento de ferramentas inovadoras inspiradas em conceitos teóricos impulsiona a produtividade tanto no ambiente académico como no mundo dos negócios.

A conversão de conceitos teóricos em ferramentas computacionais no âmbito do Autogarch package para o R é descrita. A estimação de modelos GARCH com base em séries temporais financeiras e a utilização de funções extratoras para classificar os modelos utilizando os critérios de informação subsequentes são algumas das funcionalidades disponíveis.

Os desenvolvimentos de código produzidos pelo autor estão disponíveis para utilização e modificação no seguinte repositório.

Palavras-Chave: Modelos GARCH, Séries Temporais, Estimação Automática, Funções Extratoras, R.

Classifcação JEL: C52, C87

Abstract

The technology-driven innovation in the financial industry creates demand for software-based solutions. Developing innovative tools based on theoretical concepts is the key for increased productivity in business and academic contexts.

The conversion of these theoretical concepts to computational tools in the Autogarch package for R is described, enabling the fitting of univariate GARCH models to financial time series and using extractor functions to rank models by the subsequent information criterion are some of the available functionalities. Additional features available in the Autogarch package will also be briefly discussed.

Code developments produced by the author are available for use and modification in the following repository.

Key Words: GARCH Models, time series, automatic fitting, extractor functions, R.

JEL Classification: C52, C87

Contents

1. Introduction

The development purpose of the Autogarch package is to automate the fitting and selection of GARCH models that best suit the characteristics of the data in analysis. The increasing popularity of R and the hands-on experience of the author in simplifying otherwise complex problems using this tool, served as inspiration to propose this package framework.

By optimizing the process, the end-user will be enriched with faster access to processed information about their models which, consequently, will increase the accuracy and efficiency of the decision-making process.

1.1 Research Context

The classic problem of applied econometrics is to determine how much a variable responds to the change in another variable, the ordinary least squares (OLS) estimation introduced by Legendre (1805) and co-credited to Gauss (1809) is the most appropriate tool to deal with these questions. The limitations of OLS arise when the homoskedasticity assumption is violated because it seeks to minimize residuals and generate the smallest standard errors possible, this violation will result in biased standard errors that in their turn will lead to incorrect conclusions regarding the significance of the regression coefficients.

Volatility is a key element to estimate market risk, pricing of derivatives, portfolio management and risk assessment. Forecasting volatility has vital importance for a financial institution, awareness of the current level of volatility of managed assets and its behavior in the future gives valuable insights about the overall enterprise risk level, especially for institutions that engage in derivatives trading which are typically leveraged. Quantitative forecasts provide estimates of future market trends, empirical evidence shows that future events are not totally unpredictable and experts in this field are devoted to creating techniques for volatility forecasting with the introduction of several models.

The findings in the volatility modeling field with the presentation of Autoregressive conditional heteroskedasticity (ARCH) by Engle (1982) and generalized autoregressive conditional heteroskedasticity (GARCH) by Bollerslev (1986) provided a statistical framework to analyze and forecast volatility. The commonly used tool by applied econometricians to deal with questions that arise from volatility are the ARCH/GARCH models, instead of treating heteroskedasticity as a hurdle that will impede modelling, the ARCH and GARCH models treat it as variance to be modelled. Since the introduction of the classic ARCH/GARCH models, various extensions have been proposed with more sophisticated characteristics, such as

threshold models that try to capture the asymmetry of the news impact and others that use different distributions than the normal to try and account for the skewness and excess kurtosis observed in the data, equity returns distributions are best described as being leptokurtic (excess kurtosis > 0).

1.2 Literature Review

Technological evolution results in an increasing adherence of statisticians to programming languages and statistical software as a fundamental tool to perform an accurate, rigorous, and detailed statistical analysis. This phenomenon empowered practitioners to come up with their own innovations enabling a revolution in the way that statistics are taught and applied. The advent of big data poses a constant challenge that propelled statistical breakthroughs even further, dealing with complex data sets in terms of volume, diversity of variables and the speed at which information is collected requires expertise in computational areas.

Statistics plays a major role in extracting meaningful, accurate data from a data set that grants the identification of patterns by translating a research question with the application of complex techniques and models. Statistical methodology is critical to make inferences as it helps the practitioner in handling problems such as biasedness, missing data, eliminating redundancies, and using effective data visualization methods.

Collection of literary evidence to substantiate the definition, planning, and execution of this project followed a structure that in a first stage was focused on R statistical software and the major advancements of the previous decades regarding package development. Some of the authors contemplated in this stage include Cook and Wickham (2008) who propose that progress in academia will come through the support of statistical computing, as it bridges statistics with other disciplines, R packages system is accessible for everybody to develop and implement groundbreaking research and the developmental process has shortened from years to months. Recognizing that statistical education must be supported with data management and programming skills is crucial, the expectation is that graduate students will have computational skills at par to their mathematical skills.

Innovative automation opportunities were analyzed by Staniak and Biecek (2019) who reviewed several tools for automated exploratory data analysis. Two groups of packages were examined, the first group automates exploratory data analysis and the second group of packages present data summaries. The problems identified include working with imperfect data, packages

that are not prepared to deal with constant variables, error messages that are not user friendly, new standards and conventions for package creation should be implemented.

We must contemplate three different roles for creating a statistics application, developer, statistician, and the end-user. Throughout the year's statisticians have evolved into programmers but the majority still use Microsoft Excel as their user interface of choice, the big disadvantage of this approach is that more advanced computations fall out of the Microsoft Excel spectrum. For many professionals R is the solution for this problem but the complexity of learning the syntax of a whole new programming language and of the R interface is intimidating for new users. Since 1998, there have been various projects to integrate R and Microsoft Excel with the development of various add-ins. Baier, *et al*. (2011) intend to contribute for the further integration of R and Microsoft Excel by replicating the rpart package in Microsoft Excel, this package provides us tools for analyzing complex data structures.

The orderbook package provides a wide range of tools for analyzing, visualize and simulate data associated with order books. Kane, et al. (2011) provide a framework for processing a high volume of orders for a specific financial instrument and create graphical representations allowing the user to draw more substantiated conclusions. Order book data is the fundamental input for active trading, simulations can give a trader valuable insight about the trend and value of a security by allowing faster volume-weighted average price (VWAP) estimations.

The second stage of this literary research was focused on studying the effectiveness of incorporating programming languages and computational skills courses in the curriculum of university programs. Analyzing the most important skills for a business school student to dominate nowadays and their level of computational knowledge is a very good starting point for designing innovative solutions and out of the box teaching methods that empower students to be confident with their skillset and applying it out of the academical context and in the corporate environment.

A study about the level of computer expertise of 140 newly accepted business school students, examined if they have skills to require exemption of an introductory course of computer fundamentals was performed by Wallace and Clariana (2005). College students have had experiences with technology since kindergarten and use it on an everyday basis, this reality created the assumption that students have the level of computer mastery required to face the challenges of a business school and many colleges eliminated computer literacy courses. This article defends that curriculum decisions must be based on data instead of mere assumptions,

to test the hypothesis that the implementation of an introductory computer literacy course plays an important role in the development of skills the authors gave an online computer proficiency test to students before and after receiving computer instruction.

The conclusions are the following, students revealed a significant improvement in the average results in the Microsoft Excel module after taking the introductory course and the assumption that first year students already possess the minimum required skills fails. The authors present their experience trying to teach business major students programming and describe their innovative approach to teaching this topic. There is a positive correlation between the number of hours in the lab practicing concepts and the ability of business students to absorb the knowledge. Programming languages are very diverse and have different spectrums we need to have in mind which type of task we are expecting our students to perform before deciding on which language is best suited to be taught. There is a clear-cut need to approach students with statistical computing by removing unnecessary hurdles and reinventing the curriculum of undergraduate statistics courses was supported by creating three types of commands, formulas, functions, and extractors with a clear and concise syntax making it user/student friendly. The goal identified by Pruim, et al. (2017) is to reduce the number of frameworks that a student must learn releasing cognitive capacity to be used in the learning process of introductory statistics.

Lastly, articles that delve deeper into statistical models were reviewed to obtain a better understanding about the purpose of modelling and model selection. Traditional statistical models are applied as a tool to deal with uncertainty which allows us to make inferences about stochastic variables and helps describing causal relationships. To make accurate conclusions from our models it is very important to select the model that fits our data the best and evaluate the goodness of fit. The information criteria were developed to evaluate if the model is approximate to the true distribution of the data. Technological advancements and the wide availability of computers have contributed to the construction of models that apply to complex data and stochastic processes. Statisticians have different points of view when it comes to defining the true purpose of modelling, for (Akaike, 1985) the goal is to predict new data as precisely as possible and for some practitioners the purpose is to distinguish the true distribution of our data. It all depends on the objective of our modelling efforts because the model that is best for predictions might not be the best for deducing the true distribution. In recent years statistical models are also viewed as vehicles to extract information that support our analysis (Konishi & Kitagawa, 2008). Model selection is a sensitive subject because models that are too

simple and can lead to wrongful predictions, on the other hand, a model that is too complex may be inflexible. Different information criteria lead to different conclusions generating confusions about which criteria to apply and in which situation. The more we study about each criteria our decisions will consequently be more diligent and reasonable. The Akaike Information Criteria (AIC) and the Bayesian Information Criteria (BIC) assess the goodness-of-fit of a model and adjust for overfitting providing a balance between the sensitivity of the predictions and the ability to apply this model to another dataset (Dziak, et al., 2019).

1.3 Research purpose and structure

The introduction of a variety of computational tools that allow the user to estimate several types of univariate GARCH models with different distributions will yield a valuable collection of data. The aim of the Autogarch package is to use that data to rank the models based on their respective information criterion and test the hypothesis that modelling using different distributions other than the normal will minimize them. Additional methods for importing time series from Microsoft Excel, computing returns, and estimating betas will be available. This document will discuss the specific features of the package and their implementation.

This research will be structured as following, the theoretical notions that sustained the package development will be described. The starting point is the in-depth analysis of the different GARCH models available, posteriorly the different underlying distributions used for model estimation are also scrutinized. Data analytic decomposition, statistical, conditional volatility estimations, return simulations are some of the concepts featured. Furthermore, the conclusions that stem from the elaboration of this research are identified, potential research opportunities are raised, with prospective applications for similar projects in the academic and business environment being suggested as ‘food for thought’.

1.4 Methodology

The methodology of a quantitative research project involves the collection of numerical data, processing and analyzing the existence of relevant patterns and causality effects. Experimental research is used to evaluate the impact of the change in an independent variable on a dependent variable *ceteris paribus*, in the context of this research the independent variable is the underlying conditional distribution of a GARCH model. This experiment assesses whether changing the underlying distribution has material impact in the subjacent GARCH model. Will the model make more accurate variance forecasts? How will the use of leptokurtic distributions on stock prices compare to using the normal distribution?

Methodology development has two fundamental drivers, answering the research questions and the creation of a user-friendly statistical package, to this effect a software development process that is based on a deliberate, structured, and methodical workflow was applied which requires each stage of development from embryonic to maturity to be executed in a rigid, sequential, and accurate environment. Prototypes are incomplete versions of the software that serve the purpose of testing new features in a stand-alone context of the package, it allows the developer to reduce the probability of syntax errors, bugs, and glitches by segmenting the project into smaller pieces. A proof-of-principle prototype will validate the feasibility of some fundamental functional features without all the functionalities of the final product, the next step after getting validation for the projected features is to collect data on the end-user by deploying a user experience prototype. The *ethos* of a developer is to deserve the trust of users by identifying that there is constant room for improvement and tailoring new functionalities to their needs.

2. Applied Development Concepts

A well-defined theoretical scope is the base for the development process as it defines the raw materials that will be embedded in our code, consequently defining the functionality range and our target audience. This chapter features the Autogarch process, and the subsequent underlying theoretical concepts used as the building block of the Autogarch package for R.

2.1 Logarithmic Returns

Given that x_t (t = 1, …, T) is a time series of stock and/or indices. The logarithmic returns of this time series are calculated as following:

$$R_{\log} = \ln\left(\frac{V_f}{V_i}\right) \tag{1}$$

where V_i is the initial price and V_f is the future price of an asset.

2.2 Simple Returns

Considering the time series described above. The simple returns of this time series are obtained using this method:

$$R = \frac{V_f - V_i}{V_i} \tag{2}$$

where V_i is the initial price and V_f is the future price of an asset.

2.3 Beta

Beta is a measure of systematic risk utilized in the capital asset pricing model (CAPM) of William Sharpe (1964) and John Lintner (1965) and it is considered a key breakthrough in asset pricing theory as it delivers intuitive reasoning on risk measures and the widely discussed risk-return relationship. Using the market volatility as a benchmark, the beta allows make sustained inferences regarding the volatility of a security or portfolio.

The mathematical formulation of this concept is:

$$\beta_p = \frac{\mathrm{Cov}(r_p, r_b)}{\mathrm{Var}(r_b)} \tag{3}$$

where r_p are the security or portfolio returns and r_b are the market returns.

2.4 Introduction of the GARCH class of models

Given that the time series of simple or logarithmic returns x_t (t = 1, …, T) that results from the process described in section 2.1 and 2.2 has specific properties such as:

1. Conditional mean of X given Y = y is defined as:

$$\mu_{X|Y} = E[X|y] = \sum_x x\, g(x|y) \tag{4}$$

2. Conditional variance of X given Y = y is:

$$\sigma^2_{X|y} = E[X^2|y] - \mu^2_{X|y} = \left[\sum_x x^2\, g(x|y)\right] - \mu^2_{X|y} \tag{5}$$

Conventional econometric models consider that the variance is constant, this means that when we compare each individual error with its predicted value the variance error persists at the same value. However, this assumption is rejected when we apply these models to financial time series, it is widely assumed that the volatility of financial time series follows a time-varying function.

2.4.1 ARMA Model

In general, models for time series data can have multiple formats that represent a wide variety of stochastic processes. Two of the most popular linear time series models in the literature are the Autoregressive (AR) and Moving average (MA). The Autoregressive moving average model was proposed by Peter Whittle (1951) and it is a combination of the two models previously described, it aims to provide a simplified and easily implemented description of a stochastic process.

An AR(p) model has p autoregressive terms:

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t, \tag{6}$$

where ε_t is white noise. This is like a multiple linear regression but with lagged values of y_t as predictors. Autoregressive models are notorious for their flexibility and their handling of a wide range of time series patterns.

An MA(q) model has q order moving terms:

$$y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}, \tag{7}$$

where ε_t is white noise. Since we do not observe the values of ε_t, it is not a regression in the usual sense.

Hence, the ARMA(p,q) model consists of the combination of AR(p) and MA(q) models with p autoregressive terms and q moving-average terms:

$$y_t = \mu + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} \tag{8}$$

The ARMA (p, q) main purpose is to model the conditional mean by attempting to capture the momentum and mean reversion effects observed in the markets through the autoregressive part of the model and the shock effects observed in the white noise terms through the moving average part of the model which can be thought of as unexpected events e.g. surprise earnings, political turmoil.

2.4.2 ARCH Model

This concept was developed by Robert F. Engle in 1982, ARCH models played a major role in demonstrating that the volatility clusters, which is relevant evidence for investors whose holding period spills over different time periods. Engle identified that econometric models had room for improvement by substituting the constant volatility assumption for a conditional volatility assumption, and by acknowledging that past observations impact future observations. Before ARCH, the *modus operandi* of practitioners consisted in recognizing that volatility differed from period to period, but the constant volatility assumption would persist because there was no modelling option that provided a conditional volatility assumption.

The ARCH model is stated as:

$$\begin{gathered} y_t = x_t' b + \epsilon_t \\ \sigma_t^2 = \alpha_0 + \alpha_1 \epsilon_{t-1}^2 + \cdots + \alpha_q \epsilon_{t-q}^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i \, \epsilon_{t-i}^2 \end{gathered} \tag{9}$$

where $\alpha_0 > 0$ and $\alpha_i \geq 0$, i > 0.

2.4.3 GARCH Model

Generalized autoregressive conditional heteroskedasticity (GARCH) was introduced by Tim Bollerslev (1986), it adopts the assumption that the variance of the error term follows an ARMA process. GARCH models are particularly useful when the error term is heteroskedastic, that is, the variance of the error term is not constant and there is an unpredictable pattern of variation. Hence, if models that assume constant variance are used on heteroskedastic data,

which macroeconomic and financial data typically is, the inferences that can be made from the model are biased.

The notation of the GARCH model originally developed by Bollerslev (1986) is given by:

$$y_t = x_t' b + \epsilon_t$$

$$\epsilon_t \mid \psi_{t-1} \sim \mathcal{N}(0, \sigma_t^2) \tag{10}$$

$$\sigma_t^2 = \omega + \alpha_1 \epsilon_{t-1}^2 + \cdots + \alpha_q \epsilon_{t-q}^2 + \beta_1 \sigma_{t-1}^2 + \cdots + \beta_p \sigma_{t-p}^2 = \omega + \sum_{i=1}^{q} \alpha_i \epsilon_{t-i}^2 + \sum_{i=1}^{p} \beta_i \sigma_{t-i}^2$$

GARCH models place superior weight upon more recent observations than in past observations, therefore, it is believed that recent observations will have greater influence in future observations.

2.4.4 GJR-GARCH Model

The Glosten, Jagannathan and Runkle-GARCH (GJR-GARCH) (1993) model differs from the original GARCH model because it considers asymmetrical shocks, this means that the sign of the shock is a function of its own sign. The definition of the GJR-GARCH (1,1) model is the following:

$$\sigma_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \delta_1 I(\varepsilon_{t-1} < 0) \varepsilon_{t-1}^2 + \beta_1 \sigma_{t-1}^2 \tag{11}$$

Where $I(\varepsilon_{t-1} < 0)$ is an indicator function of events commonly denominated as dummy variables, whose value is one if the shock is negative, otherwise the value is null.

The model encapsulates the asymmetrical nature of financial time series by including this indicator function of events. The predictive model is depicted as follows:

$$\sigma_{t+1}^2 = \alpha_0 + \alpha_1 \varepsilon_t^2 + \delta_1 I(\varepsilon_t < 0) \varepsilon_t^2 + \beta_1 \sigma_t^2 \tag{12}$$

2.4.5 EGARCH Model

The exponential generalized autoregressive conditional heteroskedastic (EGARCH) model by Nelson (1991) was developed to address the limitations of the classic GARCH model in capturing the asymmetric influence of returns. In other words, the evolution of the volatility trend is dependent on the trend that the asset returns follow. As returns tend to be consistently negative and the prices of an asset decreases, the volatility will consistently register higher levels. In contrast, if the performance of an asset is continuously positive through a given period, as the asset prices rise the volatility will drop consistently to lower recorded values.

The exponential GARCH may generally be specified as:

$$\log_e(\sigma_t^2) = \left(\omega + \sum_{j=1}^{m} \zeta_j \, v_{jt}\right) + \sum_{j=1}^{q} \left(\alpha_j z_{t-j} + \gamma_j \left(|z_{t-j}| - E|z_{t-j}|\right)\right) + \sum_{j=1}^{p} \beta_j \log_e\left(\sigma_{t-j}^2\right) \quad (13)$$

2.4.6 IGARCH Model

The traditional GARCH context shows evidence of a conditional volatility process which is highly persistent but not covariance-stationary, this suggests that a model that consider shocks will have a permanent effect on volatility is more suitable. This is a feature of the integrated GARCH model (IGARCH) developed by Engle and Bollerslev (1986). Before choosing between the standard GARCH process and the IGARCH, practitioners need to detect the presence of structural breaks. An IGARCH (1,1) model is specified as following:

$$a_t = \sigma_t \epsilon_t, \quad \sigma_t^2 = \alpha_0 + \beta_1 \sigma_{t-1}^2 + (1 - \beta_1) a_{t-1}^2 \quad (14)$$

A core feature of the IGARCH models is the persistent impact of past squared shocks on a_t^2 (Tsay, 2010).

2.4.7 APARCH Model

The asymmetric power ARCH model of Ding, Granger and Engle (1993) was developed to express characteristics of financial time series such as fat-tails, excess kurtosis, and both Taylor and leverage effects, named after Taylor (1986) who observed sample autocorrelation and concluded that the sample autocorrelation of absolute returns is regularly higher than that of squared returns.

The general structure is expressed as follows:

$$y_t = x_t \xi + \varepsilon_t \quad t = 1,2 \ldots . T$$

$$\sigma_{\mathrm{t}}^{\delta} = \omega + \sum_{j=1}^{q} \alpha_j \left(|\varepsilon_{t-j}| - \gamma_j \varepsilon_{t-j}\right)^{\delta} + \sum_{i=1}^{p} \beta_i \left(\sigma_{t-i}\right)^{\delta} \quad (15)$$

$$\varepsilon_{\mathrm{t}} = \sigma_{\mathrm{t}} \mathrm{z}_{\mathrm{t}}, \mathrm{z}_{\mathrm{t}} \sim \mathrm{N}(0,1)$$

$$k\left(\varepsilon_{\mathrm{t-j}}\right) = \left|\varepsilon_{\mathrm{t-j}}\right| - \gamma_{\mathrm{j}} \varepsilon_{\mathrm{t-j}}$$

2.4.8 Component GARCH Model

The component model of Engle and Lee (1993) is constituted by two additive GARCH (1,1) components. The first one is a temporary component, in contrast, the other is considered a long-run component. Volatility component models are the topic of numerous research studies that attempt to study their capacity to model complex dynamics and try to find evidence to sustain the belief of the scientific community, that they can cope with structural breaks or non-

stationarities in asset price volatility. Given q_t as the long-run component of conditional variance, the component model is specified as follows:

$$\sigma_t^2 = q_t + \sum_{j=1}^{q} \alpha_j \left(\varepsilon_{t-j}^2 - q_{t-j}\right) + \sum_{j=1}^{p} \beta_j \left(\sigma_{t-j}^2 - q_{t-j}\right)$$
$$q_t = \omega + \rho q_{t-1} + \phi(\varepsilon_{t-1}^2 - \sigma_{t-1}^2) \tag{16}$$

2.5 Underlying distribution of the model

Equity return distributions are usually leptokurtic and negatively skewed, this evidence leads to the rejection of traditional GARCH process by Bollerslev (1986) assumption of normal distribution. Hence, it is imperative that we find the distribution that best captures the specific characteristics of our data. The Autogarch package facilitates this search for the best suited distribution by estimating the same model with nine different distributions. The theoretical context and characteristics of these distributions will be discussed in this section.

2.5.1 The normal distribution

The normal distribution originally proposed by Gauss (1809), commonly known as the bell curve or the Gaussian curve, is the most popular distribution in statistical and econometric fields. The first two moments of the distribution (mean and variance) describe it in full. Standard deviation controls the dispersion of the distribution, a smaller standard deviation indicates tight clustering around the mean, on the other hand, a higher standard deviation points to more spreading around the mean.

The probability density function stated as:

$$f(x) = \frac{e^{\frac{-0.5(x-\mu)^2}{\sigma^2}}}{\sigma\sqrt{2\pi}} \tag{17}$$

Key properties of the normal distribution include symmetry around the mean, zero excess kurtosis which points to a mesokurtic distribution, and the mean, mode and median are all equal.

2.5.2 Skew-Normal Distribution

Despite how popular the normal distribution is, modelling financial time series using only the normal distribution has proven to be a difficult task due to asymmetry and fat-tails which leads to the conclusion that using this distribution is not the best option in this case. Given this, the skew-normal distribution introduced by O'Hagan and Leonard (1976) is a

generalization of the normal distribution that includes an allowance for asymmetries in the data set.

Let us consider a random variable x with the following probability density function:

$$f(x) = 2\phi(x)\Phi(\alpha x) \tag{18}$$

where α is a random constant,

$$\phi(x) = \exp(-x^2/2)/\sqrt{2\pi}$$

$$\Phi(\alpha x) = \int_{-\infty}^{\alpha x} \phi\,(t)dt$$

representing the Gaussian density function and the distribution function. The component α controls the shape of the density function, which has some specific characteristics, if $\alpha = 0$ there is no skewness and we have the density of the normal distribution, as α increases the skewness of the distribution also increases, a sign change of α will cause a change of the density to the reverse side of the y-axis.

2.5.3 Student's t-distribution

The student's t-distribution was first discovered by W.S. Gosset (1908) while working for an Irish brewery, the paper was published under the pseudonym Student, hence the name of the distribution. This distribution plays a major role in several statistical analyses, for example, the student's t statistical significance test between two sample means and to assess the statistical significance of the regressors in a linear regression model. The shape of the distribution identical to the normal distribution shape, except for heavier tails which means that the occurrence of extreme values is more prevalent. The theoretical representation of the distribution is as follows:

$$f(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\beta\nu\pi}\Gamma\left(\frac{\nu}{2}\right)}\left(1 + \frac{(x-\alpha)^2}{\beta\nu}\right)^{-\left(\frac{\nu+1}{2}\right)} \tag{19}$$

Skewed generalization of the student's t-distribution established by Fernandez and Steel (1998) is also available in the Autogarch package and it is a special case of the Generalized error distribution described below.

2.5.4 Generalized error distribution

The generalized error distribution (GED) is also a candidate for the description of financial returns. It is an "error" distribution that is a generalization of the normal distribution, that has a multivariate form, contains a parametric unbounded kurtosis and special

cases that resemble the normal distribution. The most appropriate use of this distribution is when it is believed that there is special interest in the errors around the mean.

Three main parameters define this distribution that is part of a wider family of exponential distributions, conditional density is represented by:

$$f(x) = \frac{\kappa e^{-0.5\left|\frac{x-\alpha}{\beta}\right|^{\kappa}}}{2^{1+\kappa^{-1}}\beta\Gamma(\kappa^{-1})} \tag{20}$$

where α, β and κ represent the location, dispersion, and shape of the distribution. Since we are in the presence of a symmetric distribution, the location parameter represents the mean, mode, and the median. Generalized error distribution also has a skewed variant proposed by Ferreira and Steel (2006).

2.5.5 Generalized hyperbolic distribution

Generalized hyperbolic (GH) distributions were suggested by Barndorff-Nielsen (1978), their application in Finance was studied by Eberlein and Keller (1995), who constructed stochastic processes based on this distribution. Some key characteristics of this distribution include semi-heavy kurtosis which makes it an attractive distribution for modelling in the economic field, especially financial markets, and risk management, it is a natural adaptation to the most efficient volatility model by substituting the generalized inverse Gaussian.

Its probability density function is defined as a blend between the normal mean-variance relationship and the generalized inverse gaussian distribution given by:

$$GH(\lambda, \alpha, \beta, \delta, \mu) := N(\mu + \beta y, y) \circ GIG\left(\lambda, \delta, \sqrt{\alpha^2 - \beta^2}\right) \tag{21}$$

where the most popular five parameters describe the location and scale (μ, δ, respectively), α and β control the shape of the distribution, while the values of λ will describe notorious cases, for example, $\lambda = 1$ will refer to the generalized hyperbolic distribution and $\lambda = 0.5$ will describe the generalized inverse Gaussian distribution.

The parameter restrictions of the generalized hyperbolic distribution imply that these mandatory constraints will have to be satisfied:

$$\lambda, \mu \in \mathbb{R} \quad \text{and} \quad \begin{array}{l} \delta \geq 0, 0 \leq |\beta| < \alpha \text{ if } \lambda > 0 \\ \delta > 0, 0 \leq |\beta| < \alpha \text{ if } \lambda = 0 \\ \delta > 0, 0 \leq |\beta| \leq \alpha \text{ if } \lambda < 0 \end{array} \tag{22}$$

2.5.6 Normal-inverse Gaussian distribution

The normal-inverse Gaussian distribution (NIG) was proposed by Blaesild (1981), it is derived from the GED distribution, and it is integrating part of the curriculum in mathematical finance master's and PhD programs in every major university throughout the world. The applications of this distribution in the finance field include modelling stochastic volatility, estimating expected tail loss and value at risk (VaR). The distribution has the capacity to model symmetric and asymmetric data sets with the possibility to accommodate distended left and right tails. The tails of the NIG distribution are classified as semi-heavy, that is, the tails are much heavier than the normal distribution, but they might still not to be able to deal with extreme cases of tail heaviness. The distribution is controlled by two parameters δ and γ that define scale and shape, respectively. As δ increases the density function will flatten and the range of values of the density is smaller. In contrast, the higher the value of γ, the probability density function will also achieve new highs.

The NIG probability density function is described as:

$$f_{IG}(x, \delta, \gamma) = \frac{\delta}{\sqrt{2\pi}} x^{-3/2} \exp\left(\delta\gamma - \frac{\delta^2 x^{-1} + \gamma^2 x}{2}\right) \tag{23}$$

2.5.7 Johnson's SU-Distribution

The Johnson's SU-distribution is the product of a research conducted by N. L. Johnson (1949) with the objective of proposing a transformation of the normal distribution. The procedure of N. L. Johnson (1949) was originally centered around the moments extracted from the observed data and he used a graphical calculator. In the next years, algebraic transformations were proposed to increase accuracy and replace the graphical calculator, as the computer era started flourishing, algorithms were developed in the FORTRAN computer to fit a Johnson's SU-distribution curve and the distribution was permanently attached to the field of computational statistics. It incorporates a family of four distributions, the normal distribution, the lognormal distribution, the SB, and SU distribution, where "B" stands for bounded, and "U" stands for unbounded. The distribution is considered one of the most flexible in adapting to asymmetries and kurtosis in the error distribution.

Despite its flexibility and simplicity, applied research using this distribution in the econometric field were scarce until the beginning of the twenty first century, when the distribution started being more frequently applied to multivariate GARCH models.

The probability density function commonly denoted as:

$$f(x) = \frac{\delta \exp\left(-\frac{1}{2}(\gamma + \delta \sinh^{-1} z)^2\right)}{\xi\sqrt{2\pi}\sqrt{z^2+1}} \tag{24}$$

where $z = \frac{x-\lambda}{\xi}$,

and γ , δ are shape parameters with $\delta > 0$ being one of the constraints, λ and ξ controls the location and scale parameters, respectively. With $\xi > 0$ being the last constraint.

2.6 Model selection criteria

After the subsequent calculation of returns, model specification and fitting different types of GARCH models with different underlying distributions, the last step of the Autogarch process is to sort the models according to the value of information criterion and the maximum likelihood estimator (MLE).

2.6.1 Akaike information criterion

The Akaike information criterion (AIC) assesses the goodness-of-fit of the model to our data without overfitting. The AIC penalizes the model for overly complex fitting, there is not much use for the AIC as a stand-alone indicator, to take full advantage of its use it must be compared with the AIC of a rival model, hence it is used as a model selection tool. The lower the AIC score of a model, it is considered that the model achieves a better balance between the capacity to fit the data properly and to avoid over-parametrization and complex model structures.

The AIC is named after its proponent, Hirotugu Akaike, who developed it in the early 1970's. Given a statistical model, let *k* be the count of parameters estimated by the model and $\hat{L}$ the maximum value of the likelihood function, then AIC is given as follows:

$$\text{AIC} = 2k - 2\ln(\hat{L}) \tag{25}$$

2.6.2 Baesyan information criterion

The Bayesian information criterion (BIC) also known as the Schwarz information criterion as it derives from a paper by Gideon Schwarz (1978), is used to select among formal econometric models. BIC penalizes unnecessary model complexity more than the AIC and it has been considered a too liberal of a criterion, in some cases it selects models with many spurious explanatory variables and an extended family of Bayes information criterion has been proposed. The extended Bayes information criterion have proved especially useful in variable

selection in situations where there are many explanatory variables, and the sample size is small. Generally, the BIC is defined as:

$$\mathrm{BIC} = k\ln(n) - 2\ln(\hat{L}) \tag{26}$$

2.6.3 Shibata information Criterion

Shibata (1976) demonstrated through his empirical research that the AIC tends to choose overly complex models. Given this, the Shibata information criterion (SIC) was designed to penalize the overfitting. Formally, it is represented as:

$$SIC = \frac{-2LL}{N} + \log_e\left(\frac{(N+2m)}{N}\right) \tag{27}$$

2.6.4 Hannan-Quinn information criterion

The Hannan-Quinn information criterion (HQC), named after the proponent authors (Hannan & Quinn, 1979), was designed to have the slowest growing penalty term for complexity compared to the AIC and the BIC. The HQC was originally defined as:

$$\mathrm{HQC} = -2L_{\max} + 2k\ln(\ln(n)) \tag{28}$$

Unlike the AIC, the HQC is not asymptotically efficient, essentially it will require a higher number of observations to reach desired performance. An efficient estimator is characterized by small variance or small error, suggesting that there is small deviation between the estimated value and the observed value.

2.6.5 Maximum likelihood estimator

The maximum likelihood estimator (MLE) determines the best hypothesis among a very broad set of hypotheses. The MLE has various desirable properties, especially in the case of large samples, this includes consistency, in other words, as the sample size increases the estimator becomes more concentrated around the mean, and the properties of the log likelihood surface, given the shape of the surface at the maximum likelihood estimate, if the surface is flat then the variance is greater than if the surface has an accentuated curvature, thus the variance will be smaller. The likelihood function is the density function deemed the function of θ:

$$\mathrm{L}(\theta \mid \mathrm{x}) = \mathrm{f}(\mathrm{x} \mid \theta), \theta \in \Theta \tag{29}$$

The MLE is generally defined as:

$$\hat{\theta}(\mathrm{x}) = \mathrm{argmaxL}(\theta \mid \mathrm{x}) \tag{30}$$

3. The Autogarch Package: development notes

Firstly, the data selection process is discussed, posteriorly a detailed description of the process followed by our software-based solution ensues. This chapter walks the reader through the reasoning behind the fundamental developmental decisions that shaped the Autogarch package.

3.1 Data Selection

To develop, test and maintain the Autogarch package we need to use financial time series. The source is the Yahoo Finance Application Programming Interface (API), by using quick snippets of code it is possible to obtain a very high volume of data on a structured dataset in a matter of seconds. Microsoft Excel linked data types feature is an alternative way to access the Yahoo Finance API without coding, it allows the user to collect information ranging from the closing price of the stock to the number of employees of the corporation under analysis.

To take full advantage of the abilities of the Autogarch package, the dataset should contain time series of stocks prices and indices points. Specific characteristics such as the data frequency (daily, weekly, monthly) and the sampling period are defined by the user. In the developmental stage of the Autogarch package, the dataset used for development purposes contains daily information on five companies (Amazon, Facebook, Netflix, Ford, General Electric) that are New York Stock Exchange listed and the S&P 500 (GSPC), the sampling period is from January of 2016 to July of 2020, a total of 1142 observations. The package is designed to handle stock prices data independently of the sampling period, however, potential alternatives include the use of high frequency data (intra-day, hourly) as algorithmic trading becomes an increasingly popular topic and the use of data from an historically important period that marked economical and financial history (1980's oil glut, dotcom bubble, subprime mortgage crisis). In the mature stage of development, the Bloomberg terminal will be an alternative data source, access to a wider range of equity time series from all around the world through a flexible platform for finance professionals that need real-time data and analytics is important to test and adjust functionalities of the package.

3.2 Autogarch package

To take full advantage of the package functionalities the user must import to R a time series of stocks and/or indices either by importing a Microsoft Excel file that contains a time series, or by using the quantmod (Ryan & Ulrich, 2020) package which makes automatic calls to the Yahoo Finance application programming interface (API) to obtain a time series.

For the beginner level user, the data importing process can be the first hurdle, although RStudio provides a comprehensive user framework for this task. The development process tried to give a response to this issue by triggering a pop-up window that allows the user to choose the file that contains the time series interactively through a call to the **`import_file`** function,

```
> args(import_file)

function (variable_name = "df", sheet_name = NULL)
```

- variable_name (default = "df". The name of the variable to be assigned to the Global Environment.)
- sheet_name (default = NULL. Applicable when the user intends to read data from a specific sheet given a file with multiple sheets.)
- Link to Source Code

Upon completion of this step, to estimate our models, the user must use the corresponding imported time series and compute the returns of the stocks and/or indices. Since there is a lot of debate between academic researchers and finance practitioners about the appropriate use of simple or logarithmic returns and research by (Miskolczi, 2017) shows that the riskiness of an asset depends on the return type utilized for estimation, the Autogarch package empowers users with the capability of using either method (see section 3.1.1 and 3.1.2).

Using the **`returns`** function will yield a data frame with the desired returns estimation:

```
> args(returns)

function (x, simple = FALSE, view = TRUE)
```

- x (input data frame containing a time series.)
- simple (boolean variable, FALSE logarithmic returns (3.1.1), TRUE simple returns (3.1.2).)
- view (boolean variable, TRUE opens a tab with the output data frame.)
- Link to Source Code

The resulting data frame of the rentability calculation is the raw material for all the remainder processes that can be executed with the Autogarch package. To estimate the betas the user must specify four input parameters, the respective data frame of returns of the companies for which they aim to estimate this metric, the start date, and the end date of the estimation period, finally the market ticker to be used as a benchmark. (see section 3.1.3).

This set of parameters defined by the package to control the workflow of the **beta** method are the following:

```
> args(beta)
function (x, mkt_ticker, start_date, end_date)
```

- x (a data frame with the estimated returns.)
- mkt_ticker (a string identifying the column header of the market returns.)
- start_date (a string with the start date for the estimation period.)
- end_date (a string with the end date for the estimation period.)
- Link to Source Code

For model specification and estimation purposes we give the user the possibility to use six different types of GARCH models (see section 3.2) and nine different types of underlying distributions (see section 3.3) allowing the user to automatically perform a grand total of fifty-four different model estimations.

The specification of the model parameters is a vital step of modelling. By applying the `init_spec` function the user will specify the parameters of the models that will be fitted in a later stage to the returns data obtained in previous steps.

```
> args(init_spec)

function (model, all = FALSE)
```

- model (a string or vector of strings identifying specific models, available options listed below.)
- all (boolean variable, TRUE all the available models are specified, FALSE only the user inputs in the model argument are specified.)
- Link to Source Code

A posteriori, we can proceed with model fitting, given that all the required inputs have been obtained (returns, model parameters specified). The **`init_fit`** function gives the user the ability to fit the model, using the model specification variables already stored in memory by the **`init_spec`** method and the previously estimated returns.

```
> args(init_fit)

function (x, spec_rm = TRUE)
```

- x (a data frame with the estimated returns.)
- spec_rm (boolean variable, True removes the model spefication variables stored in memory, False does not remove the model specification variables.)
- Link to Source Code

The resulting data from all these estimations can be used for numerous financial, numerical, and statistical analyses. In this case, the data that is used by the Autogarch package are the information criteria and the maximum likelihood estimator (MLE) (see section 2.4), the user can aggregate these indicators for all the models in a single data frame.

Combining the model classification criteria in a data frame gives the user a visual of which models from the estimated above best suit the analysis that is being conducted. The **`agg_models`** method materializes the will synthetize model classification criteria in a single easily accessible data structure.

```
> args(agg_models)
function (models, all, view = TRUE)
```

- models (a string or vector of strings identifying specific models.)
- all (boolean variable, TRUE all the available models are specified, FALSE only the user inputs in the model argument are specified.)
- view (boolean variable, TRUE opens a tab with the output data frame.)
- Link to Source Code

	Akaike	Bayes	Shibata	Hannan-Quinn	LogLik
sGARCH.ged.fit	-6.967188	-6.936270	-6.967262	-6.955512	3981.781
sGARCH.jsu.fit	-6.986536	-6.951201	-6.986633	-6.973192	3993.819
sGARCH.nig.fit	-6.986514	-6.951179	-6.986612	-6.973171	3993.806

Figure 3-1 agg_models function example output

Lastly, taking the output data frame of the **`agg_models`** method as an input the **`rank_models`** will print the final output of the Autogarch package to the console, informing the user about which model and respective underlying distribution achieved minimized the information criteria.

```
> args(rank_models)
function (info_criteria_data)
```

- data frame populated with the information criteria of previously estimated models
- Link to Source Code

4. Empirical applications to stock returns

Data is the fundamental end-user input that sets the Autogarch package in motion, allied with a structured development method it is crucial to give the user as much output as possible for their input, meaning that the amount of information generated by the Autogarch process will give the end-user a swift, comprehensive, and precise depiction of the data in analysis.

4.1 Data analytic decomposition

Descriptive statistics have the power of summarizing and quantitively describe the characteristics of a given data set. Using the logarithmic returns calculation of the Autogarch package as a starting point we have produced measures of central tendency and dispersion to accurately describe the data used for development. The first contact with the characteristics of our data is in the pre-processing stage, where we aim to prepare the data for the modelling stage, which is regarded as the most important, fundamentally the models estimated will reflect some of these embedded features. Practitioners must acquire deep knowledge about their data, it will come handy when it is time to make judgement calls, it is on this stage that the Autogarch package steps in, facilitating these decisions by providing the end-user with valuable information. Raw data is very difficult to analyze, trends and patterns identification is challenging to perform.

	FB	NFLX	AMZN	F	GE
Beta	0.93	0.62	0.59	1.07	1.27

Figure 4-1 Beta estimation for the first semester of 2020

Using the logarithmic returns as an input for the beta function included in the Autogarch package we provide automated estimates of the beta for each stock in analysis. We can see that the first semester of 2020 was atypical due to the Covid-19 pandemic outbreak, usually when we look at the beta of these five companies listed in Figure 4-1 we will see a higher beta from those included in the info-tech sector (FB, NFLX, AMZN), in contrast, here we see that due to the pandemic-related uncertainty, there was a shift of investor confidence from the theoretically more volatile companies in the info-tech sector to the companies in more defensive sectors such as electricity (GE) and automobile manufacturing (F) resulting on an inversion of the usual beta paradigm.

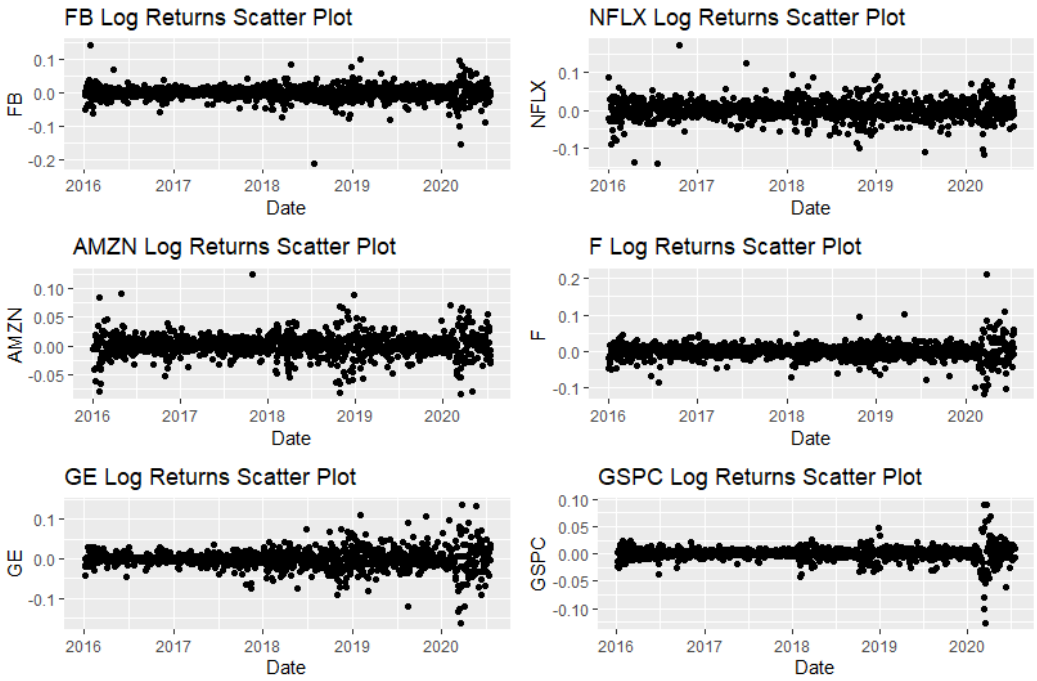

Figure 4-2 Logarithmic Returns Scatter Plots

As expected, we can observe a degree of correlation between each individual stock and the market. Another remark can be made regarding volatility clustering, more volatile periods tend to coincide in time between the assets targeted by our analyses, in this case, the volatility clusters that standout occurred in January 2019 and March 2020, the first is related with a market correction and the latter is connected to the negative shock of the Covid-19 pandemic. Volatility clustering phenom can also be observed by analyzing the average of the returns, which on the long run tends to zero, and across all the scatter plots we can see a clear pattern of long periods where average returns are close to zero interrupted by casual volatility spikes.

It is easy to make the connection from volatility clustering to modelling, it is in models that practitioners place their confidence to help them anticipate a volatile period, which will allow them to sail through uncertain periods unscathed and come out of them on top. The underlying distribution of a model is the key for this equation, modelling our data to the correct distribution will give our model the capacity to pinpoint volatile periods, and to be able to check which distribution is more suitable, we start by visualizing the density curve of the stocks and indices under analysis.

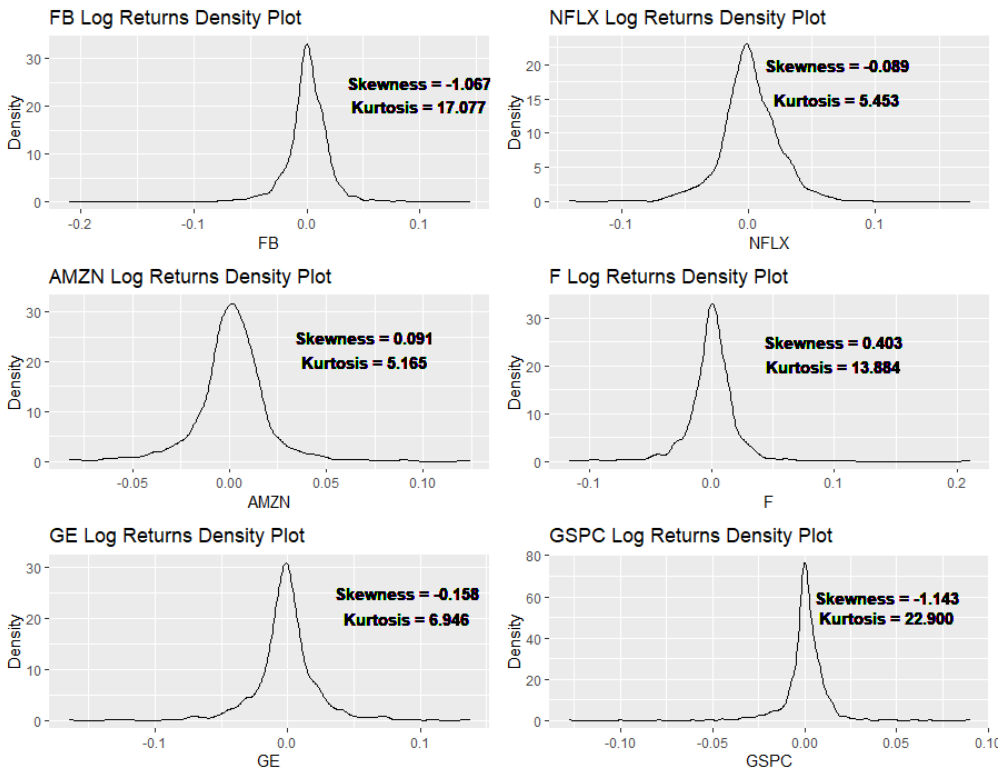

Figure 4-3 Logarithmic Returns Density Plots

The staple characteristics of the normal distribution are symmetry and mesokurtic tails however, equity returns disrupt this *status quo*. Figure 4-3 corroborates this statement; we can see that all the stocks and indices in analysis are either negatively or positively skewed and leptokurtic (kurtosis > 3). This complex and intriguing dynamic of stocks has been the attracting factor for researchers from a plethora of fields, the two problems mentioned add to the complexity but there are many more whose relevance deserves to be the subject of research. Fat-tails indicate that it is possible to win or lose much more than the normal distribution indicates, it is known that this distribution has an history of underestimating large events and therefore it is considered unreliable for risk analysis.

For example, events like the 1987 stock market crash commonly known as the 'Black Monday', where the S&P 500 (GSPC) fell by more than twenty standard deviations is impossible if the market returns strictly followed the Gaussian law, causes of this crash are still to this day not definite, stocks were on an outstanding bull-run at the time. In the aftermath of this crash, regulations were introduced to prevent similar events from occurring in the future, circuit breakers were implemented, giving exchanges the power to halt trading when the market declines below a previously determined threshold on a single trading day, the rationale being that this forced trading hiatus will ease panic selling and negative impacts will be contained.

4.1.1 Jarque-Bera test

The most popular test for normality testing purposes is the one developed by (Jarque & Bera, 1987) that tests the hypothesis that the skewness and kurtosis of the sample data match the normal distribution. The JB test statistic is formulated as following:

$$JB = \frac{n}{6}\left(S^2 + \frac{1}{4}(K-3)^2\right) \tag{31}$$

The test statistic always assumes positive values, the further from zero is the value of the test statistic it indicates that the sample data does not have a normal distribution. The null hypothesis is the aggregation of two conditions, the skewness and the excess kurtosis being simultaneously equal to zero, the third and fourth central moments of a distribution, respectively.

	JB TEST	P Value
FB	14140	2.20E-16
NFLX	1423.2	2.20E-16
AMZN	1277	2.20E-16
F	9235.6	2.20E-16
GE	2310.1	2.20E-16
^GSPC	25280	2.20E-16

Figure 4-4 Output of the Jarque- Bera Test

The table above presents the results of the Jarque-Bera test applied to the returns of our data in analysis. There is statistical evidence that neither of the stocks and indices objected in this research have an underlying distribution that corresponds to the normal, as the null hypothesis is rejected at a 5% significance level for every test conducted.

4.2 Applied model estimation

Research evidence states that, in general, GARCH (1, 1) models perform better in modelling the volatility of stock returns than other GARCH (p, q) models. In the development process of the Autogarch package it was assumed that (1,1) models are best suited to model the volatility of the data in analysis. In the past decades, researchers have switched their focus from model estimation to model optimization, and intrinsic properties of the models have been thoroughly decomposed, in this specific case, developmental efforts were focused on analyzing the change in the conditional distribution of our previously estimated models.

A volatility model performs to the best of its abilities when it can replicate the specificities of the data while maintaining a flexible and simple structure. GARCH models with a higher order of p and q are sometimes diagnosed with overfitting problems, when the model can reproduce a set of data with a degree of accuracy close to perfection it will most likely have difficulty in fitting new observations added to the dataset, or even fail terribly at predicting upcoming observations.

Applying alternate conditional distributions to GARCH models is often seen as a response to the overfitting dilemma caused by recursive increases in lag order, producing models for conditional volatility and simply changing their underlying distributions for distributions that detach themselves from some of the 'unquestionable truths' of the Gaussian curve often yields desirable results.

Summing up, many modified versions of GARCH models are being constantly researched and developed, the main objective being the improvement of the effectiveness of volatility forecasts. Although there are a lot of backers for these alternate models that believe they are the best tools for volatility forecasting presently, studies have also found that despite constant sharpening of the GARCH model the coefficient of determination (R^2) achieved by these models is impedingly low (Poon & Granjer, 2003). For example, (Andersen & Bollerslev, 1998) demonstrated that the R^2 of GARCH (1,1) models tend to $\frac{1}{K}$ where K represents the kurtosis, as it was previously discussed the kurtosis for financial time series is above three, which means that for returns that are non-Gaussian the volatility forecast performance is worse than for Gaussian returns which have a kurtosis of exactly three.

4.2.1 Estimated conditional variance

Subsequently to the estimation of the GARCH models, it is easy to retrieve a time series of conditional volatilities and build graphical representations that corroborate the evidence that volatility is a variable of time and allow the identification of volatility clusters. This analysis plays a vital role in identifying lucrative investment opportunities as expected returns are conditional on volatility. Conditional return volatility, commonly defined as conditional standard deviation is obtained by $\sqrt{\sigma^2}$ and it is represented by a time series just like the one that was obtained for the returns.

The estimate for the S&P 500 conditional volatility presented below is an industry-wide agreed upon performance gauge for the United States stock market.

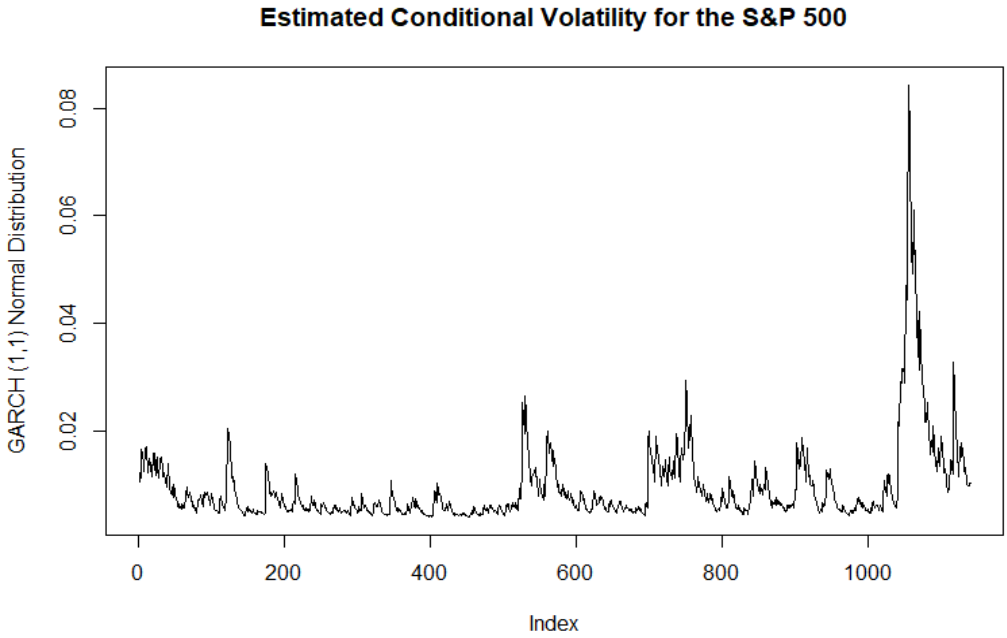

Figure 4-5 Estimated conditional volatility for the S&P 500

4.2.2 Simulation of S&P 500 returns

The ability to precisely assess the possible returns of a portfolio is one of the most sought out skills by the financial industry, the most popular method is to use historical data and consider all the past outcomes to try and predict the future behavior of an asset. Over the years, several variations of this approach have been developed, they differ in complexity and popularity, from which the Monte Carlo simulation stands out. Volatility and returns are stochastic variables, as such, practitioners have attained satisfactory results using the Monte Carlo method, which consists in random sampling of inputs to virtually represent this problem, achieving a wide array of outcomes for the statistical problem at hand. Practical applications for Monte Carlo simulations include, determining the expected value of a portfolio at the retirement date of a clients, and determining the ideal asset allocation of their portfolio. It is common knowledge that assets prices are not random, its features resemble a random walk, which indicates that price trends or previous up or down movements are ineffective in the attempt to draw conclusions about future price expectations.

Using the previously estimated by the Autogarch process GARCH (1,1) model with normal distribution, a thousand simulations of the S&P 500 returns were performed yielding the result presented below.

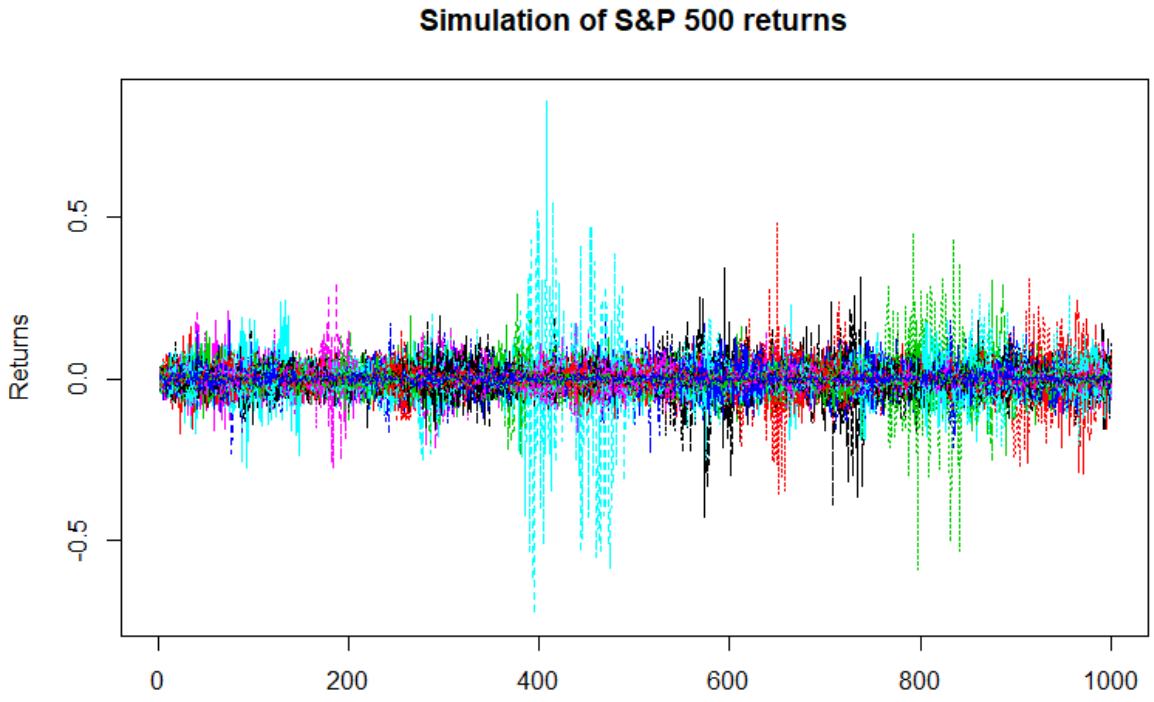

Figure 4-6 Thousand simulations of S&P 500 returns

4.2.3 Information criteria

Model estimation indirectly results in the estimation of various metrics, such as coefficients, residuals, and the information criteria, which are common target of research studies. Although these metrics are different in their calculation, estimation methods, or interpretation, they were all developed and are currently used to achieve a common objective, which is the monitoring of model performance and analysis of subsequent further enhancements that will optimize the usefulness of this tool.

The analysis conducted by the Autogarch process allows the end-user to evaluate if there is a change in the underlying distribution of a GARCH model then, *ceteris paribus*, the performance of this model should improve, for instance, by minimizing the value of the intrinsic information criteria and apply this logic to disregard by direct comparison other non-performing models.

This methodology applied to a GARCH (1,1) model with an estimation based on S&P 500 data resulted in a lower information criterion when using underlying distributions that are known alternatives to the Gaussian curve, for instance the skewed student's-t distribution and the skew-normal distribution was minimized. The results of this essay are presented below.

	Akaike	Bayes	Shibata	Hannan-Quinn	LogLik
sGARCH.sstd.fit	-6.982	-6.947	-6.982	-6.969	3991.152
sGARCH.std.fit	-6.973	-6.943	-6.974	-6.962	3985.352
sGARCH.snorm.fit	-6.897	-6.866	-6.897	-6.886	3941.917
sGARCH.norm.fit	-6.856	-6.829	-6.856	-6.846	3917.180

Figure 4-7 Estimated Information Criteria and Log-Likelihood

5. Conclusion and further development opportunities

Over the years, quantitative methods association to statistical software evolved in a fast, innovative, and exciting fashion. In contrast, traditional methods of teaching, applying, and researching are being continuously revolutionized, as they prove to be either limited or inefficient. The exponential increase in computational power of the previous four decades has yet to be fully taken advantage of by practitioners, to achieve this goal the focus is to be set on education and research. Development of groundbreaking teaching resources will essentially equip educators and students to push the excellency of their *alma mater* one step further, and research is the main driving factor for these developmental efforts while the academic environment can play the fundamental role of a controlled testing scenario.

Financial markets are catalysts for transformational technology, from Merrill Lynch structuring a 23,000 miles wire network in the 1940's to be used for orders to the flash crash of the 2010's that intensified the discussion about algorithmic trading, it is all about fast access to information and acting on that information first than your competitors. Since most of the time information comes in raw form and datasets are quite 'noisy', it is not just about who gets the information first but also about who is able to implement an effective data processing operation that allows conclusions to be drawn in a matter of minutes. Given this highly competitive environment, that gives a literal validation to the expression 'time is money', the Autogarch package positions itself amongst many other software solutions that aim to automate theoretical concepts, more specifically in the volatility modelling field.

This research substantiates the premise that alternative distributions haver a better performance when it comes to capturing all the specificities of a financial dataset, the normal distribution often does a poor job at capturing asymmetric and leptokurtic disturbances. As for model optimization, the process of the Autogarch package proposes to optimize modelling by fitting the same set of data to a plethora of different GARCH models and distributions. Fitting through this process helped us find the desired model for all the stocks used for development purposes, we concluded that across all estimated models the distributions that predominantly minimized the information criteria were the student's-t and the skewed student's-t.

There are endless opportunities for research in this field, either by adding on features to the current version of the Autogarch package or by starting from scratch with another practical concept. However, the current pattern when it comes to automated software solutions shows that packages are very fragmented, meaning that each package tends to work on a narrow scope

with a very specific set of capabilities. It would be interesting to see this work replicated on a macro scale with a broader range of action where we could recursively rely on the software tool to perform any type of data processing, calculation, estimation, and graphical representation desired by the end-user. Scaling to broader packages does not necessarily mean that it must be followed by a step up in complexity, often it is the simplest and more expedite solutions that are valued the most by the end-user, focusing on substituting simple repetitive tasks with software-based solutions is a great starting point for the developer and the end-user, fundamentally it can represent marginal improvements in efficiency and effectiveness, reducing the probability and the intensity of human error.

Taking the curriculum of a bachelor's degree as a starting point with the objective of integrating the respective theoretical concepts on this aggregational tool ranging from more trivial calculations to more advanced concepts would be a very challenging prospect, that could be an ideal fit given the current *status quo* where distance learning is becoming the educational standard. Given that corporations and academic institutions have some common goals, it is in the best interest of both parties that students are as ready as possible to tackle the hurdles of an ever evolving social, political and economic context, we can infer that opening communication channels between the educational and business side is the path to achieve this common goal as this can result in incisive adaptations of the curriculum to accommodate business processes and give students an hands-on contact with them at an embryonic stage of their academic development.

6. Bibliography

Ahamed, L. (2009). *Lords of Finance: The Bankers Who Broke the World.* Penguin Books.

Akaike, H. (1985). Prediction and entropy. *A Celebration of Statistics*, 1-24.

Aldrich, J. (1997). R.A. Fisher and the making of maximum likelihood 1912-1922. *Statist. Sci.*, 162 - 176. doi:https://doi.org/10.1214/ss/1030037906

Andersen, T. G., & Bollerslev, T. (1998). ANSWERING THE SKEPTICS: YES, STANDARD VOLATILITY MODELS DO PROVIDE ACCURATE FORECASTS. *International Economic Review*, 885-905.

Ardia, D., & Hoogerheide, L. F. (2010). Baesyan estimation of the GARCH(1,1) model with student-t innovations. *The R Journal*, 41-47. doi:https://doi.org/10.32614/RJ-2010-014

Baier, T., Neuwirth, E., & Meo, M. D. (2011). Creating and deploying an application with (R)excel and R. *The R Journal*, 5-11. doi:https://doi.org/10.32614/RJ-2011-011

Barndorff-Nielsen, O. (1978). Hyperbolic Distributions and Distributions on Hyperbolae. *Scandinavian Journal of Statistics*, 151-157.

Bischoff, F., & Rodrigues, P. P. (2020). tsmp: An R package for time series with matrix profile. *The R Journal*, 76-86. doi:https://doi.org/10.13140/RG.2.2.13040.30726

Black, F. (1993). Beta and return. *The Journal of Portfolio Management*, 8-18.

Blaesild, P. (1981). Multivariate Distributions of Hyperbolic Type. *Statistical Distributions of Hyperbolic Type*, 45-66.

Bodie, Z., Kane, A., & Marcus, A. J. (2004). *Essentials of Investments.* New York: McGraw-Hill/Irwin.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 307-327. doi:https://doi.org/10.1016/0304-3076(86)90063-1

Cook, D., & Wickham, H. (2008, November). Comment. *Technometrics*, 442-443.

Ding, Z., Granger, C., & Engle, R. (1993). A long memory property of stock market returns and a new model. *Journal of Empirical Finance, 1*(1), 83-106.

Dziak, J. J., Coffman, D. L., Lanza, S. T., & Li, R. (2019). Sensitivity and specificity of information criteria. *Briefings in Bioinformatics*, 553-565. doi:https://doi.org/10.1101/449751

Eberlein, E., & Keller, U. (1995). Hyperbolic distributions in finance. *Bernoulli*, 281-299.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica*, 987-1008. doi:https://doi.org/10.2307/1912773

Engle, R. F., & Bollerslev, T. (1986). Modeling the persistence ofconditional variances. *Econometric Reviews*, 1–50.

Fama, E. F., & French, K. R. (1992). The cross-section of expected stock returns. *The journal of finance*, 427-465. doi:https://doi.org/10.2307/2329112

Fernandez, C., & Steel, M. F. (1998). On Bayesian Modeling of Fat Tails and Skewness. *Journal of the American Statistical Association*, 359-371.

Ferreira, J. T., & Steel, M. F. (2006). On describing multivariate skewed distributions: A directional approach. *The Canadian Journal of Statistics*, 411-429.

Gauss, C. F. (1809). Theoria motvs corporvm coelestivm in sectionibvs conicis Solem ambientivm.

Glosten, L. R., Jagannathan, R., & Runkle, D. E. (1993). On the Relation between the Expected and the Volatility of the Nominal Excess Return on Stocks. *The Journal of Finance*, 1779-1802.

Godfrey, A. J. (2013). Statistical software from a blind person's perspective. *The R Journal*, 73-79. doi:10.32614/RJ-2013-007

Gosset, W. S. (1908). The probable error of a mean. *Biometrika*, 1-25.

Hannan, E. J., & Quinn, B. G. (1979). The Determination of the order of an autoregression. *Journal of The Royal Statistical Society*, 190-195.

Hentschel, L. (1995). All in the family Nesting symmetric and asymmetric GARCH models. *Journal of Financial Economics, 39*(1), 71-104. doi:https://doi.org/10.1016/0304-305X(94)00821-H

Jarque, C. M., & Bera, A. K. (1987). A test for normality of observations and regression residuals. *International Statistical Review*, 163-172.

Johnson, N. L. (1949). Systems of Frequency Curves Generated by Methods of Translation. *Biometrika*, 149-176.

Jr, W. H. (2014). *Catching Lightning in a Bottle: How Merrill Lynch Revolutionized the Financial World.* Wiley.

Kane, D., Liu, A., & Nguyen, K. (2011). Analyzing an electronic order book. *The R Journal*, 64-58.

Konishi, S., & Kitagawa, G. (2008). *Information criteria and statistical modeling.* New York: Springer.

Lee, G. G., & Engle, R. F. (1993). A Permanent and Transitory Component Model of Stock Return Volatility.

Legendre, A.-M. (1805). Nouvelles méthodes pour la détermination des orbites des comètes.

Lintner, J. (1965). The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets. *The Review of Economics and Statistics*, 13-37.

Lu, Y., & Kaine, D. (2013). Performance attribution for equity portfolios. *The R Journal*, 53-61.

Miskolczi, P. (2017). Note on simple and logarithmic return. Applied Studies in Agribusiness and Commerce. *Applied Studies in Agribusiness and Commerce*, 127-136. doi:https://doi.org/10.19041/APSTRACT/2017/1-2/16

Nelson, D. B. (1991). Conditional Heteroskedasticity in Asset Returns: A New Approach. *Econometrica*, 347-370. doi:10.2307/2938260

Ngo-ye, T., & Park, S.-H. (2014). Motivating business major students to learn computer programming- A case study . *Southern Association for Information Systems Conference.* Macon.

O'Hagan, A., & Leonard, T. (1976). Bayes estimation subject to uncertainty about parameter constraints. *Biometrika*, 201-203.

Poon, S.-H., & Granjer, C. W. (2003). Forecasting Volatility in Financial Markets: A Review. *Journal of Economic Literature*, 478-539.

Pruim , R., Kaplan, D. T., & Horton, N. J. (2017). The mosaic package: Helping students to "think with data" using R. *The R Journal*, 77-102. doi:https://doi.org/10.32614/RJ-2017-024

Reuven, Y. R., & Dirk, P. K. (2016). *Simulation and the Monte Carlo Method.* New York: John Wiley & Sons.

Ryan, J. A., & Ulrich, J. M. (2020). quantmod: Quantitative Financial Modelling Framework.

Sardá-Espinosa, A. (2019). Time-series clustering in R using the dtwclust package. *The R Journal*, 1-22.

Sax, C., & Steiner, P. (2013, December). Temporal disaggregation of time series. *The R Journal*, 80-87.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, 461-464.

Shang, H. L. (2011). rainbow: An R package for visualizing functional time series. *The R Journal*, 54-49. doi:https://doi.org/10.32614/RJ-2011-019

Shang, H. L. (2013). ftsa: An R package for analyzing functional time series. *The R Journal*, 64-62. doi:https://doi.org/10.32614/RJ-2013-006

Sharpe, W. F. (1964). CAPITAL ASSET PRICES: A THEORY OF MARKET EQUILIBRIUM UNDER CONDITIONS OF RISK. *The Journal of Finance*, 425-442.

Shibata, R. (1976). Selection of the order of an autoregressive model by Akaike's information criterion. *Biometrika*, 117-126.

Staniak, M., & Biecek, P. (2019). The landscape of R packages for automated exploratory data analysis. *The R Journal*, 347-369.

Taleb, N. N. (2010). *The Black Swan: Second Edition: The Impact of the Highly Improbable.* Random House.

Taylor, S. (1986). *Modelling Financial Time Series.* New York: John Wiley.

Tsay, R. S. (2010). *Analysis of Financial Time Series.* Hoboken: John Wiley & Sons, Inc.

Wallace, P., & Clariana, R. B. (2005). Perception versus reality - determining business student's computer literacy skills and need for instruction in information concepts and technology. (E. Cohen, Ed.) *Journal of Information Technology Education, 4*, 142-150. doi:https://doi.org/10.28945/269

Whittle, P. (1951). *Hypothesis testing in time series analysis.* Uppsala: Almqvist & Wiksells Boktryckeri AB.

Wickham, H. (2010). A layered grammar of graphics. *Journal of Computational and Graphical Statistics*, 3-28. doi:https://doi.org/10.1198/jcgs.2009.07098

Wickham, H. (2014). Tidy Data. *Journal of Statistical Software, 59*. doi:https://doi.org/10.18637/jss.v059.i10

Zagaglia, P. (2013). PIN: Measuring asymmetric information in financial markets with R. *The R Journal*, 80-86. doi:https://doi.org/10.32614/RJ-2013-008

7. Annex A

7.1 R Documentation

import_file {Autogarch} R Documentation

Import File

Description

Import a time series from excel

Usage

```
import_file(variable_name, sheet_name = NULL)
```

Arguments

variable_name	A string to define name of the variable assigned to Global Environment
sheet_name	A string to define a specific sheet to import data from

Examples

```
import_file(variable_name = "df")

import_file(variable_name = "df",sheet_name = "Sheet1")
```

returns {Autogarch} R Documentation

Returns

Description

A function to calculate the logarithmic or simple returns of a time series dataframe

Usage

```
returns(x, simple = FALSE, view = TRUE)
```

Arguments

x	A time series dataframe
simple	TRUE - Calculates simple returns FALSE - calculates logarithmic returns
view	TRUE opens a tab to view the output dataframe FALSE does not open a tab to view the output dataframe

Examples

```
returns(df, simple = FALSE, view = TRUE)

returns(df, simple = TRUE, view = FALSE)
```

beta {Autogarch} R Documentation

Beta

Description

Estimates the beta given a data frame of returns

Usage

```
beta(x, mkt_ticker, start_date, end_date)
```

Arguments

`x`	A dataframe of returns
`mkt_ticker`	A string with the column header of the market returns
`start_date`	A string with the start of the estimation period
`end_date`	A string with the end of the estimation period

Examples

```
beta(log_returns,mkt_ticker = "X.GSPC", start_date = "2020-01-01", end_date = "2020-06-01")
```

init_spec {Autogarch} R Documentation

Initialize Model Specificication

Description

Initializes model specification and stores them as variables

Usage

```
init_spec(model, all = FALSE)
```

Arguments

`model`	A vector of strings to define a specific model to be specified
`all`	TRUE indicates that the user wishes to specify all available models, model argument will be disregarded FALSE indicates that the user wants to specify a defined set of models, model argument will be considered

Examples

```
init_spec(all = TRUE)

init_spec(c("gjrGARCH", "apARCH"), all = FALSE)
```

init_fit {Autogarch} R Documentation

Initialize Fitting

Description

Initializes fitting given returns data

Usage

```
init_fit(x, spec_rm = TRUE)
```

Arguments

`x`	The returns of a stocks or a market index
`spec_rm`	TRUE deletes the model specification variables FALSE does not delete the model specification variables

Examples

```
init_fit(log_returns$AMZN, spec_rm = TRUE)

init_fit(smpl_returns$AMZN, spec_rm = FALSE)
```

agg_models {Autogarch} R Documentation

Aggregate Models

Description

Calculates the information criteria for all estimated models and aggregates them in a dataframe

Usage

```
agg_models(models, all, view = TRUE)
```

Arguments

`models`	A vector of strings defining specific models
`all`	TRUE all models will be considered, models argument will be disregarded FALSE specific models given in the models argument will be considered
`view`	TRUE opens a tab with the output dataframe FALSE does not open a tab with a specific dataframe

Examples

```
agg_models(all = TRUE, view = TRUE)

agg_models(models = c("gjrGARCH.snorm.fit", "apARCH.norm.fit"),all = FALSE, view = TRUE)

agg_models(all = TRUE, view = FALSE)

agg_models(models = c("gjrGARCH.snorm.fit", "apARCH.norm.fit"),all = FALSE, view = FALSE)
```

Rank Models

Description

Prints to console the information criteria that was minimized and the model and conditional distribution that achieved minimization

Usage

```
rank_models(info_criteria_data)
```

Arguments

`info_criteria_data`	A dataframe with the calculated information criteria. The output of agg_models function.

Examples

```
rank_models(info_criteria)
```

8. Annex B

8.1 Links to Source Code

Main directory - https://github.com/scogli/Autogarch

import_file - https://github.com/scogli/Autogarch/blob/main/import_file.R

returns - https://github.com/scogli/Autogarch/blob/main/returns.R

beta - https://github.com/scogli/Autogarch/blob/main/beta.R

init_spec - https://github.com/scogli/Autogarch/blob/main/init_spec.R

init_fit - https://github.com/scogli/Autogarch/blob/main/init_fit.R

agg_models - https://github.com/scogli/Autogarch/blob/main/agg_models.R

rank_models - https://github.com/scogli/Autogarch/blob/main/rank_models.R