Measuring the uncertainty in Value-at-Risk & Expected Shortfall through Bootstrapping - An Empirical analysis

A dissertation submitted by **16457** to the Department of Finance, the London School of Economics and Political Science, in part completion of the requirements for Course FM403: Management and Regulation of Risk.

September 2019

10502 words

Abstract

Studies on estimation uncertainty in Value-at-Risk (VaR) and Expected Shortfall (ES) are mostly theoretical. This paper then focuses on the empirical application of a non-parametric method as proposed by Christoffersen et al. (2004) to account for and measure the uncertainty in the risk estimates. I found no considerable improvements in the risk estimates even after accounting for the uncertainty. Next, uncertainty in the risk estimates increases as one look further into the tail; ES, while theoretically superior to VaR, is measured with substantially more uncertainty. Furthermore, my results indicate that the importance of the length of the backtesting window cannot be overlooked - a small window of 250 may lead to a higher likelihood of choosing a wrong model. More importantly, there seems to be a trade-off between model adequacy and estimates accuracy - the Filtered Historical Simulation produced the most accurate out-of-sample forecasts, but the estimates were the most uncertain.

Acknowledgments

This journey would not have been possible if not for the continuous support from my family; this dissertation is dedicated to them. Huge thanks to my supervisor, Dr Domingos for nudging me in the right direction and for the help he gave for my rather unfortunate incident.

Contents

List of Tables

List of Figures

1. Introduction

The notion of associating risk with volatility undoubtedly came from Harry Markowitz and Andrew Donald Roy.[1] In 1952, Markowitz and Roy separately proposed a portfolio selection technique that was based on the expected value and variance of the portfolio - the optimal portfolio was not one that had the largest absolute return, but rather the one that produced the maximum return for a given level of risk. The next ten years or so saw the emergence of the Capital Asset Pricing Model[2] that was built on Markowitz and Roy proposed technique - an asset pricing model that incorporates an asset's sensitivity to market risk.

Financial Risk Management, however, was only popularised in the 1970s, through the work of Fischer Black, Myron Scholes and Robert Merton (herein referred to as BSM). The authors proposed a pricing model for which the fair value of an option can be easily calculated. The widespread adoption of the BSM model, an increased in the usage of derivatives to manage risk, and the burgeoning of technological innovation in the industry then spurred the exponential growth of the derivatives market in the early 1980s. However, up until then, risk measurement, in particular volatility, had never been the focal point. It was not until the market crash on Black Monday in 1987 that the industry saw the need for accurate and reliable risk measurement. Consequently, by the early 1990s, at the recommendation of the Group of Thirty (G30)[3], most major financial institutions had established an independent risk function to oversee the market and credit risks undertaken by the front-office.

While the concept of VaR did not originate from JP Morgan, it was through their public release of RiskMetrics that revolutionised the use of VaR to measure and quantify risk. Furthermore, amendments to the 1988 Basel 1 accord that incorporated market risk capital charge and the permission of internal-models resulted in a rampant adoption of VaR across the industry. However, RiskMetrics was always criticised for its unrealistic assumption that returns are normally distributed. By then, several empirical stylised facts had already been established. Fama (1963) and Fama (1965) argued that returns

[1] Markowitz (1999) acknowledged that Roy "has an equal share of the honour of being called Father of Modern Portfolio Theory"

[2] See Sharpe (1964) and Treynor (1961)

[3] Specifically in their famous G30 report - Derivatives: Practice and Principles. Arguably, it was also in that report that the term Value-at-Risk (VaR) had appeared officially for the first time in the industry

are not normally distributed and that the distribution exhibit leptokurtosis. More important, volatility was found to be time-varying - a direct contradiction of the explicit assumption in the BSM model - and it tends to cluster. The inadequacy of RiskMetrics then shifted the industry's attention to conditional volatility models. Owing to their effectiveness in incorporating volatility dynamics and modelling time series, these models and their subsequent extensions became the workhorse of the industry to forecast volatility and estimate VaR.

The following ten years saw the reliance on VaR grew larger, resulting in the development of a plethora of VaR models, and also methodologies to assess the accuracy of VaR estimates - a procedure known as backtesting. However, with attention, comes scrutiny. VaR is often criticised for its theoretical deficiencies, with non-coherence arguably the most detrimental. Furthermore, the collapse of Long Term Capital Management and the 2008 global financial crisis saw the reliability of VaR came under heavy fire - academia argued for the replacement of VaR with Expected Shortfall (ES). In their seminal work, Artzner et al. (1999) presented definitive evidence that showed VaR is not necessarily sub-additive and proposed ES as an alternative to VaR.[4] However, Yamai and Yoshiba (2005) recommended that usage of VaR and ES should be complementary and not mutually exclusive. Regardless of choice between the two, risk measurement relies on statistical modelling and that inevitably introduces uncertainty and estimation error.

While backtesting could address the adequacy of an assumed risk model, it does not relay any information that pertains to the estimation uncertainty. According to Spierdijk (2016), there are three sources of uncertainty in QML-based VaR[5]: parameter uncertainty in the estimation of conditional returns and volatility; sampling uncertainty in the estimation of empirical quantile; and correlation of the uncertainty between the first two. Fortunately, from Statistics-101, we know that uncertainty could be measured with a confidence interval. However, given that the "true" value and distribution of the risk measures are unknown to us, construction of the interval would generally require: making an explicit distributional assumption for returns; deriving the asymptotic behaviour of the risk measures; using resampling or simulation techniques. An accurate interval along with the point estimate could prevent unnecessary capital reserves (Christoffersen et al. (2004)) and would improve the reliability of the risk management process (Danielsson and Zhou (2016)). Therefore, it should be of high interest to practitioners on constructing an interval, regardless of the approach used.

The remaining part of this paper is structured as follow: Literature Review that first provides a review of the existing literature on VaR specifically focusing on the performance of different estimation method, followed by review of the literature on methods to construct intervals and the measurement and comparison of uncertainty in

[4]See Section 3.10 for explanation of non-coherence and sub-additivity

[5]See Section 3.3.3 for explanation of QML-based VaR

risk estimates; Theoretical Framework section that presents the theory of the conditional volatility models, risk measures and backtesting framework; Methodology section that describes in detail the steps to implements the bootstrapping procedures as introduced in Christoffersen et al. (2004); Data that looks at the underlying properties of the data used and statistical tests that justify the use of conditional volatility models; Empirical Analysis section that consists of 2 types of analysis - In-sample that focus on the in-sample properties of the volatility models; Out-sample, which can be further split into 2 subsections - Point Forecast and Prediction Interval; and Conclusion that summarises the results of the analysis and potential extension.

2. Literature Review

According to Manganelli and Engle (2001), VaR estimation method can be categorised into three categories: Non-parametric, Parametric and Semi-parametric. Historical Simulation (HS) is a non-parametric method that does not impose any distribution on returns. Pérignon and Smith (2010) reported that 73% of banks[1] used HS because of its simplicity, ease of implementation and its ability to capture non-linear relationship between financial assets. However, Danielsson et al. (1998) found that HS-VaR estimates have considerable uncertainty because of the method's sensitivity to the size of the estimation window (herein referred to as T). The authors looked at simulated data from US equity and found that HS-VaR was generally over-estimated. While it was reasonably accurate at 95% confidence level, the problem worsens at higher confidence level - no violations were found at 99.95% level. Inui et al. (2005) reported similar results. They reported "considerable positive bias" in HS-VaR when the distribution of portfolio exhibits leptokurtosis. Moreover, they found that the bias-ness increases with confidence level but decreases with T. Moreover, Pritsker (2006) looked at the performance of HS during the Black Monday and found that the VaR estimates did not react to the market crash, and the particular problem was not detected during backtesting.

While HS has its own merits, the main criticism is its inability to account for time-varying volatility. To account for time-varying volatility and its persistence, Bollerslev (1986) introduced the Generalised Auto-Regressive Conditional Heteroscedasticity (GARCH) model, an extension of Engle (1982), which incorporates both past shocks and volatility. In an extension, Engle and Bollerslev (1986) proposed the Student-t (herein referred to as t-dist) variant of the GARCH to account for non-normality. The out-of-sample performance of GARCH models in estimating VaR has been thoroughly studied: Angelidis et al. (2004) looked at five major stock indices and compared the performance of normal, t-dist and generalised error distribution. They found that the t-dist models performed better than the normal models. Similar conclusion was also reached in Cerović Smolović et al. (2017).

Another well-documented stylised fact in financial assets is Leverage Effect. Black

[1] 60 large banks over a 10 year period

(1976) found that stocks are more volatile after a negative shock than a positive shock of the same magnitude. Many extensions of the GARCH models such as EGARCH (Nelson (1991)) and GJR (Glosten et al. (1993)) have been proposed to account for leverage effect. These extensions are known as the asymmetric models, and many authors have reported superior performance of these models than the symmetric ones in estimating VaR at different confidence level. Brooks and Persand (2003) looked at several equity indices in Asia and compared EGARCH against GARCH and found that EGARCH performed better than GARCH. Miletic and Miletic (2015) compared the HS against parametric method. In particular, the authors looked at RiskMetrics, symmetric and asymmetric GARCH models in selected emerging economies in Europe during the 2008 crisis. They reported superior performance of the asymmetric GARCH models against the remaining models. In a more comprehensive study, Brownless et al. (2011) looked at several FX currencies, sector and international equity indices. They compared numerous asymmetric (GJR, EGARCH, APARCH) against GARCH and found that the performance of GJR was well ahead of the pack over the extended period of volatility surge during the 2008 crisis and that the performance was consistent across different asset classes and Ts.

The semi-parametric method aims to pool the strengths of GARCH models and HS. Many models have been proposed, and the most popular ones are those that reflect changes in market condition onto the risk estimates quickly. In particular, Barone-Adesi et al. (1999) proposed the Filtered Historical Simulation (FHS) - a 2 step procedure in which a volatility model is first used to forecast volatility and residuals (also known as standardised returns) are obtained by dividing the returns with volatility estimates[2], after which VaR is computed as the multiplication of the volatility with the empirical quantile of the residuals. Many authors have reported that FHS generally performed very well, especially at 99%. Omari (2017) looked at USD/KES currency and compared the FHS against HS and parametric method (RiskMetrics, GARCH & GJR), with normal and t-dist for several confidence levels. He found that FHS (t-GARCH filter) and parametric GJR (with t-dist) to have the best conditional coverage performance and that HS performed the worst. Kuester et al. (2006) looked at the NASDAQ index and compared the performance of HS against GARCH and FHS, with several error distributions. They found that the FHS method produced the most robust VaR estimates and the skewed-t-GARCH filter produced the most accurate VaR. Similar findings that FHS outperformed GARCH and HS were also reached in Pritsker (2006), Brandolini and Colucci (2012) and Adcock et al. (2012).

In comparison, lesser attention is paid to the topic of estimation uncertainty. Jorion (1996) and Dowd (2000) were the first few authors that attempted to measure estima-

[2]This process is also known as filtering. For instance, assuming a t-dist for a GARCH model will produce a t-GARCH filter

tion uncertainty. Jorion (1996) proposed to quantify the uncertainty using two methods - standard error of a normally distributed quantile estimate and standard error of the sample standard deviation. He found that the VaR's standard error from the former method was twice that of from the latter method. On the other hand, Dowd (2000) proposed the construction of an interval to measure the uncertainty. He simulated normally distributed data and imposed normality to construct the interval. He found that the interval widths were large and concluded that T is critical to estimation uncertainty- the width of the interval and bias-ness of the estimates decreases as T increases. However, while their proposed methods are easy to implement, the premise of their proposed methodology is in direct violation of the non-normality found in returns.

Given the ability of GARCH models to incorporate excess kurtosis[3], Christoffersen et al. (2004) extended Pascual et al. (2006)[4] to a build prediction interval for VaR and ES. The authors proposed to use the independent and identically distributed (herein referred to as i.i.d) bootstrap procedure on the model residuals - a procedure known as residual resampling with replacement - to construct pseudo returns. They investigated the performance of their proposed method on two types of simulated data - t-dist and normal. For the t-dist, it was further categorised into high and low persistence, both of which were characterised by the estimated parameters values from GARCH models. They constructed intervals for several commonly used VaR and ES methods that included HS, normal-GARCH, FHS (using t-GARCH filter) and Hill method, and reported the following: HS produced very inaccurate point forecast and the interval generated had poor coverage rate and were the widest; performance of normal model was comparable to HS for most cases, with the exception of normally distributed simulated data; point forecast of FHS and Hill method were considerably more accurate, and interval constructed had good coverage rate (close to nominal rate) but were generally wide.[5]

According to Spierdijk (2016), even though the proposed method has good properties, it would breakdown in the absence of finite fourth moment for the model error. Consequently, Spierdijk (2016) proposed using a subsample bootstrap[6] to construct the prediction interval. The constructed intervals had good coverage rates across different assumed distributions and were robust to different estimation window and subsample size. More important, they found that Christoffersen et al. (2004) did not fail as expected and even fared well for symmetric error distributions. It did, however, broke down in the presence of skewed error distributions. In a more analytical approach, Chan et al.

[3]It can easily be shown that the unconditional excess kurtosis of a GARCH (1,1) is >0

[4]The authors proposed a non-parametric method to construct prediction intervals for GARCH volatility

[5]Coverage rate here refers to the proportion of number of times the true VaR and ES were within the constructed intervals. Nominal refers to the confidence level.

[6]Unlike the residual resampling, subsample bootstrap refers to k-out-of-n resampling, where k<n

(2007) relied on Extreme Value Theory (EVT) and proposed 2 methods of constructing the intervals - asymptotic normality of the risk measures and data tilting method (see Hall and Yao (2003) for more information). Gao and Song (2008) derived the limiting distribution of the risk measures based on the GARCH residuals and constructed the intervals using asymptotic variance.

The construction of the interval then allows us to compare the uncertainty between VaR and ES. While different methods have been proposed, similar conclusions were reached. Christoffersen et al. (2004) found that the ES estimates across all methods were more inaccurate than the corresponding VaR estimates. Looking at three categories of the non-parametric risk measures: estimation error, risk factors decomposition & optimisation, Yamai et al. (2002) pitted the VaR against ES. In particular, they simulated data with fat-tail and normal characteristics and ran a Monte-Carlo simulation to obtain a confidence interval and standard deviation. They reported the following findings: at 95% confidence level, the uncertainty[7] in VaR and ES were comparable under normality. However, deviation from normality resulted in larger estimation error in ES than VaR; the uncertainty in ES was almost five times that of the VaR at 99%. They attributed their findings to the larger sensitivity of ES to tail events, as compared to VaR that ignores the shape of the tail. Danielsson and Zhou (2016) looked at the robustness of the risk measures with small T (T=300) using simulated data from t-dist. They employed EVT to derive the asymptotic variance of both risk measure under HS, and their results indicated that at the same confidence level, the standard deviation of ES was at least two times that of VaR. Similar findings that ES estimates are more uncertain were also reported in Nieto and Ruiz (2010).

With all the above in mind, next, I pen down the source of my motivation and the aim of my analysis. Christoffersen et al. (2004) were the first few authors that attempted to address the uncertainty of both VaR and ES. Unlike the methods proposed by Dowd (2000) and Jorion (1996), the proposed method by Christoffersen et al. (2004) avoids any distributional assumption, incorporates volatility dynamics and is easy to implement. However, the study by Christoffersen et al. (2004) was theoretical. The authors did not implement their proposed method on empirical data. Hence, inspired by their intuition and motivated by their satisfactory results, I would like to extend their paper to perform empirical analysis, and also extend their proposed method onto the asymmetric GJR-GARCH model. Hence, the aim of this paper is twofold. First, a comparison of the three VaR estimation approach and the use of the proposed method in Christoffersen et al. (2004) to assess the magnitude of improvement for both VaR and ES estimates after accounting for estimation uncertainty. Second, to provide an empirical analysis and comparison of the degree of uncertainty in VaR and ES. Also, Danielsson and Zhou (2016) reported that uncertainty increases as one look deeper into the tail region. Hence,

[7]Defined here as the width of the interval

the analysis will also be based on two different significance level: p=0.01 and 0.05[8], to investigate if the degree of improvement would be larger when looking further into the tail, and also 2 different backtesting window size (herein referred to as m) - to illustrate the impact of the size on the reliability of the backtesting results.

[8]For p=0.01 (0.05), the corresponding confidence level is 99% (95%)

3. Theoretical Framework

3.1 Returns

Denoting P_t and R_t as the closing price and returns of an asset on day t, R_t can be defined as

$$R_t = \log(\frac{P_t}{P_{t-1}}) \tag{3.1}$$

where Equation 3.1 is also known as the continuously compounded returns or logarithmic returns and will be used throughout the remaining sections of my analysis.

3.2 Generalised Autoregressive Conditional Heteroskedaticity models

3.2.1 Symmetric - GARCH(p,q)

In this paper, R_t is modelled according to the following

$$R_t = \sigma_t \epsilon_t, \qquad \text{for } t = 1, 2....T \tag{3.2}$$

where ϵ_t is the model error and assumed to be i.i.d, with mean zero, variance of unity and distribution function F. Specifically, I will focus on 2 cases for which F is symmetric - a standard normal and standard Student's t with v degrees of freedom (herein referred to as dof), i.e. $\sqrt{\frac{v}{v-2}}\epsilon_t \sim t_v$ (Christoffersen et al. (2004)).

The importance of F cannot be overlooked. According to Asem (2007), the role of F is twofold. First, the assumed distribution underpins the construction of the likelihood functions, and second, it dictates the distribution of the computed risk measure, given the estimated conditional volatility. While any potential issues arising from the latter can be readily resolved by FHS, a misspecified distribution may result in inefficient parameter estimates. However, Bollerslev and Wooldridge (1992) showed that conditional on a correct specification of the first two moments, parameters will still be consistent even with normality assumption. These estimates are known as Quasi Maximum Like-

lihood (QML) estimates. Following that, Asem (2007) showed that even though the QML estimates are less efficient than Maximum Likelihood (ML) estimates, the QML VaR forecasts are not worse off because the loss of parameters efficiency is not reflected on the VaR forecast.

R_t is said to follow a GARCH (p,q) process if it satisfies not only Equation (3.2) but also

$$\sigma_t^2 = \omega + \sum_{i=1}^{p} \alpha_i R_{t-i}^2 + \sum_{j=1}^{q} \beta \sigma_{t-j}^2 \tag{3.3}$$

where $\omega > 0$, $\alpha_i \geq 0$, $\beta_j \geq 0$, and $\sum_{j=1}^{max(p,q)} (\alpha_j + \beta_j) < 1$ ensures stationarity of R_t. α smoothens the volatility estimates, while β dictates the responsiveness of the model to new information. The simplest form of the GARCH model is the GARCH (1,1), which, despite its simplicity, is most commonly used due to it being parsimonious.

The GARCH (1,1) can be specified as

$$\sigma_t^2 = \omega + \alpha R_{t-1}^2 + \beta \sigma_{t-1}^2 \tag{3.4}$$

The 1-day ahead volatility forecast can be computed recursively using the following formulation

$$\hat{\sigma}_{T+1}^2 = \hat{\omega} + \hat{\alpha} R_T^2 + \hat{\beta} \sigma_T^2 \tag{3.5}$$

3.2.2 Asymmetric - GJR-GARCH(p,q)

R_t is said to follow a GJR-GARCH (p, q) process if it satisfies not only Equation (3.2), but also

$$\sigma_t^2 = \omega + \sum_{i=1}^{p} (\alpha_i + \gamma I_{t-i}) R_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2 \tag{3.6}$$

$$I_{t-i} = \begin{cases} 1, \text{ if } R_{t-i} < 0 \\ 0, \text{ otherwise} \end{cases} \tag{3.7}$$

where $\omega > 0, \alpha_i \geq 0, \beta_j \geq 0,$ and $\sum_{j=1}^{max(p,q)} (\alpha_j + \frac{\gamma_j}{2} + \beta_j) < 1$ ensures stationarity of R_t. Typically γ is found to be >0, which represents a larger impact to volatility.

Due to similar reasoning as GARCH, the GJR-GARCH (1,1) is also most commonly used. For $R_{t-1} < 0$, a GJR-GARCH (1,1) can be defined as

$$\sigma_t^2 = \omega + (\alpha + \gamma) R_{t-1}^2 + \beta \sigma_{t-1}^2 \tag{3.8}$$

Notice that the GJR-GARCH simply collapse back to GARCH for $R_{t-1} \geq 0$. The 1-day

ahead volatility forecast can be computed recursively using

$$\hat{\sigma}^2_{T+1} = \hat{\omega} + (\hat{\alpha} + \hat{\gamma} I_t) R^2_T + \hat{\beta} \hat{\sigma}^2_T \tag{3.9}$$

where I_t refers to the indication function as defined in Equation (3.7).

Finally, the estimated parameters will denoted as $\hat{\theta} : (\hat{\omega}, \hat{\alpha}, \hat{\beta}, \hat{\gamma}, v)$, where $\hat{\gamma}$ for GJR models and v for t-dist models.

3.3 Risk Measures

3.3.1 Value-at-Risk (VaR)

In layman terms, VaR is simply the worst possible loss that will not be exceeded $(1 - p)100\%$ [1] of the time, given a horizon. Next, the formulations in the remaining sections will be based on VaR and ES as a positive number.

Mathematically, VaR at the time t, for a given probability p is defined as the value such that

$$prob(R_t \leq -VaR^p_t) = p \tag{3.10}$$

Equation (3.10) can be interpreted as the probability of observing a return that is smaller than VaR is p. As mentioned earlier, Artzner et al. (1999) presented four desirable properties, known as the four axioms of coherence, that a risk measure should possess, namely Monotonicity, Translation Invariance, Homogeneity, and Sub-additivity. The authors showed that VaR fails to satisfy sub-additivity[2] unless normality is assumed. Furthermore, as a quantile measure, VaR ignores the shape of the tail of the distribution. By ignoring the shape, it can provide a misleading interpretation of risk. For instance, 2 portfolios can have same VaR but different tail fatness. That is, the portfolio with a fatter tail has a higher probability of incurring a larger loss, i.e. higher risk than the other.

3.3.2 Expected Shortfall

ES can be thought of as the average of returns that are lesser than VaR. Mathematically, ES is defined as

$$ES^p_t = -E[R_t | R_t \leq -VaR^p_t] \tag{3.11}$$

[1] p is typically 0.05 and 0.01

[2] Denoting $\psi(.)$ as a risk measure, sub-additivity refers to the property where $\psi(X + Y) \leq \psi(X) + \psi(Y)$. That is, a combined portfolio with two assets cannot be riskier than the sum of the two individual assets. The violation directly contradicts diversification effect

Artzner et al. (1999) showed that ES resolves the issue of non-subadditivity. Moreover, according to Danielsson (2011), ES takes into consideration the shape of the tail distribution due to the expectation taken in Equation (3.11). Given Equation (3.2), VaR_t^p and ES_t^p can be further simplified and expressed as

$$VaR_t^p = -\sigma_t F_p^{-1} \equiv -\sigma_t q_{1,p} \tag{3.12}$$

and

$$ES_t^p = -\sigma_t E[R_t | R_t \leq -F_p^{-1}] \equiv -\sigma_t q_{2,p} \tag{3.13}$$

where F_p^{-1} refers to the inverse distribution of ϵ_t. Next, 3 categories of estimation methods that will be used in this paper are introduced.

3.3.3 Non-Parameteric

Historical Simulation

As a non-parametric method, VaR is simply the p-th percentile of the empirical distribution of past returns - past returns are first sorted in ascending order, and the VaR estimate corresponds to p-th percentile and is given by

$$HS - VaR_{T+1}^p = -Q_p\{R_t\} \tag{3.14}$$

where $Q_p\{.\}$ is the p-th percentile of empirical distribution of past returns $\{R_t, t = 1, 2....T\}$. On the other hand, ES is given by

$$HS - ES_{T+1}^p = -\frac{1}{\#(R_t \leq -VaR_{T+1}^p)} \sum_{R_t \leq -VaR_{T+1}^p} (R_t) \tag{3.15}$$

where the denominator refers to the total count of observed returns that are lesser than or equals to HS-VaR. While this method avoids any distributional assumption, it relies on a rigid assumption that returns are i.i.d and assigns an equal weight of 1/T to each data point in T. By assigning equal weights, it implies that observations that are far out in the past have the same impact as more recent observations. Moreover, as will be shown later, it is extremely sensitive to T. If T is too small, there might be insufficient tail risk events being included. On the other hand, VaR estimates do not react to changes in the market conditions, i.e. unresponsive. Finally, it disregards any sorts of volatility dynamics. Given the limitations above, it is, therefore, more appropriate to consider GARCH models that incorporate volatility dynamics. In particular, the method described below employs the GARCH class of models to estimate VaR and ES in 2 steps:

Step 1. Estimate the models and obtain $\hat{\theta}$ through QML. Compute the volatility forecast $\hat{\sigma}^2_{T+1}$ recursively using Equation (3.5) and (3.8).

Step 2. Denoting $\hat{q}_{1,p}$ and $\hat{q}_{2,p}$ as the corresponding estimates of $q_{1,p}$ and $q_{2,p}$, $\hat{VaR}_{T+1}$ and $\hat{ES}^p_{T+1}$ can be estimated using

$$\hat{VaR}^p_{T+1} = -\hat{\sigma}^2_{T+1}\hat{q}_{1,p} \tag{3.16}$$

$$\hat{ES}^p_{T+1} = -\hat{\sigma}^2_{T+1}\hat{q}_{2,p} \tag{3.17}$$

where $\hat{q}_{1,p}$ and $\hat{q}_{2,p}$ are determined using the remaining 2 methods: Parametric and Semi-Parametric, both of which are described below.

3.3.4 Parametric

Conditional Normal

An explicit assumption of standard normal distribution on ϵ_t gives the following estimates of $\hat{q}_{1,p}$ and $\hat{q}_{2,p}$

$$\hat{q}_{1,p} = \Phi^{-1}(p) \tag{3.18}$$

$$\hat{q}_{2,p} = -\frac{\phi(\Phi^{-1}(p))}{p} \tag{3.19}$$

where $\Phi^{-1}(.)$ and $\phi(.)$ denote the inverse distribution and density function of standard Normal respectively. Equation (3.19) is multiplied by -1 to make ES a positive number.

Student-t

On the other hand, if ϵ_t is assumed to follow a standardised student-t with v dof, $\hat{q}_{1,p}$ and $\hat{q}_{2,p}$ are given by

$$\hat{q}_{1,p} = \sqrt{\frac{v-2}{v}} t_v^{-1}(p) \tag{3.20}$$

and

$$\hat{q}_{2,p} = -\sqrt{\frac{v-2}{v}} \frac{g_v(t_v^{-1}(p))}{1-p} \left(\frac{v + (t_v^{-1}(p))^2}{v-1} \right) \tag{3.21}$$

where $t_v^{-1}(.)$ and $g_v(.)$ denote the inverse distribution and density of standardised student-t respectively.

3.3.5 Semi-Parametric

Filtered Historical Simulation

As mentioned earlier, several variants of FHS have been suggested and extensively studied, In my analysis, the method proposed by Barone-Adesi et al. (1999) and as described in Christoffersen et al. (2004) is used. The authors proposed that the estimation of $\hat{q}_{1,p}$ and $\hat{q}_{2,p}$ be done by using the empirical distribution of the residuals $\hat{\epsilon}_t$. Assuming a conditional volatility model, $\hat{q}_{1,p}$ and $\hat{q}_{2,p}$ are given by

$$\hat{q}_{1,p} = F_p\left(\{\hat{\epsilon}_t - \bar{\hat{\epsilon}}_t\}_{t=1}^{T}\right) \tag{3.22}$$

$$\hat{q}_{2,p} = \frac{1}{\#(e \leq \hat{q}_{1,p})}\left(\sum_{e \leq \hat{q}_{1,p}} (e)\right) \tag{3.23}$$

where $\bar{\hat{\epsilon}}_t = \frac{\sum_{t=1}^{T} \hat{\epsilon}_t}{T}$ and $e = \hat{\epsilon}_t - \bar{\hat{\epsilon}}_t$. Note that, centering of $\hat{\epsilon}_t$ is necessary given Equation (3.2)[3]. It is, however, not required if returns are assumed to follow a different type of specification, for instance $R_t = \mu + \sigma_t \epsilon_t$.

Consequently, $\hat{VaR}^p_{T+1}$ and $\hat{ES}^p_{T+1}$ can be computed using

$$\hat{FHS\text{-}VaR}^p_{T+1} = -\hat{\sigma}_{T+1}\hat{q}_{1,p} \tag{3.24}$$

$$\hat{FHS\text{-}ES}^p_{T+1} = -\hat{\sigma}_{T+1}\hat{q}_{2,p} \tag{3.25}$$

3.4 Backtesting

According to Jorion et al. (2007), backtesting is a statistical testing framework that can be used to evaluate the reliability and adequacy of risk models and is done through the comparison of the ex-ante risk forecasts with the ex-post realised returns. In the Basel 1996 amendments, banks, or more general, financial institutions were allowed to adopt the Internal Model Approach - the usage of in-house models to compute VaR/ES. This power of discretion, albeit subject to approval, means regulators must verify the accuracy of bank's risk forecasts, to prevent banks from falsely reporting their VaR estimate and minimise capital charge.

[3]According to Christoffersen et al. (2004), centering so that the mean of $\hat{\epsilon}_t$ is 0, aligning with the model assumption

3.4.1 VaR

The wide spectrum of methodologies established by researchers can be generally categorised into: Frequency, Independence, Duration and Magnitude based tests. Frequency-based tests such as Kupiec Proportion of Failure (POF) test are based on violation ratios; Independence tests such as the Conditional Coverage test attempt to detect any serial dependency between the violations; Duration-based tests focus on the time between violations (see Christoffersen and Pelletier (2004)). Magnitude-Based tests pay special attention to the variance between realised returns and VaR estimates (see Lopez et al. (1999)).

For my analysis, I will focus on the frequency and independence based tests that are used to infer on the accuracy of VaR models as duration and magnitude based tests are generally used for selection of the best models i.e. the VaR models have already been established as valid and accurate.

Unconditional Coverage

A violation is said to have occurred, if the realised return is lower than the ex-ante risk forecast. Formally, VaR violations can be defined using an Indicator Function, I_t

$$I_t = \begin{cases} 1, \text{ if } R_t < -\hat{VaR}_t^p \\ 0, \text{ otherwise} \end{cases} \tag{3.26}$$

where $\hat{VaR}_t^p, t = T+1, ...T+m$ denotes the 1-day ahead VaR forecast. Furthermore, $\{I_t\}_{t=T+1}^{T+m}$, a series of ones and zeroes, also known as a "hit" sequence, can be thought of as a sequence of Bernoulli-distributed random variables, that is $I_t \sim Bern(p)$.

The Unconditional Coverage (UC) property can be defined as

$$prob(I_t = 1) = p \tag{3.27}$$

and can be tested using the POF test. First developed by Kupiec in 1995, it tests for the statistical significance of the UC property for any VaR models. Essentially, it compares the empirical coverage rate, $\hat{p}$ that is computed as the total number of observed violation divided by the size of m, against p, the assumed coverage rate.[4] Next, the UC test is presented.

For a random variable that is $\sim Bern(p)$, its likelihood function can be written as :

$$L(p) = \prod_{t=T+1}^{T+m} (1-p)^{1-I_t}(p)^{I_t} = (1-p)^{V0}(p)^{V1} \tag{3.28}$$

[4]That is, $\hat{p} = \frac{\sum_{t=T+1}^{T+m} I_t}{m}$

where V0 and V1 are the sum of 0s and 1s in the backtesting window. Replacing p with the observed $\hat{p}$, where $\hat{p}$

$$L(\hat{p}) = \prod_{t=T+1}^{T+m} (1-\hat{p})^{V0}(\hat{p})^{V1} \tag{3.29}$$

A likelihood ratio test can be conducted, with the test statistic LR_{UC} calculated as

$$LR_{UC} = -2log(L(p/L(\hat{p})) \overset{asym}{\sim} \chi^2_{(1)} \tag{3.30}$$

where LR_{UC} is asymptotically chi-square distributed with 1 dof. While relatively straightforward, the POF test has its drawback. Kupiec (1995) found that the POF test requires a large sample size to produce reliable test results. For a small sample size, such as the BASEL requirement of 250, the particular test has very low power. Furthermore, it fails to capture any dependence structure between the violations, i.e. violations cluster. The clustering of violations is undesirable as it means that the probability of observing a violation tomorrow, given a violation today, will be bigger than p.

Conditional Coverage

Building on the framework established by Kupiec (1995), Christoffersen (1998) proposed the Conditional Coverage (CC) test that can be separated into two stages. In the first stage, the independence property is evaluated, after which, the CC property is examined with a joint test of the UC and Independence (herein referred to as IND) property.

According to Berkowitz et al. (2011), the coverage rate, p, can be defined as the conditional probability of the occurrence of the violations. That is,

$$prob(I_t = 1|\Omega_t) = p \tag{3.31}$$

where Ω_t denotes the information set available at time t.

Independence

Assuming the "hit" sequence is dependent over time, Christoffersen (1998) proposed that it could be describe by a discrete-time Markov chain with the following transition probability matrix

$$\Pi_1 = \begin{bmatrix} 1-p_{01} & p_{01} \\ 1-p_{11} & p_{11} \end{bmatrix} \tag{3.32}$$

where $p_{ij} = prob(I_t = j|I_{t-1} = i)$ is the probability of observing $I_t = j$ given $I_{t-1} = i$. For instance, p_{01} is the probability of observing no violation today ($I_t = 0$) given a violation yesterday ($I_{t-1} = 1$) , while p_{11} refers to the probability of observing a violation today

conditional on a violation yesterday. The likelihood function of this process, given m observations is :

$$L(\Pi_1) = (1-p_{01})^{V_{00}} p_{01}^{V_{01}} (1-p_{11})^{V_{10}} p_{11}^{V_{11}} \tag{3.33}$$

where V_{ij} is the total number of observations where j follows i [5].

The maximum likelihood estimate is then given by

$$\hat{\Pi}_1 = \begin{bmatrix} \frac{v_{00}}{v_{00}+v_{01}} & \frac{v_{01}}{v_{00}+v_{01}} \\ \frac{v_{10}}{v_{10}+v_{11}} & \frac{v_{11}}{v_{00}+v_{10}} \end{bmatrix} \tag{3.34}$$

Assuming that the violations are independent across time, then the probability of observing one tomorrow is not dependent on observing one today. Thus, we have $p_{01} = p_{11} = p$, with a estimated transition matrix of

$$\hat{\Pi}_0 = \begin{bmatrix} 1-\hat{p} & \hat{p} \\ 1-\hat{p} & \hat{p} \end{bmatrix} \tag{3.35}$$

where

$$\hat{p} = \frac{v_{01} + v_{11}}{v_{00} + v_{10} + v_{01} + v_{11}} \tag{3.36}$$

The null hypothesis of independence, $p_{01} = p_{11}$ can then be tested using the likelihood ratio test with the following statistic

$$LR_{ind} = -2\left[L(\hat{\Pi}_0)/L(\hat{\Pi}_1)\right] \overset{asym}{\sim} \chi^2_{(1)} \tag{3.37}$$

Note that $L(\hat{\Pi}_0)$ in Equation (3.37) corresponds to $L(\hat{p})$ in the UC test.

Conditional Coverage

Finally, both test are combined to test for the CC property. The test statistic can be calculated as

$$LR_{cc} = LR_{uc} + LR_{ind} \overset{asym}{\sim} \chi^2_{(2)} \tag{3.38}$$

As noted in Campbell et al. (2005) and as will be shown later, the joint test should be not chosen over individual test for the evaluation of individual property because the joint test has lesser power in detecting a violation of a specific property. For instance, a model may ace the UC test, fails the IND test but pass the CC test. Model selection based solely on results from CC test will then result in selecting a wrong model. Finally, the backtesting will be conducted at 95% confidence level.

[5] For $v_{11} = 0$, $L(\Pi_1)$ can be calculated as $(1-p_{01})^{V_{00}}(p_{01})^{V_{01}}$

3.4.2 Expected Shortfall

Normalized Shortfall

Denoting NS as the normalized shortfall, it can be computed using

$$NS_t = \frac{R_t}{ES_t} \tag{3.39}$$

where ES_t refers to the ES on day t. Given Equation (3.11), the expectation of R_t, given a violation is

$$\frac{E[R_t | R_t \leq -VaR_t]}{ES_t} = 1 \tag{3.40}$$

Denoting $\bar{NS}$ as the average NS, the null hypothesis of $\bar{NS} = 1$ can be tested. According to Danielsson (2011), it is much trickier to create a formalised test to examine the statistical significance of the normalised shortfall as the test would be required to test for the accuracy of both risk measures simultaneously. Henceforth, the $\bar{NS}$ computed in my analysis serves only as a point of comparison between models, rather than the adequacy of the models themselves.

4. Methodology

4.1 Bootstrap Historical Simulation Risk Measure

In this section, the method proposed in Christoffersen et al. (2004) for the non-parametric VaR method is presented. The authors proposed to use the i.i.d bootstrapping - a n-out-of-n resampling with replacement technique that makes no assumption about the distribution of the underlying data - on the historical returns.

Step 1. From the original return series $\{R_t, t = 1, 2....T\}$, resample with replacement to generate bootstrap returns $\{R_t^*, t = 1, 2....T\}$

Step 2. Estimate the 1-day ahead $HS - VaR_{T+1}^p$ and $HS - ES_{T+1}^p$ using Equation (3.14) and (3.15)

Step 3. Repeat step 1 and 2 for B=1000 times and obtain 1000 estimates of $HS \hat{-} VaR_{T+1}^p$ and $HS \hat{-} ES_{T+1}^p$. The VaR and ES used for backtesting are computed as the average of all the bootstrap estimates, i.e.

$$\overline{HS - VaR}_{T+1}^p = \frac{\sum_{i=1}^{1000} HS \hat{-} VaR_{T+1}^{p(i)}}{B} \tag{4.1}$$

$$\overline{HS - ES}_{T+1}^p = \frac{\sum_{i=1}^{1000} HS \hat{-} ES_{T+1}^{p(i)}}{B} \tag{4.2}$$

Step 4. Denoting $q_\alpha(.)$ as the α percentile of the bootstrap empirical distribution of $HS \hat{-} VaR_{T+1}^p$, the $100(1-\alpha)\%$ prediction interval is

$$[q_{\alpha/2}\{\hat{VaR}_{T+1}^{*p(i)}\}_{i=1}^B, q_{1-\alpha/2}\{\hat{VaR}_{T+1}^{*p(i)}\}_{i=1}^B] \tag{4.3}$$

The prediction interval for $\hat{ES}_{T+1}^{*p}$ is obtained similarly.

Critics of the method have argued that as empirical analysis have long established that returns exhibit dependency over time, the use of such method on returns is invalidated because it requires returns to be i.i.d. To that end, Christoffersen et al. (2004)

proposed to replace returns with the model residuals that are in theory i.i.d in the population.

4.2 Bootstrap GARCH-Based Risk Measures

The method of re-sampling from the residuals is outlined below.

Step 1. Estimate each model and obtain the parameters $\hat{\theta}$ for the respective conditional volatility models.

Step 2. Obtain the residuals $\hat{\epsilon}_t = \frac{R_t}{\hat{\sigma}_t}$ and compute the centered residuals via $\hat{\epsilon}_t - \bar{\hat{\epsilon}_t}$

Step 3. Generate a series of pseudo returns $\{R_t^*, t = 1, 2....T\}$ recursively using the following formulation

For GARCH

$$\hat{\sigma}_t^{2*} = \hat{\omega} + \hat{\alpha} R_{t-1}^{*2} + \hat{\beta}\hat{\sigma}_{t-1}^{2*} \tag{4.4}$$

For GJR

$$\begin{aligned} &\text{If } R_{t-1}^* < 0 \\ &\quad \hat{\sigma}_t^{2*} = \hat{\omega} + (\hat{\alpha} + \hat{\gamma}) R_{t-1}^{*2} + \hat{\beta}\hat{\sigma}_{t-1}^{2*} \\ &\text{else} \\ &\quad \hat{\sigma}_t^{2*} = \hat{\omega} + \hat{\alpha} R_{t-1}^{*2} + \hat{\beta}\hat{\sigma}_{t-1}^{2*} \end{aligned} \tag{4.5}$$

where $R_t^* = \hat{\sigma}_t^* \epsilon_t^*$ for $t = 1, 2....T$; $\hat{\sigma}_1^{2*} = \hat{\sigma}_1^2 = \frac{\hat{\omega}}{1-\hat{\alpha}-\hat{\beta}}$ for the GARCH; $\hat{\sigma}_1^{2*} = \hat{\sigma}_1^2 = \frac{\hat{\omega}}{1-\hat{\alpha}-\hat{\beta}-\frac{\hat{\gamma}}{2}}$ for the GJR. Note that ϵ_t^* are random draws with replacement from the empirical distribution of $\hat{\epsilon}_t$.

Step 4. Re-estimate the models with pseudo return series R_t^* using QML and obtain the bootstrap parameters, $\hat{\theta}^*$.

Step 5. Compute the 1-day ahead GARCH volatility forecast $\hat{\sigma}_{T+1}^{2*}$ recursively using :

$$\hat{\sigma}_{T+1}^{2*} = \hat{\omega}^* + \hat{\alpha}^* R_T^2 + \hat{\beta}^* \hat{\sigma}_T^{2*} \qquad \text{for } t = 1, 2....T \tag{4.6}$$

where $\hat{\sigma}_1^{2*}$ here is computed as $\frac{\hat{\omega}^*}{1-\hat{\alpha}^*-\hat{\beta}^*}$.

Step 6. For GJR, for $R_t \geq 0$, use Equation (4.6), else

$$\hat{\sigma}_{T+1}^{2*} = \hat{\omega}^* + (\hat{\alpha}^* + \hat{\gamma}^*) R_T^2 + \hat{\beta}^* \hat{\sigma}_T^{2*} \qquad \text{for } t = 1, 2....T \tag{4.7}$$

where $\hat{\sigma}_1^{2*}$ here is computed as $\frac{\hat{\omega}^*}{1-\hat{\alpha}^*-\hat{\beta}^*-\frac{\hat{\gamma}^*}{2}}$

Step 7. $\hat{q}_{1,p}^*$ and $\hat{q}_{2,p}^*$ are the bootstrap variant of $q_{1,p}$ and $q_{2,p}$. For instance, in the conditional normal models, $\hat{q}_{1,p}^*$ = Equation (3.18) and $\hat{q}_{2,p}^*$ = Equation (3.19). The same applies to the t-dist models. Consequently, for the parametric method, there are no estimation uncertainty in the estimation of $\hat{q}_{1,p}^*$ and $\hat{q}_{2,p}^*$.

The FHS method, on the other hand, requires the obtaining of the bootstrap residuals, $\hat{\epsilon}_t^*$. With the bootstrap residuals, we can now compute both the $\hat{q}_{1,p}^*$ and $\hat{q}_{2,p}^*$. For instance,

$$\hat{q}_{1,p}^* = \hat{Q}_p\left(\{\hat{\epsilon}_t^* - \bar{\hat{\epsilon}}_t^{\,*}\}_{t=1}^T\right) \tag{4.8}$$

where $\hat{Q}_p(.)$ denotes the p-th quantile of the empirical distribution of the bootstrap residuals. $\hat{q}_{2,p}^*$ can be computed using Equation (3.23), via replacing the residuals with the bootstrap version.

Step 8. Compute the corresponding $\hat{VaR}_{T+1}^p$ and $\hat{ES}_{T+1}^p$ for each method.

Step 9. Repeat steps 3 and 4 under the Bootstrap HS method to obtain the average and prediction interval for each risk measure.

Notice that, while $\hat{\sigma}_{T+1}^{2*(B)}$ will be different for all B = 1,..1000 replicates, its value is based on the bootstrap parameters and on the original returns series $\{R_t, t = 1, 2...T\}$, rather than the pseudo returns R_t^*. According to Pascual et al. (2006), through doing so, the bootstrap volatility forecast will be small when actual returns (losses) are small, and large when actual returns (losses) are large.

Finally, step 4 attempts to account for any possible estimation error in $\hat{\sigma}_{T+1}^2$, $\hat{q}_{1,p}$ and $\hat{q}_{2,p}$ for the semi-parametric method and only $\hat{\sigma}_{T+1}^2$ for the parametric method, through the replacement of $\hat{\theta}$ with its bootstrap version $\hat{\theta}^*$.

5. Data

The data in my analysis consists of 3018 observations of daily closed prices of the S&P500 index from 02/07/2007- 28/06/2019. As mentioned earlier, log-returns as per defined in Equation (3.1) are used and are multiplied by 100 to avoid any potential optimisation issues.

There will be two types of analysis in this paper. First, in-sample data properties are studied, after-which, the total sample size is split into two windows - estimation window (T) and backtesting window (m). Instead of focusing on T, this analysis will focus on m. Specifically, the analysis will be based on m=250 and m=1250. The decision to use the former is based on the BASEL requirement while looking at a 5-year backtesting window size would look at the performance of the considered methods in several periods of extended volatility and shocks[1]. Consequently, for m=250 (1250), T will consist of 2768 (1768)[2] data points and will be used for the estimation of all the models. Thereafter, both the forecast of volatility and risk measures will be computed.

Next, the models are re-estimated daily with fixed rolling windows. That is, the window (i.e. both start and end point) is rolled one day forward, producing 250 (1250) out-of-sample risk forecasts.

5.1 Descriptive Statistic

Table 5.1: Descriptive Statistics

Mean	Standard Deviation	Skewness	Excess Kurtosis	JB	p value
0.000214	0.012523791	-0.34717	10.56305	14105.57	0.001

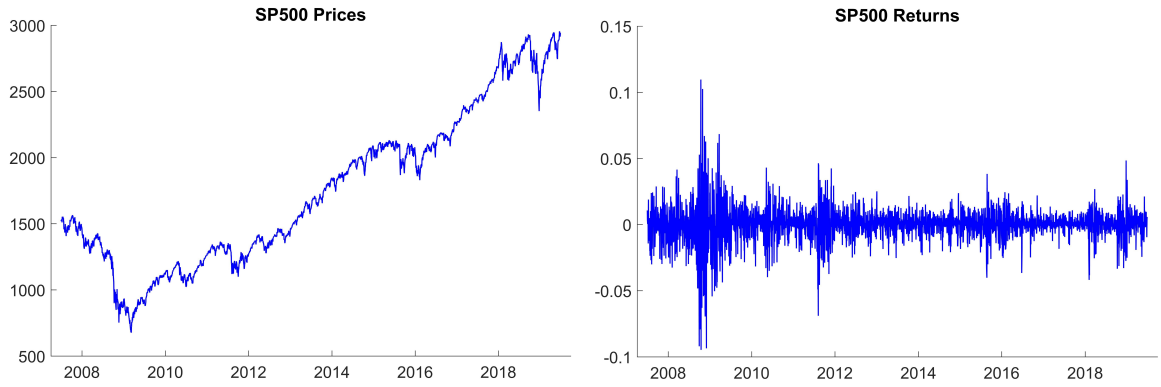
The statistics are based on the original log-returns (without the multiplication of 100

First, notice that the sample daily mean is close to 0, which justifies the use of the specification of returns per defined Equation (3.2), in which $E[R_t] = 0$. As mentioned

[1]For e.g. the equity market selloff between 2015-2016 and also Dec 2018 that was coined the worst December since Great Depression

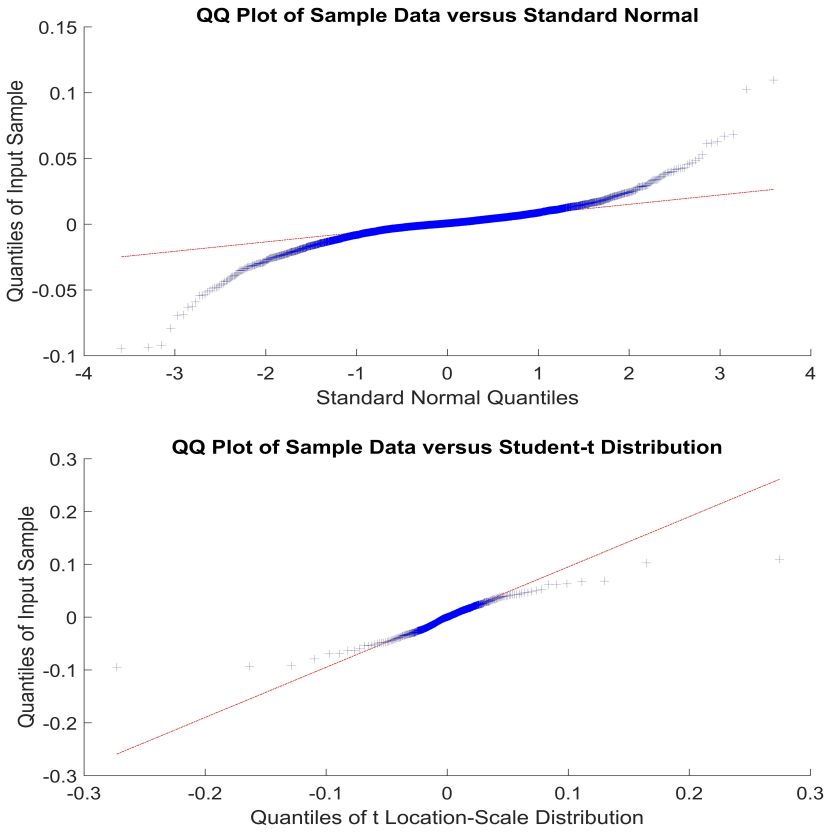
[2]Both are in accordance to the recommendation in NG and Lam (2006). The authors investigated the impact of T on GARCH models, and found that for T<700, ML may produce 2 or more "optimal solutions", and recommended T be $\geq$1000

Figure 5.1: SP500 Prices and Returns



earlier, a correct specification of the first two moments underpins the validity of QML. Next looking at Figure 5.1, we can see evidence of volatility clustering in the returns, especially during the 2008 crisis.

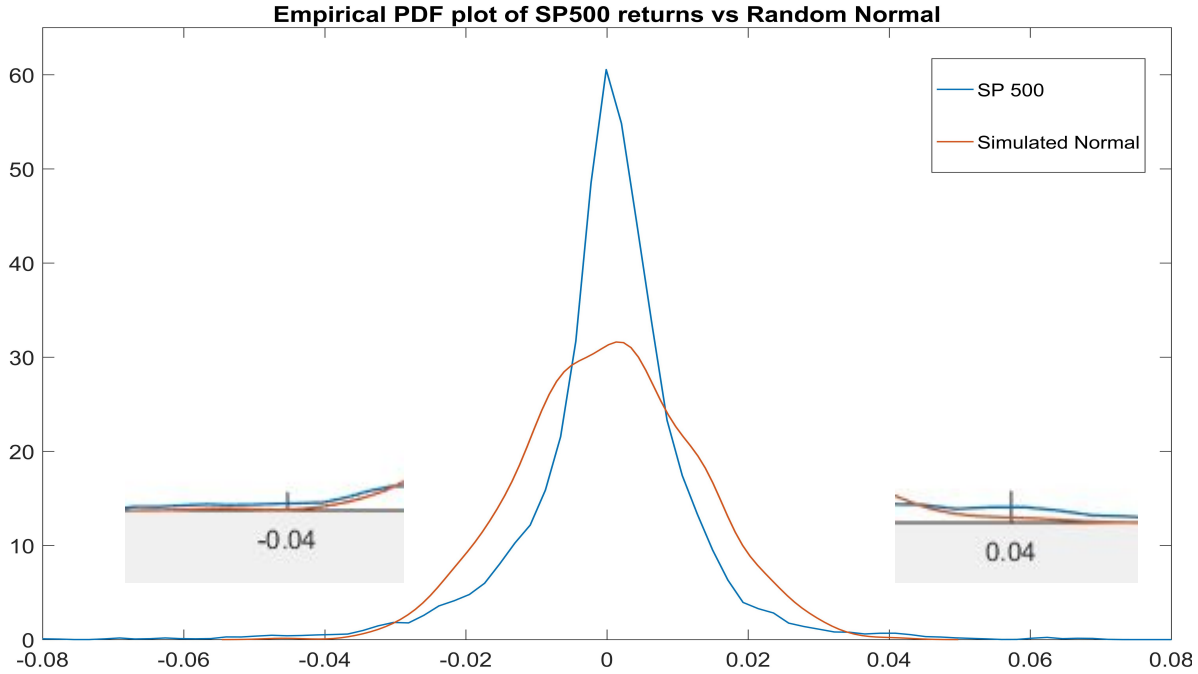
Figure 5.2: QQ plot of returns with normal & student-t distribution



Next, the excess kurtosis is >10, implying that the distribution is leptokurtosis (fat-tail) and suggesting that the normality assumption does not hold in the data. To this end, two types of test were conducted. In the first test, the graphical approach: Quantile Quantile (QQ)-plot and Density Plot, is employed. In Figure 5.2, notice that both the

tails seem to fit nicely on the straight line in the bottom plot and bents away in the top plot, suggesting that the distribution of returns seems to follow the t-dist, rather than normal. For the density plot, random numbers from the normal distribution with the data's sample mean and variance are first generated. Next, the density function of the data is then plotted against the density of the simulated data. Figure 5.3 plots the density plot, together with a zoom-in on both the tails. Clearly the tails are fatter than the normal distribution with the same first and second moment.

Figure 5.3: Density Plot



For the second test, the JB test for normality was conducted. As shown in Table 5.1, the p-value from JB test is <0.01, suggesting strong evidence to reject the null hypothesis of normality.

Next, Figure 5.4 plots the ACF function of SP500 returns and squared returns. The left plot suggests returns are not autocorrelated (fluctuates around 0) up to 20 lag. On the other hand, it is clear from the right plot that R_t^2 exhibit strong linear dependency. To confirm the hypothesis, a formal statistical test - Ljung Box (LB) test, is conducted to test for joint significance of autocorrelation in squared returns. According to Tsay (2014), the power of LB test depends on the number of lags used for the test. The author suggested optimal number of lags $\approx \ln(n)$ for better power performance. Using n=3020, the optimal number is ≈8. Hence, the null hypothesis corresponds to $H_0 : \rho_1 = \rho_2 = ...\rho_8 = 0$. From the p-value in Table 5.2, the null hypothesis of no serial autocorrelation is rejected, indicating the presence of conditional heteroscedasticity and the suitability of using GARCH models to model volatility.

Figure 5.4: ACF plot of SP500 returns and squared returns

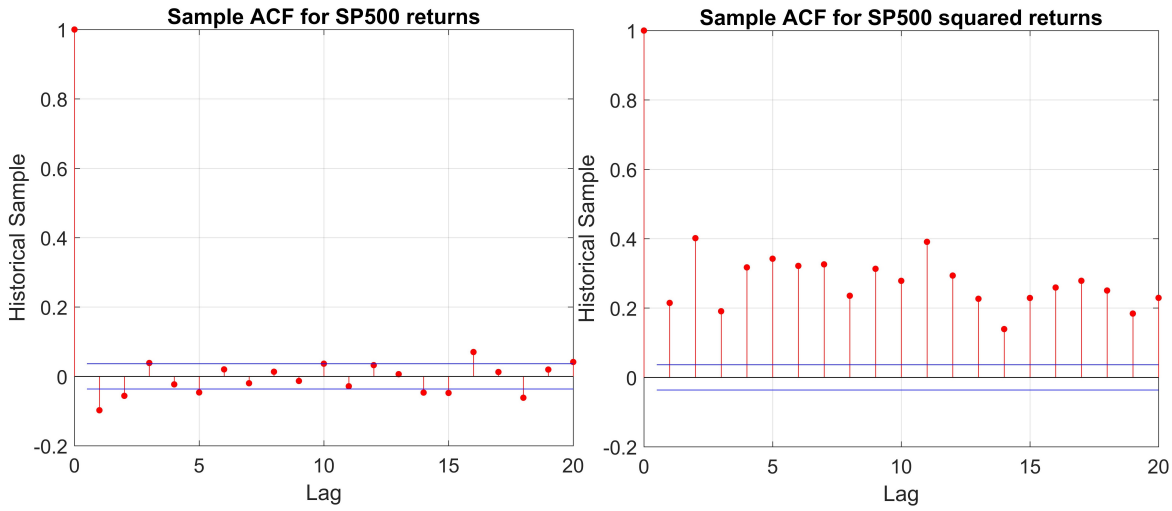


Table 5.2: LB test

p value
0.000

6. Analysis

6.1 In Sample Estimation

6.1.1 Models Diagnosis

In this section, the in-sample performance of the GARCH models is assessed using two different methods - parameters significance test and residual analysis. Other methods such as goodness-of-fit measures are used for model comparison and selection (see McMillan et al. (2000)), while Information Criterions are typically used for choosing the appropriate lag length, p and q.

Table 6.1: Estimated Model Parameters

Parameters	**GARCH**	**GJR**	**t-GARCH**	**t-GJR**
$\hat{\boldsymbol{\omega}}$	0.0239 (0.0000)	0.0268 (0.0000)	0.0133 (0.001187)	0.0218 (0.0000)
$\hat{\boldsymbol{\alpha}}$	0.1300 (0.0000)	1.10e-08 (0.0000)	0.12822 (0.0000)	2.09e-08 (0.0008)
$\hat{\boldsymbol{\beta}}$	0.8527 (0)	0.8644 (0.0000)	0.8716 (0)	0.8626 (0.0000)
$\hat{\gamma}$		0.2298 (0)		0.2640 (0)
dof			5.279	5.828

The numbers in the brackets correspond to the p-value from the t-test

In Table 6.1, the estimated parameters of the models are presented. The close to 0 p-values from the t-test indicate strong evidence of the statistical significance of the parameters from each model.

Likelihood Ratio

In addition to standard t-test that can be used to determine the statistical significance of estimated parameters, Likelihood ratio test can be conducted for nested models to examine the significance of additional parameters added in the model. For instance, the GARCH (1,1) is a nested version of the GJR (1,1), with the number of restriction = 1. Denoting L_U and L_R as the unrestricted and restricted log-likelihood, the LR test statistic is

$$LR = -2(L_R - L_U) \tag{6.1}$$

where $LR \sim \chi^2$(number of restrictions).

Table 6.2: Likelihood Ratio Test

Unrestricted	Restricted	Log L	LR statistic	Restrictions	p-value
GJR	**GARCH**	-4023.85	159.66	1	0
t-GARCH	**GARCH**	-4020.56	166.23	1	0
t-GJR	**GARCH**	-3952.49	302.37	2	0

Last column refers to the p-value from the LR test

Looking at the p-value from Table 6.2, we can reject the null hypothesis of both the unrestricted and restricted models are equal and conclude that all of the unrestricted models are better than the GARCH. Furthermore, if we replace the restricted model to be either GJR or t-GARCH, and unrestricted to be t-GJR, we can see that t-GJR is also significantly different and better than either of the models.

Residual Analysis

A correct model used should produce residuals that are in line with the underlying model assumption. For instance, assuming $e_t \sim N$ should produce $\hat{e}_t \sim N$ as well. According to Andersen et al. (2005), if $\hat{e}_t$ is normally distributed, then conditional normality assumption can be considered valid. More often than not, however, $\hat{e}_t$ is found to be non-normal. Having said that, in Andersen et al. (2005), they found the following - $\hat{e}_t$ fits better to the normal distribution than returns; the left tail of $\hat{e}_t$ is heavier, as compared to its right tail.

Figure 6.1: QQ plot of normal models residuals

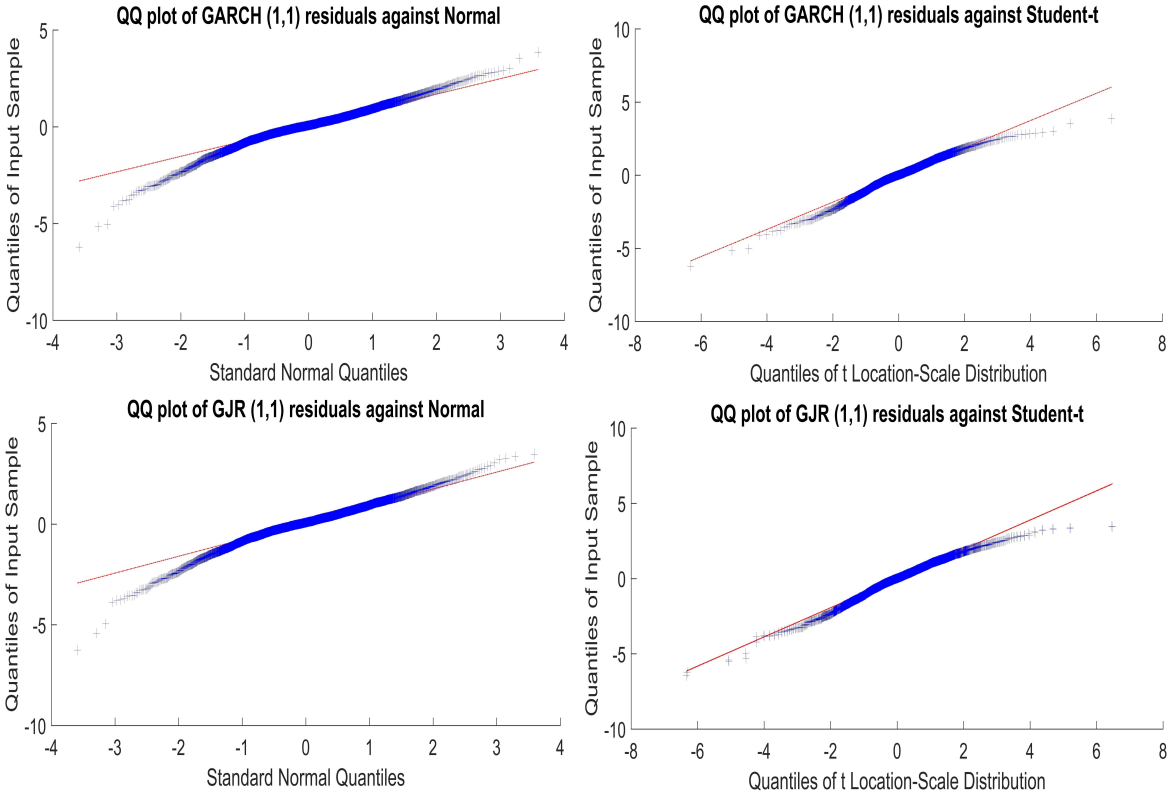


Figure 6.1 and Figure 6.2 show the QQ plot of residuals from the normal and t-dist models respectively. The results from both Figures coincide with the findings in Andersen et al. (2005). Looking at the residuals from all models, we can see that the right tail conforms closer to the normal distribution, while the t-dist seems to fit better to the left tail, i.e. heavier left tail. The p-value from JB test in Table 6.3 further confirms that the residuals are non-normal.

Figure 6.2: QQ plot of t-models residual

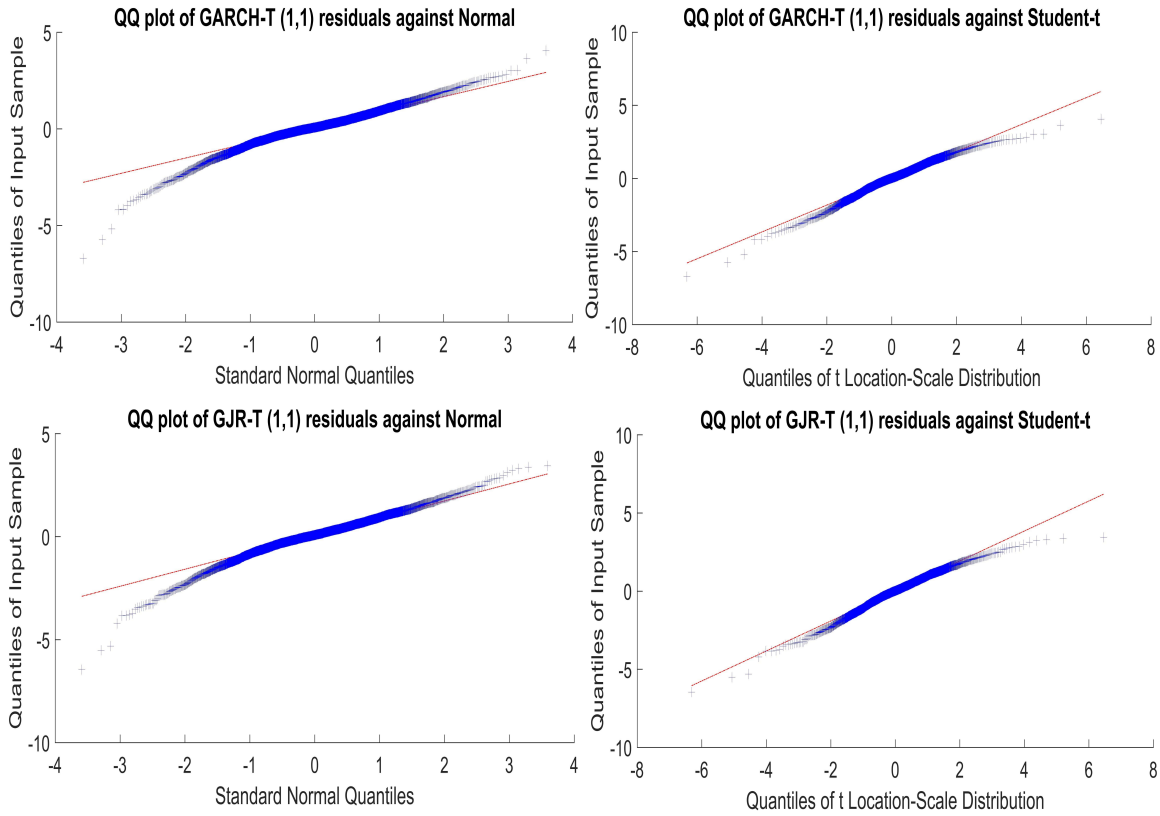


Figure 6.3: ACF plot of normal models residuals

Next, looking at Figure 6.3 and 6.4 , we can see that no significant autocorrelation was found in both residuals and squared residuals of all models. The LB test results from Table 6.3 further confirms the hypothesis - we can't reject the null hypothesis of no autocorrelation.

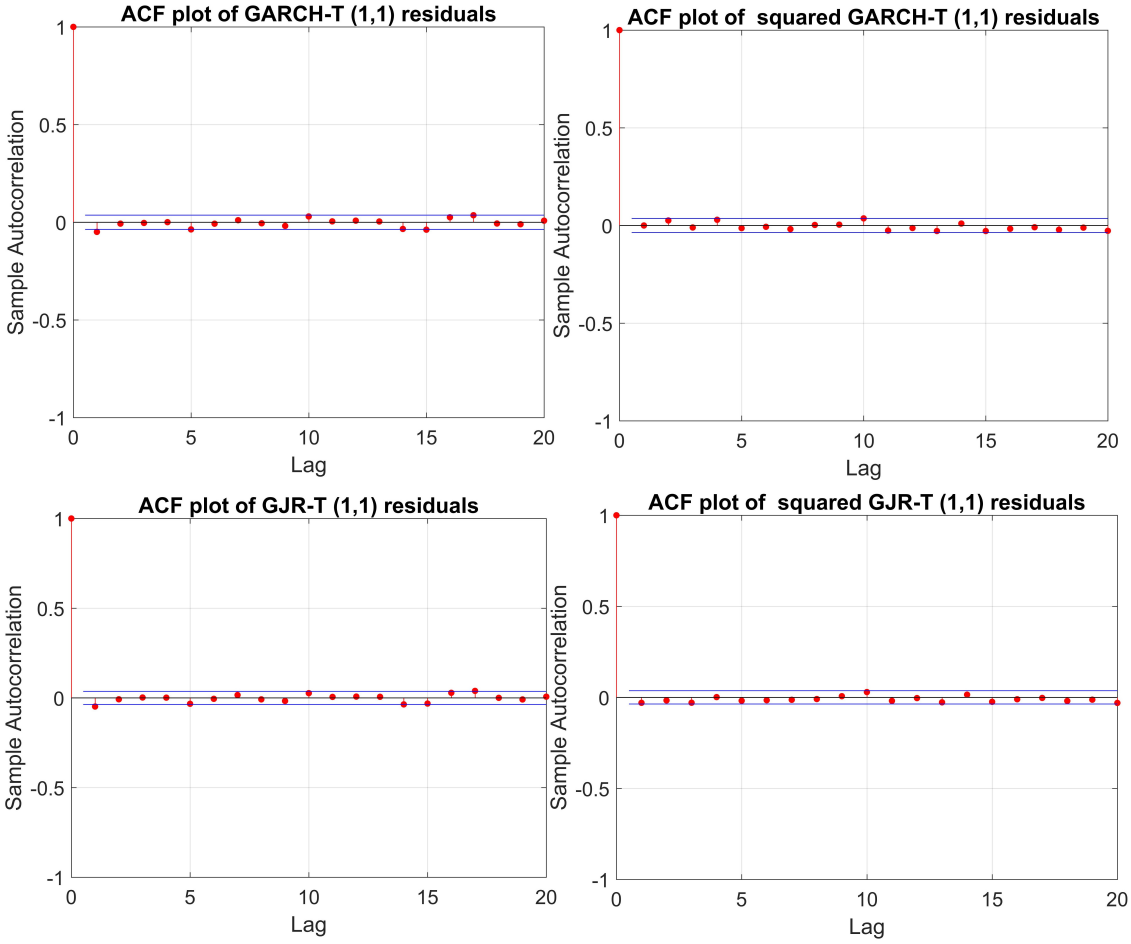
Table 6.3: Statistical test for residuals

Model	JB test	LB test
GARCH	0.001	0.296
GJR	0.001	0.473
t-GARCH	0.001	0.255
t-GJR	0.001	0.332

Column 1 and 2 contain the p-value from JB and LB test respectively

Summarising the results, first, the non-normality of returns and residuals indicates that even the assumption of conditional normality for measuring risk will be inadequate.

Figure 6.4: ACF plot of t-models residual



However, the absence of autocorrelation in the residuals and squared residuals means that all 4 models have performed well to capture the volatility dynamics. Furthermore, it justifies the use of the i.i.d bootstrapping on an i.i.d series such as the residuals. Next, an initial examination of the models indicates that the t-GJR would be the best model in estimating VaR and ES. Finally, the results illustrate why FHS tends to perform well - the preservation of fat-tail property within the residuals means FHS would not face the issue of underestimation.

6.2 Out of Sample Analysis

6.2.1 Historical Simulation

In this section, the length of the estimation window is varied to illustrate the impact it has on VaR and ES.

Table 6.4 contains the backtesting results from HS. 6 rolling Ts were considered. First, notice the strong negative correlation between T and the total number of violations

for both values of p. When T increased by tenfold, total violations fell by more than 90%, confirming the hypothesis that HS is incredibly sensitive to the length of estimation window. For p=0.05, with the exception of T=1250, the remaining windows passed the UC test. However, none of the windows passed the remaining test. For p=0.01, all windows passed the UC test. In cases of T=2000 and T=2500, both passed the IND and CC test. The remaining windows failed both the tests. The results found above coincide with Pritsker (2006). He found that the HS method had relatively good performance in UC test, albeit very poor VaR estimates. He then went on to conclude that the UC test was ineffectual as a form of backtesting.

Table 6.4: Historical Simulation Backtesting Results

				VaR			**ES**
	T	**# Vio**	**VR**	**UC**	**IND**	**CC**	**NS**
5%	250	143	1.0332	0.6898	0.0000	0.0002	1.1043
	500	109	0.8658	0.1141	0.0000	0.0000	1.0671
	750	94	0.8289	0.0544	0.0000	0.0000	1.0078
	1250	67	0.7579	0.0149	0.0000	0.0000	0.9877
	2000	39	0.7662	0.0748	0.0000	0.0000	0.9492
	2500	23	0.8880	0.5515	0.0021	0.0073	0.8726
	T	**# Vio**	**VR**	**UC**	**IND**	**CC**	**NS**
1%	250	36	1.3006	0.1287	0.0111	0.0125	1.1825
	500	21	0.8340	0.3888	0.0005	0.0015	1.1431
	750	18	0.7937	0.3056	0.0071	0.0158	1.0684
	1250	12	0.6787	0.1495	0.0019	0.0029	1.0226
	2000	6	0.5894	0.1539	0.7896	0.3491	0.9383
	2500	2	0.3861	0.1087	0.9008	0.2741	0.7590

The top panel corresponds to p=0.05 while the panel below corresponds to p=0.01. The 2nd column represents the total number of violations; 3rd column is the computed Violation Ratio. Column 4-6 contain the p-value from the respective backtests. The final column is the normalized shortfall

Figure 6.5 and 6.6 further illustrates the serious shortcomings of HS. Figure 6.5 plots the risk measures for large T. As shown by the almost flat lines, it is clear that the VaR and ES estimates were very unresponsive, especially for T=2500 and 2000. In contrast, the VaR estimates were extremely volatile in Figure 6.6. The problem was most severe for T=250 and for 99% VaR and ES. The volatility that we see at the beginning was most probably due to the sudden drop of relatively large losses that occurred during the 2008 crisis. A relatively calmed market between 2013-2015 stabilised the risk estimates. However, when volatility returned, the variation in the risk estimates increased. Hence, with such poor and unreliable results, HS will be excluded in the subsequent analysis. Instead, the i.i.d-HS will be used as the benchmark.

Figure 6.5: HS-VaR and ES for large T

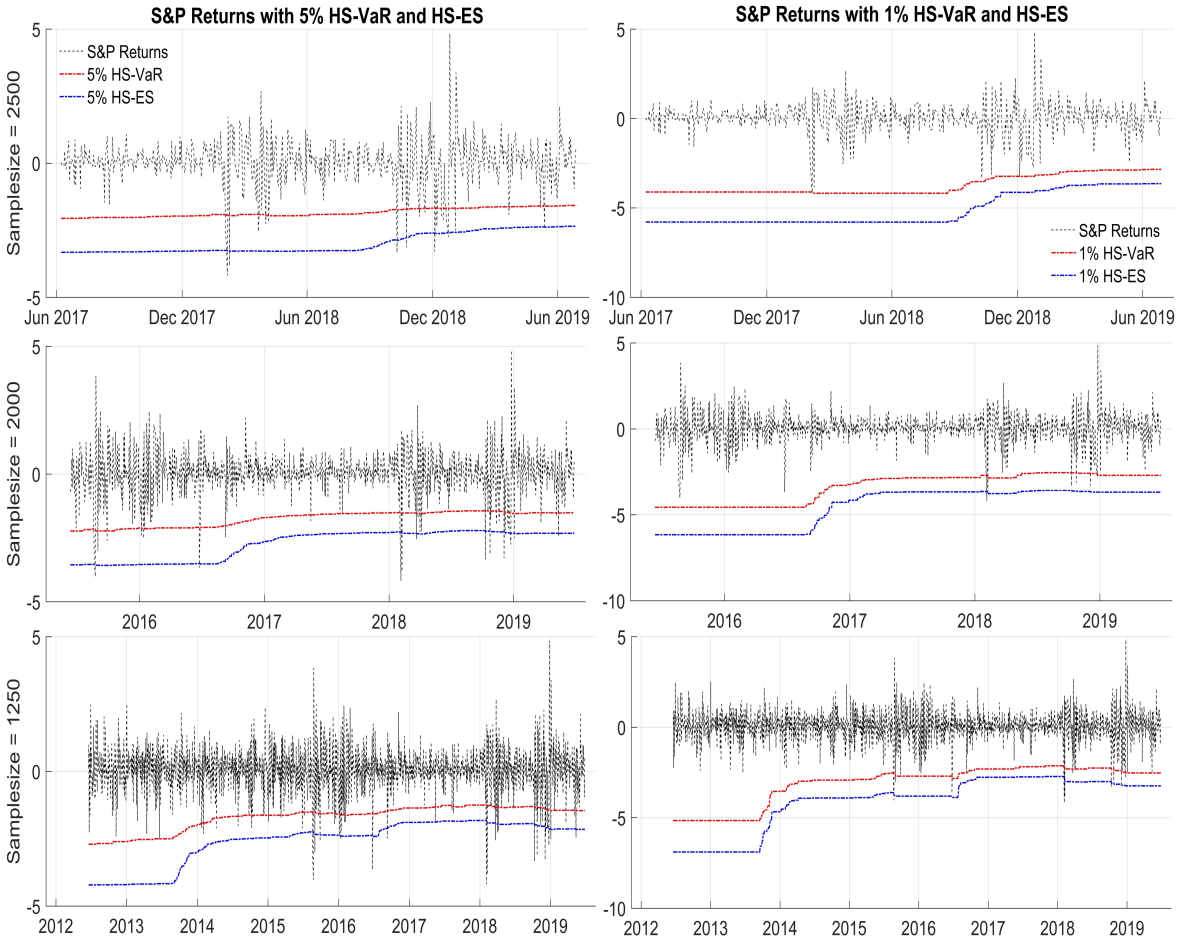
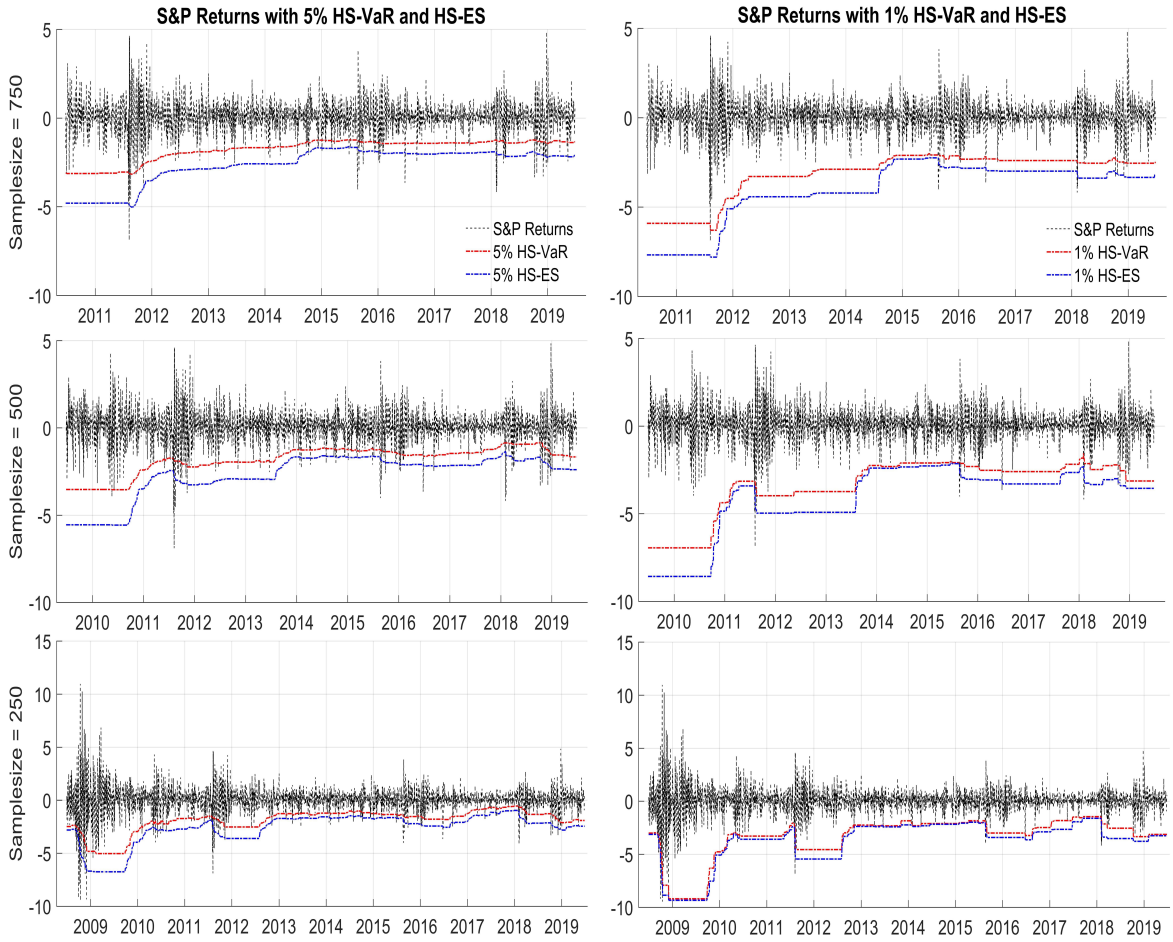


Figure 6.6: HS-VaR and ES for small T



6.2.2 GARCH based models

Point Prediction

In this section, the backtesting results of all the GARCH based models, together with the benchmark model, are presented. For the FHS method, GARCH (GJR) filters will refer to the residuals from GARCH (GJR) models (both normal and t-dist), while t-filters will refer to residuals from t-dist models (both GARCH and GJR).[1]

p=0.05

First, looking at m=250 in Table 6.5, notice the upward bias of the VaR estimates, i.e. there are fewer violations than expected (VR <1). In particular, the FHS method seems to have a higher tendency to overestimate risk. Consequently, the FHS method has a poorer performance in the UC test, albeit both methods were able to pass the test.

[1]For instance, residuals from student-t GJR model will be referred to as t-GJR filter

Table 6.5: 5% VaR and ES Backtesting Results

Table 6.5: 5% VaR and ES Backtesting Results

		m = 250						m = 1250					
		VaR					ES	VaR					ES
Boot	**Methods**	**# Vio**	**VR**	**UC**	**IND**	**CC**	**NS**	**# Vio**	**VR**	**UC**	**IND**	**CC**	**NS**
N	Parametric												
	GARCH	9	0.7200	0.2860	0.4113	0.4039	1.3920	56	0.8960	0.3909	0.0468	0.0960	1.2184
	GJR	9	0.7200	0.2860	0.4113	0.4039	1.3896	57	0.9120	0.4690	0.3964	0.5370	1.1764
	t-GARCH	10	0.8000	0.4529	0.4009	0.5302	1.2509	57	0.9120	0.4690	0.0552	0.1225	1.1172
	t-GJR	10	0.8000	0.4529	0.3602	0.4965	1.2474	63	1.0080	0.9483	0.3165	0.6043	1.0702
	FHS												
	GARCH	8	0.6400	0.1632	0.4661	0.2901	1.2180	46	0.7360	0.0249	0.1131	0.0231	1.0934
	GJR	7	0.5600	0.0828	0.5245	0.1814	1.2957	41	0.6560	0.0030	0.1999	0.0054	1.1049
	t-GARCH	8	0.6400	0.1632	0.4661	0.2901	1.2599	44	0.7040	0.0114	0.2806	0.0228	1.1180
	t-GJR	8	0.6400	0.1632	0.4661	0.2901	1.2403	43	0.6880	0.0075	0.2518	0.0145	1.0949
Y	Parametric												
	GARCH	9	0.7200	0.2860	0.4113	0.4039	1.3909	56	0.8960	0.3909	0.0468	0.0960	1.2143
	GJR	9	0.7200	0.2860	0.4113	0.4039	1.3878	55	0.8800	0.3208	0.7071	0.5693	1.1861
	t-GARCH	11	0.8800	0.6571	0.4941	0.7173	1.2018	57	0.9120	0.4690	0.0552	0.1225	1.1078
	t-GJR	10	0.8000	0.4529	0.3602	0.4965	1.2462	62	0.9920	0.9482	0.5953	0.8666	1.0697
	FHS												
	GARCH	7	0.5600	0.0828	0.5245	0.1814	1.2908	46	0.7360	0.0249	0.1131	0.0231	1.0929
	GJR	7	0.5600	0.0828	0.5245	0.1814	1.2965	41	0.6560	0.0030	0.1999	0.0054	1.1068
	t-GARCH	8	0.6400	0.1632	0.4661	0.2901	1.2474	45	0.7200	0.0170	0.3113	0.0348	1.1058
	t-GJR	8	0.6400	0.1632	0.4661	0.2901	1.2389	44	0.7040	0.0114	0.2806	0.0228	1.0875
	HS	11	0.8800	0.6571	0.0783	0.1923	0.7929	45	0.7200	0.0170	0.0000	0.0000	1.0069

On the left panel are the results from m=250. 1st column - N: without bootstrap, Y: with bootstrap; 3rd column represents the total number of violations; 4th column is the computed Violation Ratio; Column 5-7 are the p-value from the various backtest. The final column is the normalized shortfall. On the right panel are the results from m=1250. The layout is the same as the left panel

Moving on to IND and CC test, both methods have a relatively better performance where all models passed both tests.

Next, the results for m=1250 are more interesting. Looking first at the UC test, things are significantly better for the parametric method. All models passed the test, with t-GJR being the best performer. On the other hand, none of the filters under the FHS method passed the UC test. Similar to the case of a smaller m, risk is overestimated. This could be attributed to the residual's left tail being fatter than that of the returns'. Next, moving onto IND test, the GARCH models under the parametric method fared poorly and failed the test, while the GJR models performed better. For the FHS method, all filters passed the IND test. However, the lacklustre performance of the FHS method in the UC test resulted in all filters failing the CC test. In contrast, all models under the parametric method passed the CC test. The results from the GARCH models illustrates the pitfall of focusing solely on CC test - passed the CC test but with undetected clustered violations. Finally, the HS method performed poorly unsurprisingly. The unresponsiveness resulted in consecutive violations, thus failing all the tests.[2]

Moving onto ES, the results from NS for the two backtesting window are conflicting. Looking at m=250, the NS for both methods are at least 20% higher than it should be. In particular, the normal-models under the parametric method underestimated ES by around 40%. With a larger backtesting window, now most of the NS are only around 10% larger; Expanding the backtesting window seems to improve the ES estimates.

Lastly, looking at the bootstrap results, there are not much improvements to the VaR estimates from both methods. However, looking specifically at m=1250, we can see that there are slight improvements in the performance of the t-GJR model and filter for VaR.

p=0.01

Looking further into the tail reveals different results in Table 6.6. First, for m=250, risk appears to be severely underestimated for the parametric method - there are twice as many violations as expected. Despite that, all the models passed the UC test. The situation is reversed for the FHS. With an expected violation of 2.5, the FHS method produced between 2 and 3 violations for all the filters. For the remaining 2 test, all the models under both methods passed, with FHS outperforming the parametric method again. In contrast, the HS recorded 0 violations - risk is severely overestimated.

Now, the situation is slightly different for m=1250. The first 3 models under the parametric method failed the UC test, while only the t-GJR produced satisfactory results and passed the test. In contrast, the FHS method produced much better results

[2]Note that, even though the HS method passed all 3 tests for m=250, the reliability of the results is questionable

Table 6.6: 1% VaR and ES Backtesting Results

Table 6.6: 1% VaR and ES Backtesting Results

		m = 250						m = 1250					
		VaR					ES	VaR					ES
Boot	**Methods**	**# Vio**	**VR**	**UC**	**IND**	**CC**	**NS**	**# Vio**	**VR**	**UC**	**IND**	**CC**	**NS**
N	Parametric												
	GARCH	6	2.4000	0.0594	0.5862	0.1458	1.2896	24	1.9200	0.0037	0.0095	0.0005	1.2353
	GJR	6	2.4000	0.0594	0.5862	0.1458	1.2872	20	1.6000	0.0499	0.0389	0.0173	1.2659
	t-GARCH	6	2.4000	0.0594	0.5862	0.1458	1.0337	21	1.6800	0.0277	0.0041	0.0014	1.0318
	t-GJR	5	2.0000	0.1619	0.6508	0.3393	1.1063	16	1.2800	0.3403	0.1974	0.2765	1.1164
	FHS												
	GARCH	2	0.8000	0.7419	0.8572	0.9320	1.3925	16	1.2800	0.3403	0.0147	0.0324	1.0682
	GJR	3	1.2000	0.7580	0.7868	0.9194	1.2008	14	1.1200	0.6757	0.1451	0.3169	1.0787
	t-GARCH	2	0.8000	0.7419	0.8572	0.9320	1.4448	15	1.2000	0.4908	0.0110	0.0312	1.0902
	t-GJR	3	1.2000	0.7580	0.7868	0.9194	1.2058	14	1.1200	0.6757	0.1451	0.3169	1.0816
Y	Parametric												
	GARCH	6	2.4000	0.0594	0.5862	0.1458	1.2872	24	1.9200	0.0037	0.0095	0.0005	1.2289
	GJR	6	2.4000	0.0594	0.5862	0.1458	1.2848	20	1.6000	0.0499	0.0389	0.0173	1.2631
	t-GARCH	6	2.4000	0.0594	0.5862	0.1458	1.0250	20	1.6000	0.0499	0.0030	0.0018	1.0265
	t-GJR	5	2.0000	0.1619	0.6508	0.3393	1.1015	16	1.2800	0.3403	0.1974	0.2765	1.1042
	FHS												
	GARCH	2	0.8000	0.7419	0.8572	0.9320	1.3917	16	1.2800	0.3403	0.0147	0.0324	1.0724
	GJR	3	1.2000	0.7580	0.7868	0.9194	1.2024	14	1.1200	0.6757	0.1451	0.3169	1.0843
	t-GARCH	2	0.8000	0.7419	0.8572	0.9320	1.4143	15	1.2000	0.4908	0.0110	0.0312	1.0889
	t-GJR	3	1.2000	0.7580	0.7868	0.9194	1.2044	13	1.0400	0.8877	0.1221	0.2996	1.1090
	HS	0	0.0000	0.0250	1.0000	0.0811	0.0000	10	0.8000	0.4615	0.6879	0.7034	0.9889

On the left panel are the results from m=250. 1st column - N: without bootstrap, Y: with bootstrap; 3rd column represents the total number of violations; 4th column is the computed Violation Ratio; Column 5-7 are the p-value from the various backtest. The final column is the normalized shortfall. On the right panel are the results from m=1250. The layout is the same as the left panel

- all filters passed the UC test. For the remaining 2 test, the t-GJR under both methods passed both tests. The remaining models from both methods failed the CC test, driven mainly by their poor performance in the IND test. Surprisingly, HS was the best performer in the CC test.

Looking at NS, notice that the parametric method produced rather consistent results across two backtesting window - the normal models underestimated ES by around 30%, while the t-dist models performed much better. However, the results from the FHS method are conflicting again. With a smaller m, the ES estimates were considerably smaller than it should be. In particular, the GARCH-filter underestimated ES by around 40%, while the GJR-filter were off by around 20%. Expanding the backtesting size generated different results. Now, we see that the underestimation problem is less severe (around only 10%) across all filters.

Moving onto the bootstrap results, we see that there are again no improvements to the VaR estimates from both methods. For both m=250 and 1250, both methods yield the same coverage as before without bootstrapping. However, we see again that the t-GJR filter benefited again from bootstrapping.

Prediction Interval

In this section, the prediction intervals (measured at 95% confidence level) and the corresponding properties are presented. Hereinafter, the term "width" will be used to describe the width of the prediction intervals for VaR and ES, while the phrases "width difference" and "SE difference" will be used to describe the difference in the width and Standard Error between VaR and ES.

p=0.05

Table 6.7 contains the result from p=0.05. First, looking at m=250, for the HS method, the width and SE of ES are twice as large than the width and SE of VaR. Looking at the parametric method, the width difference from the normal models is around 0.5, while the width difference from the t-models is two times larger. The SE difference is small - between 1-2% for all models.

Next, for the FHS, notice that the width difference and SE difference from the normal-filters are two times larger than those of from the normal models under the parametric method, while they are largely the same between t-models and t-filters. However, the width difference and SE difference between the filters are around the same. This could be attributed to the observation of the residuals from all models following the t-dist. Furthermore, both VaR and ES under the FHS method were estimated with more uncertainty than the parametric method as we see an increase in the width and SE across all filters. Looking across all the methods, we can see that the HS has the

Table 6.7: 5% VaR and ES Prediction Interval Properties

Table 6.7: 5% VaR and ES Prediction Intervals Properties

		m = 250						m = 1250					
Parametric	**Models**	**Lower Bound**	**Upper Bound**	**Width**	**Width Diff**	**SE**	**SE Diff**	**Lower Bound**	**Upper Bound**	**Width**	**Width Diff**	**SE**	**SE Diff**
N	VaR	1.7829	2.1247	0.3418	0.2975	0.0931	0.0704	1.5555	1.8804	0.3249	0.2566	0.0823	0.0662
	ES	2.9062	3.5456	0.6393		0.1635		2.3842	2.9657	0.5815		0.1485	
Y													
	GARCH	1.4054	1.6005	0.1951	0.0496	0.0498	0.0127	1.2390	1.4568	0.2178	0.0553	0.0555	0.0141
	GARCH	1.7624	2.0071	0.2447		0.0625		1.5538	1.8269	0.2732		0.0696	
	GJR	1.4172	1.6092	0.1920	0.0488	0.0490	0.0124	1.2516	1.4672	0.2156	0.0548	0.0549	0.0139
	GJR	1.7772	2.0180	0.2408		0.0614		1.5695	1.8399	0.2704		0.0688	
	t-GARCH	1.3674	1.5204	0.1530	0.0926	0.0388	0.0235	1.2093	1.4096	0.2003	0.1179	0.0512	0.0301
	t-GARCH	1.9370	2.1827	0.2456		0.0624		1.6870	2.0053	0.3182		0.0813	
	t-GJR	1.3991	1.5523	0.1531	0.0926	0.0393	0.0242	1.2328	1.4169	0.1841	0.1049	0.0471	0.0270
	t-GJR	1.9453	2.1910	0.2458		0.0634		1.6901	1.9791	0.2890		0.0741	
Semi													
	GARCH	1.4714	1.7419	0.2705	0.0985	0.0687	0.0254	1.3130	1.6111	0.2982	0.0938	0.0762	0.0236
	GARCH	2.0840	2.4530	0.3690		0.0940		1.8137	2.2057	0.3920		0.0999	
	GJR	1.4963	1.7487	0.2524	0.1084	0.0636	0.0284	1.3340	1.6281	0.2941	0.0921	0.0754	0.0229
	GJR	2.0878	2.4486	0.3608		0.0920		1.8180	2.2042	0.3862		0.0983	
	t-GARCH	1.4756	1.7189	0.2433	0.1004	0.0619	0.0252	1.3075	1.6004	0.2929	0.0947	0.0754	0.0235
	t-GARCH	2.1200	2.4637	0.3437		0.0870		1.8218	2.2094	0.3877		0.0988	
	t-GJR	1.5105	1.7440	0.2336	0.1036	0.0600	0.0258	1.3384	1.6142	0.2759	0.0937	0.0705	0.0236
	t-GJR	2.1232	2.4603	0.3371		0.0858		1.8345	2.2041	0.3695		0.0941	

On the left panel are the results from m=250. The parametric column states the type of the method - N: HS, Y: Parametric and Semi: FHS; Entries under the models column correspond to VaR first, then ES; Width column is calculated as the difference between the upper and lower bound; Width difference column is the difference between the width; SE stands for Standard Error, which is computed as the standard deviation of the all bootstrap estimates; SE difference is the difference between the SE of VaR and ES. On the right panel are the results from m=1250. The layout is the same as the left panel.

worst performance across all considered metrics, with the measured uncertainty of the risk estimates and between the risk estimates easily two times larger than the remaining methods'.

Moving onto m=1250, the HS again produced terrible results - the width difference is around 0.26. However, it seems to benefit from a smaller estimation window, which might be due to the sudden drop of extreme events - the width and SE for both the risk measure and the width difference and SE difference are smaller than before. Next, for the parametric method, similar to the case of a smaller m, the normal models were estimated with lesser uncertainty - the width difference and SE difference from the t-models are around two times larger than the width difference and SE difference from the normal-models. The results from FHS are largely similar to before - an increase in the width for both risk measure; similar width difference and SE difference between the filters. Moreover, we see again that FHS generated more uncertain estimates as compared to the parametric method.

Figure 6.7: Width Difference (zoom-in) between 5% VaR and ES

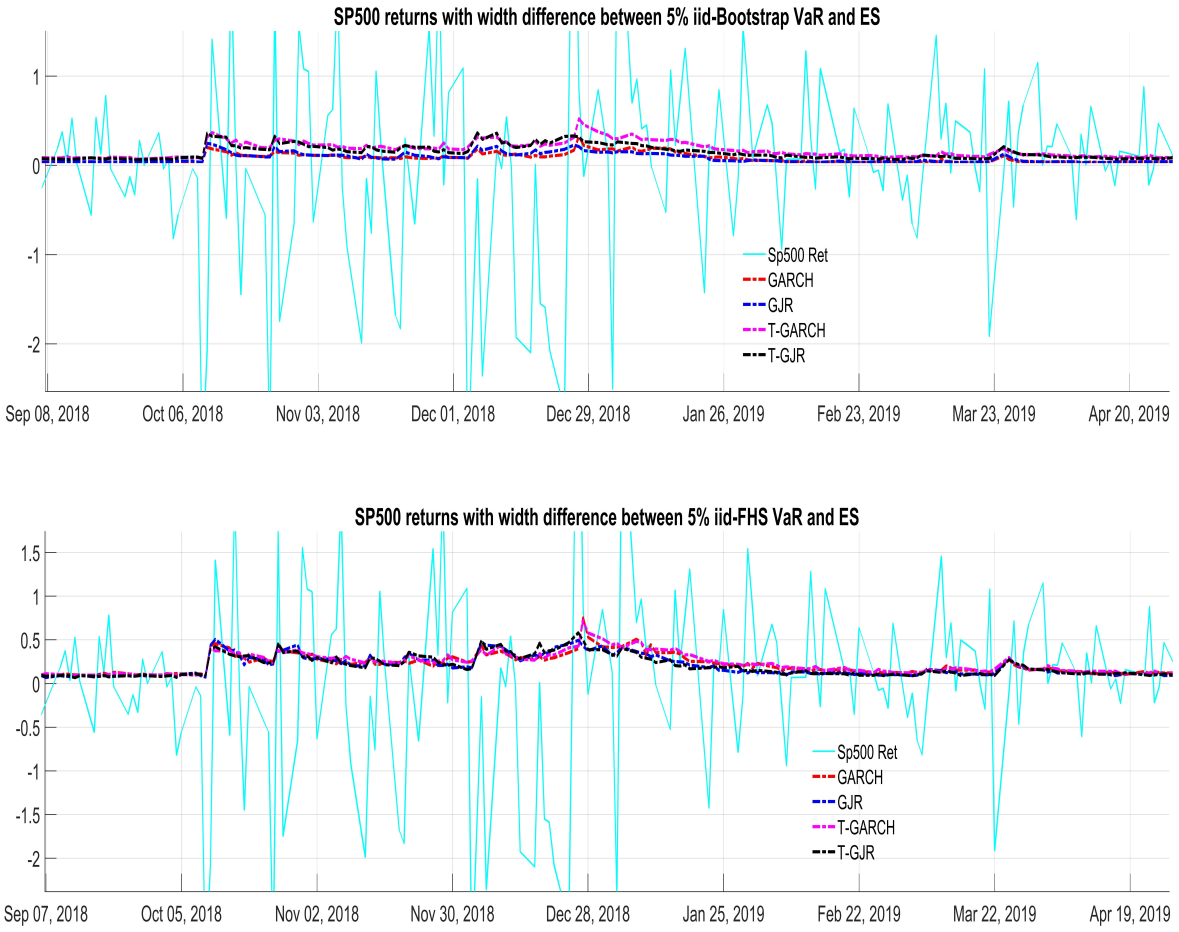
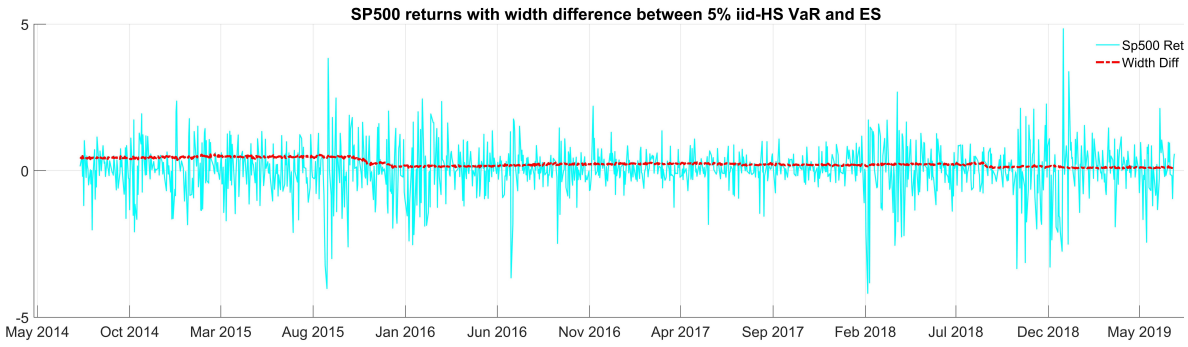


Figure 6.7 plots the width difference for each model and is intentionally zoom-in

for better illustration[3]. Looking at the top panel, we can see that the width difference generated by all the models were similar during calm periods (t-models were marginally larger). However, when the market was hit with shocks in Oct 2018, the disparity between the t and normal models widened. Looking at the period of extended volatility in Dec 2018 revealed certain interesting insights. First, when the market was hit by large negative shocks, t-GJR generated the largest difference. However, t-GARCH produced the largest difference when the market rebounded in Jan 2019. The results are slightly different under the FHS. Looking at the same periods, we now see a similar width difference across all filters after Oct 2018, albeit larger than before. Moreover, the market rebound saw the GARCH models producing a larger difference than GJR, with t-GARCH generating the largest difference. Looking at Figure 6.8, we can see that the width difference between HS VaR and ES remained roughly the same during both calm and volatile periods.

Figure 6.8: iid-HS Width Difference between 5% VaR and ES



p=0.01

Moving onto p=0.01 results in Table 6.8, first, notice that the width and SE of both risk measure under HS almost tripled as compared to before. Moreover, the width difference and SE difference increased as well. For the parametric method, the width of both risk measure increased across all models. In particular, the width from the t-models is two times larger than that of the t-models at p=0.05. Moreover, there was also a threefold increase in the width difference and SE difference as well. However, it remained the same for the normal-models. Next, looking at all filters, the width and SE of both risk measure under FHS are now wider and larger than those of the parametric method (twice as large). More importantly, the width of both risk measure doubled, and the width difference and SE difference tripled when comparing the two significance level.

For m=1250, we see again that there is a slight improvement in the width and SE for both risk measure from the HS method, but with a larger width difference and SE

[3]See Figure 8.1 for the original plot and Figure 8.3 for the plot for m=250. Since both backtesting window produced similar results, only the plot for m=1250 is shown here; the analysis applies to both

Table 6.8: 1% VaR and ES Prediction Interval Properties

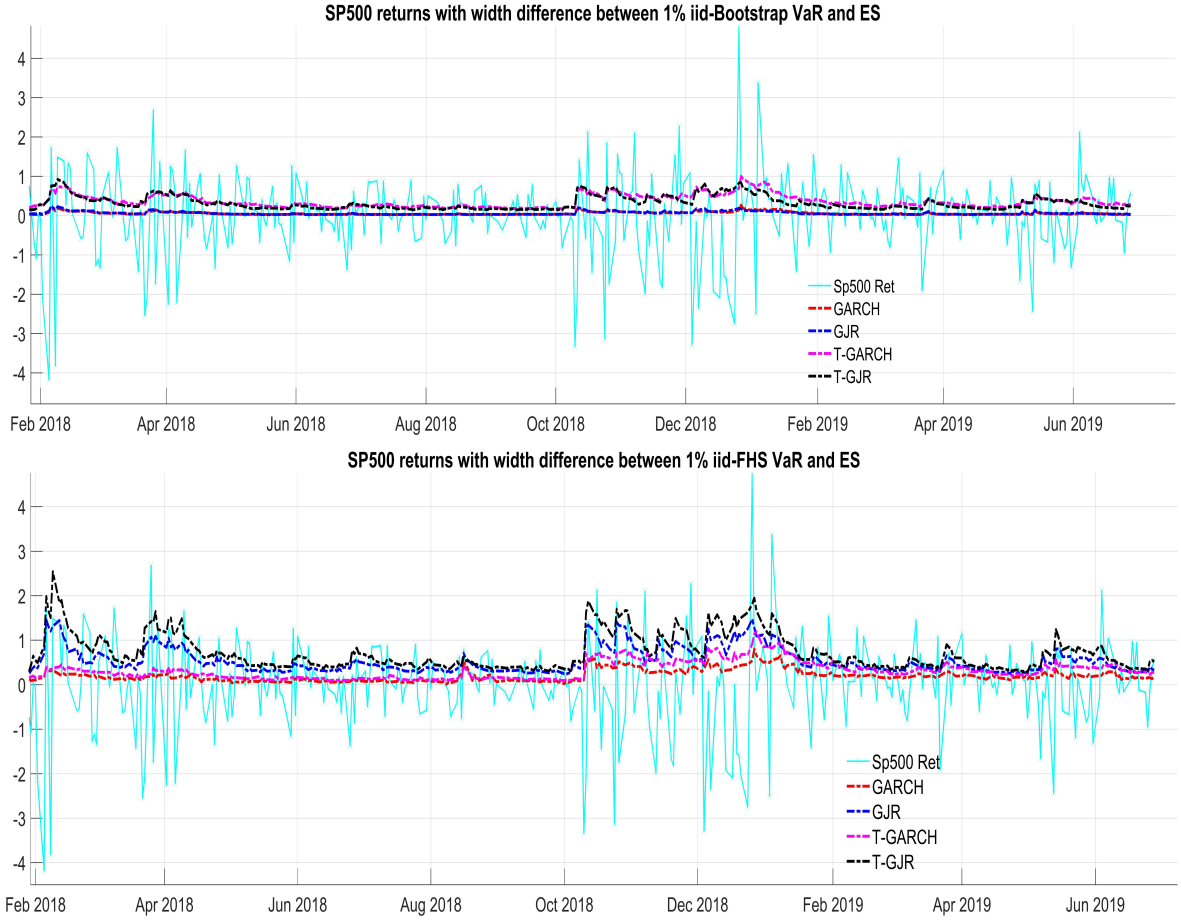
Table 6.8: 1% VaR and ES Prediction Intervals properties

		m = 250						m = 1250					
Parametric	**Models**	**Lower Bound**	**Upper Bound**	**Width**	**Width Diff**	**SE**	**SE Diff**	**Lower Bound**	**Upper Bound**	**Width**	**Width Diff**	**SE**	**SE Diff**
N	VaR	3.3745	4.6490	1.2745	0.4219	0.3315	0.1015	2.8067	3.7228	0.9161	0.4983	0.2336	0.1286
	ES	4.7514	6.4478	1.6964		0.4330		3.6285	5.0429	1.4144		0.3622	
Y													
	GARCH	1.9876	2.2636	0.2760	0.0402	0.0704	0.0103	1.7523	2.0604	0.3081	0.0449	0.0785	0.0114
	GARCH	2.2772	2.5934	0.3162		0.0807		2.0076	2.3606	0.3530		0.0900	
	GJR	2.0044	2.2759	0.2716	0.0396	0.0693	0.0101	1.7701	2.0751	0.3050	0.0444	0.0776	0.0113
	GJR	2.2964	2.6075	0.3111		0.0793		2.0280	2.3774	0.3494		0.0889	
	t-GARCH	2.2482	2.5448	0.2966	0.2459	0.0753	0.0625	1.9522	2.3342	0.3820	0.2717	0.0976	0.0693
	t-GARCH	2.8925	3.4350	0.5425		0.1378		2.4675	3.1211	0.6536		0.1670	
	t-GJR	2.2506	2.5486	0.2980	0.2287	0.0768	0.0582	1.9496	2.2979	0.3484	0.2377	0.0893	0.0605
	t-GJR	2.8425	3.3692	0.5268		0.1350		2.4209	3.0070	0.5861		0.1498	
Semi													
	GARCH	2.4147	2.9048	0.4901	0.2262	0.1272	0.0557	2.0785	2.6430	0.5646	0.1024	0.1489	0.0207
	GARCH	2.8800	3.5963	0.7163		0.1829		2.4540	3.1209	0.6669		0.1696	
	GJR	2.4085	2.9070	0.4984	0.2505	0.1243	0.0671	2.0684	2.5895	0.5211	0.2368	0.1314	0.0628
	GJR	2.8617	3.6106	0.7489		0.1913		2.4243	3.1822	0.7580		0.1942	
	t-GARCH	2.4343	2.9335	0.4992	0.3276	0.1304	0.0810	2.0802	2.6469	0.5667	0.1429	0.1484	0.0324
	t-GARCH	2.9882	3.8149	0.8268		0.2114		2.4795	3.1891	0.7096		0.1807	
	t-GJR	2.4241	2.9393	0.5151	0.2793	0.1249	0.0784	2.0790	2.5851	0.5061	0.2894	0.1268	0.0774
	t-GJR	2.9424	3.7368	0.7945		0.2033		2.4561	3.2516	0.7955		0.2042	

On the left panel are the results from m=250. The parametric column states the type of the method - N: HS, Y: Parametric and Semi: FHS; Entries under the models column correspond to VaR first, then ES; Width column is calculated as the difference between the upper and lower bound; Width difference column is the difference between the width; SE stands for Standard Error, which is computed as the standard deviation of the all bootstrap estimates; SE difference is the difference between the SE of VaR and ES. On the right panel are the results from m=1250. The layout is the same as the left panel.

difference. Looking at the parametric method, we have the same situation as before with a smaller m. The situation for FHS is also similar to a smaller m, except for the GARCH-based filters - smaller width difference and SE difference in comparison to before.

Figure 6.9: Width Difference (zoom-in) between 1% VaR and ES



Looking at the top panel of Figure 6.9 (zoom-in version) and 8.2, now we see that the width difference from the t-models were uniformly larger than the normal models and that the gap widened significantly during periods of shocks. However, similar to before, a large positive shock resulted in the t-GARCH producing a larger difference than t-GJR, while the opposite holds for large negative shocks. On the other hand, the situation is more complicated for FHS. Looking at the bottom plot in Figure 8.2, first, notice that the width difference between the models were similar for the period between Jun 2014 - Jun 2016. After which, the GJR filters generated larger width difference. An extended volatile period in 2018 caused the t-GJR filter to produce the largest difference for both positive and negative shocks. In contrast, the width difference from GARCH filters were significantly smaller. Furthermore, looking at the bottom right plot in Figure 6.11, we can see that the large increase in the width difference was due to a significant widening of the interval for ES. Looking at Figure 6.10, for the HS, the width difference was around the same during volatile periods but smaller during calm periods.

Figure 6.10: iid-HS Width Difference between 1% VaR and ES

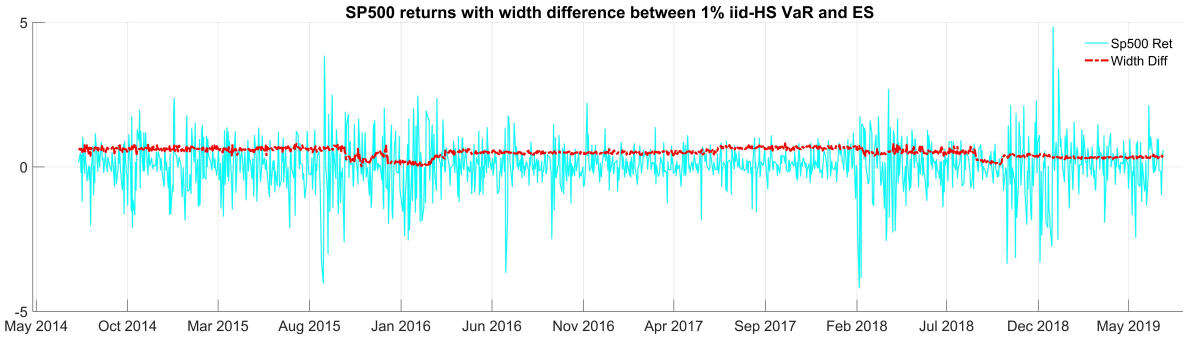


Figure 6.11: Width of the interval for 1% t-models VaR and ES

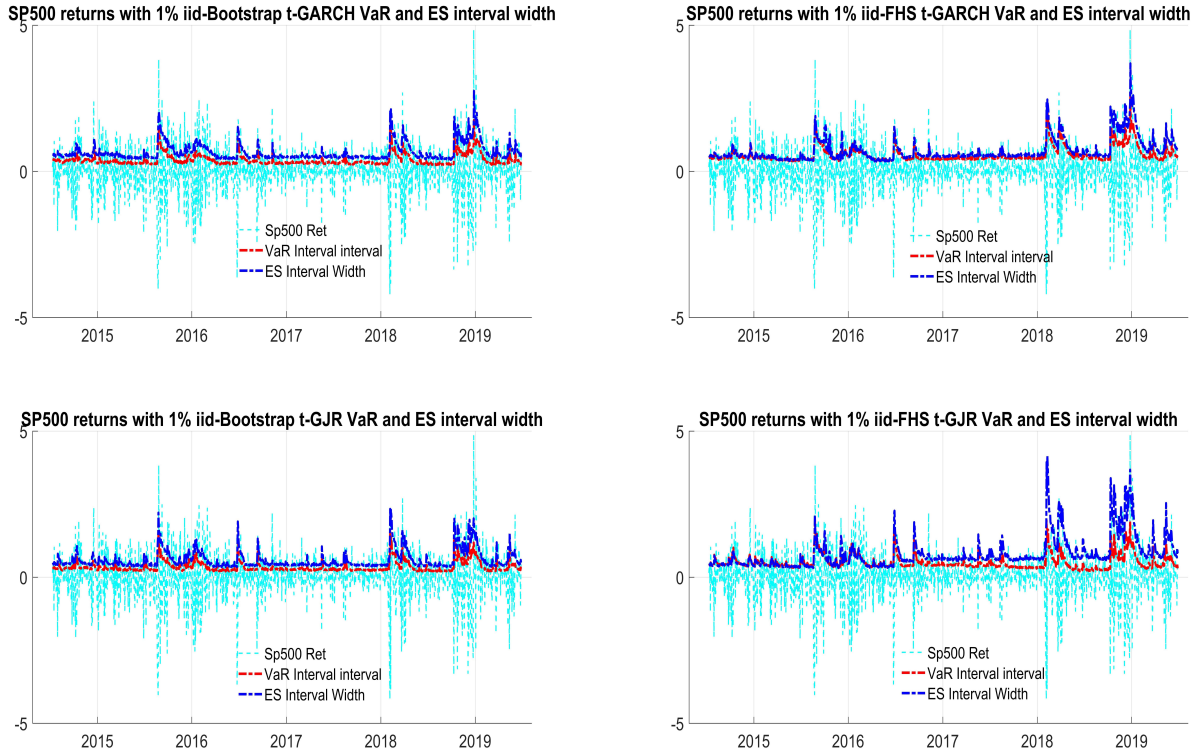
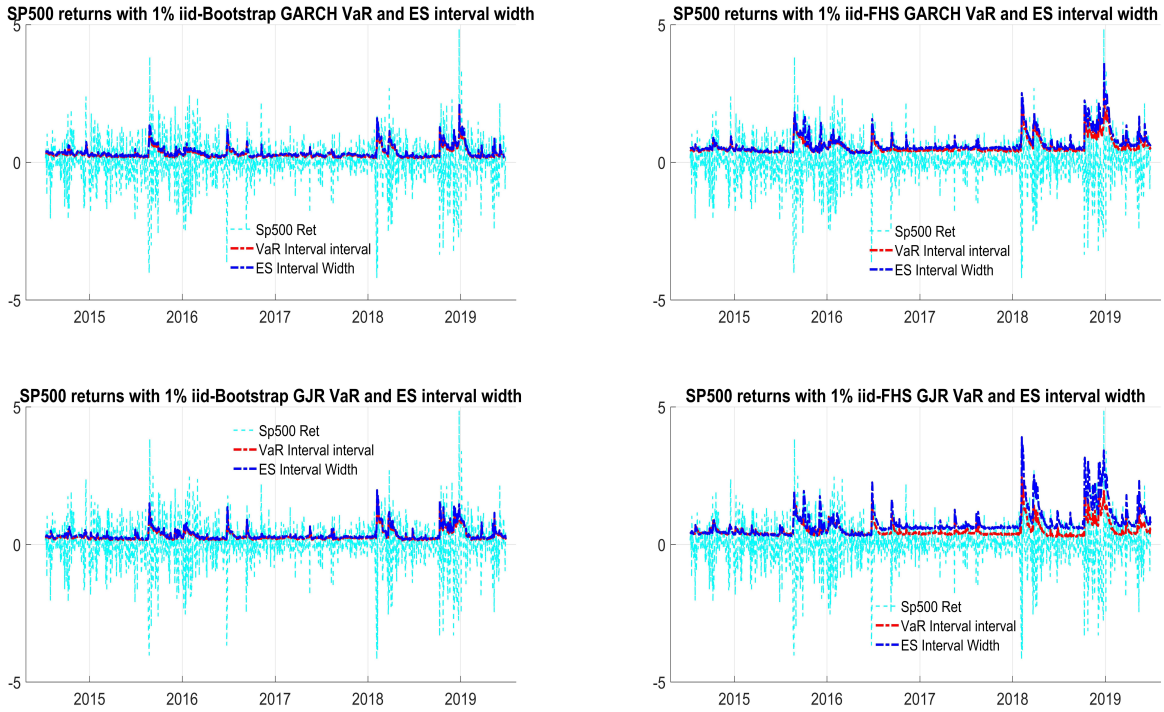


Figure 6.12: Width of the interval for 1% normal-models VaR and ES



7. Conclusion

The aim of this analysis was twofold - first, to compare the performance of various estimation method and to assess the degree of improvement of the risk estimates after accounting for estimation uncertainty. Second, to quantify and compare the uncertainty of risk estimates. The results I obtained from implementing the proposed method by Christoffersen et al. (2004) on empirical data were largely similar to the theoretical studies that were reviewed earlier. First, HS is a very unreliable method and the reliance on HS among practitioners is disconcerting given so; The most accurate risk model would be the one that reflects the most empirical properties found in returns - t-GJR model and the corresponding filter is the best performer. Second, HS produces extremely uncertain estimates as shown by the large SE and wide intervals for both VaR and ES at both significance level. Third, regardless of the risk measure, uncertainty in the risk estimates increases as one looks further into the tail. Lastly, across the board, ES was estimated with larger uncertainty as compared to VaR, and the difference in uncertainty is larger for t-dist than normal. The situation is also exacerbated when looking at a higher confidence level, with HS producing the largest width difference and SE difference across the considered methods.

My analysis, however, also revealed some unexpected results. First, the importance of the length of the backtesting window cannot be overlooked. The disparity in the backtesting performance of the considered methods for different m seems to indicate an increased likelihood of committing a type II error[1] if a small m is used. For instance, for m=250, the GARCH models seem to be an adequate choice, producing accurate and reasonable VaR and ES estimate. However, the performance took a deep dive when m was expanded. On the other hand, the t-GJR produced consistent and good results across methods and different m. Second, estimation uncertainty is negligible in simpler models regardless of the confidence level; increased complexity in the model specification does benefit from bootstrapping. Last, but more important, there seems to be a trade-off between the model adequacy and estimates accuracy when looking deeper into the tail. While the GJR filters delivered a solid performance during backtesting, the risk estimates were far more uncertain, and the uncertainty between VaR and ES

[1] A type II error in this case refers to the situation of accepting a false model as true

skyrocketed during volatile periods, driven mainly by a larger increase in the uncertainty of ES. On the other hand, while the estimates from the GARCH models and filters had a poor out-of-sample forecasting performance, they were measured with more certainty.

Finally, my findings could due in part to the use of an adequate estimation window size (T). Hence, the same analysis could be conducted for a much smaller T, for instance T=250, to investigate for any improvements. Also, while the results above are consequential to the choice of models and risk measures, further analysis should be first carried out to understand the volatility of the risk estimates exhibited by the GJR residuals. Lastly, further studies can be conducted on other symmetric error distributions such as the Generalised Error Distribution or Generalised Pareto Distribution - models that are based on those distributions are also known as EVT models. Several authors (see McNeil (1999) and Omari et al. (2017)) have reported better backtesting performance of the models than the models considered in this paper. Hence, it will be insightful to investigate if the risk estimates from the EVT models exhibit the same level of uncertainty as the FHS, given that both are classified as semi-parametric.

8. Appendix

Width Difference between VaR and ES (m=1250)

Figure 8.1: Width Difference between 5% VaR and ES (m=1250)

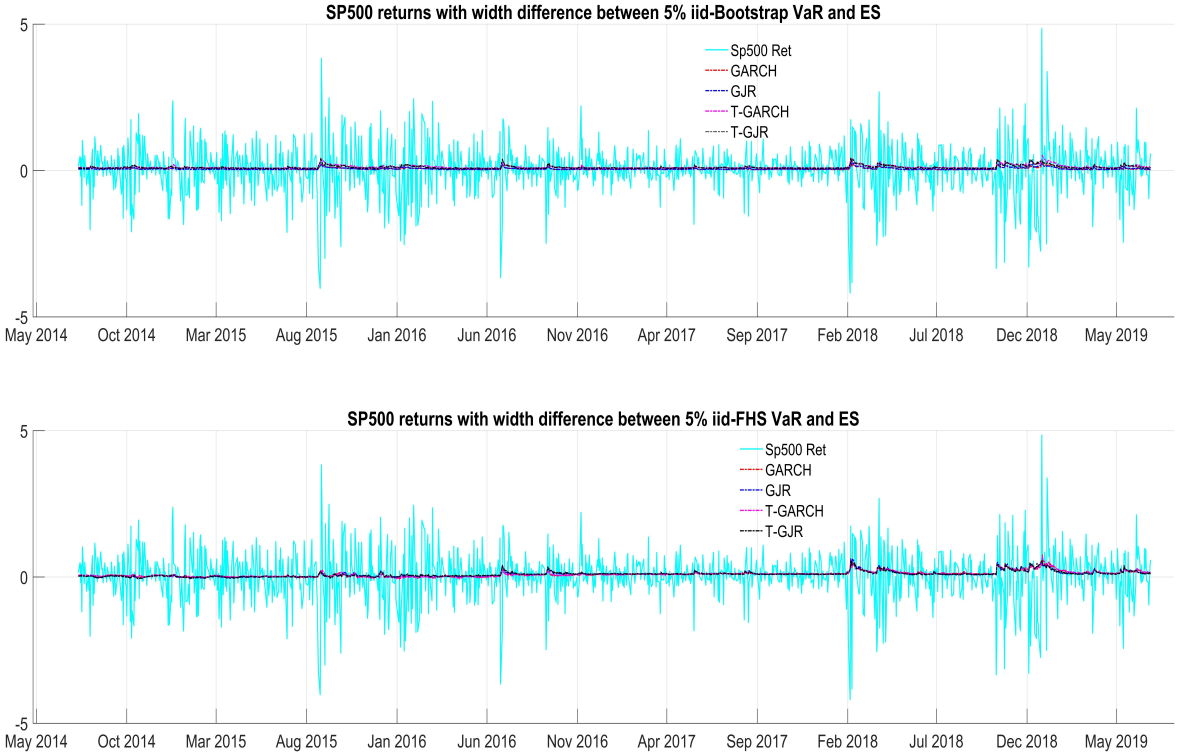
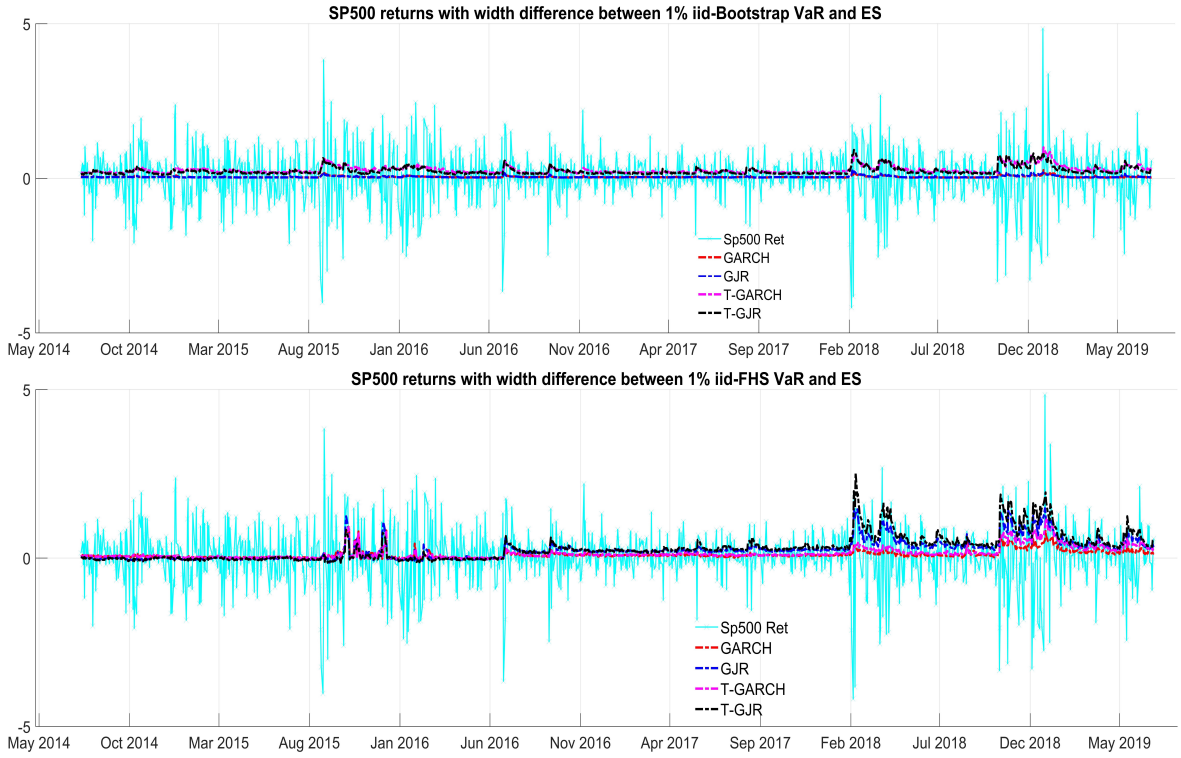


Figure 8.2: Width Difference between 1% VaR and ES (m=1250)


SP500 returns with width difference between 1% iid-Bootstrap VaR and ES
5
0
-5
May 2014
Oct 2014
Mar 2015
Aug 2015
Jan 2016
Jun 2016
Nov 2016
Apr 2017
Sep 2017
Feb 2018
Jul 2018
Dec 2018
May 2019
Sp500 Ret
GARCH
GJR
T-GARCH
T-GJR

SP500 returns with width difference between 1% iid-FHS VaR and ES
5
0
-5
May 2014
Oct 2014
Mar 2015
Aug 2015
Jan 2016
Jun 2016
Nov 2016
Apr 2017
Sep 2017
Feb 2018
Jul 2018
Dec 2018
May 2019
Sp500 Ret
GARCH
GJR
T-GARCH
T-GJR

Width Difference between VaR and ES (m=250)

Figure 8.3: Width Difference between 5% VaR and ES (m=250)

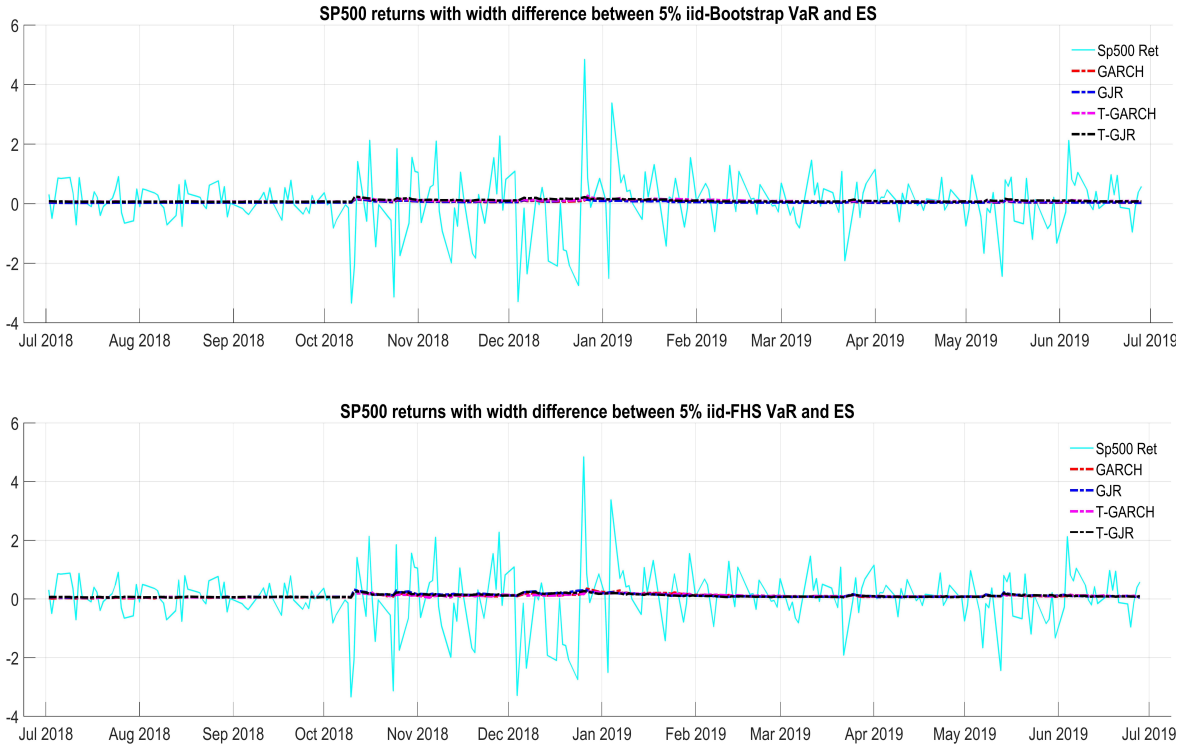


Figure 8.4: Width Difference between 1% VaR and ES (m=250)

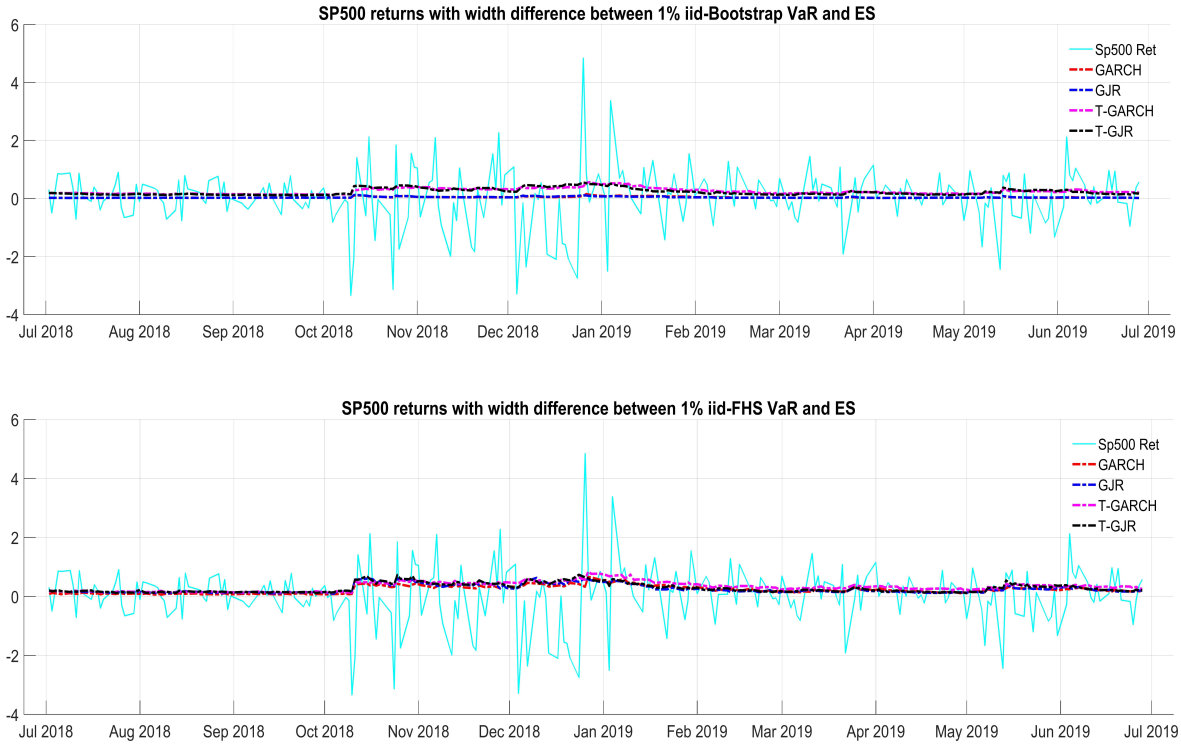
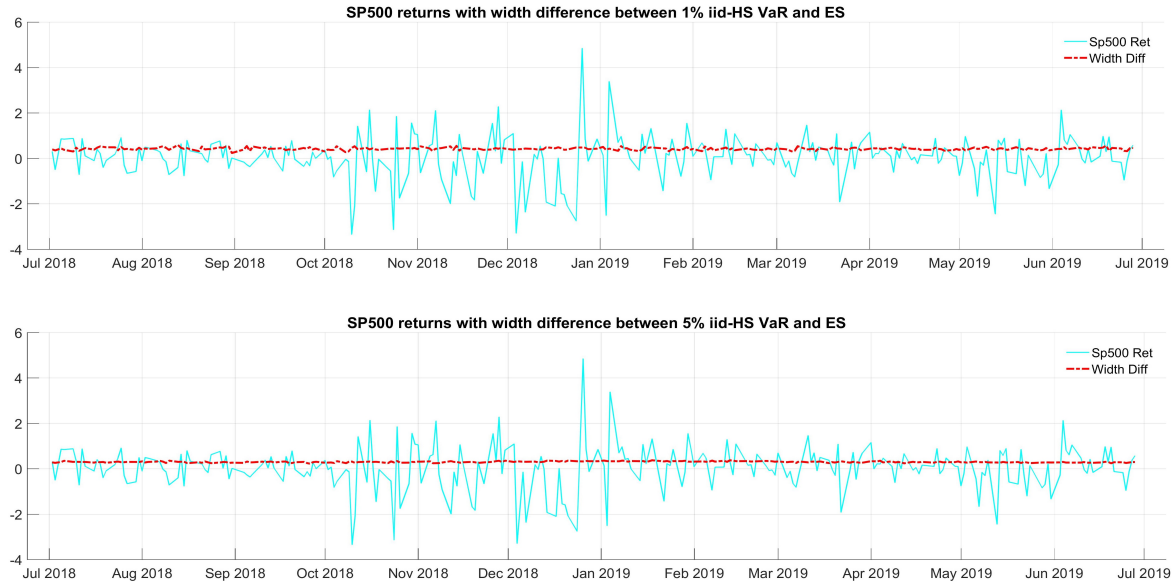
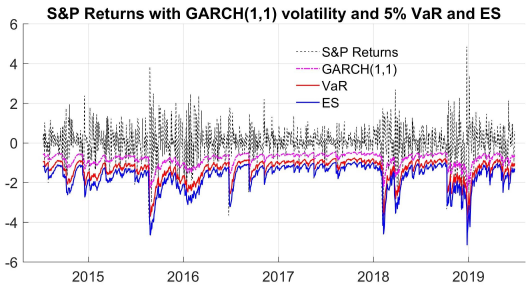


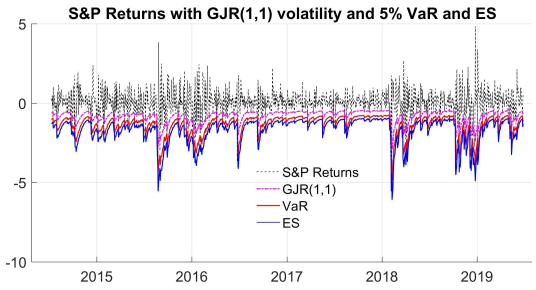
Figure 8.5: iid-HS Width Difference between VaR and ES (m=250)

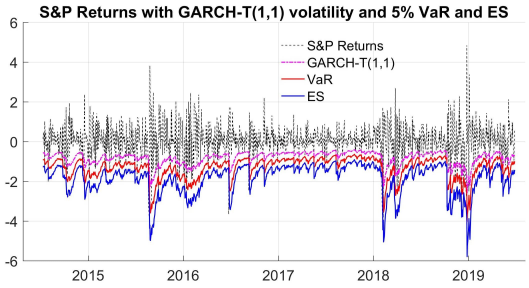


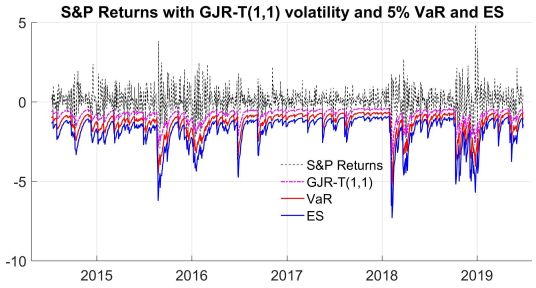
Risk Estimates plot without bootstrap (m=1250)

Parametric 95% Risk Estimates

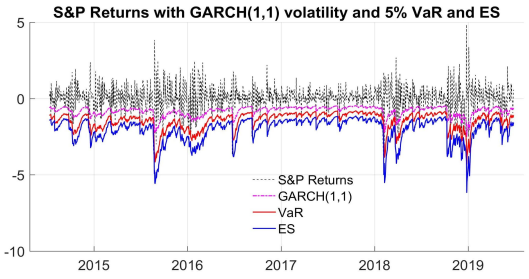

S&P Returns with GARCH(1,1) volatility and 5% VaR and ES
S&P Returns
GARCH(1,1)
VaR
ES
6
4
2
0
-2
-4
-6
2015
2016
2017
2018
2019

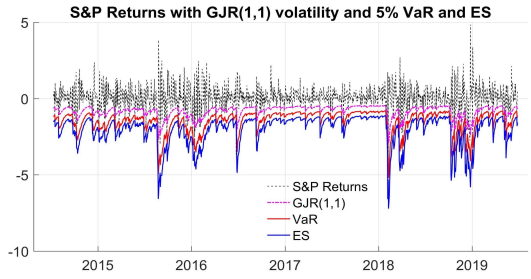

S&P Returns with GJR(1,1) volatility and 5% VaR and ES
S&P Returns
GJR(1,1)
VaR
ES
5
0
-5
-10
2015
2016
2017
2018
2019

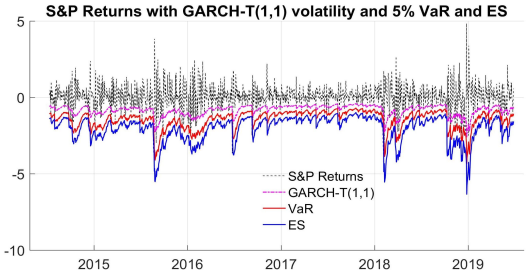

S&P Returns with GARCH-T(1,1) volatility and 5% VaR and ES
S&P Returns
GARCH-T(1,1)
VaR
ES
6
4
2
0
-2
-4
-6
2015
2016
2017
2018
2019

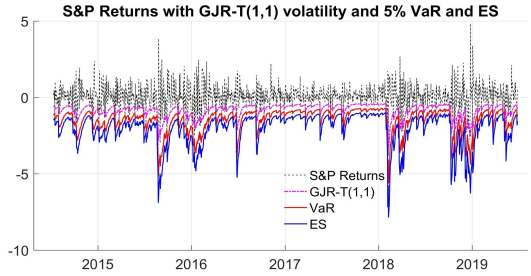

S&P Returns with GJR-T(1,1) volatility and 5% VaR and ES
S&P Returns
GJR-T(1,1)
VaR
ES
5
0
-5
-10
2015
2016
2017
2018
2019

FHS 95% Risk Estimates

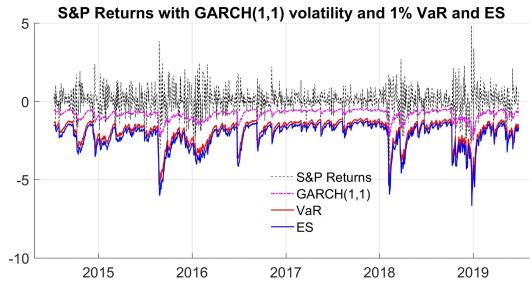


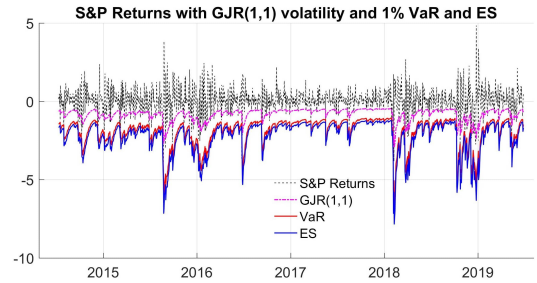


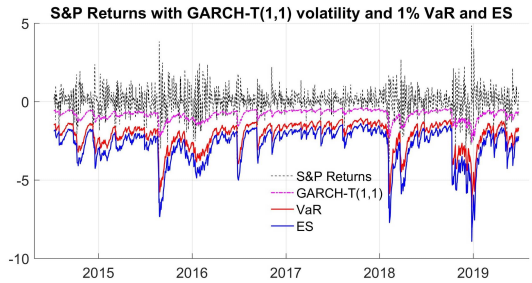


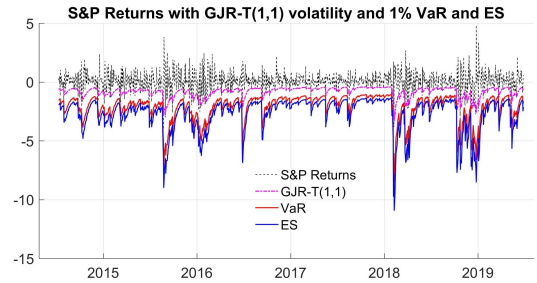


Parametric 99% Risk Estimates

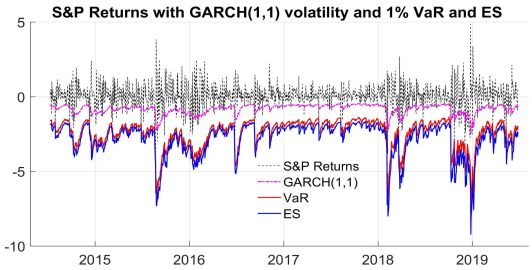


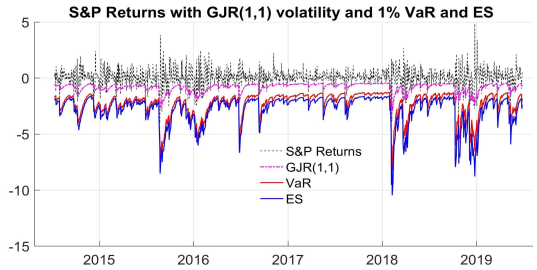


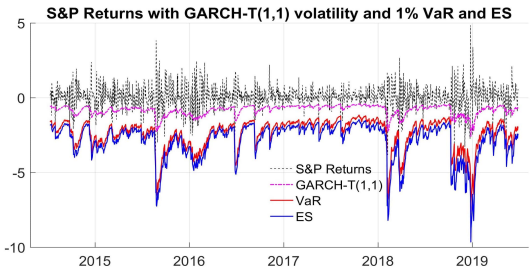


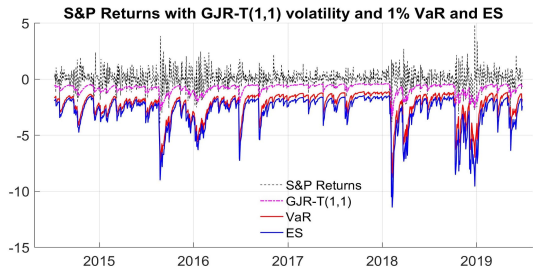


FHS 99% Risk Estimates



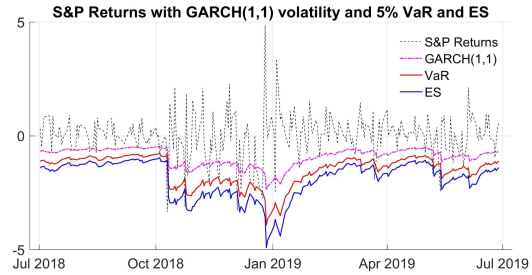


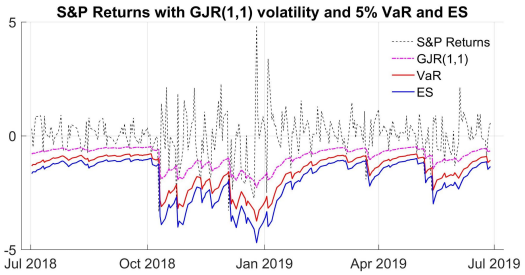


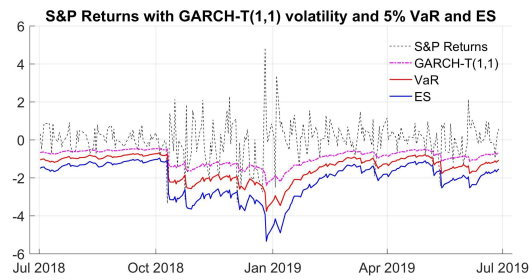


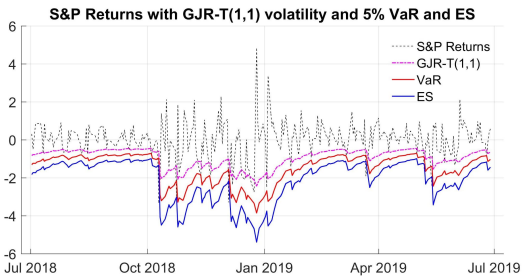
Risk Estimates plot without bootstrap (m=250)

Parametric 95% Risk Estimates

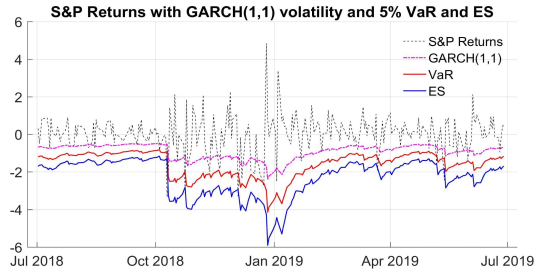


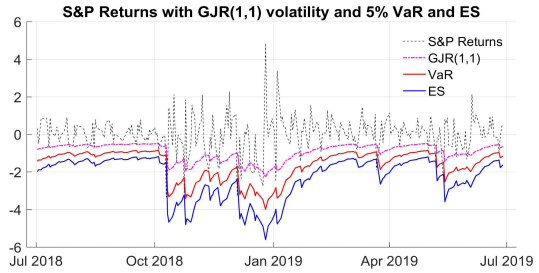


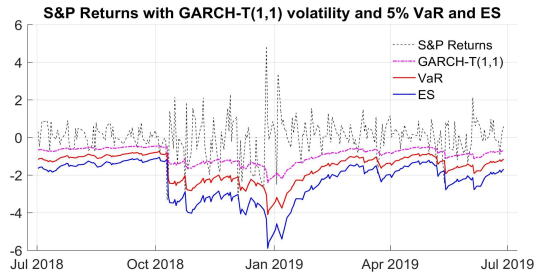


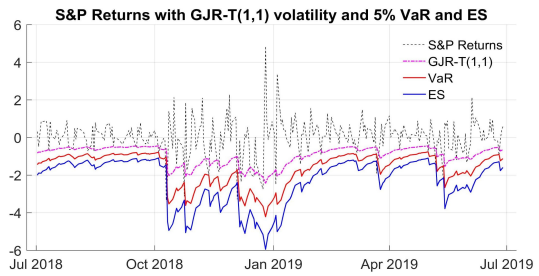


FHS 95% Risk Estimates

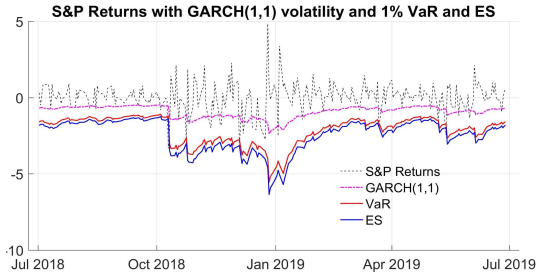


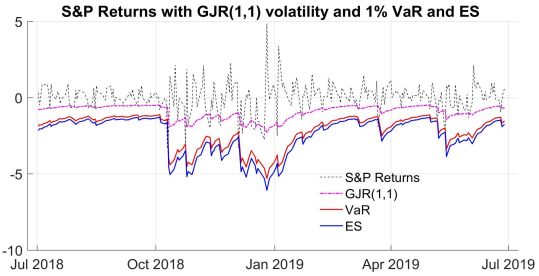


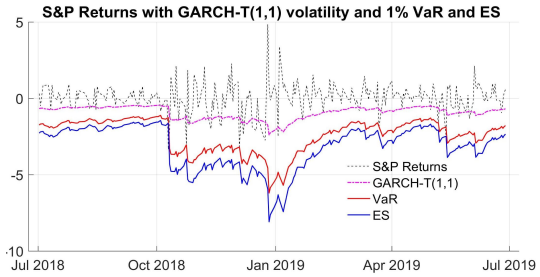


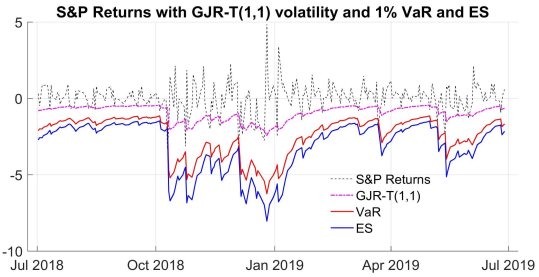


Parametric 99% Risk Estimates

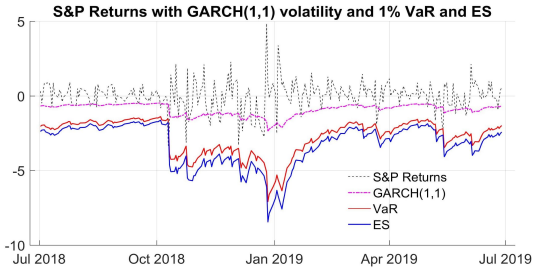


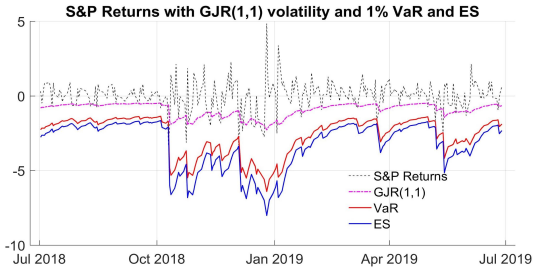


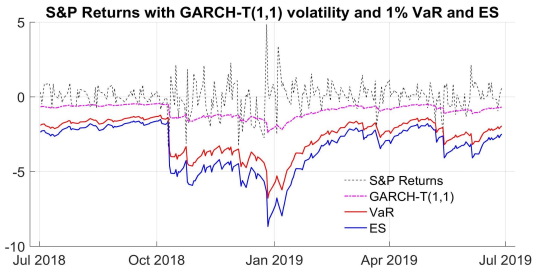


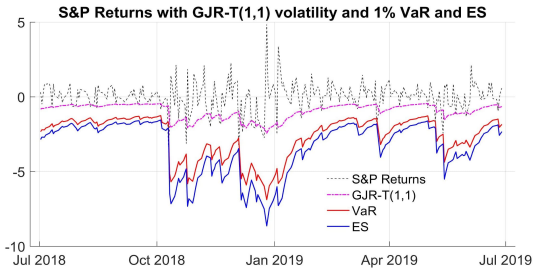


FHS 99% Risk Estimates



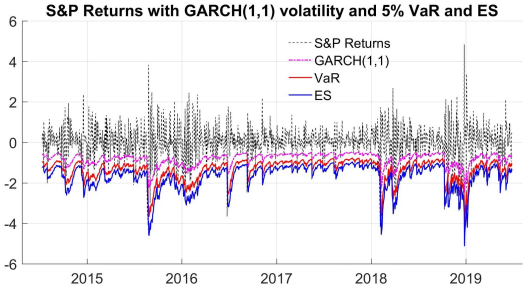


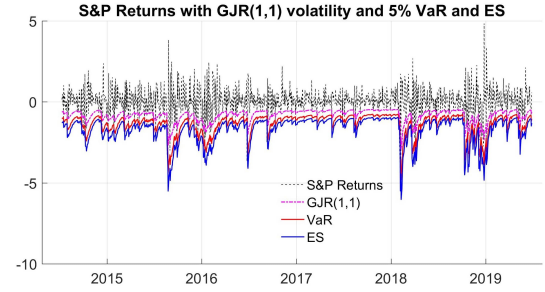


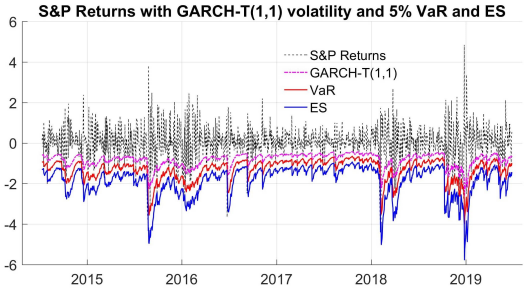


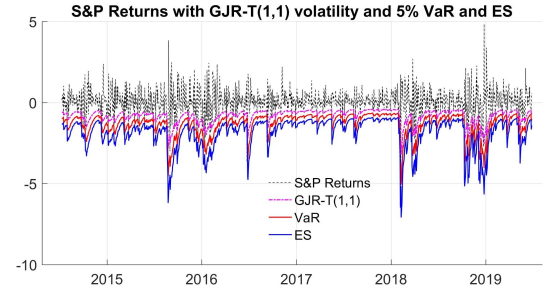
Bootstrap Risk Estimates plot(m=1250)

Parametric 95% Risk Estimates

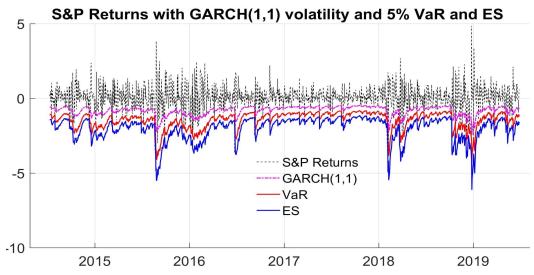


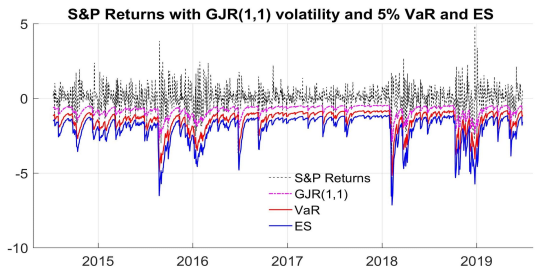


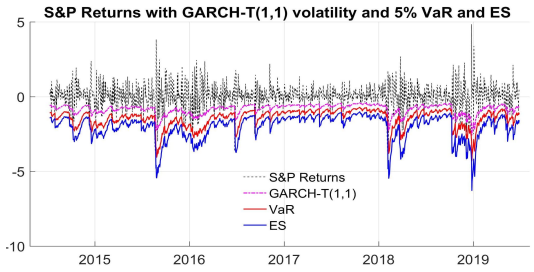


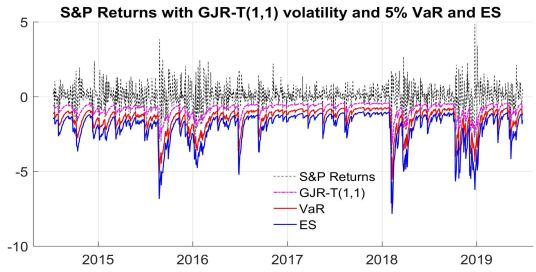


FHS 95% Risk Estimates

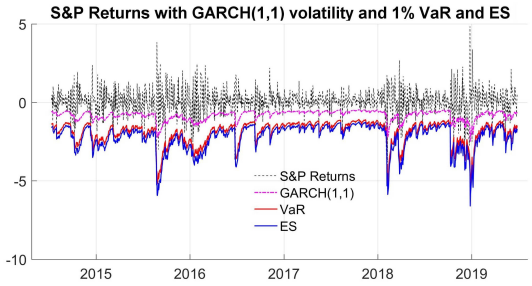


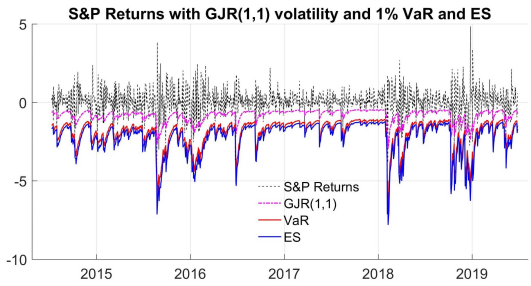


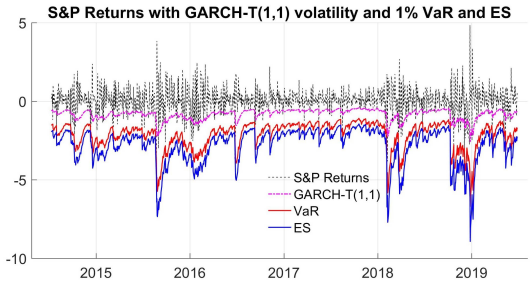


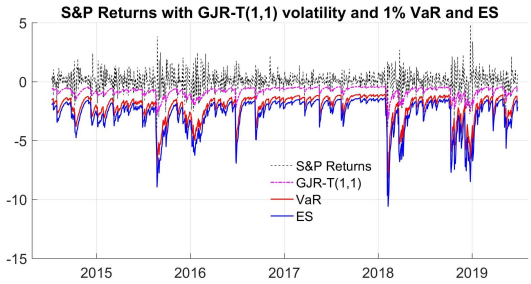


Parametric 99% Risk Estimates

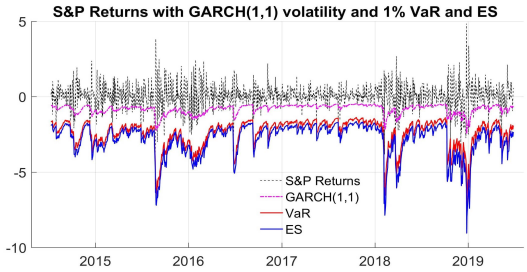


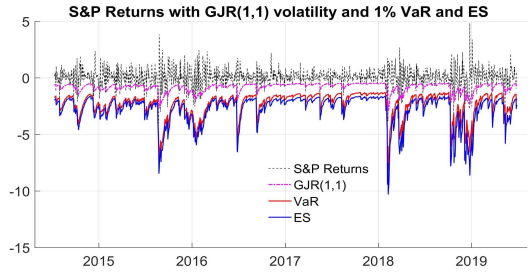


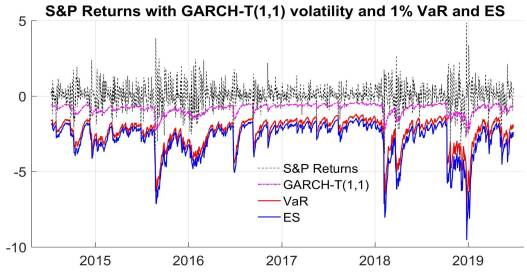


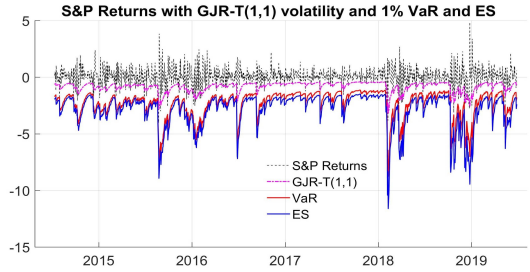


FHS 99% Risk Estimates



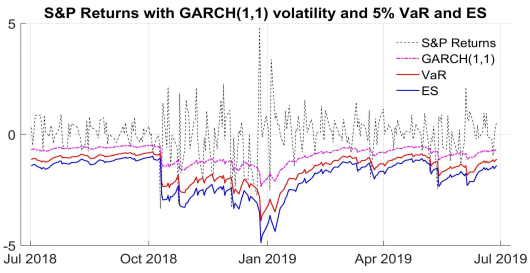


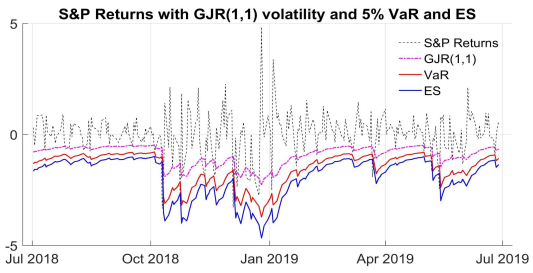


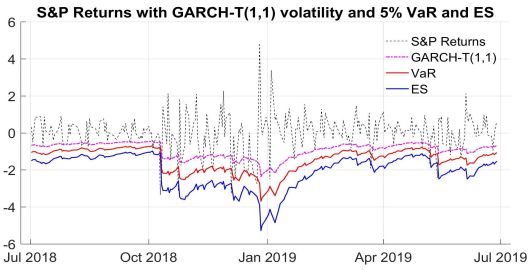


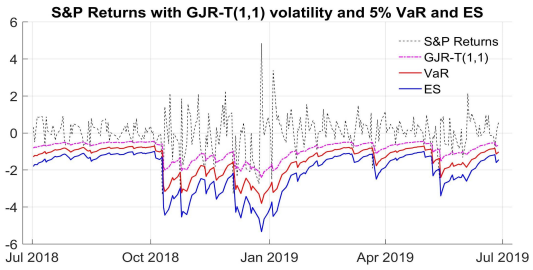
Bootstrap Risk Estimates plot (m=250)

Parametric 95% Risk Estimates

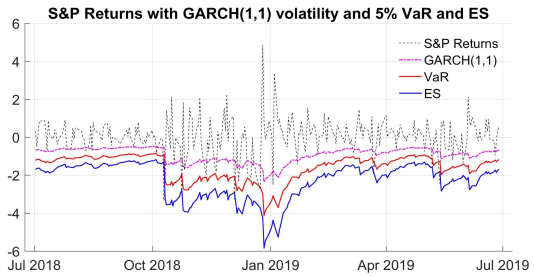


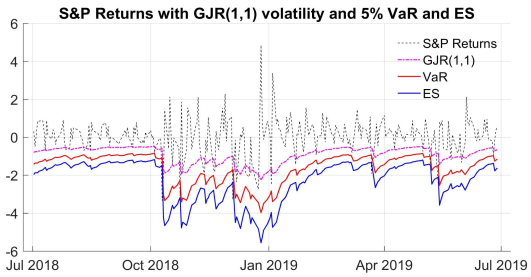


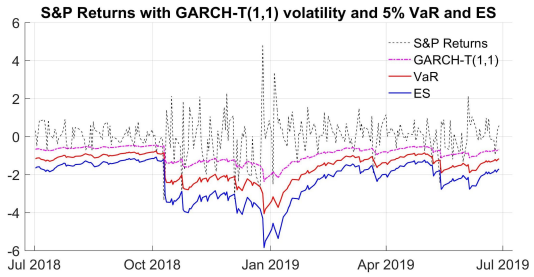


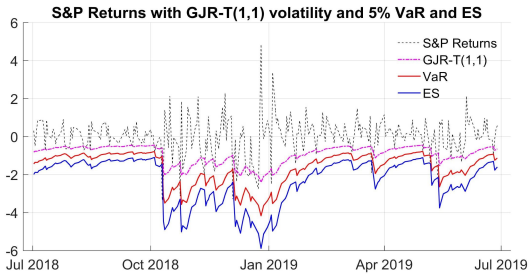


FHS 95% Risk Estimates

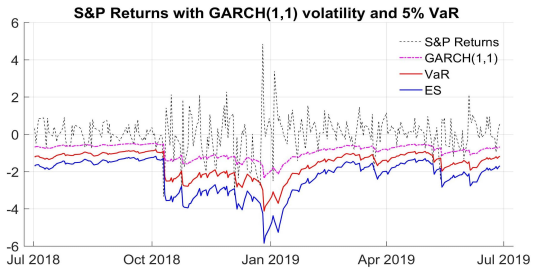


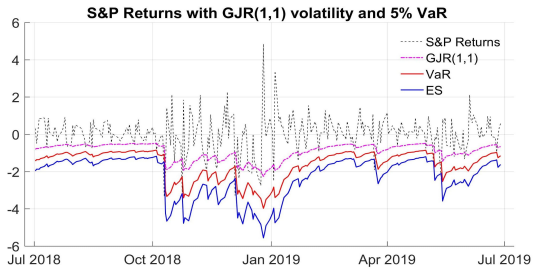


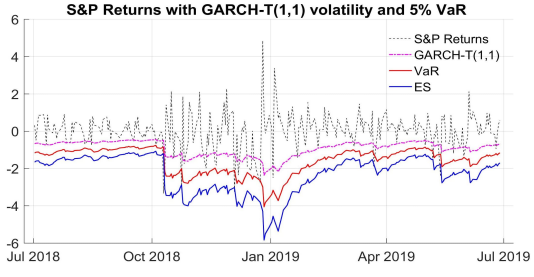


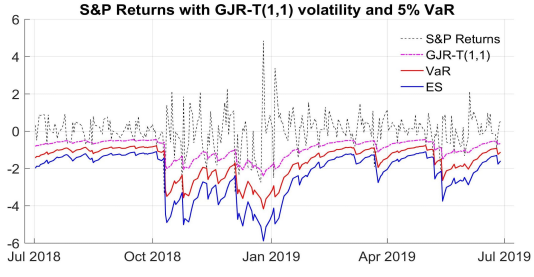


Parametric 99% Risk Estimates

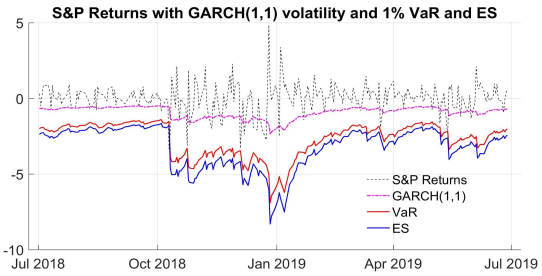


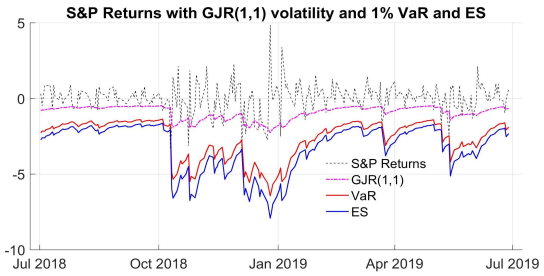


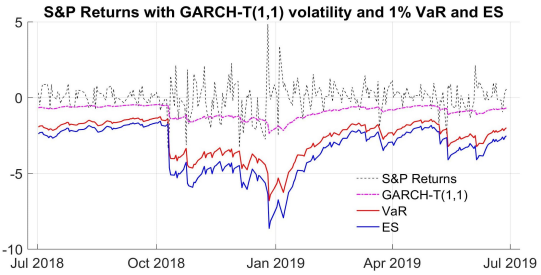


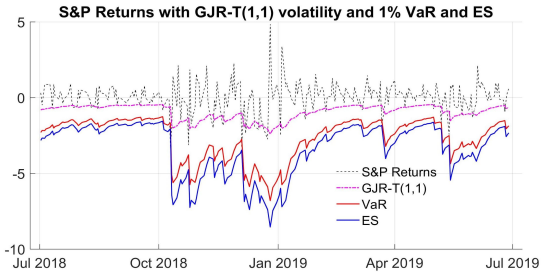


FHS 99% Risk Estimates



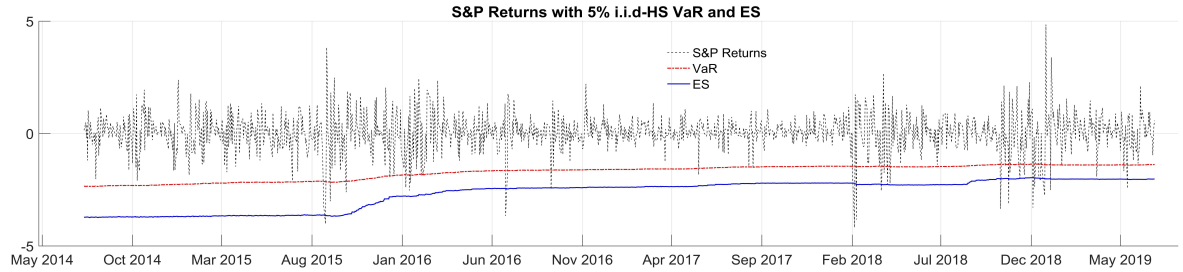


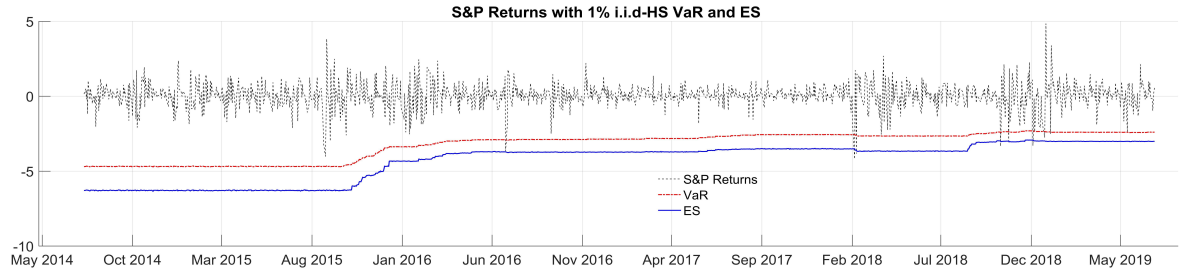




iid-HS Risk Estimates plot (m=1250)

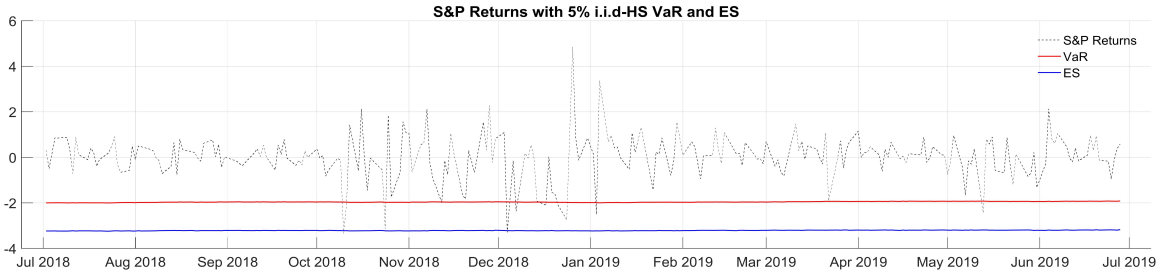
iid-HS Risk Estimates

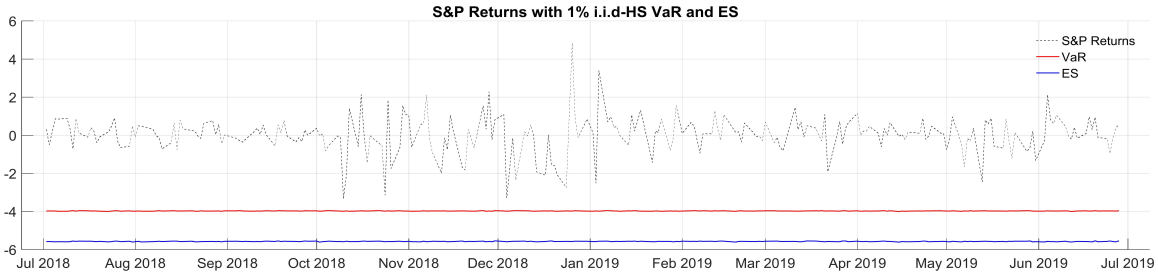




iid-HS Risk Estimates plot (m=250)

iid-HS Risk Estimates


S&P Returns with 5% i.i.d-HS VaR and ES
S&P Returns
VaR
ES
6
4
2
0
-2
-4
Jul 2018
Aug 2018
Sep 2018
Oct 2018
Nov 2018
Dec 2018
Jan 2019
Feb 2019
Mar 2019
Apr 2019
May 2019
Jun 2019
Jul 2019


S&P Returns with 1% i.i.d-HS VaR and ES
S&P Returns
VaR
ES
6
4
2
0
-2
-4
-6
Jul 2018
Aug 2018
Sep 2018
Oct 2018
Nov 2018
Dec 2018
Jan 2019
Feb 2019
Mar 2019
Apr 2019
May 2019
Jun 2019
Jul 2019

Bibliography

Adcock, C. J., Areal, N., and Oliveira, B. (2012). Value-at-risk forecasting ability of filtered historical simulation for non-normal garch returns. In *Midwest Finance Association 2013 Annual Meeting Paper*.

Andersen, T. G., Bollerslev, T., Christoffersen, P. F., and Diebold, F. X. (2005). Volatility forecasting. Technical report, National Bureau of Economic Research.

Angelidis, T., Benos, A., and Degiannakis, S. (2004). The use of garch models in var estimation. *Statistical methodology*, 1(1-2):105–128.

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

Asem, E. (2007). Misspecified likelihood function and value-at-risk. *The Journal of Risk*, 9(3):101.

Barone-Adesi, G., Giannopoulos, K., and Vosper, L. (1999). Var without correlations for portfolios of derivative securities. *Journal of Futures Markets*, 19(5):583–602.

Berkowitz, J., Christoffersen, P., and Pelletier, D. (2011). Evaluating value-at-risk models with desk-level data. *Management Science*, 57(12):2213–2227.

Black, F. (1976). Studies of stock market volatility changes. *1976 Proceedings of the American Statistical Association Bisiness and Economic Statistics Section*.

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

Bollerslev, T. and Wooldridge, J. M. (1992). Quasi-maximum likelihood estimation and inference in dynamic models with time-varying covariances. *Econometric reviews*, 11(2):143–172.

Brandolini, D. and Colucci, S. (2012). Backtesting value-at-risk: a comparison between filtered bootstrap and historical simulation. *The Journal of Risk Model Validation*, 6(4):3.

Brooks, C. and Persand, G. (2003). Volatility forecasting for risk management. *Journal of forecasting*, 22(1):1–22.

Brownless, C., Engle, R., and Kelly, B. (2011). A practical guide to forecasting through calm and storm. *The Journal of*.

Campbell, S. D. et al. (2005). A review of backtesting and backtesting procedures.

Cerović Smolović, J., Lipovina-Božović, M., and Vujošević, S. (2017). Garch models in value at risk estimation: empirical evidence from the montenegrin stock exchange. *Economic research-Ekonomska istraživanja*, 30(1):477–498.

Chan, N. H., Deng, S.-J., Peng, L., and Xia, Z. (2007). Interval estimation of value-at-risk based on garch models with heavy-tailed innovations. *Journal of Econometrics*, 137(2):556–576.

Christoffersen, P., Gonçalves, S., et al. (2004). *Estimation risk in financial risk management*. CIRANO.

Christoffersen, P. and Pelletier, D. (2004). Backtesting value-at-risk: A duration-based approach. *Journal of Financial Econometrics*, 2(1):84–108.

Christoffersen, P. F. (1998). Evaluating interval forecasts. *International economic review*, pages 841–862.

Danielsson, J. (2011). *Financial risk forecasting: the theory and practice of forecasting market risk with implementation in R and Matlab*, volume 588. John Wiley & Sons.

Danielsson, J., Hartmann, P., and de Vries, C. (1998). The cost of conservatism. *Risk*, 11(1):101–103.

Danielsson, J. and Zhou, C. (2016). Why risk is so hard to measure.

Dowd, K. (2000). Assessing var accuracy. *Derivatives Quarterly*, 6(3):61–63.

Engle, R. F. and Bollerslev, T. (1986). Modelling the persistence of conditional variances. *Econometric reviews*, 5(1):1–50.

Fama, E. F. (1963). Mandelbrot and the stable paretian hypothesis. *The journal of business*, 36(4):420–429.

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

Gao, F. and Song, F. (2008). Estimation risk in garch var and es estimates. *Econometric Theory*, 24(5):1404–1424.

Glosten, L. R., Jagannathan, R., and Runkle, D. E. (1993). 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.

Hall, P. and Yao, Q. (2003). Data tilting for time series. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):425–442.

Inui, K., Kijima, M., and Kitano, A. (2005). Var is subject to a significant positive bias. *Statistics & probability letters*, 72(4):299–311.

Jorion, P. (1996). Risk2: Measuring the risk in value at risk. *Financial analysts journal*, 52(6):47–56.

Jorion, P. et al. (2007). *Financial risk manager handbook*, volume 406. John Wiley & Sons.

Kuester, K., Mittnik, S., and Paolella, M. S. (2006). Value-at-risk prediction: A comparison of alternative strategies. *Journal of Financial Econometrics*, 4(1):53–89.

Kupiec, P. (1995). Techniques for verifying the accuracy of risk measurement models. *The J. of Derivatives*, 3(2).

Lopez, J. A. et al. (1999). Methods for evaluating value-at-risk estimates. *Economic review*, 2:3–17.

Manganelli, S. and Engle, R. F. (2001). Value at risk models in finance.

Markowitz, H. M. (1999). The early history of portfolio theory: 1600–1960. *Financial analysts journal*, 55(4):5–16.

McMillan, D., Speight, A., and Apgwilym, O. (2000). Forecasting uk stock market volatility. *Applied Financial Economics*, 10(4):435–448.

McNeil, A. J. (1999). Extreme value theory for risk managers. *Departement Mathematik ETH Zentrum*.

Miletic, M. and Miletic, S. (2015). Performance of value at risk models in the midst of the global financial crisis in selected cee emerging capital markets. *Economic research-Ekonomska istraživanja*, 28(1):132–166.

Nelson, D. B. (1991). Conditional heteroskedasticity in asset returns: A new approach. *Econometrica: Journal of the Econometric Society*, pages 347–370.

NG, H. R. and Lam, K.-P. (2006). How does sample size affect garch models? In *9th Joint International Conference on Information Sciences (JCIS-06)*. Atlantis Press.

Nieto, M. R. and Ruiz, E. (2010). Bootstrap prediction intervals for var and es in the context of garch models.

Omari, C. O. (2017). A comparative performance of conventional methods for estimating market risk using value at risk.

Omari, C. O., Mwita, P. N., and Waititu, A. G. (2017). Using conditional extreme value theory to estimate value-at-risk for daily currency exchange rates.

Pascual, L., Romo, J., and Ruiz, E. (2006). Bootstrap prediction for returns and volatilities in garch models. *Computational Statistics & Data Analysis*, 50(9):2293–2312.

Pérignon, C. and Smith, D. R. (2010). The level and quality of value-at-risk disclosure by commercial banks. *Journal of Banking & Finance*, 34(2):362–377.

Pritsker, M. (2006). The hidden dangers of historical simulation. *Journal of Banking & Finance*, 30(2):561–582.

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *The journal of finance*, 19(3):425–442.

Spierdijk, L. (2016). Confidence intervals for arma–garch value-at-risk: The case of heavy tails and skewness. *Computational Statistics & Data Analysis*, 100:545–559.

Treynor, J. L. (1961). Market value, time, and risk. *Time, and Risk (August 8, 1961).*

Tsay, R. S. (2014). Financial time series. *Wiley StatsRef: Statistics Reference Online*, pages 1–23.

Yamai, Y. and Yoshiba, T. (2005). Value-at-risk versus expected shortfall: A practical perspective. *Journal of Banking & Finance*, 29(4):997–1015.

Yamai, Y., Yoshiba, T., et al. (2002). Comparative analyses of expected shortfall and value-at-risk: their estimation error, decomposition, and optimization. *Monetary and economic studies*, 20(1):87–121.