THE GRADUATE UNIVERSITY FOR ADVANCED STUDIES, SOKENDAI

DOCTORAL THESIS

Quantitative Risk Management Using Extreme Value Theory

Author:
Hibiki KAIBUCHI

Supervisor:
Prof. Yoshinori KAWASAKI

A thesis submitted in fulfillment of the requirements
for the degree of Doctor of Philosophy

in the

School of Multidisciplinary Sciences
Department of Statistical Science
The Graduate University for Advanced Studies, SOKENDAI

September 2022

THE GRADUATE UNIVERSITY FOR ADVANCED STUDIES, SOKENDAI

Abstract

School of Multidisciplinary Sciences

Department of Statistical Science

Doctor of Philosophy

Quantitative Risk Management Using Extreme Value Theory

by Hibiki KAIBUCHI

This thesis considers about quantitative risk management using Extreme Value Theory (EVT). We specifically focus on the use of EVT to study extreme financial market risk, which is the risk of losses arising from movements in market prices, from a quantitative point of view. This is because quantitative risk management has now become a standard requirement for all financial institutions due to an increase in number of extreme market risk events, especially post 1980s. Such events include the Black Monday of 1987, the Dot-Com Bubble of 2000, the Global Financial Crisis of 2007-2008, and the recent COVID-19 recession of 2020. Extreme market events are rare but have high severity. The risk stemming from these extreme events is called tail risk, which contributes to the propagation of deep and unpredictable financial crises. Tail risk is clearly related to extreme events and hence the use of EVT is natural and effective.

The standards of quantitative risk management are laid down by Basel Committee on Banking Supervision (BCBS). Financial institutions are asked to estimate specific risk measures so that they can protect themselves against future extreme market catastrophes. Risk measures can be understood as providing a risk assessment in the form of capital amount that are set aside to absorb unexpected future losses. Recently, the BCBS announced a change in the risk measure used for capital requirements in internal market risk models, moving from the Value-at-Risk (VaR) to the Expected Shortfall (ES). VaR is defined as a measure of the potential losses on a portfolio of financial instruments resulting from market movements over a given time horizon and for a probability level. Similarly, ES is a measure of the mean of the losses exceeding VaR at a given probability level. The amendment is driven by the fact that VaR could not predict or cover the extreme losses during the turbulence of 2007-2008 crisis and mathematically does not satisfy the important coherence property.

It is no surprise that the switching from VaR to ES has generated many reactions from both the practical sector and the academic sector as evidenced by the numerous literatures. The backtesting approach established by the BCBS, which tests the accuracy of ES estimates, is causing the problem. More specifically, financial institutions now face the paradox of using ES for computing their market risk capital requirements and using VaR for backtesting ES. For this reason, both estimation and

backtesting of VaR are still important nowadays because sensible ES estimates are based on correctly specified VaR estimates by the definition of ES. This was the motivation for the proposal of a two-step bias-reduced conditional EVT approach called GARCH-UGH for the estimation of one-step ahead dynamic extreme VaR. At the same time, there has not been sufficient investigation to establish the superiority of a certain estimator of ES relative to the others in the literature and no particular type of ES model is prescribed in the framework of the BCBS. We thus considered the estimation of dynamic extreme ES based on our proposed GARCH-UGH approach and the use of the first-order asymptotic equivalence between VaR and ES. Moreover, we also tackled an urgent problem of which ES backtesting methods can be used in practice as we can expect that upcoming regulations will require financial institutions to backtest ES without using VaR backtesting methods.

We tackle the question of estimating the VaR of loss return distribution at extreme levels, which is an important question in financial applications, both from operational and regulatory perspectives. In particular, the dynamic estimation of extreme VaR given the recent past has received substantial attention because the occurrence of extreme financial events has increased since 1980s. Moreover, an accurate estimation of VaR is still essential in practice even if the BCBS changed the risk measure for the calculation of capital requirements from VaR to ES. This is because sensible estimation of ES is based on correctly specified VaR estimates. We propose here a new two-step bias-reduced estimation methodology for the estimation of one-step ahead dynamic extreme VaR, called GARCH-UGH (Unbiased Gomes-de Haan), whereby financial returns are first filtered using an AR-GARCH model, and then a bias-reduced estimator of extreme quantiles is applied to the standardized residuals. We analyze the performance of our approach on four financial time series, which are the Dow Jones, NASDAQ and Nikkei stock indices, and the Japanese Yen/British Pound exchange rate. Our results indicate that the GARCH-UGH estimates of the dynamic extreme VaR are more accurate than those obtained either by historical simulation, conventional AR-GARCH filtering with Gaussian or Student-t innovations, or AR-GARCH filtering with standard extreme value estimates, both from the perspective of in-sample and out-of-sample traditional VaR backtestings, which are the unconditional and conditional coverage tests. The numerical results

of comparative VaR backtesting, which is based on the Diebold-Mariano test, also support the use of the GARCH-UGH approach by yielding definitive answers to the cases when GARCH-UGH and GARCH-EVT approaches are either all accepted, or all rejected in the traditional VaR backtestings. In addition, our bias-reduction procedure will be designed to be robust to departure from the independence assumption, and as such will be able to handle residual dependence present after filtering in the first step. Our finite-sample results also illustrate that the GARCH-UGH method leads to one-step ahead extreme conditional VaR estimates that are less sensitive to the choice of sample fraction, and hence mitigates the difficulty in selecting the optimal number of observations for the estimations. Finally, the computational cost of GARCH-UGH is lower than that of conventional GARCH-EVT: the extreme value step in the GARCH-UGH method is semiparametric with an automatic and fast recipe for the estimations of the one-step ahead extreme conditional VaR, while the competing GARCH-EVT method is based on a parametric fit of the Generalized Pareto Distribution to the residuals using Maximum Likelihood Estimation.

We also extend the GARCH-UGH approach used in dynamic extreme VaR estimation to the dynamic extreme ES estimation by means of the asymptotic equivalence between quantile (VaR) and ES. This is motivated by the fact that there has not been sufficient investigation to establish the superiority of a certain estimator of ES relative to the others in the literature and no particular type of ES model is prescribed in the framework of the BCBS. Our results show that the GARCH-UGH approach produces more accurate ES estimates than those obtained by basic estimation methods, both from the perspective of traditional and comparative ES backtestings. We use the exceedance residual test, the conditional calibration test and the expected shortfall regression test for traditional backtestings, and Diebold-Mariano test again based on the joint elicitability of VaR and ES for comparative backtesting. When compared to other EVT-type methods, comparative backtestings with chosen two scoring functions result in a good agreement with the GARCH-UGH approach being the best estimator of ES, while traditional backtestings are not always in line with the superiority of our proposed approach.

In contrast to the estimation of dynamic extreme ES where most of the existing

methods including the ones we referred and proposed for VaR estimation can easily be adapted to the ES, such adaptions are not straight-forward for backtesting ES estimates. Based on the strict definition of backtesting, we understand that a backtesting for specific risk measure should only require its estimates and realized returns as input variables. In contrast to the VaR, fulfilling this definition for ES is very difficult task because ES is strongly related to the VaR through its definition and joint elicitability. As in every statistical method, each of different ES backtesting methods has its strengths and weaknesses. We thus strongly suggest adopting a two-stage backtesting framework, i.e., the use of both traditional and comparative backtestings for risk measures that will enhance the regulatory framework for financial institutions by providing the correct incentives for accuracy of risk measure estimates. More precisely, the comparative backtesting methods can be used by financial institutions internally to select better performing methods among competing alternatives when traditional backtestings methods do not yield definitive answers as competitive methods are all accepted, or all rejected. Supplementing with comparative backtestings is essential, and hence can adequately quantify the risks even though they still have some drawbacks to consider for the practical use, e.g., there exists no optimal scoring function with any theoretical guarantee. We think that the major challenge of the regulations of BCBS in the implementation of the ES as a risk measure for market risk is the unavailability of simple tools for its evaluation. We also believe that the findings of the estimation and backtesting of risk measures for tail risks in volatile financial market given in Chapter 3 and 4 would be useful for developing regulatory framework of the BCBS and monetary policies aimed at mitigating tail risks.

Acknowledgements

First of all, I would like to express my great gratitude to my supervisor, Dr. Yoshinori Kawasaki. Even though Extreme Value Theory, which is my man interest, was not his specialty, he accepted me as a PhD student and supervised all through my research during my PhD course. I could not finish the work without his considerable encouragement and endless support. I am grateful to Dr. Takaaki Shimura, who is my second supervisor, for teaching me the probabilistic aspects of Extreme Value Theory and giving me an opportunity to join the Cooperative Research Symposium "Extreme Value Theory and Applications" at the Institute of Statistical Mathematics. I would also express my special thanks to Dr. Gilles Stupfler, ENSAI and CREST, who provided me with the solid foundation not only in Extreme Value Theory but also as a researcher and also contributed as a co-author of the work in Chapter 3 (GARCH-UGH: a bias-reduced approach for dynamic extreme Value-at-Risk estimation in financial time series). He was my supervisor during my Master's course at the University of Nottingham and since then he supervised me informally.

Furthermore, I am grateful to the member of the dissertation committee, Dr. Hideatsu Tsukahara, Dr. Yoshiyuki Ninomiya, Dr. Shogo Kato and Dr. Takuma Yoshida for giving me helpful advice and judging my PhD degree. Finally, I would like to thank my family for their understanding, and sincere and continuous support.

Contents

List of Figures

List of Tables

List of Abbreviations

ADF	**A**ugmented **D**ickey-**F**uller
AR	**A**uto**R**egressive
BCBS	**B**asel **C**ommittee on **B**anking **S**upervision
CC	**C**onditional **C**alibration
D-M	**D**iebold-**M**ariano
ER	**E**xceedance **R**esidual
ES	**E**xpected **S**hortfall
ESR	**E**xpected **S**hortfall **R**egression
EVI	**E**xtreme **V**alue **I**ndex
EVT	**E**xtreme **V**alue **T**heory
GARCH	**G**eneralized **A**uto**R**egressive **C**onditional **H**eteroskedasticity
GEV	**G**eneralized **E**xtreme **V**alue
GPD	**G**eneralized **P**areto **D**istribution
HAC	**H**eteroscedasticity and **A**utocorrelation **C**onsistent
HS	**H**istorical **S**imulation
J-B	**J**arque-**Bera**
LR	**L**ikelihood **R**atio
MS	**M**edian **S**hortfall
POF	**P**roportion **O**f **F**ailures
POT	**P**eaks-**O**ver-**T**hreshold
QMLE	**Q**uasi-**M**aximum **L**ikelihood **E**stimation
QRM	**Q**uantitative **R**isk **M**anagement
RVaR	**R**ange **V**alue-**a**t-**R**isk
UGH	**U**nbiased **G**omes-de**H**aan
VaR	**V**alue-**a**t-**R**isk

I would like to dedicate this thesis to my family and friends for their endless and thoughtful support and encouragement.

Chapter 1

General Introduction

1.1 Quantitative risk management

In this thesis the focus is on the use of Extreme Value Theory (EVT) to study extreme (financial) market risk, which is the risk of losses arising from movements in market prices, from a quantitative point of view. This is because quantitative risk management has now become a standard requirement for all financial institutions due to an increase in number of extreme market risk events, especially post 1980s. Extreme market events are rare but have high severity. The risk stemming from these extreme events is called tail risk, which contributes to the propagation of deep and unpredictable financial crises. Tail risk is clearly related to extreme events, and hence the estimation of risk measure heavily relies on accurate estimation of a tail of the underlying distribution. Let us consider the model of risk measure called Value-at-Risk (VaR) based on an assumption of Gaussian data distribution as an example. These models ignore the tail risk and therefore tend to underestimate it at volatile period and overestimate at normal period.

The use of EVT is urgently necessary in risk management, particularly in relation to tail risks, as the occurrence of extreme financial market events are increasingly frequent in the post-1980s. Such events include the Black Monday of 1987, the Dot-Com Bubble of 2000, the Global Financial Crisis of 2007-2008, and the recent COVID-19 recession of 2020. A non-exhaustive chronology of extreme market events, i.e. financial crises, is given in Table A1 of Chakraborty et al. (2021). Prior to the Global Financial Crisis began in mid-August 2007, the potential for sudden extreme fluctuations in volatility of financial market variables had been severely underestimated

and the prevalence of tail risks was largely neglected. Orlowski (2012) states that it was a major contributing factor to the unprecedented scale of this crisis. Regarding the global pandemic of COVID-19, Zhang et al. (2020) find that the rapid spread of COVID-19 has created an unprecedented level of rise in risk, causing financial institutions to suffer against massive losses in a very short period of time. They also report that the standard deviation of daily returns of S&P500 was increased from 0.0069 to 0.0268 within one month at the outbreak of COVID-19. Abuzayed et al. (2021) study the extreme risk spillover among global financial markets and their result show that there exists a high degree of integration in the extreme risk of the stock market system during the COVID-19 outbreak. Li et al. (2022) use the risk measures called VaR and find that COVID-19 indeed increased the risk exposure of equity markets, with the US showing the greatest average impact among other countries. Therefore, the measurement of extreme market risk is essential for financial institutions. At the same time, we have to admit that for measuring extreme risk the tail part of a underlying loss distribution is difficult to estimate, and hence involves substantial model uncertainty.

The standards of quantitative risk management are laid down by Basel Committee on Banking Supervision (BCBS). Financial institutions are asked to estimate specific risk measures so that they can protect themselves against future extreme market catastrophes. Risk measures can be understood as providing a risk assessment in the form of capital amount that are set aside to absorb unexpected future losses. A review of extreme market risk measures (for example, Sharma 2012; Chakraborty et al. 2021; He et al. 2022) reveal that risk measures have improved over the past three decades from the naive standard deviation of price returns to the recent VaR and Expected Shortfall (ES). These two risk measures are the heart of this thesis (appear everywhere especially in Chapter 3 and 4), which are main tools for quantifying the riskiness that is implied by the variability of losses and the tails of their distribution.

In practice there was a lively debate of which risk measure would be best in regulatory framework over the last or two decades. The debate mainly focused on aforementioned VaR and ES. Recently, the BCBS announced a change in the risk measure used for capital requirements in internal market risk models, moving from the VaR to the ES. In other words, the quantitative standards made under Basel II (Basel

Committee on Banking Supervision 2009) was amended by Basel III (Basel Committee on Banking Supervision 2019). To check what has been changed, see for example Sharma (2012). The amendment is driven by the fact that VaR could not predict or cover the losses during the turbulence of 2007-2008 crisis, revealing its major drawback unfortunately. Moreover, VaR mathematically does not satisfy the important property called coherence given in Artzner et al. (1999), while ES does (discussed in Subsection 1.2.2). The purpose of Basel Accord III is to cover the shortcomings that regulations failed to capture during 2007-2008 crisis.

It is no surprise that the switching from VaR to ES has generated many reactions from both the practical sector and the academic sector as evidenced by the numerous literature of ES given in Section 4.2. The backtesting approach established by Basel Committee on Banking Supervision (2019), which tests the accuracy of ES estimates, is causing the problem. More specifically, financial institutions now face the paradox of using ES for computing their market risk capital requirements and using VaR for backtesting ES. For this reason, both estimation and backtesting of VaR are still important nowadays because sensible ES estimates are based on correctly specified VaR estimates. This was the motivation for the proposal of a bias-reduced conditional EVT approach called GARCH-UGH in Chapter 3. Finally, an introduction of ES brought some new challenges regarding the backtesting and also the choice of estimation methods, discussed in Chapter 4.

1.2 Basic concepts in risk management

In this section, we briefly describe a collection of empirical observations of financial data, properties of risk measures, VaR, ES and other alternative risk measures exist in the literature. These materials are fundamental concepts used in the quantitative risk management.

In this thesis, we denote asset prices, e.g., a stock, index, and exchange rate, by p_t where t refers to a day, but can indicate any frequency, e.g., hourly, weekly, monthly and yearly.The financial data we consider are the negative daily log-returns and expressed by $X_t = -\log(p_t/p_{t-1})$ in Subsection 3.2.1. The log-returns are also

defined as continuously compounded returns in Danielsson (2011). They play an important role in the background of many financial calculations.

1.2.1 The stylized facts of financial time series

The stylized facts of financial time series are gathered from empirical observations and inference drawn from these observations. See Cont (2001), Danielsson (2011) and McNeil et al. (2015) for a comprehensive study. They apply to most, if not all, financial time series including the daily negative log-returns on indices and exchange rates. For a single time series of financial returns, some stylized facts are as follows: heavy tails, absence of autocorrelations and volatility.

Heavy tails - empirical data show that financial returns exhibit very large (i.e. extreme) positive or negative returns, which are very unlikely to be observed if a normal distribution is implied. Financial returns are hence leptokurtic. The normal distribution also cannot capture the skewness of financial returns. The formal definition of heavy tails can be found in Section 2.2 and Subsection 3.2.3 in terms of the regular variation and tail quantile function, respectively.

Absence of autocorrelations - autocorrelations of financial returns are generally insignificant but they exhibit strong serial dependence, especially in their second moment (Ergen 2015).

Volatility - volatility, the standard deviation of returns, appears to vary over time. Financial returns exhibit a phenomenon known as volatility clustering. It is the tendency for extreme returns to be followed by other extreme returns together, so that we observe many days of high volatility followed by many days of low volatility. Danielsson and de Vries (1997) state that the changing volatility is often absent from monthly financial returns, but the property of heavy tails does not fade.

1.2.2 Risk measure and its properties

There is no universal definition of risk and its measure. It is natural to measure risk in terms of probability distributions and useful to represent the risk of an asset as a single number, which is comparable with other assets. A risk measure η is defined on some spaces of random variables (r.v.s). In this thesis, we only consider risk

measures that are law-invariant; that is, two real-valued r.v.s. X_1 and X_2 satisfy

$$P(X_1 \leq x) = P(X_2 \leq x), \quad x \in \mathbb{R} \Rightarrow \eta(X_1) = \eta(X_2).$$

Therefore, risk measures are mappings from spaces of probability distributions of r.v.s to real numbers. The desirable properties of risk measures as follows: coherence, robustness, elicitability and backtestability. We now explain them briefly in order.

Coherence - Artzner et al. (1999) propose that a risk measure η is coherent if it satisfies the following axioms that are monotonicity, homogeneity, translation invariance and subadditivity.

- η is monotonic if for all X_1 and X_2 it holds that $X_1 \leq X_2 \Rightarrow \eta(X_1) \leq \eta(X_2)$.
- η is homogenous if for all X and $c \geq 0$ it holds that $\eta(cX) = c\eta(X)$.
- η is translation invariant if for all X and $c \in \mathbb{R}$ it holds that $\eta(X + c) = \eta(X) + c$ (see Axiom 2.19 in McNeil et al. 2015, p.73).
- η is subadditive if for all X_1 and X_2 it holds that $\eta(X_1 + X_2) \leq \eta(X_1) + \eta(X_2)$.

The other commonly used axioms for risk measures in the literature are given in He et al. (2022). The subadditivity axiom fails to hold for VaR in general, so VaR is not a coherent risk measure. The lack of subadditivity contradicts the notion that there should be a diversification merit associated with merging portfolios. In other words, the total risk on a portfolio should not be greater than the sum of the risks of the consistent parts of the portfolio (Artzner et al. 1999). On the other hand, ES is a coherent risk measure. The controversies about whether VaR really violates the subadditivity axiom or not are discussed in for example Danielsson (2011) and He et al. (2022). They find that VaR is subadditive in the special case when financial returns are normally distributed and only violates when the tails are extremely heavy.

Robustness - He et al. (2022) state that a risk measure η is robust if it can accommodate model misspecification and has statistical robustness regarding the changes in data. The formal definition of robustness in the sense of the Wasserstein distance is given in Emmer et al. (2015) and other definition is given in Cont et al. (2010), which relates more to the sensitivity to outliers in the data sample than to mere

measurement errors. Ergen (2015) tests the robust performance of the model of the conditional VaR estimation called GARCH-EVT (McNeil and Frey 2000), which is the model we will discuss in Chapter 3 and 4.

Elicitability - Gneiting (2011) states that a risk measure η, i.e. a statistical functional, is elicitable if it admits a strictly consistent scoring, i.e. loss, function. A scoring function is strictly consistent for a risk measure η if the risk measure can be obtained by minimizing the expected value of the score. The formal and mathematical definition of elicitability is presented in Subsection 4.5.1. Elicitability is a helpful decision-theoretic framework for the determination of optimal point forecasts. It can be used to compare in a natural way (yet not the only way) the performance of different estimation methods of risk measures VaR and ES. Hence, elicitability is crucial for what is called the comparative backtesting, which will be explained in Section 3.5 and 4.5.

Some commonly used scoring functions are given in Emmer et al. (2015). The mean functional is elicited by the squared error, an alternative risk measure called expectile (see Bellini and Di Bernardino 2015 and Ziegel 2016) is elicited by the weighted squared error, the median is elicited by the absolute error and the VaR, i.e. quantile, is elicited by the weighted absolute error. However, Gneiting (2011) has pointed out that ES is not elicitable. This had a big influence and initiated active new researches on the feasibility of backtesting ES (see Chapter 4 for a detailed discussion).

Backtestability - There was a myth that a lack of elicitability means the risk measure, for example ES, cannot be backtested. The highly influential papers Acerbi and Szekely (2014) and Emmer et al. (2015) have argued that elicitability is not relevant for backtesting risk measures but rather for comparing the performance of different estimation methods. We thus think that a lack of elicitability is not a hurdle to backtest ES. Indeed, ES cannot be backtested through any scoring functions but there is no reason why it could not be done using another methods that do not exploit the eliciability or use the idea called joint elicitability of VaR and ES (see again Chapter 4 for a detailed discussion).

1.2.3 Value-at-Risk (VaR)

In this thesis, we consider that the generic financial position X is a real-valued random variable, that is the random profit if $X > 0$ or loss if $X < 0$. As mentioned previously, we normally focus on the daily negative log-returns.

The most widely known risk measure is Value-at-Risk (VaR), first devised by J.P. Morgan on the aftermath of the Black Monday of 1987. VaR is defined as a measure of the potential loss on a portfolio of financial instruments resulting from market movements over a given time horizon and for a given distribution of historical returns and probability level, i.e. confidence level. VaR is merely a quantile of the loss distribution. Mathematically, for a probability level $\tau \in (0,1)$, generally τ tends to 1, the τth quantile (VaR) of a distribution F is

$$\text{VaR}_\tau(X) = q_\tau = \inf\{x \in \mathbb{R} : F(x) \geq \tau\}.$$

Based on above definition, it is known that $100(1-\tau)\%$ of losses will be higher than the VaR q_τ at level τ, but the VaR alone cannot give any further information about the size of these large losses. As mentioned in Section 1.1 with VaR being non-coherent, these two weaknesses pushed the Basel Committee to recommend calculating the Expected Shortfall (ES) as an alternative to the VaR. A detailed discussion of pros and cons of VaR with ES is given in Section 4.1.

From a regulatory point of view, Basel III Accord (Basel Committee on Banking Supervision 2019) provides the backtesting requirements of ES, which is computed on daily basis, using VaR. It says that at trading desk (i.e. a group of traders), backtesting must compare one-day ahead VaR measure, which is calibrated to the most recent 12 month's data and equally weighted, at both $\tau = 0.975$ and 0.99, using at least one year of current observations. Moreover, it states that financial institutions may also implement backtesting VaR for probability levels other than $\tau = 0.99$, or may perform other statistical tests not set out in the standard. This was the motivation of the research in Chapter 3 that is the proposal of a bias-reduced approach for dynamic extreme VaR estimation called GARCH-UGH.

1.2.4 Expected Shortfall (ES)

An alternative risk measure replacing VaR is the ES (also known as Conditional VaR), which is a measure of the mean of potential (extreme) losses X exceeding the $\text{VaR}_\tau(X)$ at a given probability level τ. Mathematically, for $\tau \in (0,1)$, generally τ tends to 1, the τth ES is defined as

$$\text{ES}_\tau(X) = e_\tau = \text{E}[X \mid X > \text{VaR}_\tau(X)] = \frac{1}{1-\tau}\int_\tau^1 \text{VaR}_s(X)ds.$$

Unlike VaR, which only contains information on one point quantile itself, ES contains information from the whole right tail (supposing negative returns). On the other hand, the major drawback of ES is its difficulty to be backtested because it lacks the elicitability property, mentioned in Subsection 1.2.2. Again, a detailed discussion of pros and cons of ES is given in Section 4.1.

From a regulatory point of view, Basel III Accord (Basel Committee on Banking Supervision 2019) sets the ES as the new risk measure for the purpose of calculating market risk capital requirements, replacing VaR. In calculating ES, financial institutions must use a probability level $\tau = 0.975$. Notice that the probability level for ES is lower than for the VaR due to the fact that ES is systematically greater than VaR based on the definition. Keeping a probability level $\tau = 0.99$ leads to overly conservative ES. Furthermore, it is stated that no particular type of ES model is prescribed in the framework, which is the motivation of the research in Chapter 4.

1.2.5 Other risk measures

Other risk measures that can be considered as alternatives to VaR and ES are the followings: median shortfall, range VaR and expectile. Note that methods of both estimation and backtesting of these measures are not considered in this thesis.

Median Shortfall (MS) - In contrast to ES, which is the mean of loss X exceeding VaR, Median Shortfall (MD) is the median of the loss size conditional on that the loss exceeds VaR at level τ. It is introduced by Kou et al. (2013) to mitigate the nonelicitability of ES. Mathematically, for a probability level $\tau \in (0,1)$, the MS of X is defined as

$$\text{MS}_\tau(X) = \text{VaR}_{(1+\tau)/2}(X).$$

Based on the equation above, one can notice that the MS does not quantify the size of large losses beyond $\text{VaR}_{(1+\tau)/2}$, which means that it has a same problem as VaR. Conversely, MS is simple to backtest because we can use the methods of backtesting VaR directly while ES is difficult to backtest. MS is also statistically robust as it uses the median of the tail distribution instead of the mean and the median is robust. Actually He et al. (2022) argue that the MS is a better option than the ES for calculating the capital requirements in the Basel Accords due to above nice properties.

Range Value-at-Risk (RVaR) - Cont et al. (2010) propose the risk measure called RVaR (also known as interquantile expectation) that is an interpolation between VaR and ES. It takes account of a trade-off between ES's sensitivity to the potential losses and VaR's robustness. For a probability level $0 < \tau_1 < \tau_2 < 1$, $\text{RVaR}_{\tau_1,\tau_2}$ is defined as the average of all VaR_τ between τ_1 and τ_2. Mathematically, Fissler and Ziegel (2021) show that

$$\text{RVAR}_{\tau_1,\tau_2}(X) = \frac{1}{\tau_2 - \tau_1} \int_{\tau_1}^{\tau_2} \text{VaR}_\tau(X) d\tau$$

if $\tau_1 < \tau_2$ and VaR_{τ_2} if $\tau_1 = \tau_2$. Note that RVaR is not elicitable like ES and its backtesting is expected to be difficult.

Expectile - Daouia et al. (2018) explain that the notion of expectiles is a least square analogue of quantiles. As Koenker and Bassett (1978) defined the quantiles as the minimization framework of an absolute error loss, Newey and Powell (1987) introduced the expectiles as the minimizers of a quadratic loss, for a probability level $\tau \in (0,1)$:

$$ep_\tau(X) = \underset{x \in \mathbb{R}}{\text{argmin}} \mathbb{E}[\xi_\tau(X - x) - \xi_\tau(X)],$$

where $\xi_\tau(x) = \mid \tau - \mathbb{1}(x \leq 0) \mid x^2$ with the indicator function $\mathbb{1}$. It is given in Bellini and Di Bernardino (2015) that $ep_\tau(X) = \mathbb{E}[X]$ when $\tau = \frac{1}{2}$ so that the expectiles can be seen as the generalization of the mean just as the quantiles generalize the median. Regarding the quantitative risk management, using the expectile with $\tau = 0.99855$ leads to comparable risk measures as $\text{VaR}_{0.99}$ and $\text{ES}_{0.975}$ under the standard normal distribution. Unlike VaR and ES, the expectiles are only elicitable and coherent risk measures, and are possible to quantify extreme risks exceeding the VaR (Ziegel 2016). While expectiles can be treated as alternative risk measures to VaR and ES, they can be used to estimate VaR and ES using an asymptotic equivalence between

expectiles (Taylor 2008; Daouia et al. 2020; Daouia et al. 2021). Moreover, sample expectiles produce a class of smooth curves as functions of the probability level τ unlike sample quantiles (Daouia et al. 2018). On the other hand as it can be seen from the definition of expectiles, it lacks a visual and intuitive interpretation unlike VaR and ES. Emmer et al. (2015) state that there is no sufficient evidence to justify replacement of ES by expectiles in applications. It is now clear that expectiles are becoming increasingly popular in the finance literature.

1.3 Methodology of VaR and ES estimations

We list the methodologies for the VaR and ES estimation based on different statistical and volatility models often used in the literature. They are categorized into four broad types as follows: nonparametric, parametric, semiparametric and EVT, where in fact EVT appears in other three categories. We did not use all methods but some are used in the applications in Chapter 3 and 4. See Section 3.2 and 4.3.1 for our proposed estimation methods of VaR and ES, and Section 3.3 and 4.3.2 for other methods used in the applications. We also clarify the difference between conditional and unconditional estimation methods, which is further discussed in Section 2.3 in relation to the use of EVT in finance. Conditional means an effect of volatility of financial time series is considered and for unconditional it is not considered. We believe that it is necessary to incorporate dynamic changes in the market to reflect the most updated risk level for more accurate estimation of risk measures VaR and ES.

Nonparametric - The most widely used nonparametric VaR method is historical simulation (HS), which is also most popular in the industry. It requires no assumption in underlying distribution and estimates by simply finding the quantile of the empirical distribution of historical financial returns. Apart from being the simplest method, HS has the advantage that it is able to model the heavy tails and hence derives VaR estimates that perform better than Gaussian approaches (Hendricks 1996). However, VaR estimates based on HS have high standard error, especially for a high (or extreme) probability level representing tail events in market. This finding agrees with our results in the applications.

Parametric - The most common distributional assumption for financial returns in the estimation of (either conditional or unconditional) VaR is normality. However, methods based on normal distribution severely underestimate extreme VaR and ES because they fail to account for heavy tails.

This fact led to the improvement of parametric estimations in three ways. Firstly, it is to consider the suitable conditional volatility modeling that are Generalized Autoregressive Conditional Heteroskedasticity (GARCH)-type (Bollerslev 1986; Jalal and Rockinger 2008; Danielsson 2011; Furió and Climent 2013; Youssef et al. 2015; Zhao et al. 2019) and Realized Volatility models (Corsi 2009; Bee et al. 2016; Degiannakis and Potamia 2017). Secondly, it is to take account of asymmetric and heavy-tailed distributions (Ergen 2015; Righi and Ceretta 2015). It is known that Student's t distribution fits the financial data better in the tails than the normal distribution. Empirical studies from aforementioned papers find that VaR and ES estimates based on the skewed Student's t distribution are more accurate than the estimates based on the normal and Student's t distribution. Lastly, it is to improve the performance of the estimation methods with higher conditional moment, i.e. skewness and kurtosis, by using Corner Fisher expansion and Gram Charlier expansion (Wong and So 2003; So and Wong 2012).

Semiparametric - The commonly used approaches in this field are the EVT, the Filtered Historical Simulation (FHS) and the Conditional Autoregressive VaR (CAViaR). The EVT provides a theoretical and practical foundation for modeling extreme events statistically (Coles 2001; Beirlant et al. 2004; de Haan and Ferreira 2006; Reiss and Thomas 2007). The FHS (Barone-Adesi et al. 2002) extends the idea of volatility adjustments to HS, which is in fact a hybrid method combining HS and Monte Carlo simulation (Novales and Garcia-Jorcano 2019). The CAViaR (Engle and Manganelli 2004) directly estimates the dynamics of risk measures instead of constructing under some specific distributional assumption, even the empirical one (Righi and Ceretta 2015).

EVT - EVT considers only the tails of the distribution of financial returns without making any specific assumption concerning the center of the distribution. Unconditional EVT methods and the use of EVT in finance are explained in Chapter 2.

Conditional EVT methods of VaR and ES estimations such as the influential two-step GARCH-EVT framework (McNeil and Frey 2000; McNeil et al. 2015) and our proposed approach (Kaibuchi et al. 2022) are discussed in Chapter 3 and 4.

1.4 Backtesting

Backtesting of risk measures is of the same importance as their estimation. From a regulatory point of view, financial institutions must make sure that the models they used to estimate risk measures VaR and ES are accurate. According to Basel Committee on Banking Supervision (2019), backtesting is the process of comparing the ex-ante estimates of risk measures with the ex-post realized financial returns to assess the conservation of risk measurement systems. Strict definition is also give in Bayer and Dimitriadis (2020b), which basically state that backtesting for specific risk measure is only allowed to require estimates of this risk measure as input variables besides the realized returns. This strict version of definition will be relevant to ES (see Chapter 4) and not VaR. See also for example, Danielsson (2011) and McNeil et al. (2015), as a reference for backtestings used in practice.

We discuss the backtestings from two different aspects as follows: approaches (direct, based on elicitability and indirect) and types (traditional and comparative). In this section, we only introduce the concepts and details are reserved for Chapter 3 and 4.

Direct backtesting - This approach examines whether the estimates of risk measures VaR and ES under a certain model match with the unknown true values of risk measures or not. More precisely, we test whether the point estimates of risk measures are acceptable or not. For backtesting VaR, the tests are often based on a hit sequence of VaR violations $I_t = \mathbb{1}\{x_t > \hat{q}_\tau(X_t)\}$. If a VaR estimation method is accurate, then the sequence (I_t) should approximately be an independent sequence of Bernoulli variables with success probability $p = 1 - \tau$. Based on such observations, the unconditional likelihood ratio coverage test is proposed by Kupiec (1995), also known as the proportion of failures (POF) test, which tests the distributional assumption. In the similar manner, Basel Committee on Banking Supervision (2019) adopted the traffic light approach, which counts the number of VaR violations and

classifies the model into three backtesting zones, distinguished by colors into a hierarchy responses. Christoffersen (1998) proposes independence and conditional coverage, i.e. joint, tests based on a first-order Markov process model to check independence property and both properties as the name suggests, respectively. However, there have been no methods for direct backtesting ES in the existing literature (He et al. 2022).

Backtesting based on elicitability - This approach is based on the forecast evaluation framework based on the Diebold-Mariano (D-M) test (Diebold and Mariano 1995; Bellini et al. 2019). D-M test is the test of null hypothesis of no difference in the accuracy of two competing forecasts of risk measures by means of realized scoring functions. Since scoring functions are used, risk measures to be assessed must be elicitable, which means that VaR can be backtested and ES cannot be backtested by this approach.

Indirect backtesting - This approach is indirect in the sense that it either examines auxiliary quantities that are closely related to specific risk measure, e.g., ES, or backtests by means of the joint elicitability, i.e., coelicitability (see Section 4.5.1) of a collection of risk measures, e.g., a pair of VaR and ES. The former kind of approach has been proposed for backtesting ES in the literature because VaR can be backtested easily based on the VaR violations. These tests require the entire return distribution, tail return distribution, the cumulative violation process (Wong 2008; Acerbi and Szekely 2014; Costanzino and Curran 2015; Du and Escanciano 2015; Löser et al. 2018), use multiple probability level for a risk measure (Emmer et al. 2015; Kratz et al. 2018; Basel Committee on Banking Supervision 2019), or utilize VaR and the volatility in addition to ES (McNeil and Frey 2000; Righi and Ceretta 2013; Nolde and Ziegel 2017). Strictly speaking, this type of indirect backtesting should not be regarded as a suitable backtesting approach for ES because a rejection of the null hypothesis of above tests do not necessarily imply that the ES estimates are wrong and simply implies that some auxiliary quantities are wrong. This belief is supported by the strict definition of backtesting given in Bayer and Dimitriadis (2020b).

The latter is actually the D-M test using the scoring functions of a pair of VaR and ES, thanks to Fissler and Ziegel (2016) who have shown that the pair (VaR, ES) are joint elicitable with respect to the set of loss distributions. This observation has led

researchers to uncover new type of backtesting that evaluates the accuracy of VaR and ES jointly.

Traditional backtesting - This backtesting can be viewed as a model verification test, and includes direct backtesting and the former kind of indirect backtesting. Hence, traditional backtestings are concerned with assessing some optimality property of a set of risk measure estimates and not suited to compare different estimation methods for risk measures.They perform a statistical tests for the null hypothesis:

$$H_0 : \text{The risk measure estimates are correct.}$$

If H_0 is not rejected, then the risk measure estimates are deemed to be adequate. Generally, the traditional backtestings with the hypothesis H_0 are not relevant to elicitability of the risk measure and are not aimed at model comparison and ranking.

Comparative backtesting - This backtesting is better suited for model comparison on the basis of forecasting accuracy, and includes the D-M tests based on elicitability and joint elicitability. In order to compare the estimation performance of two models, say the competing and benchmark models, and decide which one is better, we can use the comparative version of the traffic light approach (i.e. three-zone approach) in the Basel III for the VaR (Basel Committee on Banking Supervision 2019) proposed by Fissler and Ziegel (2016) and Nolde and Ziegel (2017). In this comparative backtesting, we consider the two following two hypotheses:

H_0^- : The competing model predicts at least as well as the benchmark model,

H_0^+ : The competing model predicts at most as well as the benchmark model.

The null hypothesis H_0^- is an analogue of H_0 but adapted to a comparative setting. The other hypothesis H_0^+ is more conservative in the sense that a backtest is passed if we can reject H_0^+. By this hypothesis, we can explicitly control the type I error of accepting an inferior competing model over a benchmark model.

In this thesis, we use both traditional and comparative backtestings to check the accuracy of our proposed approach compared to other basic and EVT-type approaches in the VaR and ES estimations, given in Chapter 3 and 4, respectively. We believe that the use of both backtestings for risk measures, i.e. adopting a two-stage

backtesting framework, will enhance the regulatory framework for financial institutions by providing the correct incentives for accuracy of risk measure estimates. Furthermore, we classify backtestings as traditional and comparative throughout this thesis following the classification given in Nolde and Ziegel (2017).

1.5 Motivation for this thesis

In this section, we introduce the motivation for using EVT for quantitative risk management, estimating dynamic extreme VaR in Chapter 3 and ES in Chapter 4.

The occurrence of extreme financial market events are increasingly frequent in the post 1980s. Traditional estimation methods for VaR and ES, for example, HS and normal distribution, suffer from major drawbacks that were revealed during the turbulence of extreme financial crises. They tend to underestimate the risk measures severely and the prevalence of tail risks was largely neglected. The accurate estimation of extreme market risk had been receiving a lot of attention in quantitative risk management since it is difficult to model unexpected extreme events that usually lie outside the domain of available financial observations. Hence, the use of EVT in the estimation of risk measures is natural and this motivated the development of an alternative approach of VaR and ES estimations called GARCH-UGH.

Although Basel Committee on Banking Supervision (2019) changed the risk measure for capital requirements in the internal market risk model from VaR to ES, estimation of VaR is still needed in practice. This is because sensible estimation of ES is based on correctly specified VaR estimates by the definition of ES, i.e. ES is the mean of losses exceeding VaR. The more accurate the VaR estimates are, the more accurate the ES estimates. Moreover, we think that it is necessary to incorporate dynamic changes in the market to reflect the most updated risk level. We therefore considered a novel conditional EVT method for extreme VaR estimation in Chapter 3.

With regard to estimation of ES, there has not been sufficient investigation to establish the superiority of a certain estimator relative to the others in the literature. In addition, no particular type of ES model is prescribed in the framework of Basel Committee on Banking Supervision (2019). We thus proposed a novel approach of

dynamic extreme ES estimation, which is based on our proposed GARCH-UGH approach and the use of asymptotic equivalence between VaR (quantile) and ES, in Chapter 4. Regarding the backtesting of ES, Basel Committee on Banking Supervision (2019) still demands financial institutions to use traditional VaR backtesting for ES. At the same time, we can expect that upcoming regulations will require them to backtest ES without using VaR backtesting method. We also tackled an urgent problem of which ES backtesting methods can be used in the practice.

1.6 Outline of this thesis

In Chapter 2, we briefly describe the statistical aspects of Extreme Value Theory focusing on the tail estimation methods for heavy-tail (i.e. Pareto) distributions that are the cornerstone of the use of EVT in finance. We look at the important concepts of EVT such as extreme value index, extreme value condition and second-order condition. For tail estimation methods, we rely on the heavy-tail property and estimate extreme value index, extreme quantile (VaR) and second-order parameter, which is required for bias-reduction procedures. In particular, we focus on the famous Hill estimator (Hill 1975), Weissman quantile estimator (Weissman 1978), Peaks-Over-Threshold method using the generalized Pareto distribution (Pickands 1975) and second-order parameter estimator by Gomes et al. (2002) (actually given in Chapter 3) for the purpose of introducing our EVT-type method for VaR and ES estimations. We also review both unconditional and conditional VaR and ES methods based on EVT with the limitations in finance.

In Chapter 3, we tackle the question of estimating the VaR of loss return distribution at extreme levels, which is an important question in financial applications, both from operational and regulatory perspectives. In particular, the dynamic estimation of extreme VaR given the recent past has received substantial attention because the occurrence of extreme financial events has increased since 1980s including the Black Monday of 1987, the Dot-Com Bubble of 2000, the Global Financial Crisis of 2007-2008, and the recent COVID-19 recession of 2020. Moreover, accurate estimation of VaR is still essential in practice even if Basel Committee on Banking Supervision (2019) changed the risk measure for the calculation of capital requirements from

VaR to ES. This is because sensible estimation of ES is based on correctly specified VaR estimates by the fact that ES is the mean of losses exceeding VaR. We propose here a new two-step bias-reduced estimation methodology for the estimation of one-step ahead dynamic extreme VaR, called GARCH-UGH (Unbiased Gomes-de Haan), whereby financial returns are first filtered using an AR-GARCH model, and then a bias-reduced estimator of extreme quantiles is applied to the standardized residuals. We analyze the performance of our approach on four financial time series, which are the Dow Jones, NASDAQ and Nikkei stock indices, and the Japanese Yen/British Pound exchange rate. Our results indicate that the GARCH-UGH estimates of the dynamic extreme VaR are more accurate than those obtained either by historical simulation, conventional AR-GARCH filtering with Gaussian or Student-t innovations, or AR-GARCH filtering with standard extreme value estimates, both from the perspective of in-sample and out-of-sample traditional VaR backtestings, which are the unconditional coverage test by Kupiec (1995) and conditional coverage test by Christoffersen (1998). The numerical results of comparative VaR backtesting, which is based on the Diebold-Mariano test (Diebold and Mariano 1995), also support the use of the GARCH-UGH approach by yielding definitive answers to the cases when GARCH-UGH and GARCH-EVT approaches are either all accepted, or all rejected in the traditional VaR backtestings. In addition, our bias-reduction procedure will be designed to be robust to departure from the independence assumption, and as such will be able to handle residual dependence present after filtering in the first step. Our finite-sample results will also illustrate that the GARCH-UGH method leads to one-step ahead extreme conditional VaR estimates that are less sensitive to the choice of sample fraction, and hence mitigates the difficulty in selecting the optimal number of observations for the estimations. Finally, the computational cost of GARCH-UGH is lower than that of conventional GARCH-EVT: the extreme value step in the GARCH-UGH method is semiparametric with an automatic and fast recipe for the estimations of the one-step ahead extreme conditional VaRs, while the competing GARCH-EVT method is based on a parametric fit of the Generalized Pareto Distribution (GPD) to the residuals using Maximum Likelihood Estimation.

In Chapter 4, we extend the GARCH-UGH approach used in dynamic extreme VaR estimation to the dynamic extreme ES estimation by means of the asymptotic

equivalence between quantile (VaR) and ES. This is motivated by the fact that there has not been sufficient investigation to establish the superiority of a certain estimator of ES relative to the others in the literature and no particular type of ES model is prescribed in the framework of the Basel Committee on Banking Supervision (2019). Our results show that the GARCH-UGH approach produces more accurate ES estimates than those obtained by basic estimation methods, both from the perspective of traditional and comparative ES backtestings. We use the exceedance residual test (McNeil and Frey 2000), the conditional calibration test (Nolde and Ziegel 2017) and the expected shortfall regression test (Bayer and Dimitriadis 2020b) for traditional backtestings, and Diebold-Mariano test again based on the joint elicitability of VaR and ES for comparative backtesting. When compared to other EVT-type methods, comparative backtestings with chosen two scoring functions result in a good agreement with the GARCH-UGH approach being the best estimator of ES, while traditional backtestings are not always in line with the superiority of our proposed approach.

Chapter 5 is the conclusion of this thesis. In summary, we believe that the findings of both estimation and backtesting of risk measures for tail risks in financial extreme market given in Chapter 3 and 4 would be useful for developing regulatory framework of the BCBS and monetary policies aimed at mitigating tail risks. We strongly suggest adopting a two-stage backtesting framework, i.e., the use of both traditional and comparative backtestings for risk measures that will enhance the regulatory framework for financial institutions by providing the correct incentives for accuracy of risk measure estimates. More precisely, the comparative backtesting methods can be used by financial institutions internally to select better performing methods among competing alternatives when traditional backtestings methods do not yield definitive answers as competitive methods are all accepted, or all rejected. Supplementing with comparative backtestings is essential, and hence can adequately quantify the risks even though they still have some drawbacks to consider for the practical use, e.g., there exists no optimal scoring function with any theoretical guarantee.

Chapter 2

Statistical aspects of Extreme Value Theory (EVT)

2.1 Extreme Value Theory

Extreme Value Theory (EVT) is a branch of probability theory associated with limiting laws for extreme values in large samples. Statistically, it uses only extreme event data, focuses on analyzing the tail regions of distributions and allows the extrapolation beyond available data to forecast unforeseen extreme events. On the other hand, many statistical models focus on the whole distribution at the expense of less consideration in the tails. Thus, EVT could potentially provide better risk measure estimates in quantitative risk management. EVT has a long and successful history. We referred to many excellent books including Beirlant et al. (2004), de Haan and Ferreira (2006) and Resnick (2007) for a detailed theory of extreme values, Coles (2001) and Reiss and Thomas (2007) for the practical applications, and Embrechts et al. (1997), Danielsson (2011) and McNeil et al. (2015) for the applications in finance.

In this chapter, we briefly describe the statistical aspects of EVT focusing on the tail estimation methods for heavy-tail, i.e., Pareto, distributions that are the cornerstone of the use of EVT in finance. We look at the important concepts of EVT such as extreme value index, extreme value condition and second-order condition. For tail estimation methods, we rely on the heavy-tail property, and estimate extreme value index, extreme quantile (VaR) and second-order parameter, which is required for bias-reduction procedures. In particular, we focus on the famous Hill estimator (Hill 1975), Weissman quantile estimator (Weissman 1978), Peaks-Over-Threshold

method using the generalized Pareto distribution (Pickands 1975) and second-order parameter estimator by Gomes et al. (2002) (actually given in Chapter 3) for the purpose of introducing our EVT-type method for estimations of VaR and ES. We also review both unconditional and conditional estimation methods of VaR and ES based on EVT with the limitations of EVT in finance.

2.1.1 Asymptotic model formulation

In Chapter 2, we consider i.i.d. random variables $X_1, ..., X_n$ representing financial losses with finite mean μ, finite variance σ^2 and common distribution function F. We also assume that the second moment $E(X^2) < \infty$ and the underlying distribution function F is continuous and strictly increasing to avoid an overly mathematical treatment. Later in Chapter 3 and 4 we relax the assumption of independence and consider a time series of dependent losses, which will be negative log-returns for a stock, index or exchange rate.

The EVT has been developed in parallel with the central limit theory and in fact the two theories have some similarities. The central limit theory is concerned with the limiting behaviour of the partial sums $S_n = X_1 + X_2 + \cdots + X_n$ as $n \to \infty$. Hence, the probabilistic problem is expressed as follows:

$$P\left(\frac{S_n - n\mu}{\sqrt{n\sigma^2}} \leq x\right) \to \Phi(x) \quad \text{as} \quad n \to \infty,$$

where Φ is the distribution function of the standard normal distribution

$$\Phi(x) = \frac{1}{2\pi} \int_{-\infty}^{x} e^{-u^2/2} du.$$

On the other hand, the EVT is concerned with the limiting behaviour of the sample extremes, either $\max(X_1, X_2, ..., X_n)$ or $\min(X_1, X_2, ..., X_n)$ as $n \to \infty$. In this thesis, we will always consider the maximum and it is defined by

$$X_{n,n} = \max(X_1, X_2, ..., X_n).$$

Of course, we can study the minimum due to the relation

$$\min(X_1, X_2, ..., X_n) = -\max(-X_1, -X_2, ..., -X_n).$$

We are interested in the possible limit distributions of the maximum $X_{n,n}$ in the similar way the central limit theory is derived for the partial sums S_n.

In theory we can derive the asymptotic distribution of the maximum $X_{n,n}$ as $n \to \infty$ exactly using an assumption of the sample:

$$\begin{aligned} P(X_{n,n} \leq x) &= P(X_1 \leq x, ..., X_n \leq x) \\ &= P(X_1 \leq x) \times \cdots \times P(X_n \leq x) \\ &= F^n(x), \quad \forall n. \end{aligned} \tag{2.1}$$

We now proceed by looking at the asymptotic behaviour of F^n in (2.1). Let x^* be the right endpoint of F, which is given by $x^* := \sup\{x : F(x) < 1\}$. We know that the maximum, which corresponds to the extreme event occurs near the upper end of F, thus the behaviour of the maximum $X_{n,n}$ is related to the right tail of the distribution F near the right endpoint. This means that

$$X_{n,n} \xrightarrow{P} x^*, \quad \text{as} \quad n \to \infty,$$

where $\xrightarrow{P}$ indicates the convergence in probability. Using the above idea, we now obtain that $F^n(x) \to 0$ as $n \to \infty$ for $x < x^*$ and $F^n(x) = 1$ for $x \geq x^*$. This case is known as a degenerate limit distribution that is non-informative. In order to avoid this difficulty, a normalization of $X_{n,n}$ is needed.

Suppose we can find a sequence of positive numbers $\{a_n; n \geq 1\}$ and a sequence of numbers $\{b_n; n \geq 1\}$ such that $(X_{n,n} - b_n)/a_n$, the sequence of normalized maxima, then it converges in distribution as follows:

$$P\left(\frac{X_{n,n} - b_n}{a_n} \leq x\right) = F^n(a_n x + b_n) \to G(x) \quad \text{as} \quad n \to \infty, \tag{2.2}$$

where G is a non-degenerate distribution function. This development of the model is the cornerstone of EVT so it appears in many literature, for instance, Coles (2001),

Beirlant et al. (2004) and de Haan and Ferreira (2006).

From this asymptotic distribution of $X_{n,n}$ in (2.2), there are two problems to consider. The first problem is known as the extremal limit problem. We want to identify the class of non-degenerate distributions G that can appear as a possible limit in (2.2). It has been solved by by Fisher, Tippett and Gnedenko (see Section 2.1.2). The second problem is the domain of attraction problem. For each of those limit distributions found in the extremal limit problem, we want to find necessary and sufficient conditions on the underlying distribution functions F such that (2.2) holds. This is briefly touched in Section 2.1.4 as it is not our main interest in this thesis.

2.1.2 Fisher-Tippett-Gnedenko Theorem

To answer the extremal limit problem, we will now look at the theoretical foundation of EVT known as the Fisher-Tippett-Gnedenko three-types theorem.

Theorem(Fisher-Tippett-Gnedenko theorem, Embrechts et al. 1997; Coles 2001; Beirlant et al. 2004; de Haan and Ferreira 2006). If there exists a sequence of of positive numbers $\{a_n; n \geq 1\}$ and a sequence of numbers $\{b_n; n \geq 1\}$ such that

$$P\left(\frac{X_{n,n} - b_n}{a_n} \leq x\right) \to G(x) \quad \text{as} \quad n \to \infty,$$

where G is a non-degenerate distribution function, then G must belong to one of the following three distribution functions ($\alpha > 0$)

- Fréchet: $\Phi(x) = \begin{cases} 0, & \text{if} \quad x \leq 0, \\ \exp(-x^{-\alpha}), & \text{if} \quad x \geq 0, \end{cases}$
- Weibull: $\Psi(x) = \begin{cases} \exp(-|x|^{\alpha}), & \text{if} \quad x \leq 0, \\ 1, & \text{if} \quad x \geq 0, \end{cases}$
- Gumbel: $\Lambda(x) = \exp(-\exp(-x)), \quad -\infty < x < \infty.$

This theorem states that the normalized maxima $(X_{n,n} - b_n)/a_n$ converge in distribution to one of the three types of distributions that are Fréchet, Weibull and Gumbel. The remarkable characteristic of this theorem is that the three types of distributions are the only possible limits for the distribution of the normalized maxima, regardless of the underlying distribution function F. It is in the sense that it provides an extreme value analogue of the central limit theory explained in section 2.1.1. Note

that the proof of this theorem is omitted, which is given in Beirlant et al. (2004) using the idea of transformation from the convergence in distribution to the convergence of expectations.

2.1.3 Extreme value index (EVI)

Three types of distributions discussed in previous section can be thought of as members of a simple one-parametric family of distribution that is known as the generalized extreme value (GEV) or the extreme value distribution. The distribution function of the standard GEV distribution is given by

$$G_\gamma(x) = \exp(-(1+\gamma x)^{-1/\gamma}), \quad \text{for} \quad 1+\gamma x > 0, \tag{2.3}$$

where $\gamma = 1/\alpha$ is a new parameter introduced. The real quantity γ is called the extreme value index (EVI) also known as the shape parameter. It is a key parameter in the whole of extreme value analysis since it indicates the heaviness of the tail, i.e., how extreme and how frequent extreme events can be under the given probability distribution. Therefore, the knowledge and understanding of γ are necessary for the tail estimators for extreme quantiles (VaR) of X (discussed in Section 2.2).

The extreme value distribution given in Equation (2.3) is generalized in the sense that the parametric form subsumes three types of distribution based on the value of γ. We can see that the sign of EVI is the dominating factor in the description of the tail behaviour of the underlying distribution F. For that reason, we will distinguish between three cases where $\gamma > 0$, $\gamma < 0$ and the intermediate case where $\gamma = 0$ and they are widely known as Fréchet-Pareto, Weibull and Gumbel cases, respectively. Let us consider the behaviour of F in its right tail for three cases separately:

- For $\gamma > 0$ (Fréchet-Pareto), the right endpoint of the distribution is infinity and an absence of positive exponential moments, i.e., they are infinite, is visible from (2.3). Tails decline by a power law. We say that a distribution is heavy-tailed if above condition is satisfied (see Foss et al. 2013), thus $1 - G_\gamma$ is heavy-tailed (also known as fat-tailed).

- For $\gamma < 0$ (Weibull), the right endpoint of the distribution is finite and $-1/\gamma$ from (2.3) so $1 - G_\gamma$ is short-tailed (also known as thin-tailed).
- For $\gamma = 0$ (Gumbel), the right endpoint of the distribution is infinity but all exponential moments are finite and an exponential decay appears. A distribution F is called light-tailed if above condition is satisfied (see Foss et al. 2013), hence $1 - G_\gamma$ is light-tailed.

In this thesis, we assume that the tails of the financial returns to be heavy-tailed ($\gamma > 0$) for VaR and ES estimations in Chapter 3 and 4. This assumption is ubiquitous in actuarial and financial risk management (see Embrechts et al. 1997 and Resnick 2007).

2.1.4 First-order condition

We present the condition required on distribution function F for the normalized maxima $X_{n,n}$ to have the limiting distribution G, which was illustrated in the Fisher-Tippett-Gnedenko theorem. This condition is known as the first-order condition or the extreme value condition. Let U be the tail quantile function defined by

$$U(x) = q\left(1 - \frac{1}{x}\right)$$

where q is the quantile function. It is the generalized inverse of $1/(1-F)$, i.e., $U(y) = F^{\leftarrow}\left(1 - \frac{1}{y}\right)$ for $y \geq 1$. Then, the distribution function F is in the domain of attraction of the extreme value distribution G_γ if and only if there exists a positive (or first-order auxiliary) function a such that,

$$\lim_{x\to\infty} \frac{U(xu) - U(x)}{a(x)} = h_\gamma(u) := \begin{cases} \dfrac{u^\gamma - 1}{\gamma}, & \gamma \neq 0, \\ \log u, & \gamma = 0, \end{cases} \qquad (\mathcal{C}_\gamma(a)) \qquad (2.4)$$

holds for all $u > 0$. Note that most of the EVI estimators that are based on a set of upper order statistics k are motivated by this condition. These estimators will be introduced in Section 2.2.1 and 2.2.4. Additionally, it is an important framework for the derivation of the second-order condition given in next section.

In this section, we also explain briefly the basic of the theory of regular variation that will show up in Section 2.2. Let ℓ be a positive measurable function on $(0, \infty)$.

We say that ℓ is regularly varying of index $\theta \in \mathbb{R}$ if

$$\lim_{x\to\infty} \frac{\ell(xu)}{\ell(x)} = u^{\theta} \quad \text{for all} \quad u > 0.$$

In the case $\theta = 0$, the function ℓ is called slowly varying. The slowly varying function thus satisfies the condition

$$\frac{\ell(xu)}{\ell(x)} \to 1 \quad \text{as} \quad x \to \infty. \tag{2.5}$$

Moreover, there is another property of the slowly varying function to mention (see Beirlant et al. 2004), which is

$$\lim_{x\to\infty} \frac{\log \ell(x)}{\log x} = 0. \tag{2.6}$$

2.1.5 Second-order condition

We are going to develop a second order condition related to $(\mathcal{C}_\gamma)$ (2.4). It is used to derive the asymptotic behaviour of several EVI estimators such as the Hill estimator (2.10) and the Pickands estimator (2.12), which lead to the derivation of the bias-reduced EVI estimators. The important asymptotically bias-reduced Hill estimator by de Haan et al. (2016) is also based on this second-order condition and used in the proposal of dynamic extreme VaR and ES estimators in Chapter 3 and 4.

Once again we start with the extreme value (or first-order) condition $(\mathcal{C}_\gamma)$, given by

$$\lim_{x\to\infty} \frac{U(xu) - U(x)}{a(x)} = \frac{u^{\gamma} - 1}{\gamma} =: h_\gamma(u),$$

for each $u > 0$. The first-order condition is concerned with the convergence in distribution of the normalized maxima $X_{n,n}$ to have the limiting distribution G whereas for the second-order condition the convergence rate is the main interest. The following second-order condition has been defined in de Haan and Ferreira (2006) and is commonly used in many literature (see for example Caeiro and Gomes 2010; Fraga Alves et al. 2003; Gomes et al. 2002; Gomes and Martins 2002; Peng 1998).

The function U is said to satisfy the second-order condition if for some positive first-order auxiliary function a and some positive or negative second-order auxiliary

function b with $\lim_{x\to\infty} b(x) = 0$,

$$\lim_{x\to\infty} \frac{\frac{U(xu)-U(x)}{a(x)} - \frac{u^\gamma-1}{\gamma}}{b(x)} = H_{\gamma,\rho}(u) := \frac{1}{\rho}\left(\frac{u^{\gamma+\rho}-1}{\gamma+\rho} - \frac{u^\gamma-1}{\gamma}\right), \tag{2.7}$$

holds for all $u > 0$, where $\rho \leq 0$ is a second-order parameter and $|b(x)|$ is of regular variation with index ρ. This ρ controls the speed of convergence in $(\mathcal{C}_\gamma)$. More precisely, the function b therefore controls the rate of convergence in the condition (2.4): the larger $|\rho|$ is, the faster $|b|$ converges to 0, and the smaller the error in the approximation of the right tail of U by a Pareto tail is. This makes it possible to precisely quantify the bias of the EVI and extreme quantile (VaR) estimators. Besides, it is known that the estimation of ρ is difficult. Some estimators of ρ have been proposed in the literature (see Section 2.2.4.3).

We propose our bias-reduced VaR and ES estimators that are built on the second-order estimator of Gomes et al. (2002), explained in Section 3.2.3. Note that the first-order condition (2.4) and second-order condition (2.7) are given in alternative forms by the Equations (3.4) and (3.7) in Chapter 3, respectively.

2.2 Tail estimation methods for Pareto-type distributions

We consider $X_1, X_2, ..., X_n$, n i.i.d. random variables representing financial losses with common distribution function F, and let $X_{1,n} \leq X_{2,n} \leq \cdots \leq X_{n,n}$ denote the order statistics based on the first n observations. In this section we suppose that F is of Pareto-type, i.e., the class of heavy-tailed distributions, referred to as the Fréchet maximum domain of attraction with EVI $0 < \gamma < 1$. This means that there exists a slowly varying function ℓ_F for which

$$1 - F(x) = x^{-1/\gamma}\ell_F(x), \tag{2.8}$$

where γ is a positive EVI and $x > 0$ large enough. We can present this model in terms of U equivalently

$$U(x) = x^\gamma \ell_U(x), \tag{2.9}$$

where ℓ_U is again slowly varying. Note that the condition $(\mathcal{C}_\gamma(a))$ in (2.4) is equivalent with the existence of $\ell_F(x)$ and $\ell_U(x)$ for which (2.8) and (2.9), respectively. Examples of distributions of Fréchet-Pareto type are given in Table 2.1.

Distribution	$1-F(x)$	Extreme value index (EVI)
Pareto(α)	$x^{-\alpha}$, $x>1;\alpha>0$	$\frac{1}{\alpha}$
Fréchet(α)	$1-\exp(-x^{-\alpha})$, $x>0;\alpha>0$	$\frac{1}{\alpha}$
Burr(β,τ,λ) (type XII)	$\left(\frac{\beta}{\beta+x^\tau}\right)^\lambda$, $x>0;\beta,\tau,\lambda>0$	$\frac{1}{\lambda\tau}$
Student's t $\lvert T_n\rvert$	$\int_x^\infty \frac{2\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{\omega^2}{n}\right)^{-\frac{n+1}{2}}d\omega$, $x>0;n>0$	$\frac{1}{n}$

TABLE 2.1: A list of distributions in the Fréchet domain.

In this section, we will consider the estimations of the EVI and extreme quantiles (VaR) of the distribution F. Our main interests are the famous Hill estimator (Hill 1975), Weissman quantile estimator (Weissman 1978), Peaks-Over-Threshold method using the generalized Pareto distribution (Pickands 1975) and second-order parameter estimator by Gomes et al. (2002) (actually given in Chapter 3).

An important challenge in the applications of the EVT based on k upper order statistics is choice of the tail sample fractions, i.e., how many very large or extreme values of the sample to be used in the statistical analysis. This determines whether extreme value models provide good extrapolation or not. It would be unrealistic to assume that only the maximum $X_{n,n}$ contains valuable information on the tail behaviour but lower order statistics may not contain such information. Furthermore, other reason is that choice of the optimal k affects the estimators, which will be discussed later. By choosing small k (upper order statistics), in this case few observations will be used in the tail estimation, which results in estimators with small bias but large variances. On the other hand, choosing large k (lower order statistics) causes the estimators to have small variances but large bias. Thus, the selection of optimal k is vital to balance the bias component and variance component. This will be covered further in Section 2.2.5. Finally, we clarify that we consider the order statistics $X_{n-k,n}$ with $n\to\infty$, $k=k(n)\to\infty$, and $k(n)/n\to 0$. These are called the

intermediate order statistics, which are known to be connected with EVT.

2.2.1 Hill method

Hill estimator - first estimator of EVI ($\gamma > 0$) is the Hill estimator (Hill 1975) given by

$$H_{k,n} \equiv \hat{\gamma}_k^H = \frac{1}{k}\sum_{i=1}^{k} \log X_{n-i+1,n} - \log X_{n-k,n}, \quad k = 1, ..., n-1, \tag{2.10}$$

where the term inside summation is the log of spacing between two order statistics. It plays a central role in many applications because it is derived as the maximum likelihood estimator of the Pareto distribution, that is one of the important representatives of the heavy-tailed distributions. It is inspired by the fact that the definition of a Pareto-type tail (2.8) can be re-written as

$$\lim_{t\to\infty} \frac{1-F(tx)}{1-F(t)} = \lim_{t\to\infty} \frac{(xt)^{-1/\gamma}}{t^{-1/\gamma}} \frac{\ell_F(xt)}{\ell_F(t)} = x^{-1/\gamma} \quad \text{for any} \quad x > 1,$$

due to the condition $\ell(xt)/\ell(t) \to 1$ (2.5). This means that the distribution of the relative excesses over a high threshold t conditionally on $X_i > t$ is approximately a strict Pareto distribution (see Table 2.1):

$$P\left(\frac{X}{t} > x \mid X > t\right) \approx x^{-1/\gamma} = x^{-\alpha} \quad \text{for any} \quad x > 1.$$

By letting $Y_j = X_i/t$, the likelihood equation and so the MLE for this Pareto distribution are found to obtain the Hill estimator, which is considered as the probabilistic construction in Beirlant et al. (2004).

Properties of Hill estimator - an important property of the Hill estimator is that $\hat{\gamma}_k^H$ is a consistent estimator for γ if we use the order statistics $X_{n-k,n}$ with $n, k(n) \to \infty$ and $k(n)/n \to 0$, i.e., if we use intermediate order statistics. Proof of consistency is shown in Mason (1982) using law of large numbers for sums of extreme order statistics and also in de Haan and Ferreira (2006) via a different approach. It is also asymptotically normally distributed (see for example Chapters 3 and 4 in de Haan and Ferreira 2006). However, there are some drawbacks in spite of these nice properties. They are as follows:

- Limited ranges of γ, i.e., $\gamma > 0$.

- For every choice of k, we obtain different estimates of γ. It is quite sensitive to the number of upper order statistics used in the estimation. We usually plot the estimates of γ against k to check the behaviour of the EVI estimators. In the case of $\hat{\gamma}_k^H$, the Hill plot $\{(k, \hat{\gamma}_k^H) : 1 \leq k \leq n-1\}$ is yielded. These plots typically show the large volatility, i.e., far from being constant, which makes the estimator difficult to use in practice if no further guideline is given for selection of the optimal tail sample fraction k (see Section 2.2.5). This is illustrated in Figure 2.1a.

- In many cases, $\hat{\gamma}_k^H$ overestimates the true value of γ. As a result, the substantial bias can appear (illustrated in Figure 2.1b) due to the slow convergence of the slowly varying part in the model. The slower the term $\log \ell(x)/\log x$ in (2.6) tends to zero as $x \to \infty$, the slower the ultimate linearity appears in a Pareto QQ-plot. One way to overcome this problems is to use the larger sample sizes n so that $\hat{\gamma}_k^H$ becomes unbiased for smaller values of k. Other way is the construction of the bias-reduced estimator of EVI, which is discussed in Section 2.2.4.3 and used in Chapter 3 and 4.

- Since $\hat{\gamma}_k^H$ is based on the log-transformed data, it is not-invariant with respects to the shifts of the data. This property is shared by the estimators based on log-transformed data, including the moment estimator in Section 2.2.4.1.

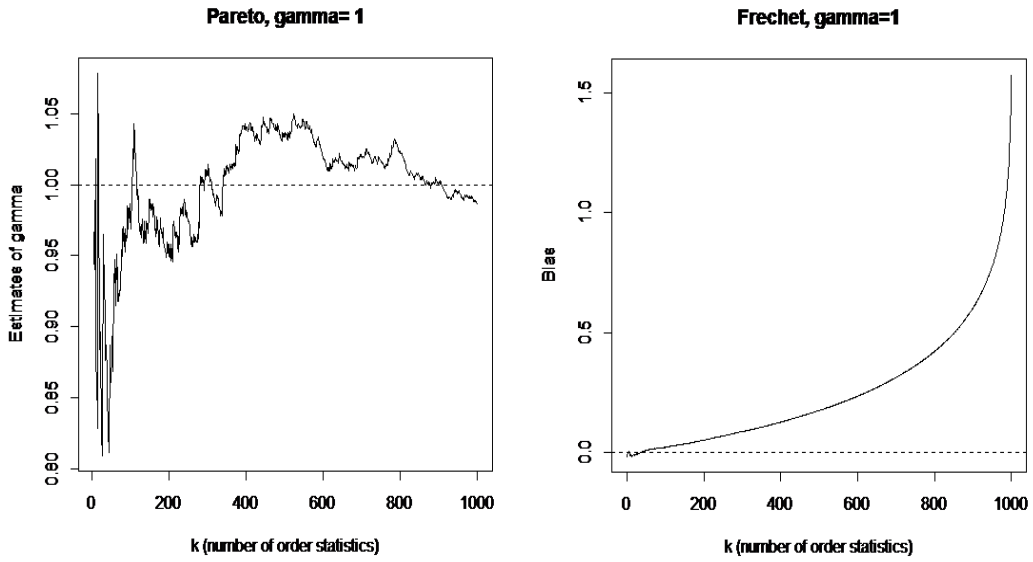
2.2.2 Weissman quantile estimator

In this section, we explain the estimation of extreme quantiles (VaR) and corresponding extreme tail probabilities via the approach proposed by Weissman (1978). For the estimation of the extreme quantile, we want to estimate the $\tau(=1-p)$th quantile based on a sample size n:

$$q_\tau \equiv Q(1-p) = \inf\{x : F(x) \geq \tau\}$$

where in fact $p = p_n$ since it depends on n. The Weissman quantile (VaR) estimator is given by

$$\hat{q}_\tau = X_{n-k,n}\left(\frac{k}{np}\right)^{\hat{\gamma}_k^H}. \tag{2.11}$$



(a) $(k, \hat{\gamma}_k^H)$ (b) $(k, \text{Bias}(k))$

FIGURE 2.1: (a) The Hill plot $(k, \hat{\gamma}_k^H)$ for simulated datasets of size $n = 1000$ from Pareto(1) and (b) the bias plot $(k, \text{Bias}(k))$ for simulated datasets of size $n = 1000$ from Fréchet(1).

We are particularly interested in the case in which a small exceedance probability p_n is outside the range of data. This allows us to estimate an extreme quantile q_τ that is to the right of all or almost all observations so that we can extrapolate outside the range of available observations. This means that p_n needs to satisfy the conditions $p_n \to 0$ as $n \to \infty$ and $np_n = o(k)$, i.e. , np_n equals a very small number. Choices of p_n are, for instance, $1/n$, $1/n^2$ and $1/(n \log n)$.

2.2.3 Peaks-Over-Threshold (POT) method

Peaks-Over-Threshold (POT) is the commonly used parametric EVT approach in finance, which is an alternative to the Hill-based estimator. When estimating VaR and ES, it essentially consists in fitting the generalized Pareto distribution (GPD) to the financial returns (see Chapter 3 and 4).

Under the i.i.d. condition, we consider the distribution function of excess $Y = X - u$ over a fixed high threshold u. The corresponding excess distribution above the threshold u is given by

$$F_u(y) = P(Y = X - u \leq y | X > u) = \frac{F(y+u) - F(u)}{1 - F(u)}, \quad y \geq 0.$$

The famous Pickands (1975) shows that the GPD occurs naturally as the limit distribution of the scaled excesses of i.i.d. random variables over high thresholds. Under this POT method, any observations that exceed a high threshold are modeled separately from non-extreme observations. The excesses Y from a fixed high threshold u follow a GPD $Y = X - u \sim GPD(\gamma, \beta)$ if

$$F_u(y) \approx GPD_{\gamma,\beta}(y) = \begin{cases} 1 - \left(1 + \dfrac{\gamma y}{\beta}\right)^{-1/\gamma}, & \gamma \neq 0, \\ 1 - \exp\left(-\dfrac{y}{\beta}\right), & \gamma = 0, \end{cases}$$

where $\beta > 0$ is a scale parameter, γ is the EVI, i.e., tail shape parameter as explained with the support $y \geq 0$ when $\gamma \geq 0$ and $0 \leq y \leq -\beta/\gamma$ when $\gamma < 0$.

In practice, we fix the number of data in the tail to be k where $k < n$ and use the proportion of tail data k/n. This effectively gives us a random threshold at the kth upper order statistic and we hence use the same setting as the Hill (2.10) and Weissman quantile estimators (2.11). For a probability level $\tau = 1 - p$, the extreme quantile (VaR) estimator by the POT approach is given by

$$\hat{q}_\tau = X_{n-k,n} + \frac{\hat{\beta}}{\hat{\gamma}}\left(\left(\frac{p}{k/n}\right)^{-\hat{\gamma}} - 1\right), \quad \hat{\gamma} \neq 0.$$

McNeil and Frey (2000) show that the EVT method based on the POT approach estimates more stable extreme quantiles than those obtained by the Hill estimator.

2.2.4 Other methods of EVI and second-order parameter estimations

2.2.4.1 EVI estimators for $\gamma \in \mathbb{R}$

Pickands estimator - one of the simplest and oldest well-known estimator for a general EVI $\gamma \in \mathbb{R}$ is the Pickands estimator from Pickands (1975):

$$\hat{\gamma}_k^P = \frac{1}{\log 2} \log\left(\frac{X_{n-\lceil k/4 \rceil+1,n} - X_{n-\lceil k/2 \rceil+1,n}}{X_{n-\lceil k/2 \rceil+1,n} - X_{n-k+1,n}}\right), \quad k = 1, ..., n. \tag{2.12}$$

It can be found through the following expansion based on the extreme value condition (2.4): since one can write $a(x) = x^\gamma \ell(x)$ with the limiting relation $a(xu)/a(x) \to$

u^γ, we find that

$$\frac{1}{\log 2}\log\left(\frac{U(4x)-U(2x)}{U(2x)-U(x)}\right) = \frac{1}{\log 2}\log\left(\frac{U(4x)-U(2x)}{a(2x)}\frac{a(2x)}{a(x)}\frac{a(x)}{U(2x)-U(x)}\right)$$
$$\to \frac{1}{\log 2}\log\left(\frac{h_\gamma(2)2^\gamma}{h_\gamma(2)}\right) = \frac{1}{\log 2}(\log 2)\gamma = \gamma, \quad x \to \infty.$$

We have shown that above expression with theoretical $U(x)$ tends to γ so replacing it by its empirical version $\hat{U}_n(x) = X_{n-\lceil n/x\rceil+1,n}$ leads to the Pickands estimator.

Due to the simplicity, weak consistency, shift and scale invariant, and asymptotic normality (see de Haan and Ferreira 2006) of $\hat{\gamma}_P$, it is quite appealing but has large asymptotic variance, which results in jagged path as a function of k in plots. In practice, we want to avoid this type of paths for better tail estimation. It is also very sensitive to the choice of the number of upper order statistics used for estimation and hence it should be not used in practice for small or moderate sample sizes, for example, when $n < 500$. To overcome this disadvantage, the refined Pickands estimator was introduced in Drees (1995).

Moment estimator - this EVI estimator generalizes the Hill estimator (2.10) for the case $\gamma > 0$ to general cases $\gamma \in \mathbb{R}$ by means of an application of a shift to the data. The resulting estimator is called the moment estimator, which has been introduced by Dekkers et al. (1989):

$$\hat{\gamma}_k^M = M_k^{(1)} + 1 - \frac{1}{2}\left(1 - \frac{(M_k^{(1)})^2}{M_k^{(2)}}\right)^{-1}, \tag{2.13}$$

where

$$M_k^{(\alpha)} := \frac{1}{k}\sum_{i=1}^{k}(\log X_{n-i+1,n} - \log X_{n-k,n})^\alpha, \quad \alpha > 0. \tag{2.14}$$

We can notice from (2.14) that $\hat{\gamma}_k^H = M_k^{(1)}$.

The moment estimator works as a diagnosis when the true value of EVI is unknown in the real data analysis due to its nice properties such as unlimited range of γ, consistency and asymptotic normality (see de Haan and Ferreira 2006 and Dekkers et al. 1989). Basically, the consistency of the moment estimator follows from

$$\frac{(M_k^{(1)})^2}{M_k^{(2)}} \xrightarrow{P} \begin{cases} \frac{1}{2}, & \text{if} \quad \gamma \geq 0, \\ \frac{1-2\gamma}{2(1-\gamma)}, & \text{if} \quad \gamma < 0, \end{cases}$$

as $k(n), n \to \infty$ and $k(n)/n \to 0$, and

$$\hat{\gamma}_k^H \xrightarrow{P} \begin{cases} \gamma, & \text{if} \quad \gamma \geq 0, \\ 0, & \text{if} \quad \gamma < 0, \end{cases}$$

since the slope of the Pareto QQ-plot when $\gamma \leq 0$ tends to zero near the higher observations. However, it usually has a high variance for small values of k and a high bias for large k, in the same way as the most of the classical EVI estimators. Besides, it is not shift invariant.

2.2.4.2 EVI estimators for specific ranges

Negative moment estimator - for $\gamma < 0$, we shall sometimes work with

$$\hat{\gamma}_k^{NM} := \hat{\gamma}_k^M - M_k^{(1)} = 1 - \frac{1}{2}\left(1 - \frac{(M_k^{(1)})^2}{M_k^{(2)}}\right)^{-1}, \tag{2.15}$$

which is called the negative moment estimator (see de Haan and Ferreira 2006 and Caeiro and Gomes 2010). We can see from this equation that the moment estimator in (2.13) is the combination of two estimators: the Hill estimator in (2.10) that is consistent for $\gamma^+ := \max(0, \gamma)$ and the negative moment estimator in (2.15) that is consistent for $\gamma^- := \min(0, \gamma)$. The remarkable feature of $\hat{\gamma}_k^{NM}$ is that it has the same asymptotic variance as $\hat{\gamma}_k^M$ when $\gamma < 0$. In Caeiro and Gomes (2010), due to this remark, an asymptotic unbiased estimator for $\gamma < 0$ based on the negative moment estimator was established. It is found that when $\gamma < 0$ this asymptotic bias-reduced estimator works well because of a smaller asymptotic bias but the same asymptotic variance as the moment estimator.

Probability-Weighted moment estimator - we review the EVI estimator, which is not based on log-transformed data. It is the probability-weighted moment estimator (PWM) of Hosking and Wallis (1987). It is derived as the one of the parameter estimations for the generalized Pareto distribution (GPD) that can be uniform, exponential or Pareto distributions depending on the shape parameters. This estimator

works for $\gamma < 1$ since moments and probability-weighted moments do not exist when $\gamma \geq 1$ for GPD. The PWM estimator consists of two statistics

$$P_n = \frac{1}{k}\sum_{i=1}^{k} X_{n-i+1,n} - X_{n-k,n},$$

and

$$Q_n = \frac{1}{k}\sum_{i=1}^{k} \frac{i}{k}(X_{n-i+1,n} - X_{n-k,n}),$$

which lead to the estimator

$$\hat{\gamma}_k^{PWM} = \frac{P_n - 4Q_n}{P_n - 2Q_n}. \tag{2.16}$$

Note that the PWM estimator is a ratio of the weighted sum of order statistics and the statistic P_n is the empirical mean excess function. Moreover, it is found by Hosking and Wallis (1987) that the scale i/k in the statistic Q_n can be replaced with $(i - 0.35)/k$, which may improve the finite sample behaviour without affecting the asymptotic behaviour of PWM estimator.

Because of its simplicity, shift and scale invariance, it is still used in many applications. However, there are some problems to discuss. Firstly, its range limitation $\gamma < 1$ does not allow us to use for strong heavy tail cases. Secondly, the convergence of $\hat{\gamma}_k^{PWM}$ to γ is different for $\frac{1}{2} < \gamma < 1$ and $\gamma < \frac{1}{2}$, and the asymptotic normality is only valid for $\gamma < \frac{1}{2}$ (see de Haan and Ferreira 2006).

2.2.4.3 Asymptotically unbiased EVI estimators with second-order parameter

In this section, four asymptotically bias-reduced estimators from Peng (1998), Hall and Welsh (1985) and Fraga Alves et al. (2003) are introduced briefly. Note that the estimator of second-order parameter by Gomes et al. (2002), which is actually used in our applications is given in Section 3.2.3. The remarkable characteristic of these estimators is that the bias is reduced even if we use a large number of upper order statistics, i.e., for large k, when other classical EVI estimators have a high bias.

We assume that the second-order condition (2.7) is satisfied, which also implies that the first-order condition (2.4) is fulfilled. An asymptotically unbiased estimator

based on the Hill estimator for $\gamma > 0$ (Peng 1998) is given by

$$\hat{\gamma}_k^{UH} = \hat{\gamma}_k^H - \frac{M_k^{(2)} - 2(\hat{\gamma}_k^H)^2}{2\hat{\gamma}_k^H \hat{\rho}_1}(1 - \hat{\rho}_1), \tag{2.17}$$

where $M_k^{(2)}$ is the statistic $M_k^{(\alpha)}$ (2.14) when $\alpha = 2$ and

$$\hat{\rho}_1 = \frac{1}{\log 2} \log \frac{M_{n/(2\log n)}^{(2)} - 2\left(M_{n/(2\log n)}^{(1)}\right)^2}{M_{n/\log n}^{(2)} - 2\left(M_{n/\log n}^{(1)}\right)^2}$$

is the estimator of the second-order parameter $\rho \leq 0$. Similarly, we introduce the unbiased estimator based on the Pickands estimator for $\gamma \in \mathbb{R}$, which is defined by

$$\hat{\gamma}_k^{UP} = \hat{\gamma}_k^P - \frac{\hat{\gamma}_k^P - \hat{\gamma}_{k/4}^P}{1 - 4^{\hat{\rho}_2}}$$

where

$$\hat{\rho}_2 = \frac{1}{\log 2} \log \frac{\hat{\gamma}_{n/(2\log n)}^P - \hat{\gamma}_{n/(4\log n)}^P}{\hat{\gamma}_{n/\log n}^P - \hat{\gamma}_{n/(2\log n)}^P}$$

is the estimator of $\rho \leq 0$. The study of the asymptotic normality and its proof are covered in Peng (1998).

We shall next deal with few other estimators of ρ, which are also built upon the statistics $M_k^{(\alpha)}$ (2.14). Hall and Welsh (1985) provide the estimator of ρ, given by

$$\hat{\rho}_3 = -\left| \log \left| \frac{1/M_{[n^{0.9}]}^{(1)} - 1/M_{[n^{0.5}]}^{(1)}}{1/M_{[n^{0.95}]}^{(1)} - 1/M_{[n^{0.5}]}^{(1)}} \right| \Big/ \log \frac{[n^{0.9}]}{[n^{0.95}]} \right|.$$

When we substitute $\hat{\rho}_3$ into the Peng's asymptotically unbiased estimator (2.17) replacing $\hat{\rho}_1$, we obtain a new estimator for γ. Gomes and Martins (2002) find that this bias-reduced estimator has a high variance, which has the same property as the original estimator using $\hat{\rho}_1$. Fraga Alves et al. (2003) also propose the estimator of second-order parameter ρ, which depends on a tuning parameter $\omega \geq 0$. It is defined as

$$\hat{\rho}_4 \equiv \hat{\rho}_k^{(\omega)} = -\left| \frac{3(T_k^{(\omega)} - 1)}{T_k^{(\omega)} - 3} \right|,$$

where

$$T_k^{(\omega)} = \begin{cases} \dfrac{(M_k^{(1)})^{\omega} - (M_{k/2}^{(2)})^{\omega/2}}{(M_{k/2}^{(2)})^{\omega/2} - (M_{k/6}^{(3)})^{\omega/3}}, & \text{if} \quad \omega > 0, \\ \dfrac{\log(M_k^{(1)}) - \frac{1}{2}\log(M_{k/2}^{(2)})}{\frac{1}{2}\log(M_{k/2}^{(2)}) - \frac{1}{3}\log(M_{k/6}^{(3)})}, & \text{if} \quad \omega = 0. \end{cases}$$

Under adequate general conditions, it is an asymptotically normal estimator of ρ, whenever $\rho < 0$. This means that it exhibits highly stable sample paths as a function of k, while other estimators show their fluctuations. In practice, it is known that ρ-estimators of Gomes et al. (2002) in Chapter 3 and Fraga Alves et al. (2003) work well in practice.

2.2.5 Optimal threshold selection

Recall that an important challenge in the implementation of EVT in practice is to choose a threshold, i.e., tail sample fraction, that will define the tail of the distribution of financial returns. In other words, it is necessary to select the number of upper observations that can be considered as the effective sample size for the extrapolation outside the range of available observations. Thus, successful practical applications of extreme values heavily depend on the determination of optimal k.

Selecting the optimal tail sample fraction k is known as the difficult task in extreme value analysis for two reasons. Firstly, there is no straightforward approach for the optimal selection: see Scarrott and MacDonald (2012) and Echaust and Just (2020) for a review of extreme value threshold. Secondly, there is a bias-variance tradeoff. By choosing a low level of k, few observations are used in estimation that results in a high level of the estimation variance. With a high level of k, the variance is reduced, but at the cost of an increasing bias.

Regarding the implementation of EVT in dynamic extreme VaR and ES estimations, the literature about threshold selection is still scarce for practical cases in which the i.i.d. condition is not appropriate. As far as we are aware, a threshold is selected as a fixed quantile of the empirical data in most cases, for example, McNeil and Frey (2000), Fernandez (2005), Youssef et al. (2015), Bee et al. (2016) and Karmakar and Paul (2019) chose the 90th quantile and Ergen (2015) chose the 92nd and 94th quantiles of the loss distribution as a threshold. In comparison with the

fixed threshold approach, Echaust and Just (2020) used four different optimal tail selection algorithms, which include the path stability method, the automated Eye-Ball method, the minimization of asymptotic mean squared error method and the distance metric method. In order to tackle the problem of selecting the optimal tail sample fraction k, we use the bias correction method of an EVI and an extreme quantile for extreme VaR and ES estimations even if we need more than the first-order condition (2.4). Nonetheless, the bias correction method is preferred because the choice of optimal k is less crucial.

2.3 Use of EVT in finance

It has been found in previous studies such as Embrechts et al. (1997), Danielsson (2011) and McNeil et al. (2015) that the heavy-tailed distributions describe sufficiently well the tail structure of financial data. In this thesis, we use EVT to estimate risk measures VaR and ES. EVT uses only extreme event data, focuses only on the tails, and allows the extrapolation beyond available data, while traditional econometric models focus on the whole distribution at the expense of less consideration in the tails. Thus, EVT could potentially provide better VaR estimates.

There have been a number of literature that discuss the use of EVT for estimating unconditional VaR. Some examples include: Danielsson and de Vries (1997), who propose the estimation of unconditional VaR by a semiparametric method combining historical simulation with parametric estimation of the tails of the return distribution; Danielsson and Morimoto (2000), who apply EVT to estimate unconditional VaR of Japanese financial data and find that the EVT method estimates better VaRs than GARCH-type models in terms of accuracy and stability; Drees (2003), who establishes the asymptotic normality of the extreme quantile estimator for a stationary β-mixing time series and applies to the NASDAQ Composite index to estimate the unconditional VaR; de Haan et al. (2016), who introduce an asymptotically unbiased estimator of extreme quantile for a β-mixing time series and apply to the Dow Jones Industrial Average index to estimate the unconditional VaR; Chavez-Demoulin and Guillou (2018), who propose an alternative asymptotically bias-corrected estimator of the extreme quantile for a β-mixing time series and apply to the S&P500 index to

estimate the unconditional VaR. For estimating unconditional ES, Righi and Ceretta (2015) evaluated unconditional quantile and expectile regression-based models for one-day ahead (daily) ES estimation.

Regarding the extensive literature of the estimations of conditional VaR and ES, they are given in Section 3.1 and Section 4.2, respectively. Most of the EVT methods to estimate conditional VaR and ES are based on the GARCH-EVT framework by the influential paper McNeil and Frey (2000), whereby financial returns are first filtered using an AR-GARCH model, and then the GPD is fitted to the standardized residuals. Note that our proposed bias-reduced EVT method called the GARCH-UGH is also influenced by the GARCH-EVT framework. Merit of using the conditional EVT model is that under a correct specification of the conditional mean and variance, the filtered residuals will be approximately i.i.d., which matches with an assumption of EVT methods.

In studies of the VaR and ES estimations, one should deal with unconditional and conditional return distributions separately. The unconditional VaR and ES are appropriate for the forecast of potential large losses in longer time horizon, for example, when we consider long-term investment decisions. Although the unconditional VaR and ES based on EVT routinely assume that financial returns are independent and identically distributed (i.i.d.), financial institutions often prefer the unconditional VaR to avoid undesirable fluctuation of risk limit widely over time for traders and portfolio managers (Danielsson and de Vries 1997). On the other hand, the conditional VaR and ES are more appropriate for short time horizon when we deal with day-to-day risks and short-term risk management by capturing the dynamics and the vital properties of financial asset returns such as volatility clustering and leptokurtosis. The conditional VaR and ES give the better understanding of the riskiness of the portfolio because the riskiness of the portfolio varies with the changing volatility. Moreover, Diebold et al. (2000) point out that the assumption of i.i.d. data for estimating the unconditional VaR based on EVT is very often violated for conditionally heteroscedastic financial time series. It is our present position that the problem of choosing an appropriate risk measure for the given situation based on theoretical and practical considerations should be separated from the statistical problem of estimating more accurate risk measures.

In this thesis, we focus only on the estimations of conditional VaR and ES. We believe that it is necessary to incorporate dynamic changes in the market to reflect the most updated risk level. Again, EVT is an accurate candidate framework to model the tails of the distribution for extreme market events and hence used in Chapter 3 and 4.

2.3.1 Limitations of EVT in finance

We finally list two limitations that must be tackled to apply EVT to financial returns to estimate dynamic extreme VaR and ES. Firstly, basic EVT models assume that financial realizations from i.i.d. samples, which are not realistic. For that, we first filter the financial returns using an AR-GARCH model and apply EVT models to the approximately i.i.d. residuals. Our bias-reduction procedure will be designed to be robust to departure from the independence assumption, and as such will be able to handle residual dependence present after filtering. Secondly, the most important parameter of EVT models, which is EVI (tail shape prameter), is sensitive to the threshold selection (optimal tail sample fraction). Our approach leads to extreme conditional VaR and ES estimates that are less sensitive to the choice of sample fraction, and hence mitigates the difficulty in selecting the optimal number of observations for estimations.

Chapter 3

Dynamic extreme Value-at-Risk estimation by GARCH-UGH

3.1 Introduction

A major concern in financial risk management is to quantify the risk associated to high-impact, low-probability extreme losses. The most widely known risk measure is Value-at-Risk (VaR), defined as a quantile of the loss distribution. Even though the Basel Committee on Banking Supervision recommends the use of VaR at high levels (see for example Basel Committee on Banking Supervision 2013), it has been criticized several times in the financial literature for two main reasons. First, the VaR only measures the frequency of observations below or above the predictor and not their magnitude: this means that, while it is known that $100(1-\tau)\%$ of losses will be higher than the VaR q_τ at level τ, the VaR alone cannot give any further information about the size of these large losses. Second, the VaR is not a coherent risk measure in the sense of Artzner et al. (1999), because it is not sub-additive in general, meaning that it does not abide by the intuitive diversification principle stating that a portfolio built on several financial assets carries less risk than a portfolio solely consisting of one of these assets. These two weaknesses pushed the Basel Committee to also recommend calculating the Expected Shortfall (or Conditional Value-at-Risk) as a complement or alternative to the VaR. In practice, this is hampered by the fact that the Expected Shortfall is not elicitable in the sense of Gneiting (2011), and therefore the development of a simple backtesting methodology for the Expected Shortfall is

not clear (see Deng and Qiu 2021 for a very recent comprehensive study of backtesting procedures for the Expected Shortfall).

This is why the accurate estimation of VaR is worth pursuing, and is our focus in this paper. An estimation of the unconditional VaR (that is, of the common distribution of returns over time, assumed to be stationary) is appropriate for the estimation of potential large levels of loss over the long term, for example with the goal of making long-term investment decisions. On the other hand, the conditional VaR is more appropriate for day-to-day and short-term risk management by capturing the dynamics and the key properties of financial asset returns such as volatility clustering and leptokurtosis. The estimation of the extreme conditional VaR, on which we focus in this paper, therefore gives a better understanding of the riskiness of the portfolio because this riskiness varies with the changing volatility. Quantifying the risk associated to extreme losses can then be done by estimating an extreme quantile of order $1 - p$, where $p = p(n) \to 0$ as the available sample size n of the data tends to infinity.

There are two main classes of methods to estimate the conditional VaR. Nonparametric historical simulation (HS) relies on observed data directly, and uses the empirical distribution of past losses without assuming any specific distribution, see Danielsson (2011) and McNeil et al. (2015). Although HS is easy to implement, the estimation of extreme quantiles using HS is difficult as the extrapolation beyond observed returns is impossible because this method essentially assumes that one of the observed returns is expected to be the next period return. By contrast, the parametric approach generally refers to the use of an econometric model of volatility dynamics such as, among many others, the Generalized Autoregressive Conditional Heteroskedasticity (GARCH) model of Bollerslev (1986). These models estimate VaRs reflecting the conditional heteroskedasticity of financial data. However, at extreme levels GARCH-type models assuming a normal distribution of the innovation variable tend to underestimate risk because this assumption is not well-suited to the estimation of heavy-tailedness in the conditional returns of financial time series.

To overcome the problems of purely nonparametric or parametric estimations of extreme VaR, McNeil and Frey (2000) propose a two-step approach combining

a GARCH-type model and Extreme Value Theory (EVT), referred to as GARCH-EVT throughout. EVT focuses only on the tails and allows the extrapolation beyond available data, while traditional econometric models focus on the whole distribution with less consideration of the tails. The key idea of the GARCH-EVT method is to estimate the dynamic extreme VaR by first filtering financial time series with a GARCH-type model to estimate the current volatility, and then by applying the EVT method to the standardized residuals for estimating the tails of the residual distribution. This approach has been widely used to estimate the extreme conditional VaR. For instance Byström (2004) and Fernandez (2005) find GARCH-EVT to give accurate VaR estimates for standard and extreme quantiles compared with GARCH-type models and unconditional EVT methods on stock market data collected across the US, Latin America, Europe and Asia. Being a two-step procedure based on GARCH-type filtering, the accuracy of the GARCH-EVT approach has been debated. Chavez-Demoulin et al. (2005) point out that estimates of the extreme conditional VaR via the GARCH-EVT approach are sensitive to the fitting of a GARCH-type model to the dataset in the first step. On the other hand, Furió and Climent (2013) and Jalal and Rockinger (2008) have concluded that there is no evidence of any difference in the final VaR estimates, regardless of the particular GARCH model selected to filter financial data.

Since the debate on filtering, several modifications of the conventional GARCH-EVT method have been suggested in the literature to provide a more accurate calculation of the residuals before applying the EVT method in the second step. Yi et al. (2014) propose a semiparametric version of GARCH-EVT based on quantile regression. Youssef et al. (2015) adapt the FIGARCH, HYGARCH and FIAPARCH models to estimate extreme conditional VaRs for crude oil and gasoline market. Bee et al. (2016) propose an approach called realized EVT where returns are pre-whitened with a high-frequency based volatility model. Zhao et al. (2019) develop hybrid time-varying long-memory GARCH-EVT models by using a variety of fractional GARCH models. To the best of our knowledge, little work has been carried out on the EVT step itself; Ergen (2015) uses the skewed t-distribution that is fitted to the standardized residuals from the GARCH step in order to recover a fully parametric specification. In the context of estimation and inference of unconditional extreme

VaR, bias correction is a key concern and has an extensive history (see *e.g.* Cai et al. 2013 for a review). de Haan et al., 2016 develop a semiparametric bias-reduced estimator of extreme *unconditional* VaR in stationary time series. However, such improvements have not been investigated so far in the specific context of *dynamic* estimation of extreme quantiles of financial time series.

This is the contribution of the present paper. More precisely, in the context of the estimation of the one-step ahead dynamic extreme VaR, we develop a new methodology called GARCH-UGH (standing for Unbiased Gomes-de Haan, after de Haan et al. 2016 and Gomes et al. 2002). The novelty in this methodology is that, instead of applying the Peaks-Over-Threshold (POT) method in the GARCH-EVT approach as in McNeil and Frey (2000), we use an asymptotically unbiased estimator, derived from the work of de Haan et al., 2016, of the extreme quantile applied to the standardized residuals from the GARCH step. We analyze the performance of our approach on four financial time series, which are the Dow Jones, NASDAQ and Nikkei stock indices, and the Japanese Yen/British Pound exchange rate. As we shall illustrate, our results indicate that GARCH-UGH provides substantially more accurate one-step ahead extreme conditional VaRs than either HS, the conventional GARCH-N method (that is, the standard GARCH specification of heteroskedasticity with normal innovations), its GARCH-t analogue where the innovations are Student-t distributed, the GARCH-EVT approach, or simple UGH without filtering, based on the performance of the in-sample and out-of-sample backtestings. In addition, our bias-reduction procedure will be designed to be robust to departure from the independence assumption, and as such will be able to handle residual dependence present after filtering in the first step. Our finite-sample results will also illustrate that the GARCH-UGH method leads to one-step ahead extreme conditional VaR estimates that are less sensitive to the choice of sample fraction, and hence mitigates the difficulty in selecting the optimal number of observations for the estimations. Finally, the computational cost of GARCH-UGH is lower than that of conventional GARCH-EVT: the extreme value step in the GARCH-UGH method is semiparametric with an automatic and fast recipe for the estimations of the one-step ahead extreme conditional VaRs, while the GARCH-EVT method is based on a parametric fit of the Generalized Pareto Distribution (GPD) to the residuals using Maximum

Likelihood Estimation.

The rest of the paper is organized as follows. Section 3.2 presents our proposed framework and methodology. Section 3.3, 3.4 and 3.5 explain the methods of basic VaR estimation approaches and backtesing approaches for the empirical analysis. Section 3.6 first describes the four financial time series used in the empirical analysis, and discusses the performance of our proposed approach through in-sample and out-of-sample traditional and comparative backtestings of one-step ahead extreme VaRs compared to the existing approaches.

3.2 The GARCH-UGH method and framework

3.2.1 Settings

Let p_t be a daily-recorded price for a stock, index or exchange rate, and let $X_t = -\log(p_t/p_{t-1})$ be the negative daily log-return on this price, i.e., financial returns. We assume that the dynamics of X_t are governed by

$$X_t = \mu_t + \sigma_t Z_t, \tag{3.1}$$

where $\mu_t \in \mathbb{R}$ and $\sigma_t > 0$ denote the (conditional) mean and standard deviation, and the innovations Z_t form a strictly stationary white noise process, that is, they are i.i.d. with zero mean, unit variance and common marginal distribution function F_Z. We assume that for each t, μ_t and σ_t are measurable with respect to the $\sigma-$algebra $\mathscr{F}_{t-1}$ representing the information about the return process available up to time $t-1$.

We are concerned with estimating extreme conditional quantiles of these negative log-returns. Recall that for a probability level $\tau \in (0,1)$, the τth unconditional quantile of a distribution F is $q_\tau = \inf\{x \in \mathbb{R} : F(x) \geq \tau\}$. Here we focus on the one-step ahead quantile, that is, the estimation of the conditional extreme quantile of X_{t+1} given $\mathscr{F}_t$, whose order τ tends to 1 as the available sample size n goes to infinity. In this case, by location equivariance and positive homogeneity of quantiles, the one-step ahead conditional quantile (or VaR) of X_{t+1} can be written as

$$q_\tau(X_{t+1} \mid \mathscr{F}_t) = \mu_{t+1} + \sigma_{t+1} q_\tau(Z), \tag{3.2}$$

where $q_\tau(Z)$ is the common τth quantile of the marginal distribution of the innovations Z_t. The problem of estimating $q_\tau(X_{t+1} \mid \mathscr{F}_t)$ can then be tackled by estimating the mean and standard deviation components μ_{t+1} and σ_{t+1} and the unconditional quantile $q_\tau(Z)$. Given estimates $\hat{\mu}_{t+1}$, $\hat{\sigma}_{t+1}$ and $\hat{q}_\tau(Z)$ of these quantities, an estimate of $q_\tau(X_{t+1} \mid \mathscr{F}_t)$ is then

$$\hat{q}_\tau(X_{t+1} \mid \mathscr{F}_t) = \hat{\mu}_{t+1} + \hat{\sigma}_{t+1}\hat{q}_\tau(Z).$$

In calculating this estimate, there are three main difficulties. First, one has to estimate μ_{t+1} and σ_{t+1}, which supposes that an appropriate model and estimation method have to be chosen. Second, the innovations Z_t are unobserved, which means that the estimation of $q_\tau(Z)$ has to be based on residuals following the estimation of μ_{t+1} and σ_{t+1}. A third difficulty is specific to our context: we wish here to estimate a dynamic extreme VaR, that is, a conditional quantile $q_\tau(X_{t+1} \mid \mathscr{F}_t)$ with τ very close to 1. In such contexts, it is well-known that traditional nonparametric estimators become inconsistent (see for example the monographs by Beirlant et al. 2004 and Embrechts et al. 1997), and adapted extrapolation methodologies have to be employed.

Our GARCH-UGH method combines estimation of the mean and standard deviation in a GARCH-type model with a flexible bias-reduced extrapolation methodology for the estimation of $q_\tau(Z)$ ($\tau \uparrow 1$) using the residuals obtained after estimation of the model structure. We describe these two steps successively below.

3.2.2 GARCH step

In order to estimate μ_{t+1} and σ_{t+1}, one should select a particular model in the class (3.1). Many different models for volatility dynamics have been used in the literature of GARCH-EVT approach, as we highlighted in our literature review in Section 3.1 (see also Danielsson 2011; McNeil et al. 2015). Here we use an AR(1) model for the dynamics of the conditional mean, and a parsimonious but effective GARCH(1,1) model for the volatility, as in the original GARCH-EVT approach; this will allow us to subsequently illustrate how improving the second, EVT-based step can result in more accurate estimates of extreme conditional VaR.

We thus model the conditional mean of the series by

$$\mu_t = \phi X_{t-1},$$

for some $\phi \in (-1, 1)$, and the conditional variance of the mean-adjusted series $\epsilon_t = X_t - \mu_t$ by

$$\sigma_t^2 = \kappa_0 + \kappa_1 \epsilon_{t-1}^2 + \kappa_2 \sigma_{t-1}^2,$$

where $\kappa_0, \kappa_1, \kappa_2 > 0$. Necessary and sufficient conditions for the stationarity of a model following GARCH(1,1) dynamics are given in Chapter 2 of Francq and Zakoïan (2010); the condition $\kappa_1 + \kappa_2 < 1$ is a simple sufficient condition guaranteeing stationarity. The model is therefore the AR(1)-GARCH(1,1) model

$$X_t = \mu_t + \sigma_t Z_t, \text{ with } \mu_t = \phi X_{t-1} \text{ and } \sigma_t^2 = \kappa_0 + \kappa_1 (X_{t-1} - \mu_{t-1})^2 + \kappa_2 \sigma_{t-1}^2. \quad (3.3)$$

In Equation (3.3), the innovations Z_t are i.i.d. with zero mean, unit variance.

In order to make one-step ahead predictions at time t, we fix a memory n so that at the end of time t, the financial data consist of the last n negative log-returns X_{t-j}, for $0 \leq j \leq n-1$. We then fit the AR(1)-GARCH(1,1) model to the data $(X_{t-n+1}, \ldots, X_{t-1}, X_t)$ using Gaussian Quasi-Maximum Likelihood Estimation (QMLE), that is, by maximizing the likelihood constructed by assuming that the innovations Z_t are i.i.d. Gaussian with zero mean and unit variance. The R package `rugarch` (Galanos and Kley, 2022) has been used for the estimation. While of course the innovations Z_t will not be Gaussian in general (and indeed in our UGH step we shall assume that they are heavy-tailed), the QMLE method yields a consistent and asymptotically normal estimator, see for example Francq and Zakoïan (2004) for a theoretical analysis. One may also put a strong heavy-tailed parametric specification on Z_t, such as assuming that they are location-scale Student distributed; this was tried in our analysis of financial log-returns but did not improve results substantially.

Let $(\hat{\phi}, \hat{\kappa}_0, \hat{\kappa}_1, \hat{\kappa}_2)$ be the Gaussian QMLE estimates. Choosing sensible starting values for $\hat{\epsilon}_{t-n}^2$ and $\hat{\sigma}_{t-n}^2$ (for example, constant values as in Section 7.1 of Francq and Zakoïan (2010)), estimates of the conditional mean and the conditional standard

deviation, $(\hat{\mu}_{t-n+1}, \ldots, \hat{\mu}_{t-1}, \hat{\mu}_t)$ and $(\hat{\sigma}_{t-n+1}, \ldots, \hat{\sigma}_{t-1}, \hat{\sigma}_t)$ respectively, can be calculated from Equation (3.3) recursively. This leads to the residuals

$$(\hat{Z}_{t-n+1}, \ldots, \hat{Z}_t) = \left(\frac{X_{t-n+1} - \hat{\mu}_{t-n+1}}{\hat{\sigma}_{t-n+1}}, \ldots, \frac{X_t - \hat{\mu}_t}{\hat{\sigma}_t} \right).$$

We end this step by calculating the estimates of the conditional mean and standard deviation for time $t+1$, which are the obvious one-step ahead forecasts, as follows:

$$\begin{aligned} \hat{\mu}_{t+1} &= \hat{\phi} X_t, \\ \hat{\sigma}_{t+1} &= \sqrt{\hat{\kappa}_0 + \hat{\kappa}_1 \hat{\epsilon}_t^2 + \hat{\kappa}_2 \hat{\sigma}_t^2}, \end{aligned}$$

where $\hat{\epsilon}_t = X_t - \hat{\mu}_t$. In summary, this first GARCH step of the method consists in fitting an AR(1)-GARCH(1,1) model to the negative log-returns at a certain past time horizon n (not too small so that the method produces reasonable results, and not too large so that the AR-GARCH model is believable over this time period), using a Gaussian QMLE, leading to forecasts $\hat{\mu}_{t+1}$ and $\hat{\sigma}_{t+1}$ and standardized residuals $\hat{Z}_{t-j}$, $0 \leq j \leq n-1$.

3.2.3 UGH step

With standardized residuals at our disposal, we can now discuss the estimation of the extreme quantile $q_\tau(Z)$ of the innovations Z_t, for $\tau \uparrow 1$. The residuals $\hat{Z}_{t-j}$, $0 \leq j \leq n-1$, approximate the true unobservable Z_{t-j}. Assume that the underlying distribution of these Z_{t-j} is heavy-tailed, that is (see Theorem 1.2.1 p.19 and Corollary 1.2.10 p.23 in de Haan and Ferreira 2006 and also Section 2.1.4)

$$\lim_{t \to \infty} \frac{U(tz)}{U(t)} = z^\gamma, \ \forall z > 0, \ \text{ where } U(t) = q_{1-t^{-1}}(Z). \tag{3.4}$$

In other words, we assume the tail of the innovations to be approximately Pareto, with the so-called extreme value index γ tuning how heavy the tail is. This assumption is ubiquitous in actuarial and financial risk management (see e.g. p.9 of Embrechts et al. 1997, p.1 of Resnick 2007 and also Section 2.1.3). This makes it possible to construct extrapolated extreme quantile estimators: the classical Weissman quantile estimator (see Weissman 1978 and Section 2.2.2) of a quantile $q_\tau(Z) = q_{1-p}(Z)$

with $p = 1 - \tau$ close to 0 (meaning, in mathematical terms, that $p = p(n) \to 0$ as $n \to \infty$) is then

$$\overline{q}_{1-p}(Z) = \left(\frac{k}{np}\right)^{\overline{\gamma}_k} Z_{n-k,n} \tag{3.5}$$

where $Z_{1,n} \leq Z_{2,n} \leq \cdots \leq Z_{n,n}$ are the order statistics from $Z_{t-n+1}, \ldots, Z_t$ and $\overline{\gamma}_k$ is a consistent estimator of γ. The tuning parameter k denotes the effective sample size for the estimation: this parameter should be chosen not too small, so that the variance of the estimator is reasonable, but also not too large so that the bias coming from the use of the extrapolation relationship (3.4) does not dominate. The most common estimator $\overline{\gamma}_k$ of γ is the Hill estimator (introduced in Hill 1975 and see Section 2.2.1):

$$\overline{\gamma}_k = \overline{\gamma}_k^H = \frac{1}{k}\sum_{i=1}^{k} \log Z_{n-i+1,n} - \log Z_{n-k,n}. \tag{3.6}$$

The Hill and Weissman estimators can be shown to be asymptotically Gaussian under suitable conditions on $k = k(n)$ (see for example Chapters 3 and 4 in de Haan and Ferreira 2006). A reasonable idea to define an estimator of $q_{1-p}(Z)$ in our context is then to use the estimators defined in Equations (3.5) and (3.6) with the order statistics of the residuals, $\hat{Z}_{n-j,n}$, in place of the unobservable $Z_{n-j,n}$.

The choice of the parameter k requires solving a bias-variance tradeoff for which there is no straightforward approach (see Section 2.2.5). Indeed, with a low k, the estimators use observations that are very informative about the extremes, but their low number results in a high variance. With a high k, the variance is reduced, but at the cost of taking into account observations that are further into the bulk of the distribution and thus carry bias. One possible way to make the choice of k easier is to work on correcting this bias. This can be done under the following so-called second-order condition on U:

$$\lim_{t\to\infty} \frac{1}{A(t)}\left(\frac{U(tz)}{U(t)} - z^\gamma\right) = z^\gamma \frac{z^\rho - 1}{\rho}, \ \forall z > 0, \tag{3.7}$$

where $\rho \leq 0$ is called the second-order parameter and A is a positive or negative function converging to 0 at infinity, such that $|A|$ is regularly varying with index ρ. See Equation (6.35) p.341 in Embrechts et al., 1997 and Section 2.1.5 and, in our parametrization, Theorem 2.3.9 p.48 in de Haan and Ferreira, 2006. The function A

therefore controls the rate of convergence in Equation (3.4): the larger $|\rho|$ is, the faster $|A|$ converges to 0, and the smaller the error in the approximation of the right tail of U by a Pareto tail is. This makes it possible to precisely quantify the bias of the Hill and Weissman estimators, and to correct for this bias by estimating the function A and the parameter ρ. This results in bias-corrected Hill and Weissman estimators for which the selection of k is typically much easier because their performance is much more stable.

Our idea in this second, UGH step is to apply such bias-corrected estimators constructed in de Haan et al. (2016) (and built on second-order parameter estimators of Gomes et al. 2002, hence the name UGH, for Unbiased Gomes-de Haan) to our residuals obtained from the GARCH step. Our estimator of ρ motivated by Gomes et al. (2002) is

$$\hat{\rho}_k^{(\alpha)} = (s^{(\alpha)})^{\leftarrow}(S_k^{(\alpha)}).$$

Here $\alpha \notin \{1/2, 1\}$ is a positive tuning parameter, $s^{(\alpha)^{\leftarrow}}$ denotes the generalized (left-continuous) inverse of the function

$$s^{(\alpha)}(\rho) = \frac{\rho^2(1-(1-\rho)^{2\alpha} - 2\alpha\rho(1-\rho)^{2\alpha-1})}{(1-(1-\rho)^{\alpha+1} - (\alpha+1)\rho(1-\rho)^{\alpha})^2},$$

and we set

$$S_k^{(\alpha)} = \frac{\alpha(\alpha+1)^2\Gamma^2(\alpha)}{4\,\Gamma(2\alpha)} \frac{R_k^{(2\alpha)}}{(R_k^{(\alpha+1)})^2},$$

$$\text{with } R_k^{(\alpha)} = \frac{M_k^{(\alpha)} - \Gamma(\alpha+1)(M_k^{(1)})^{\alpha}}{M_k^{(2)} - 2(M_k^{(1)})^2} \text{ and } M_k^{(\alpha)} = \frac{1}{k}\sum_{i=1}^{k}(\log \hat{Z}_{n-i+1,n} - \log \hat{Z}_{n-k,n})^{\alpha}.$$

The version of $\hat{\rho}_k^{(\alpha)}$ with the true innovations Z_k instead of the residuals is known to be consistent under a so-called third-order condition which further strengthens (3.7). Here we choose $\alpha = 2$ since, on fully observed data, this appears to yield the smallest mean squared error following the numerical simulations of Gomes et al. (2002). This results in the estimator

$$\hat{\rho}_k^{(2)} = \frac{-4 + 6S_k^{(2)} + \sqrt{3S_k^{(2)} - 2}}{4S_k^{(2)} - 3}$$

provided $2/3 \leq S_k^{(2)} \leq 3/4$, where

$$S_k^{(2)} = \frac{3}{4} \frac{[M_k^{(4)} - 24(M_k^{(1)})^4][M_k^{(2)} - 2(M_k^{(1)})^2]}{[M_k^{(3)} - 6(M_k^{(1)})^3]^2}.$$

[This expression corrects a typo in p.375 of de Haan et al. (2016).] It is clear that $\hat{\rho}_k^{(2)}$ does not exist if $S_k^{(2)} \notin [2/3, 3/4]$. In practice, we select the value of k in this estimator by setting

$$k_\rho = \sup \left\{ k : k \leq \min \left(m - 1, \frac{2m}{\log \log m} \right) \text{ and } \hat{\rho}_k^{(2)} \text{ exists} \right\}. \tag{3.8}$$

Here m is the number of positive observations in the sample. The intuition is that even though this estimator of ρ requires a choice of k, this choice should be different from its counterpart used in the estimation of γ, and indeed intuitively the value of k in the estimator of ρ should be rather high in order to allow the methodology to identify the bias coming from including observations belonging to the bulk of the distribution (which correspond to a high k). We then estimate ρ by $\hat{\rho}_{k_\rho} = \hat{\rho}_{k_\rho}^{(2)}$. If the set on the right-hand side of (3.8) is empty, we define $\hat{\rho}_{k_\rho} = -1$ as recommended in p.117 of Section 4.5.1 in Beirlant et al. (2004), in p.212-215 in Gomes et al., 2000 and in p.195 of Section 6.6 in Reiss and Thomas, 2007. The choice $\hat{\rho}_{k_\rho} = -1$ is an *ad hoc* compromise between bias reduction and variability of the estimators; note that the estimation of the constant ρ is known to be a difficult problem in finite samples (see Gomes et al. 2009, p.298 and Goegebeur et al. 2010, p.2638, where it is seen that estimators of ρ typically have a low rate of convergence). The parameter γ is then estimated using the residual-based version of the bias-corrected Hill estimator introduced in de Haan et al. (2016):

$$\hat{\gamma}_{k,k_\rho} = \hat{\gamma}_k^H - \frac{M_k^{(2)} - 2(\hat{\gamma}_k^H)^2}{2\hat{\gamma}_k^H \hat{\rho}_{k_\rho}(1 - \hat{\rho}_{k_\rho})^{-1}}. \tag{3.9}$$

The addend in the right-hand side corresponds to the bias correction. We now have all the necessary ingredients to define our residual-based, bias-corrected estimator

of unconditional extreme quantiles of Z:

$$\hat{q}_{1-p}(Z) = \left(\frac{k}{np}\right)^{\hat{\gamma}_{k,k_\rho}} \hat{Z}_{n-k,n} \times \left(1 - \frac{\left[M_k^{(2)} - 2(\hat{\gamma}_k^H)^2\right][1-\hat{\rho}_{k_\rho}]^2}{2\hat{\gamma}_k^H \hat{\rho}_{k_\rho}^2}\left[1 - \left(\frac{k}{np}\right)^{\hat{\rho}_{k_\rho}}\right]\right). \tag{3.10}$$

This corresponds to a slightly different version of the estimator in Section 4.3 of de Haan et al. (2016), given later by Chavez-Demoulin and Guillou (2018), who pointed out a mistake in the analysis of de Haan et al. (2016). The use of $\hat{\gamma}_{k,k_\rho}$ rather than the Hill estimator in the extrapolation will correct the bias due to the estimation of the extreme value index; the multiplier corrects the bias specifically due to the use of the Pareto distribution for the extrapolation of extreme quantiles. The versions of these estimators for fully observed data work when this data is weakly serially dependent, as shown in Chavez-Demoulin and Guillou (2018). As such, our proposed method will be robust to the presence of residual dependence after filtering and to model misspecification in the sense of Hill (2015). We shall also show that the choice of k for this estimator is not as crucial in finite samples as for the traditional Hill estimator, because this estimator has reasonably good performance across a large range of values of k.

Note that if the Z_{t-j}, $0 \leq j \leq n-1$ are independent, then Theorem 4.2 of de Haan et al. (2016) suggests that

$$\frac{\sqrt{k}}{\log(k/np)}\left(\frac{\hat{q}_{1-p}(Z)}{q_{1-p}(Z)} - 1\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{\gamma^2}{\rho^2}(\rho^2 + (1-\rho)^2)\right).$$

A standard Gaussian 95% asymptotic confidence interval for the extreme quantile $q_{1-p}(Z)$, $p \downarrow 0$ is then given by

$$\left[\hat{q}_{1-p}(Z)\left(1 \pm \frac{1.96}{\sqrt{k}/\log(\frac{k}{np})} \times \sqrt{\frac{\hat{\gamma}_{k,k_\rho}^2}{\hat{\rho}_{k_\rho}^2}(\hat{\rho}_{k_\rho}^2 + (1-\hat{\rho}_{k_\rho})^2)}\right)\right].$$

This asymptotic Gaussian confidence interval is easy to implement, but of course its validity relies on assuming that the negative log-returns are correctly filtered using the AR-GARCH model.

3.2.4 Summary and output of the GARCH-UGH method

The GARCH-UGH approach may be briefly summarized by the following two successive steps:

1. GARCH step: based on n previous observations at time t, fit an AR(1)-GARCH(1,1) model to the negative daily log-returns data using a Gaussian QMLE. Obtain $\hat{\mu}_{t+1}$ and $\hat{\sigma}_{t+1}$ using the fitted model and compute standardized residuals. See Section 3.2.2 for full details on this GARCH step, similar to the first step in McNeil and Frey, 2000.

2. UGH step: use these standardized residuals as proxies for the true unobserved innovations Z_{t-j}, $0 \leq j \leq n-1$, to construct the asymptotically unbiased tail quantile estimator $\hat{q}_{1-p}(Z)$ in (3.10). See Section 3.2.3 for full details on this UGH step which is a different, residual-based new version of the procedure of de Haan et al., 2016.

Combining the two steps results in the final GARCH-UGH estimator

$$\hat{q}_\tau(X_{t+1} \mid \mathscr{F}_t) = \hat{\mu}_{t+1} + \hat{\sigma}_{t+1}\hat{q}_\tau(Z),\ \tau = 1-p,\ p \text{ close to } 0.$$

The goal of our real data analysis is to examine the finite-sample performance of this estimator for low exceedance probabilities, that is, τ close to 1.

3.3 VaR estimation methods for comparison

In this section, we will explain other five methods for VaR estimation, which are used to compare with our proposed GARCH-UGH approach in the empirical analysis in Section 3.6. Some methods are already described in Section 1.3 briefly.

Historical simulation (HS) method - the nonparametric HS method is based on the observed data instead of making distributional assumptions about the financial returns. Past returns are used to predict future returns and hence its VaR is simply the empirical quantile of the series X_t at the desired quantile level. As mentioned in Section 3.1, HS is easy to implement but the estimation of extreme quantiles using

HS is difficult as the extrapolation beyond observed returns is impossible. Moreover, for VaR estimation using HS, the inclusion or exclusion of one or even more observations of the sample can cause large fluctuations in the VaR estimate, while no guidelines exist for assessing which estimate is better (Danielsson and de Vries 1997).

GARCH-N (normal) method - this method uses the same filtering step as explained in Section 3.2.2, but assumes in the quantile estimation step also that the innovations Z_t are i.i.d. $\mathcal{N}(0,1)$. The extreme conditional VaR is then calculated as

$$\hat{q}_\tau(X_{t+1} \mid \mathscr{F}_t) = \hat{\mu}_{t+1} + \hat{\sigma}_{t+1}\Phi^{-1}(\tau),$$

where Φ is the cumulative distribution function of the standard normal distribution. This approach is also called as normal GARCH (see Danielsson 2011 and Ergen 2015).

GARCH-t method - this method again uses the filtering step of Section 3.2.2, but assumes in the extreme quantile estimation step that the standardized residuals from the GARCH step are Student-t distributed. With the two-step framework, the extreme conditional VaR is calculated as

$$\hat{q}_\tau(X_{t+1} \mid \mathscr{F}_t) = \hat{\mu}_{t+1} + \hat{\sigma}_{t+1}T_\nu^{-1}(\tau),$$

where T_ν is the cumulative distribution function of the standard Student-t distribution with ν degrees of freedom. This corresponds to the two-step estimation method discussed on p.1017 of Ergen (2015) with skewed normal and Student-t as alternatives.

UGH (Unbiased Gomes-de Haan) method - this method applies the UGH step directly to the series X_t without filtering (see Section 3.2.3).

GARCH-EVT method - the conventional GARCH-EVT method as described in McNeil and Frey (2000). This consists, first, in the same filtering step as described in Section 3.2.2. Standardized residuals are then recorded and a Generalized Pareto distribution (GPD) is fitted using a maximum likelihood estimator, thus producing a VaR estimate $\hat{q}_\tau(Z)$. This method therefore differs from ours as far as the extreme

value step is concerned.

Under an approximately i.i.d. condition after filtering step, we consider the distribution function of excess $Y = Z - u$ over a fixed high threshold u. The corresponding excess distribution above the threshold u is given by

$$F_u(y) = P(Y = Z - u \leq y | Z > u) = \frac{F(y+u) - F(u)}{1 - F(u)}, \quad y \geq 0.$$

The famous Pickands (1975) shows that the GPD occurs naturally as the limit distribution of the scaled excesses of i.i.d. random variables over high thresholds. The excesses Y from a fixed high threshold u follow a GPD $Y = Z - u \sim GPD(\gamma, \beta)$ if

$$F_u(y) \approx GPD_{\gamma,\beta}(y) = \begin{cases} 1 - \left(1 + \frac{\gamma y}{\beta}\right)^{-1/\gamma}, & \gamma \neq 0, \\ 1 - \exp\left(-\frac{y}{\beta}\right), & \gamma = 0, \end{cases}$$

where $\beta > 0$ is a scale parameter, γ is the EVI with the support $y \geq 0$ when $\gamma \geq 0$ and $0 \leq y \leq -\beta/\gamma$ when $\gamma < 0$.

In practice, we fix the number of data in the tail to be k where $k < n$ and use the proportion of tail data k/n. For a probability level $\tau = 1 - p$, the estimator of unconditional extreme quantiles of Z is given by

$$\hat{q}_\tau(Z) = \hat{Z}_{n-k,n} + \frac{\hat{\beta}}{\hat{\gamma}}\left(\left(\frac{p}{k/n}\right)^{-\hat{\gamma}} - 1\right), \quad \hat{\gamma} \neq 0.$$

Then, the one-step ahead conditional quantile (VaR) based on GARCH-EVT is obtained by substituting $\hat{q}_\tau(Z)$ above into the Equation (3.2).

3.4 Traditional VaR backtesting

Recall that backtesting is carried out to examine the accuracy of the one-step ahead extreme conditional VaR estimators provided by each approach. In this section, we consider the traditional backtestings, which can be viewed as a model verification. They examine whether the estimates of VaR under a certain model match with the unknown true values of VaR

Traditional VaR backtesting compares the ex-ante VaR estimates $\hat{q}_\tau(X_t \mid \mathscr{F}_{t-1})$ with the ex-post realized negative log-returns in a time window W_T, with a VaR

violation at time t said to occur whenever $x_t > \hat{q}_\tau(X_t \mid \mathscr{F}_{t-1})$. Define a hit sequence of VaR violations as $I_t = \mathbb{1}\{x_t > \hat{q}_\tau(X_t \mid \mathscr{F}_{t-1})\}$. If a VaR estimation method is accurate, then the sequence (I_t) should approximately be an independent sequence of Bernoulli variables with success probability $p = 1 - \tau$. Both the distributional and independence properties are equally important. A VaR estimation method with too few VaR violations will tend to overestimate risk and therefore to be excessively conservative in financial terms, while too many VaR violations mean that risk is underestimated, leading to insufficient provision of capital and therefore potential insolvency in case of large losses. Besides, a violation of the independence property typically arises when there is a clustering of VaR violations.

One common criticism of traditional VaR backtesting is that it only takes into account the number of VaR violations and not their size, and may be misleading (Bellini et al. 2019). Moreover, Holzmann and Eulert (2014) show that traditional VaR backtesting is insensitive with respect to the increasing information.

3.4.1 Unconditional coverage test

In order to test the distributional assumption, we use the unconditional likelihood ratio coverage test proposed by Kupiec (1995), also known as Kupiec test or POF test, for Proportion Of Failures: fix a time window W_T, let $N = \sum_{t \in W_T} I_t$ be the observed number of VaR violations over W_T and p be the theoretical violation rate. The Kupiec test statistic is the likelihood ratio (LR) statistic given by

$$\mathrm{LR}_{\mathrm{uc}} = -2\log\{p^N(1-p)^{T-N}\} + 2\log\left\{\left(\frac{N}{T}\right)^N\left(1-\frac{N}{T}\right)^{T-N}\right\}.$$

Under the null hypothesis that the I_t are independent and Bernoulli distributed with success probability p, the test statistic $\mathrm{LR}_{\mathrm{uc}}$ is asymptotically χ^2 distributed with 1 degree of freedom. The Kupiec test rejects this null hypothesis with asymptotic type I error α when $\mathrm{LR}_{\mathrm{uc}} > \chi^2_{1,1-\alpha}$, where $\chi^2_{1,1-\alpha}$ is the $(1-\alpha)-$quantile of the χ^2 distribution with 1 degree of freedom. As the number of VaR violations get closer to the expected ones, the probability values from the hypothesis test increases giving more comfort for the performance of the model.

3.4.2 Independence and Conditional coverage tests

To test the independence property, we use another likelihood ratio test called the conditional coverage test, proposed in Christoffersen (1998) and also known as the Christoffersen test and joint test. This test is based on testing for first-order Markov dependence, with the test statistic being given by

$$\mathrm{LR}_{\mathrm{cc}} = -2\log\{p^N(1-p)^{T-N}\} + 2\log\{\hat{\pi}_{00}^{N_{00}}\hat{\pi}_{01}^{N_{01}}\hat{\pi}_{10}^{N_{10}}\hat{\pi}_{11}^{N_{11}}\}.$$

Here $N_{ij} = \sum_{t\in W_T} \mathbb{1}\{I_{t+1} = j, I_t = i\}$ and $\hat{\pi}_{ij} = N_{ij}/(N_{i0} + N_{i1})$. Under the null hypothesis that the sequence (I_t) is independent and identically distributed as Bernoulli with parameter p, the test statistic $\mathrm{LR}_{\mathrm{cc}}$ is asymptotically χ^2 distributed with 2 degrees of freedom, and the conditional coverage test then rejects this null hypothesis with asymptotic type I error α when $\mathrm{LR}_{\mathrm{cc}} > \chi^2_{2,1-\alpha}$ (the $(1-\alpha)-$quantile of the χ^2 distribution with 2 degrees of freedom). Strictly speaking the conditional coverage test only assesses departure from either independence or stationarity, but in fact the test statistic is the sum of the unconditional coverage test statistic $\mathrm{LR}_{\mathrm{uc}}$ and a likelihood ratio test statistic of independence $\mathrm{LR}_{\mathrm{ind}}$:

$$\begin{aligned}\mathrm{LR}_{\mathrm{cc}} &= \mathrm{LR}_{\mathrm{uc}} + \mathrm{LR}_{\mathrm{ind}}\\ \text{with } \mathrm{LR}_{\mathrm{ind}} &= -2\log\left\{\left(\frac{N}{T}\right)^N\left(1-\frac{N}{T}\right)^{T-N}\right\} + 2\log\{\hat{\pi}_{00}^{N_{00}}\hat{\pi}_{01}^{N_{01}}\hat{\pi}_{10}^{N_{10}}\hat{\pi}_{11}^{N_{11}}\}.\end{aligned}$$

The quantity $\mathrm{LR}_{\mathrm{ind}}$ is nothing but a likelihood ratio test statistic of independence versus nontrivial first-order Markov dynamics of the sequence (I_t), which rejects independence of (I_t) provided $\mathrm{LR}_{\mathrm{uc}} > \chi^2_{1,1-\alpha}$. Since $\chi^2_{2,1-\alpha} = \chi^2_{1,1-\alpha} + \chi^2_{1,1-\alpha}$, checking stationarity and independence via the pair of test statistics $(\mathrm{LR}_{\mathrm{uc}}, \mathrm{LR}_{\mathrm{ind}})$ is exactly equivalent to checking them via the unconditional and conditional coverage tests. We therefore use both the unconditional and conditional coverage tests to assess the performance of our dynamic extreme VaR estimators. This constitutes a backtesting approach in the spirit of the one suggested by the Basel Committee on Banking Supervision.

3.4.3 Other tests

The other main coverage test is the Basel Committee's Traffic Light coverage test (Basel Committee on Banking Supervision 2019), which is first introduced in 1996. It counts the number of VaR violations for a certain probability level and classifies the model into three backtesting zones, distinguished by colors into a hierarchy of responses. The light zones are as follows.

- Green zone, which suggests there is no problem with the accuracy of a bank's model.
- Yellow zone, which indicates there is no definitive conclusion of accuracy of a bank's model. It is generally deemed more likely for inaccurate models than for accurate models.
- Red zone, which corresponds to a result that there is almost certainly a problem with a bank's risk model.

The color of the zone determines the amount of additional capital charges required from green to red being the most punitive. The table of boundaries of three zones and corresponding supervisory response based on a sample 250 observations is given in p.82 of Basel Committee on Banking Supervision (2019).

While the conditional coverage test is often used in practice, it has the disadvantage that it only examines the independence of the first VaR violation, and does not test whether the number of days between the VaR violations and realized returns are independent over time. The dynamic quantile test proposed by Engle and Manganelli (2004) overcomes this problem by using both values and series of VaR violations. The concept of this test is that VaR violation at time t should not depend on VaR violations, VaR or any information set $\mathscr{F}_{t-1}$ available at time $t-1$.They use a linear regression model that links current VaR violations to past ones. Define an auxiliary hit sequence of VaR violations as $Hit_t(p) = I_t(p) - p$. In the test, the regression model is estimated with

$$Hit_t(p) = \beta_0 + \sum_{i=1}^{q} \beta_i Hit_{t-i}(p) + \beta_{q+1}\hat{q}_\tau(X_t \mid \mathscr{F}_{t-1}) + \sum_{j=1}^{n} \beta_{q+j+1} X_{jt} + \epsilon_t,$$

where X_j are explanatory variables contained in $\mathscr{F}_{t-1}$. The null hypothesis for this test is that

$$H_0 : \beta_i = 0, i = 0, 1, 2, ..., q + n + 1.$$

This H_0 means that the sequence of Hit_t is uncorrelated with any information from the set $\mathscr{F}_{t-1}$ when the 1-step ahead VaR is estimated, which implies, in particular, that the current VaR violations are uncorrelated with past VaR violations. Under the true H_0, the Wald test statistic of the dynamic quantile test (see Engle and Manganelli 2004) has an asymptotic $\chi^2_{q+n+2,1-\alpha}$ distribution. This test is useful for identifying an incorrect VaR estimation, which is not rejected by, for example, tests of Kupiec and Christoffersen.

3.5 Comparative VaR backtesting (Diebold-Mariano test)

Evaluating a sequence of risk measure estimates using a certain method is different from comparing estimation methods. Recall that the comparative backtesting is better suited for model comparison on the basis of forecasting accuracy while traditional backtesting explained in Section 3.4 is viewed as a model verification. In practice (also illustrated in Section 3.6), there are cases when traditional backtesting methods do not yield definitive answers because the estimation methods are all accepted or all rejected. The comparative backtestings enable to conduct direct comparisons of estimation methods when traditional backtestings are not working efficiently.

Recall also from Section 1.2.2 that Gneiting (2011) states a risk measure is elicitable if it admits a strictly consistent scoring function. It is strictly consistent for a specific risk measure if the risk measure can be obtained by minimizing the expected value of the score (see Section 4.5.1 for the details). Indeed, VaR is the minimizers of the expected value of an appropriate piecewise linear score (see for example, Bellini and Di Bernardino 2015 and Nolde and Ziegel 2017). In the financial literature, the existence of scoring function gives a natural way to compare the accuracy of two different estimation models, i.e., to test the comparative hypothesis, which states one estimation model is better than another, by means of the Diebold-Mariano test on the difference of two realized scores.

The idea of comparative VaR backtesting is to reject the estimation method if the realized scores of VaR is too high. In order to compare the estimation performances of two models, say competing and benchmark models, and decide which one is better, we use the comparative version of the traffic light approach (i.e. three-zone approach) in the Basel III for the VaR (Basel Committee on Banking Supervision 2019) proposed by Fissler and Ziegel (2016) and Nolde and Ziegel (2017), which is based on the Diebold-Mariano (DM) test (Diebold and Mariano 1995).

In this comparative backtesting, we consider the following two hypotheses:

H_0^- : The competing model predicts at least as well as the benchmark model,

H_0^+ : The competing model predicts at most as well as the benchmark model.

The null hypothesis H_0^- is an analogue of H_0 of traditional backtesting but adapted to a comparative setting. The other hypothesis H_0^+ is more conservative in the sense that a backtest is passed if we can reject H_0^+. By this hypothesis, we can explicitly control the type I error of accepting an inferior competing model over a benchmark model.

For a sequence of VaR estimates, $\hat{q}_{\tau,1}, \hat{q}_{\tau,2}, ..., \hat{q}_{\tau,N}$, and corresponding realized returns $x_1, x_2, ..., x_N$, the realized VaR scores $S^1_{\text{VaR}}(\hat{q}_{\tau,N}, x_N)$ are formed for a competing model. For a benchmark model, the same process is applied leading to $S^2_{\text{VaR}}(\hat{q}_{\tau,N}, x_N)$. The comparative VaR backtesting treats S_{VaR} as a loss function and forms the t-statistic based on the DM test as follows:

$$DM = \frac{\sqrt{N}\bar{d}}{\hat{\sigma}_N}, \quad \bar{d} = \frac{1}{N}\sum_{t=1}^{N}(S^1_{\text{VaR}}(\hat{q}_{\tau,t}, x_t) - S^2_{\text{VaR}}(\hat{q}_{\tau,t}, x_t)), \tag{3.11}$$

where $\bar{d}$ is the sample mean of the loss differential of VaR estimates between the competing model (Model 1) and the benchmark model (Model 2), and $\hat{\sigma}_N$ is a suitable estimate of the asymptotic standard deviation of $\bar{d}$. Under proper mixing conditions, the test statistic is asymptotically standard normal $N(0,1)$; see Diebold and Mariano (1995) and Holzmann and Eulert (2014).

An estimation of the long-run variance σ_N^2 is a delicate mission. In fact, σ_N^2 can be expressed as the spectral density of the loss differential $d_t = S^1_{\text{VaR}}(\hat{q}_{\tau,t}, x_t) -$

$S^2_{\text{VaR}}(\hat{q}_{\tau,t}, x_t)$ at frequency 0. We use the suitable heteroskedasticity and autocorrelation consistent (HAC) estimator to estimate the variance of $\bar{d}$ (3.11) as the form of autocorrelation and heteroskedasticity is unknown. See for example, Newey and West (1994), Andrews (1991) and Zeileis (2004) for the HAC estimator in the econometric literature.

The HAC estimator concerns with the estimation of covariance matrix of parameter estimators in the linear model. We consider the linear regression model

$$d_i = r_i^T \beta + u_i, \quad i = 1, ..., N,$$

where the loss differential d_i is dependent variable, r_i is the regressor that is **1** in our case with coefficient vector β and error term u_i. Under suitable regularity conditions, the coefficients β are consistently estimated by Ordinary Least Squares (OLS), giving the well-known estimator

$$\hat{\beta} = (R^T R)^{-1} R^T d,$$

with R in the form of $N \times 1$ vector in our case and their covariance matrix is expressed as

$$(\hat{\sigma}^2_N =)\text{Var}(\hat{\beta}) = \left(\frac{1}{N} R^T R\right)^{-1} \frac{1}{N} \Phi \left(\frac{1}{N} R^T R\right)^{-1}, \tag{3.12}$$

where $\Phi = \frac{1}{N} R^T \text{Var}(u) R$ is the covariance matrix of the estimating functions $V_i(\beta) = r_i(d_i - r_i^T \beta)$. In the general linear model where we assume independent and homoskedastic errors with zero mean and variance σ^2, $\text{Var}(\hat{\beta})$ can be consistently estimated by plugging in the usual OLS estimator $\hat{\sigma}^2$. In our case when the assumption of independence and/or homoskedasticity is violated (or unknown), using the OLS estimator will lead to biased estimator of $\text{Var}(\hat{\beta})$. Use of the HAC estimator solves this problem by plugging in the estimator of Φ into the Equation (3.12), which is consistent in the presence of autocorrelation and heteroskedasticity. An estimate $\hat{\Phi}$ by HAC is given as

$$\hat{\Phi} = \frac{1}{N} \sum_{i,j=1}^{N} w_{|i-j|} \hat{V}_i \hat{V}_j^T, \tag{3.13}$$

where V_i is the estimating function and $w = (w_0, w_1, ..., w_{N-1})$ is the vector of the

weights. It is assumed that the autocorrelations decrease as the lag $\ell = |i - j|$ increases, which implies the weights decrease as the lag increases. It is obvious from the Equation (3.13) that the appropriate choice of the vector of weights is essential for the estimator $\hat{\Phi}$. See different choices of weights given in Zeileis (2004). In our research, we use the linearly decaying weights by Newey and West (1994)

$$w_\ell = 1 - \frac{\ell}{L+1},$$

where L is the maximum lag. Lag ℓ is chosen via the non-parametric bandwidth selection procedure of Newey and West (1994), which automatically select the number of autocovariances to use in computing the HAC estimator. We use prewhitening filter here because the numerical experiments of Newey and West (1994) reveal that prewhitening with a first-order vector autoregression improves the size of the test statistics. In summary, plugging in $\hat{\Phi}$ by HAC, which uses the linearly decaying weights, into $\mathrm{Var}(\hat{\beta})$ (3.12) leads to our desired variance of $\bar{d}$, $\hat{\sigma}_N^2$.

We now discuss the mathematical expression of consistent scoring functions of VaR. It is given in Fissler et al. (2015) and Nolde and Ziegel (2017) that all scoring functions of the form

$$S_{\mathrm{VaR}}(q, x) = (1 - \tau - \mathbb{1}\{x > q\})G(q) + \mathbb{1}\{x > q\}G(x) \tag{3.14}$$

are consistent for VaR q_τ for a probability level $\tau \in (0, 1)$ where G is an increasing function on $\mathbb{R}$. Note that the scoring functions used in the DM test do not have to be quadratic and symmetric, and are negatively oriented that is, the smaller the better. For the comparative VaR backtesting using the DM test (3.11) in our empirical analysis in Section 3.6, we use the particular scoring functions introduced in Nolde and Ziegel (2017) although there exist a large number of choices for consistent scoring functions for VaR in the literature. A scoring function is said to be h-homogeneous if $S(cq, cx) = c^h S(q, x)$ for all q, x and the constant $c > 0$ (see Nolde and Ziegel 2017 for the discussion of positive homogeneity of the scoring function for ranking risk measures).

First scoring function of VaR is when $G(q) = q$ and $G(x) = x$ are chosen in the general form (3.14), which leads to the classical 1-homogeneous choice

$$S_{\text{VaR}}(q, x) = (1 - \tau - \mathbb{1}\{x > q\})q + \mathbb{1}\{x > q\}x. \tag{3.15}$$

Second one is to choose $G(q) = \log q$ and $G(x) = \log x$ with $q, x > 0$ leading to the alternative 0-homogeneous choice

$$S_{\text{VaR}}(q, x) = (1 - \tau - \mathbb{1}\{x > q\}) \log q + \mathbb{1}\{x > q\} \log x. \tag{3.16}$$

Financial returns exhibit heavy-tails and this might limit the choice of a suitable scoring function. We emphasize that at present there exists no particular optimal scoring function for specific risk measures with any theoretical guarantee to use in the comparative VaR backtesting. Moreover, we refer to 0-homogeneous and 1-homogeneous choices of VaR scoring functions as $h = 0$ and $h = 1$ in the empirical analysis given in Section 3.6.3, respectively.

We finally explain the decisions taken in the comparative VaR backtesting based on the DM test under the null hypotheses H_0^- and H_0^+. Under H_0^-, the comparative backtesting is passed for the competing model (Model 1) if the null hypothesis fails to be rejected. The competing model is then considered as better model than the benchmark (Model 2) in this specific situation and it simply means that this null hypothesis cannot be falsified. On the other hand, under H_0^+ the backtesting for the competing model is passed if the null hypothesis is rejected. The decisions taken under H_0^- and H_0^+ where the colors used match with the traffic light approach of the Basel Committee on Banking Supervision (2019) are shown in the Figure 1 of Fissler et al. (2015). The green zone corresponds to the case when H_0^- is not rejected and H_0^+ is rejected, which suggests that the competing model is considered as better than the benchmark model. The yellow zone is when only one of the backtestings under H_0^- and H_0^+ is passed and we cannot conclude which model performs the best. The red zone corresponds to the case when both backtestings fail to be passed, indicating a problem with the competing model. Note that we use the comparative backtesting based on the DM test for ES as well. Comparative ES backtesting is not exactly same

as the VaR version because ES is not elicitable and we have to employ the idea of joint elicitability to obtain the scoring function of (VaR, ES), discussed in Section 4.5.

3.6 Empirical analysis of four financial time series

3.6.1 Descriptive statistics and basic statistical tests

We consider historical daily negative log-returns of three financial indices and an exchange rate, all made of $n = 4000$ observations (will be used in Chapter 4 as well):

- The Dow Jones Industrial Average (DJ) from 23 December 1993 to 9 November 2009;
- The Nasdaq Stock Market Index (NASDAQ) from 30 August 1993 to 16 July 2009;
- The Nikkei 225 (NIKKEI) from 14 May 1993 to 12 August 2009;
- The Japanese Yen-British Pound exchange rate (JPY/GBP) from 2 January 2000 to 14 December 2010.

The data have been taken from the R package `qrmdata` (Hofert and Hornik, 2016) and are represented in Figure 3.1. The graphs show that these negative log-returns are extremely volatile around the 2007-2008 financial crisis, which created a succession of extreme positive and negative returns over short time horizons. A noticeable degree of volatility clustering is also detected from a visual inspection of Figure 3.1, revealing the presence of heteroskedasticity. Including a turbulent period from a financial risk management perspective is crucial in order to examine how dynamic extreme VaR estimators behave. An inspection of more recent financial data collected during the COVID-19 crisis did not reveal a more substantial degree of volatility, so we focus on the well-studied subprime crisis in order to assess the quality of our forecasts.

Descriptive statistics and basic statistical tests applied to the negative log-returns on the four financial time series are reported in Table 3.1. According to the descriptive statistics, the means of the negative log-returns of all series are close to zero, and negative log-returns are leptokurtic. The Jarque-Bera test statistics indicate that the

Gaussian distribution is not suitable for any of these series of negative log-returns. All four series pass the augmented Dickey-Fuller (ADF) test, indicating that they can be considered stationary for modeling purposes. The Ljung-Box test applied to the squared negative log-returns, with orders 1 and 10, rejects the null hypothesis of no autocorrelation, indicating the presence of substantial conditional heteroskedasticity in all series. This provides justification for our use of GARCH-type models with these data.

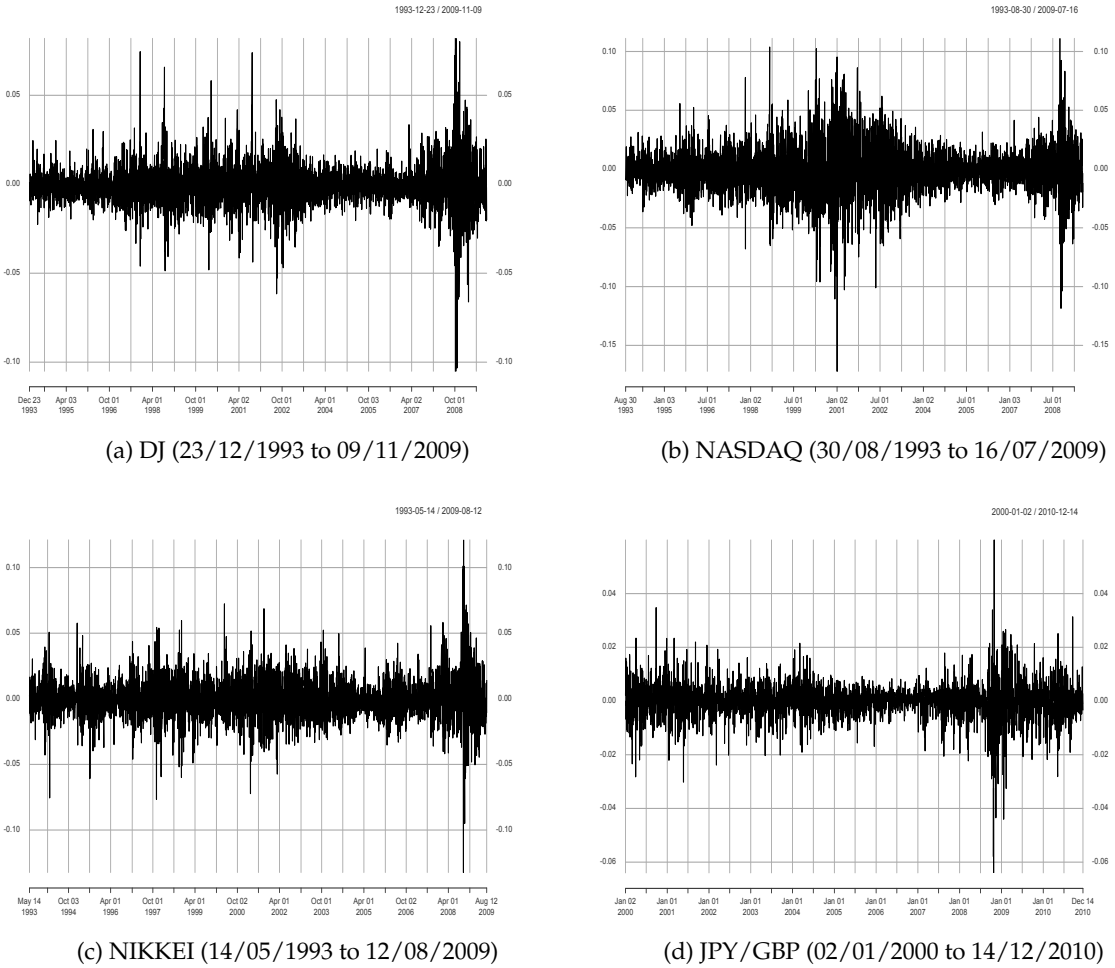
TABLE 3.1: Summary of descriptive statistics and basic statistical tests for daily negative log-returns on DJ, NASDAQ, NIKKEI and JPY/GBP.

	DJ	NASDAQ	NIKKEI	JPY/GBP
Sample size	4000	4000	4000	4000
Mean	-0.000250	-0.000355	-0.000169	-0.0000557
Median	-0.000460	-0.00123	-0.0000177	0
Maximum	0.0820	0.111	0.121	0.0600
Minimum	-0.105	-0.172	-0.132	-0.0640
Standard deviation	0.0119	0.0203	0.0155	0.00626
Skewness	0.117	-0.110	0.175	-0.586
Kurtosis	8.096	4.469	5.579	10.931
J-B test	$10933*$	$3337.3*$	$5207.9*$	$2014.6*$
	(0.0000)	(0.0000)	(0.0000)	(0.0000)
$Q(1)$	$13.159*$	$12.098*$	$6.680*$	$128.68*$
	(0.000)	(0.001)	(0.010)	(0.000)
$Q(5)$	$37.723*$	$37.62*$	$14.429*$	$146.47*$
	(0.000)	(0.000)	(0.013)	(0.000)
$Q(10)$	$50.388*$	$42.192*$	$23.023*$	$150.37*$
	(0.000)	(0.000)	(0.011)	(0.000)
$Q^2(1)$	$131.54*$	$207.21*$	$248.32*$	$275.36*$
	(0.000)	(0.000)	(0.000)	(0.000)
$Q^2(10)$	$2613.3*$	$1907.5*$	$3183.4*$	$1650.4*$
	(0.000)	(0.000)	(0.000)	(0.000)
ADF test	$-15.782**$	$-14.794**$	$-15.967**$	$-16.415**$

Notes: A kurtosis greater than 3 indicates that the dataset has heavier tails than a normal distribution. J-B stands for the Jarque-Bera test, $Q(n)$ and $Q^2(n)$ are the Ljung-Box tests for autocorrelation at lags n in the negative log-return series and squared negative log-returns, respectively. The ADF test is the augmented Dickey-Fuller stationarity test statistic without trend. The p-values are given between brackets. $**$, $*$ denote significance at 1% and 5% levels, respectively.

In this thesis, we compare six methods in total:

- The nonparametric HS method is based on the observed data, and its VaR is simply the empirical quantile of the series X_t at the desired quantile level.



(a) DJ (23/12/1993 to 09/11/2009)

(b) NASDAQ (30/08/1993 to 16/07/2009)

(c) NIKKEI (14/05/1993 to 12/08/2009)

(d) JPY/GBP (02/01/2000 to 14/12/2010)

FIGURE 3.1: Daily negative log-returns of four financial time series: DJ, NASDAQ, NIKKEI and JPY/GBP.

- The GARCH-N (normal) method uses the same filtering step as explained in Section 3.2.2, but assumes in the quantile estimation step also that the innovations Z_t are i.i.d. $\mathcal{N}(0,1)$. The extreme conditional VaR is then calculated as $\hat{q}_\tau(X_{t+1} \mid \mathscr{F}_t) = \hat{\mu}_{t+1} + \hat{\sigma}_{t+1}\Phi^{-1}(\tau)$, where Φ is the cumulative distribution function of the standard normal distribution.

- The GARCH-t method again uses the filtering step of Section 3.2.2, but assumes in the quantile estimation step that the standardized residuals from the GARCH step are Student-t distributed. This corresponds to the two-step estimation method discussed on p.1017 of Ergen (2015).

- The bias-reduced UGH method without filtering: this method applies the UGH step directly to the series X_t.

- The conventional GARCH-EVT method as described in McNeil and Frey (2000). This consists, first, in the same filtering step as described in Section 3.2.2. Standardized residuals are then recorded and a Generalized Pareto distribution is fitted using a maximum likelihood estimator, thus producing a VaR estimate $\tilde{q}_\tau(Z)$. This method therefore differs from ours as far as the extreme value step is concerned.

- The proposed GARCH-UGH method.

A comparison with the basic estimation methods (HS, GARCH-N and GARCH-t) indicates the importance of extreme value methods in the estimation of the dynamic extreme VaR. Besides, a comparison with the UGH method (without filtering) allows us to see how effective filtering is, and a comparison with the GARCH-EVT method (not featuring bias reduction) will illustrate the benefit of bias reduction at the extreme value step after filtering.

We present in-sample and out-of-sample evaluations of one-step ahead conditional VaR estimates at different τ levels and choices of k by means of traditional and comparative backtestings in Sections 3.6.2 and 3.6.3, respectively: in-sample estimation investigates the fit of the approaches to high volatile returns, while out-of-sample estimation tests how well the method predicts extreme VaR. Furthermore, supplementary simulations when the GARCH model is misspecifed and the innovations are normally distributed, i.e., not assuming heavy-tail, are given in Appendix B.

3.6.2 In-sample dynamic extreme VaR estimation and backtesting

We start by estimating in-sample one-step ahead conditional extreme VaRs $q_\tau(X_{t+1} \mid \mathscr{F}_t)$ for $\tau \in \{0.99, 0.995, 0.999\}$. For these in-sample evaluations, all methods (HS, GARCH-N and GARCH-t, UGH without filtering, GARCH-EVT without bias reduction, and our proposed GARCH-UGH method) are implemented on a fixed in-sample testing window W_T, which consists of 3000 observations; this follows advice by Danielsson (2011) which suggests that this testing window W_T should cover at least 4 years of data, or approximately 1000 observations, for a reliable statistical analysis. Specifically, we use:

- The time period from 8 December 1997 to 9 November 2009 for the Dow Jones,
- The time period from 13 August 1997 to 16 July 2009 for the NASDAQ,
- The time period from 29 May 1997 to 12 August 2009 for the Nikkei,
- The time period from 28 September 2002 to 14 December 2010 for the JPY/GBP exchange rate.

This allows us to focus on extreme VaR estimation around the 2007-2008 financial crisis, of which a consequence was a succession of extremely large negative log-returns in a very short timeframe. This should be considered a challenging problem.

In each case, we implement the three methods on these 3000 observations. The HS and UGH method work directly on the series X_t, without filtering, the estimate then being $\bar{q}_\tau(X_{t+1} \mid \mathscr{F}_t) = \bar{q}_\tau(X)$, where $\bar{q}_\tau(X)$ is the empirical τth quantile of the data for the HS method and, for the UGH method, $\bar{q}_\tau(X) = \hat{q}_\tau(X)$ is obtained as in Section 3.2.3 with the X_t in place of the $\hat{Z}_t$. By contrast, the GARCH-N, GARCH-t, GARCH-EVT and GARCH-UGH methods filter the data using an AR(1)-GARCH(1,1) model $X_t = \mu_t + \sigma_t Z_t$ with a Gaussian QMLE, and then estimate $q_\tau(Z)$ on the basis of the residuals obtained from this filtering with an approach specific to each method, before obtaining the final extreme VaR estimate by combining the AR(1)-GARCH(1,1) estimates and the estimate of $q_\tau(Z)$. In these four methods, the difference lies in how $q_\tau(Z)$ is estimated. In addition, we calculate another version of the GARCH-UGH estimate where the estimator $\hat{\rho}_{k_\rho}$ is replaced throughout by the constant -1, as mentioned in Section 3.2.3. If this other version has a number of VaR violations closer to the expected number of violations (which is known and equal to $3000(1-\tau)$ where τ is the VaR level), we retain this version.

3.6.2.1 Comparison with EVT-type methods

Results from Tables 3.2-3.5 indicate that, on the basis of in-sample validation and compared to the other two methods geared towards extreme value estimation (GARCH-EVT and UGH), the proposed GARCH-UGH approach is the most successful for

estimating one-step ahead extreme VaRs that satisfy both unconditional and conditional coverage properties. Across all samples and in terms of number of VaR violations only, in 46 out of 60 cases our GARCH-UGH approach is closest to the mark. In addition, although the unfiltered UGH estimate is somewhat reasonable in terms of number of VaR violations, it is not appropriate because it lacks responsiveness to the time-varying volatility and volatility clustering: Figure 3.2a illustrates that the non-dynamic nature of the UGH estimate leaves it unable to respond immediately to high volatility, and VaR violations tend to cluster. By contrast, the conditional VaR estimates obtained by our GARCH-UGH approach (Figure 3.2b) clearly respond to the changing volatility with no clustering of VaR violations, while bias reduction results in closer numbers of VaR violations to the expected numbers than with the conventional GARCH-EVT. Numerically, the GARCH-UGH method never fails either the Kupiec or Christoffersen tests, whereas the GARCH-EVT method fails 7 and 5 times out of 60 cases, respectively. The bias correction at the extreme value step appears to be very effective for the accurate estimation of one-step ahead dynamic extreme VaRs. It leads to results that seem less sensitive to the choice of sample fraction k than the conventional GARCH-EVT method: see Tables 3.2-3.5, where results appear to be consistently good across a large range of values of k.

3.6.2.2 Comparison with basic estimation methods

In addition, Table 3.6 shows the superiority of the GARCH-UGH approach when it is compared with the basic estimation methods (HS, GARCH-N and GARCH-t) that are commonly used by practitioners in financial risk management. Note that the number of VaR violations for GARCH-UGH shown in the Table 3.6 corresponds to when the optimal (according to Tables 3.2-3.5) sample fraction is chosen from 5% to 25% for the estimation of dynamic extreme VaR. For all cases, HS provides the same number of VaR violations as theoretically expected ones. This is because the nonparametric HS method gives the (length of testing window $\times (1-\tau)$)th ordered value in the sample as the VaR at quantile level τ from the non-updated ordered observations, which always ends up producing the same number of VaR violations as theoretically expected. Hence, in-sample HS is trivial and we exclude it from the comparison for in-sample estimation; we will see in the out-of-sample backtestings

TABLE 3.2: In-sample evaluations of one-step ahead conditional VaR estimates from 8 December 1997 to 9 November 2009 at different quantile levels for the negative log-returns of DJ index by means of the number of VaR violations, unconditional and conditional coverage tests.

Testing window	3000				
% of top obs. used	5%	10%	15%	20%	25%
DJ:					
0.999 Quantile					
Expected	3	3	3	3	3
UGH	**4**	5	**2**	**2**	**3**
	(0.583, 0.855)	(0.292, 0.569)	(0.538, 0.826)	(0.538, 0.826)	(1.000, 0.997)
GARCH-UGH	**2**	**2**	**2**	**2**	2
	(0.538, 0.826)	(0.538, 0.826)	(0.538, 0.826)	(0.538, 0.826)	(0.538, 0.826)
GARCH-EVT	**2**	**2**	**2**	**2**	4
	(0.538, 0.826)	(0.538, 0.826)	(0.538, 0.826)	(0.538, 0.826)	(0.583, 0.855)
0.995 Quantile					
Expected	15	15	15	15	15
UGH	18	18	**16**	18	20
	(0.452, 0.012)	(0.452, 0.012)	(0.798, 0.009)	(0.452, 0.012)	(0.218, 0.011)
GARCH-UGH	**15**	**14**	**14**	**15**	**15**
	(1.000, 0.927)	(0.793, 0.905)	(0.793, 0.905)	(1.000, 0.927)	(1.000, 0.927)
GARCH-EVT	13	13	13	13	13
	(0.596, 0.821)	(0.596, 0.821)	(0.596, 0.821)	(0.596, 0.821)	(0.596, 0.821)
0.99 Quantile					
Expected	30	30	30	30	30
UGH	34	34	34	36	39
	(0.472, 0.018)	(0.472, 0.018)	(0.472, 0.018)	(0.286, 0.018)	(0.114, 0.014)
GARCH-UGH	**27**	**28**	**29**	**31**	**33**
	(0.576, 0.669)	(0.711, 0.717)	(0.854, 0.741)	(0.855, 0.711)	(0.588, 0.598)
GARCH-EVT	23	23	22	22	20
	(0.180, 0.341)	(0.180, 0.341)	(0.123, 0.259)	(0.123, 0.259)	(0.050, 0.130)

Notes: The closest numbers of VaR violations to theoretically expected ones are highlighted in bold. The *p*-values for the unconditional coverage test by Kupiec (1995) and conditional coverage test by Christoffersen (1998) are given in brackets in order.

TABLE 3.3: In-sample evaluations of one-step ahead conditional VaR estimates from 13 August 1997 to 16 July 2009 at different quantile levels for the negative log-returns of NASDAQ index by means of the number of VaR violations, unconditional and conditional coverage tests.

Testing window	3000				
% of top obs. used	5%	10%	15%	20%	25%
NASDAQ:					
0.999 Quantile					
Expected	3	3	3	3	3
UGH	**3**	1	1	1	1
	(1.000, 0.997)	(0.179, 0.406)	(0.179, 0.406)	(0.179, 0.406)	(0.179, 0.406)
GARCH-UGH	4	**4**	**4**	**4**	**2**
	(0.583, 0.855)	(0.583, 0.855)	(0.583, 0.855)	(0.583, 0.855)	(0.538, 0.826)
GARCH-EVT	4	**4**	**4**	**4**	**4**
	(0.583, 0.855)	(0.583, 0.855)	(0.583, 0.855)	(0.583, 0.855)	(0.583, 0.855)
0.995 Quantile					
Expected	15	15	15	15	15
UGH	21	21	21	19	21
	(0.143, 0.295)	(0.143, 0.295)	(0.143, 0.295)	(0.320, 0.541)	(0.143, 0.295)
GARCH-UGH	**14**	**14**	**14**	**14**	**13**
	(0.793, 0.905)	(0.793, 0.905)	(0.793, 0.905)	(0.793, 0.905)	(0.596, 0.821)
GARCH-EVT	13	13	10	10	10
	(0.596, 0.821)	(0.596, 0.821)	(0.168, 0.374)	(0.168, 0.374)	(0.168, 0.374)
0.99 Quantile					
Expected	30	30	30	30	30
UGH	**32**	**33**	**33**	**35**	37
	(0.717, 0.609)	(0.588, 0.135)	(0.588, 0.135)	(0.371, 0.127)	(0.215, 0.106)
GARCH-UGH	23	23	23	**25**	**25**
	(0.180, 0.341)	(0.180, 0.341)	(0.180, 0.341)	(0.345, 0.519)	(0.345, 0.519)
GARCH-EVT	22	17	16	16	16
	(0.123, 0.259)	(0.009, 0.031)	(0.005, 0.017)	(0.005, 0.017)	(0.005, 0.017)

Notes: The closest numbers of VaR violations to theoretically expected ones are highlighted in bold. The p-values for the unconditional coverage test by Kupiec (1995) and conditional coverage test by Christoffersen (1998) are given in brackets in order.

TABLE 3.4: In-sample evaluations of one-step ahead conditional VaR estimates from 29 May 1997 to 12 August 2009 at different quantile levels for the negative log-returns of NIKKEI index by means of the number of VaR violations, unconditional and conditional coverage tests.

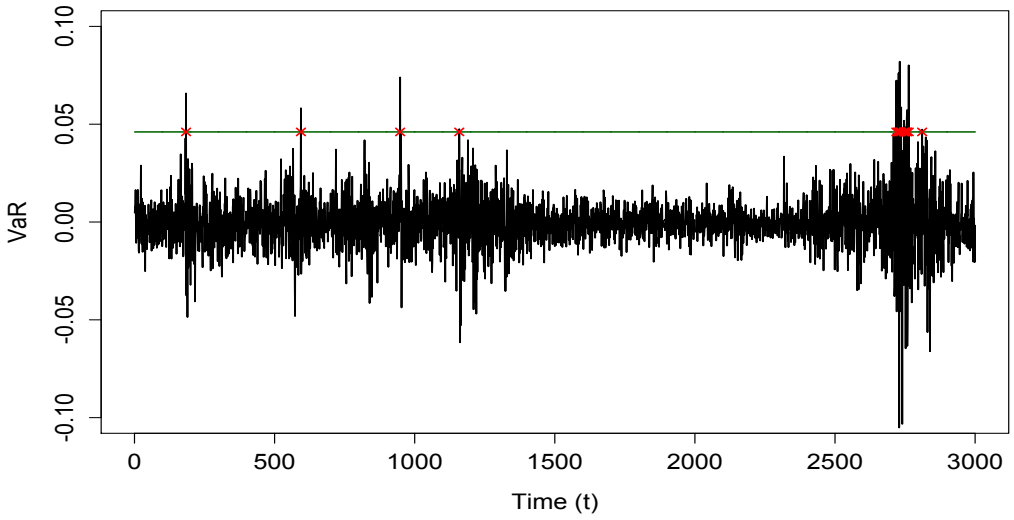
Testing window	3000				
% of top obs. used	5%	10%	15%	20%	25%
NIKKEI:					
0.999 Quantile					
Expected	3	3	3	3	3
UGH	**4**	**4**	**4**	**4**	**1**
	(0.583, 0.855)	(0.583, 0.855)	(0.583, 0.855)	(0.583, 0.855)	(0.179, 0.406)
GARCH-UGH	**4**	**2**	**4**	**4**	**1**
	(0.583, 0.885)	(0.538, 0.826)	(0.583, 0.885)	(0.583, 0.885)	(0.179, 0.406)
GARCH-EVT	5	5	5	5	**5**
	(0.292, 0.569)	(0.292, 0.569)	(0.292, 0.569)	(0.292, 0.569)	(0.292, 0.569)
0.995 Quantile					
Expected	15	15	15	15	15
UGH	**15**	**15**	**17**	18	21
	(1.000, 0.178)	(1.000, 0.178)	(0.612, 0.199)	(0.452, 0.190)	(0.143, 0.114)
GARCH-UGH	13	13	**13**	**13**	**12**
	(0.596, 0.821)	(0.596, 0.821)	(0.596, 0.821)	(0.596, 0.821)	(0.421, 0.689)
GARCH-EVT	13	12	12	12	**12**
	(0.596, 0.821)	(0.421, 0.689)	(0.421, 0.689)	(0.421, 0.689)	(0.421, 0.689)
0.99 Quantile					
Expected	30	30	30	30	30
UGH	**32**	**32**	**34**	36	38
	(0.717, 0.609)	(0.717 0.609)	(0.472, 0.562)	(0.286 0.427)	(0.159, 0.297)
GARCH-UGH	26	25	**26**	**31**	**28**
	(0.453, 0.601)	(0.345, 0.519)	(0.453, 0.601)	(0.855, 0.711)	(0.711, 0.666)
GARCH-EVT	25	24	21	19	18
	(0.345, 0.287)	(0.254, 0.430)	(0.081, 0.188)	(0.030, 0.085)	(0.017, 0.049)

Notes: The closest numbers of VaR violations to theoretically expected ones are highlighted in bold. The *p*-values for the unconditional coverage test by Kupiec (1995) and conditional coverage test by Christoffersen (1998) are given in brackets in order.

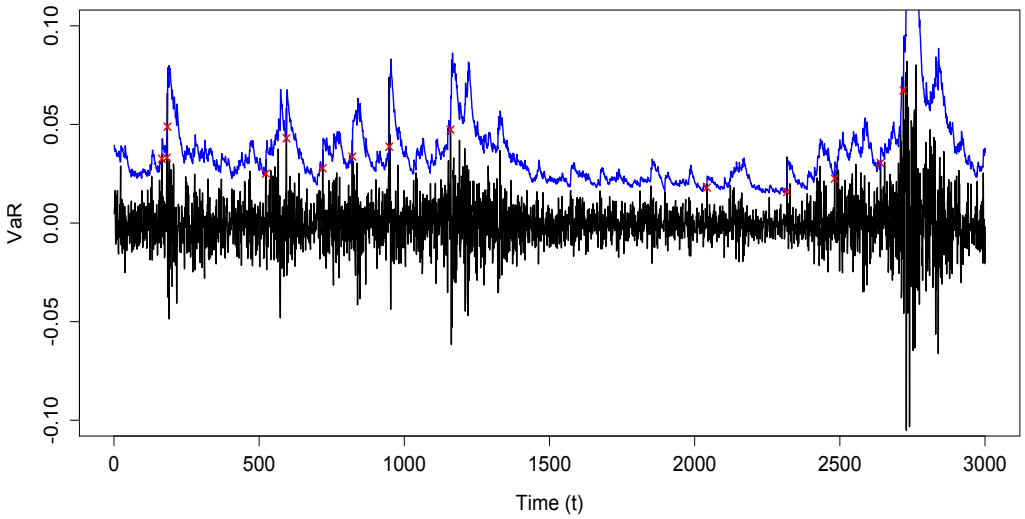
TABLE 3.5: In-sample evaluations of one-step ahead conditional VaR estimates from 28 September 2002 to 14 December 2010 at different quantile levels for the negative log-returns of JPY/GBP exchange rate by means of the number of VaR violations, unconditional and conditional coverage tests.

Testing window	3000				
% of top obs. used	5%	10%	15%	20%	25%
JPY/GBP:					
0.999 Quantile					
Expected	3	3	3	3	3
UGH	2	2	1	1	1
	(0.538, 0.826)	(0.538, 0.826)	(0.179, 0.406)	(0.179, 0.406)	(0.179, 0.406)
GARCH-UGH	**3**	2	**3**	2	2
	(1.000, 0.997)	(0.538, 0.826)	(1.000, 0.997)	(0.538, 0.826)	(0.538, 0.826)
GARCH-EVT	**3**	**3**	**3**	**3**	**3**
	(1.000, 0.997)	(1.000, 0.997)	(1.000, 0.997)	(1.000, 0.997)	(1.000, 0.997)
0.995 Quantile					
Expected	15	15	15	15	15
UGH	**16**	17	**16**	18	28
	(0.798, 0.195)	(0.612, 0.199)	(0.798 0.195)	(0.452, 0.190)	(0.003, 0.000)
GARCH-UGH	**16**	**14**	**14**	**14**	**16**
	(0.798, 0.888)	(0.793, 0.905)	(0.793, 0.905)	(0.793, 0.905)	(0.798, 0.888)
GARCH-EVT	11	11	11	11	10
	(0.277, 0.532)	(0.277, 0.532)	(0.277, 0.532)	(0.277, 0.532)	(0.168, 0.374)
0.99 Quantile					
Expected	30	30	30	30	30
UGH	38	40	41	41	46
	(0.159, 0.002)	(0.081, 0.002)	(0.056, 0.001)	(0.056, 0.001)	(0.006, 0.001)
GARCH-UGH	**31**	32	**31**	**29**	**22**
	(0.855, 0.612)	(0.717, 0.609)	(0.855, 0.612)	(0.854, 0.556)	(0.123, 0.259)
GARCH-EVT	**29**	**29**	28	24	**22**
	(0.854, 0.556)	(0.854, 0.556)	(0.711, 0.501)	(0.254, 0.430)	(0.123, 0.259)

Notes: The closest numbers of VaR violations to theoretically expected ones are highlighted in bold. The p-values for the unconditional coverage test by Kupiec (1995) and conditional coverage test by Christoffersen (1998) are given in brackets in order.



(a) UGH approach



(b) GARCH-UGH approach

FIGURE 3.2: Twelve years (8 December 1997 to 9 November 2009) of in-sample backtesting of the DJ index, and 99.5%-VaR violations by (a) the UGH approach and (b) the GARCH-UGH approach when the top 15% of observations are used for the estimation. Red cross marks denote the VaR violations.

that HS performs worse out of all estimation methods. Across all samples, in 11 out of 12 cases our GARCH-UGH approach is closest to the mark and never fails any tests, with the GARCH-N approach performing worse since it cannot capture heavy tails. The GARCH-N and GARCH-t methods are not reliable approaches for the estimation of dynamic extreme VaR because GARCH-N fails to pass the Kupiec and Christoffersen tests 6 and 5 times out of 12 cases, and GARCH-t fails 4 and 3 times respectively.

3.6.3 Out-of-sample dynamic extreme VaR estimation and backtesting

We now focus on the out-of-sample estimation (that is, prediction) of one-step ahead VaR via the same six approaches, again at level $\tau \in \{0.99, 0.995, 0.999\}$. We consider the following samples of data:

- The time period from 23 December 1993 to 9 November 2009 for the Dow Jones,
- The time period from 30 August 1993 to 16 July 2009 for the NASDAQ,
- The time period from 14 May 1993 to 12 August 2009 for the Nikkei,
- The time period from 2 January 2000 to 14 December 2010 for the JPY/GBP exchange rate.

In order to carry out this out-of-sample backtest, we adopt a rolling window estimation approach. Specifically, we first fix a testing window W_T in each case, which corresponds to the periods of time considered in our in-sample evaluation (8 December 1997 to 9 November 2009 for the Dow Jones, 13 August 1997 to 16 July 2009 for the Nasdaq, 29 May 1997 to 12 August 2009 for the Nikkei, 28 September 2002 to 14 December 2010 for the JPY/GBP exchange rate). At each time t in this testing window W_T, we use a window of length W_E of prior information in order to predict the conditional VaR on time $t+1$ (with parameter estimates updated when the estimation window changes), which is then compared to the observed log-return on day $t+1$. Various choices of W_E have been made in the literature: here we choose $W_E = 1000$ as in McNeil and Frey (2000), corresponding to approximately four years of model calibration for each prediction with stock market data, and three years with exchange rate data. Regarding the use of the GARCH-UGH method specifically,

TABLE 3.6: In-sample evaluations of one-step ahead conditional VaR estimates by basic estimation methods at different quantile levels for the negative log-returns of DJ, NASDAQ, NIKKEI indices and JPY/GBP exchange rate (time period given in Section 3.6.2) by means of the number of VaR violations, unconditional and conditional coverage tests.

	DJ	NASDAQ	NIKKEI	JPY/GBP
Testing window	3000	3000	3000	3000
0.999 Quantile				
Expected	3	3	3	3
HS	3	3	3	3
	(1.000, 0.997)	(1.000, 0.997)	(1.000, 0.997)	(1.000, 0.997)
GARCH-N	13	9	11	7
	(0.000, 0.000)	(0.005, 0.020)	(0.000, 0.002)	(0.049, 0.142)
GARCH-*t*	**2**	**4**	**2**	1
	(0.538, 0.826)	(0.583, 0.855)	(0.538, 0.826)	(0.179, 0.406)
GARCH-UGH	**2**	**4**	**4**	**3**
	(0.538, 0.826)	(0.583, 0.855)	(0.583, 0.855)	(1.000, 0.997)
0.995 Quantile				
Expected	15	15	15	15
HS	15	15	15	15
	(1.000, 0.007)	(1.000, 0.923)	(1.000, 0.178)	(1.000, 0.178)
GARCH-N	28	**16**	25	20
	(0.003, 0.008)	(0.798, 0.888)	(0.018, 0.005)	(0.218, 0.410)
GARCH-*t*	13	10	12	1
	(0.596, 0.821)	(0.168, 0.374)	(0.421, 0.689)	(0.000, 0.000)
GARCH-UGH	**15**	**14**	**13**	**16**
	(1.000, 0.927)	(0.793, 0.905)	(0.596, 0.821)	(0.798, 0.888)
0.99 Quantile				
Expected	30	30	30	30
HS	30	30	30	30
	(1.000, 0.112)	(1.000, 0.594)	(1.000, 0.594)	(1.000, 0.012)
GARCH-N	43	**27**	41	38
	(0.025, 0.044)	(0.576, 0.669)	(0.056, 0.091)	(0.159, 0.297)
GARCH-*t*	20	16	18	3
	(0.051, 0.130)	(0.005, 0.017)	(0.017, 0.053)	(0.000, 0.000)
GARCH-UGH	**29**	25	**31**	**31**
	(0.854, 0.741)	(0.345, 0.519)	(0.855, 0711)	(0.855, 0.612)

Notes: The closest number of VaR violations to the theoretically expected number is highlighted in bold, excluding historical simulation (HS). The number of VaR violations for GARCH-UGH is reported when the optimal sample fraction is selected according to Tables 3.2-3.5. The *p*-values for the unconditional coverage test by Kupiec (1995) and conditional coverage test by Christoffersen (1998) are given in brackets in order.

we retain the implementation suggested by the results of in-sample backtesting. In other words, on a given data set and for a given value of k, if we observed during in-sample backtesting that the choice $\hat{\rho}_{k_\rho} = -1$ performed better, then we retain this choice for out-of-sample estimation; otherwise, we estimate ρ as indicated in Section 3.2.3 (see Tables A.1-A.4 in Appendix A).

Figure 3.3 shows the out-of-sample estimation of EVI by GARCH-UGH and GARCH-EVT approaches using the top 15% of observations from rolling estimation windows W_E made of 1000 observations for four financial time series. It is reasonable to assume the heavy-tail of the underlying distribution for the GARCH-UGH approach as the estimates of EVI are stable between 0.2 and 0.4. On the other hand, the GARCH-EVT approach yields unstable EVI estimates ranging from 0.2 to -0.4 that are not only sensitive to the number of upper order statistics used in the estimation but also to the rolling estimation window, although it is still valid because the range of EVI for GARCH-EVT is not limited to $\gamma > 0$.

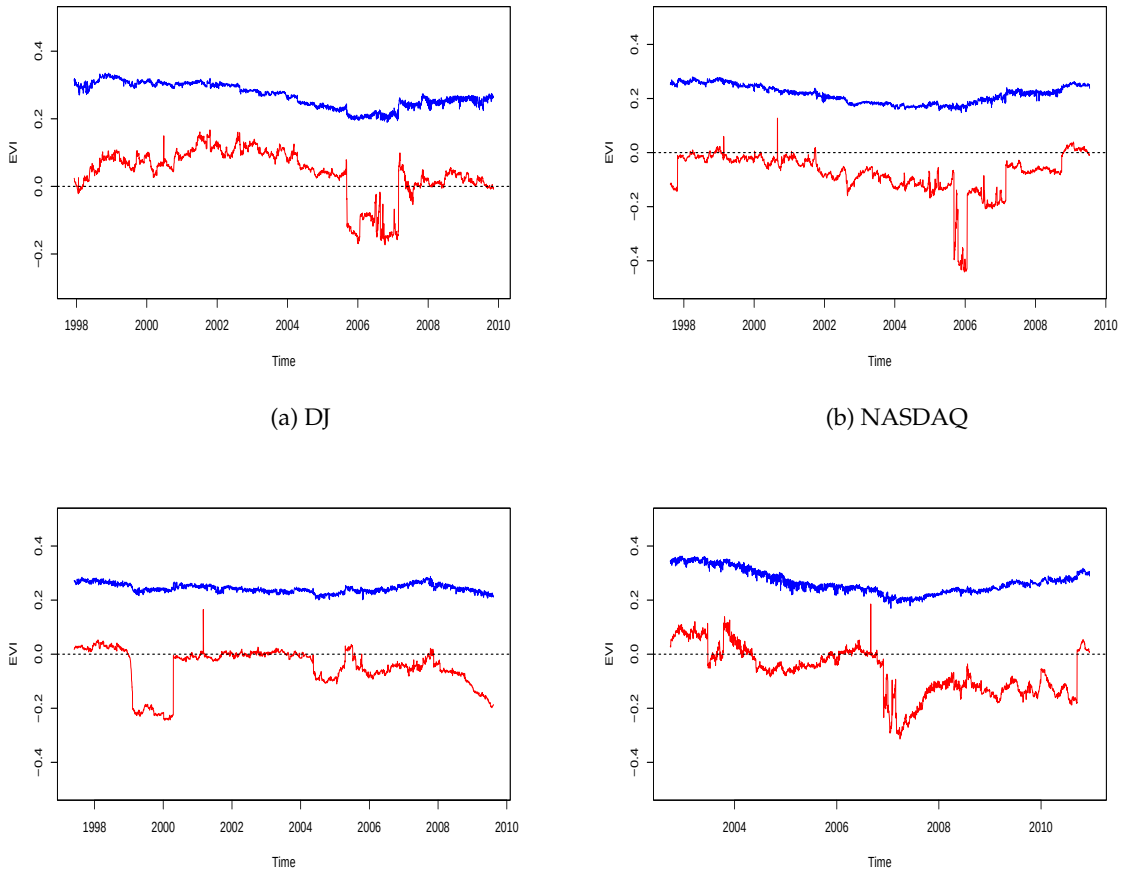
3.6.3.1 Comparison with EVT-type methods

Traditional VaR backtesting

Tables 3.7-3.10 gather the numerical results for the comparison between the GARCH-EVT, GARCH-UGH and UGH methods. It can be seen that again, the suggested GARCH-UGH approach appears to be best overall. In 47 out of 60 cases, the GARCH-UGH approach yields the closest number of VaR violations to the theoretically expected numbers, while the unfiltered UGH method fares worst. Based on the Kupiec test, the GARCH-UGH approach fails twice, whereas the GARCH-EVT and UGH fail 6 and 49 times out of 60 cases, respectively. On one occasion GARCH-UGH fails the Christoffersen test, while the GARCH-EVT and UGH methods fail 0 and 43 times out of 60 cases. GARCH-UGH typically performs better than other approaches except possibly when the top 5% and 10% of observations are used (for the choice of k); this is because the bias is not the dominating term in the bias-variance tradeoff when k is small.

Comparative VaR backtesting

Tables 3.11-3.14 display the traffic light matrices of comparative VaR backtesting given in Section 3.5 for three EVT-type methods, three quantile levels, five different



(a) DJ

(b) NASDAQ

(c) NIKKEI

(d) JPY/GBP

FIGURE 3.3: Out-of-sample estimation of extreme value index (EVI) by GARCH-UGH (blue line) and GARCH-EVT (red line) approaches using the top 15% of observations from rolling estimation windows made of 1000 observations for four financial time series: DJ, NASDAQ, NIKKEI and JPY/GBP.

TABLE 3.7: Traditional backtesting: out-of-sample evaluations of one-step ahead conditional VaR estimates by EVT-type methods from 8 December 1997 to 9 November 2009 at different quantile levels for the negative log-returns of DJ index by means of the number of VaR violations, unconditional and conditional coverage tests.

Testing window	3000				
Estimation window	1000				
% of top obs. used	5%	10%	15%	20%	25%
DJ:					
0.999 Quantile					
Expected	3	3	3	3	3
UGH	10	9	9	7	6
	(0.001, 0.006)	(0.005, 0.020)	(0.005, 0.020)	(0.049, 0.142)	(0.128, 0.310)
GARCH-UGH	**3**	**3**	**3**	**3**	**3**
	(1.000, 0.997)	(1.000, 0.997)	(1.000, 0.997)	(1.000, 0.997)	(1.000, 0.997)
GARCH-EVT	**3**	4	4	4	4
	(1.000, 0.997)	(0.583, 0.885)	(0.583, 0.885)	(0.583, 0.885)	(0.583, 0.855)
0.995 Quantile					
Expected	15	15	15	15	15
UGH	40	40	40	36	29
	(0.000, 0.000)	(0.000, 0.000)	(0.000, 0.000)	(0.000, 0.000)	(0.001, 0.001)
GARCH-UGH	**19**	**18**	**18**	**16**	**14**
	(0.320, 0.541)	(0.452, 0.676)	(0.452, 0.676)	(0.798, 0.888)	(0.793, 0.905)
GARCH-EVT	**19**	**18**	**18**	17	17
	(0.320, 0.541)	(0.452, 0.676)	(0.452, 0.676)	(0.612, 0.798)	(0.612, 0.798)
0.99 Quantile					
Expected	30	30	30	30	30
UGH	62	64	63	63	61
	(0.000, 0.000)	(0.000, 0.000)	(0.000, 0.000)	(0.000, 0.000)	(0.000, 0.000)
GARCH-UGH	**33**	35	32	**31**	**28**
	(0.588, 0.598)	(0.371, 0.433)	(0.717, 0.663)	(0.855, 0.711)	(0.711, 0.717)
GARCH-EVT	**33**	**30**	**30**	28	27
	(0.588, 0.598)	(1.000, 0.738)	(1.000, 0.738)	(0.711, 0.717)	(0.576, 0.669)

Notes: The closest numbers of VaR violations to theoretically expected ones are highlighted in bold. The p-values for the unconditional coverage test by Kupiec (1995) and conditional coverage test by Christoffersen (1998) are given in brackets in order.

TABLE 3.8: Traditional backtesting: out-of-sample evaluations of one-step ahead conditional VaR estimates by EVT-type methods from 13 August 1997 to 16 July 2009 at different quantile levels for the negative log-returns of NASDAQ index by means of the number of VaR violations, unconditional and conditional coverage tests.

Testing window	3000				
Estimation window	1000				
% of top obs. used	5%	10%	15%	20%	25%
NASDAQ:					
0.999 Quantile					
Expected	3	3	3	3	3
UGH	10	8	7	**4**	**3**
	(0.001, 0.006)	(0.017, 0.057)	(0.049, 0.142)	(0.583, 0.855)	(1.000, 0.997)
GARCH-UGH	**6**	**5**	**5**	**4**	**3**
	(0.128, 0.370)	(0.292, 0.569)	(0.292, 0.569)	(0.583, 0.855)	(1.000, 0.997)
GARCH-EVT	7	7	7	7	7
	(0.049, 0.142)	(0.049, 0.142)	(0.049, 0.142)	(0.049, 0.142)	(0.049, 0.142)
0.995 Quantile					
Expected	15	15	15	15	15
UGH	39	37	35	36	40
	(0.000, 0.000)	(0.000, 0.000)	(0.000, 0.000)	(0.000, 0.000)	(0.000, 0.000)
GARCH-UGH	20	17	**15**	**16**	**13**
	(0.218, 0.410)	(0.612, 0.798)	(1.000, 0.927)	(0.798, 0.888)	(0.596, 0.821)
GARCH-EVT	**16**	**14**	13	13	**13**
	(0.798, 0.888)	(0.793, 0.905)	(0.596, 0.821)	(0.596, 0.821)	(0.596, 0.821)
0.99 Quantile					
Expected	30	30	30	30	30
UGH	74	74	70	65	62
	(0.000, 0.000)	(0.000, 0.000)	(0.000, 0.000)	(0.000, 0.000)	(0.000, 0.000)
GARCH-UGH	34	35	**31**	**30**	**25**
	(0.427, 0.544)	(0.371, 0.490)	(0.855, 0.612)	(1.000, 0.594)	(0.345, 0.287)
GARCH-EVT	**31**	**28**	28	24	23
	(0.855, 0.612)	(0.711, 0.501)	(0.711, 0.501)	(0.254, 0.430)	(0.180, 0.341)

Notes: The closest numbers of VaR violations to theoretically expected ones are highlighted in bold. The *p*-values for the unconditional coverage test by Kupiec (1995) and conditional coverage test by Christoffersen (1998) are given in brackets in order.

TABLE 3.9: Traditional backtesting: out-of-sample evaluations of one-step ahead conditional VaR estimates by EVT-type methods from 29 May 1997 to 12 August 2009 at different quantile levels for the negative log-returns of NIKKEI index by means of the number of VaR violations, unconditional and conditional coverage tests.

Testing window	3000				
Estimation window	1000				
% of top obs. used	5%	10%	15%	20%	25%
NIKKEI:					
0.999 Quantile					
Expected	3	3	3	3	3
UGH	7	6	6	**5**	5
	(0.049, 0.142)	(0.128, 0.310)	(0.128, 0.310)	(0.292, 0.569)	(0.292, 0.569)
GARCH-UGH	**4**	**3**	**2**	**2**	**1**
	(0.583, 0.855)	(1.000, 0.997)	(0.538, 0.826)	(0.538, 0.826)	(0.179, 0.406)
GARCH-EVT	5	4	6	6	6
	(0.292, 0.569)	(0.583, 0.855)	(0.128, 0.310)	(0.128, 0.310)	(0.128, 0.310)
0.995 Quantile					
Expected	15	15	15	15	15
UGH	34	34	34	30	23
	(0.000, 0.000)	(0.000, 0.000)	(0.000, 0.000)	(0.000, 0.000)	(0.055, 0.062)
GARCH-UGH	**15**	**15**	**15**	**15**	**12**
	(1.000, 0.927)	(1.000, 0.927)	(1.000, 0.927)	(1.000, 0.927)	(0.421, 0.689)
GARCH-EVT	13	14	13	12	**12**
	(0.596, 0.821)	(0.793, 0.905)	(0.596, 0.821)	(0.421, 0.689)	(0.421, 0.689)
0.99 Quantile					
Expected	30	30	30	30	30
UGH	46	47	46	45	53
	(0.006, 0.011)	(0.004, 0.007)	(0.004, 0.007)	(0.010, 0.015)	(0.000 0.000)
GARCH-UGH	33	33	**33**	**30**	36
	(0.588, 0.598)	(0.588, 0.598)	(0.588, 0.598)	(1.000, 0.738)	(0.286, 0.365)
GARCH-EVT	**32**	**29**	**27**	27	**26**
	(0.717, 0.663)	(0.854, 0.741)	(0.576, 0.669)	(0.576, 0.669)	(0.453, 0.601)

Notes: The closest numbers of VaR violations to theoretically expected ones are highlighted in bold. The p-values for the unconditional coverage test by Kupiec (1995) and conditional coverage test by Christoffersen (1998) are given in brackets in order.

TABLE 3.10: Traditional backtesting: out-of-sample evaluations of one-step ahead conditional VaR estimates by EVT-type methods from 28 September 2002 to 14 December 2010 at different quantile levels for the negative log-returns of JPY/GBP exchange rate by means of the number of VaR violations, unconditional and conditional coverage tests.

Testing window	3000				
Estimation window	1000				
% of top obs. used	5%	10%	15%	20%	25%
JPY/GBP:					
0.999 Quantile					
Expected	3	3	3	3	3
UGH	7	7	6	**4**	**4**
	(0.049, 0.142)	(0.049, 0.142)	(0.128, 0.310)	(0.583, 0.855)	(0.583, 0.855)
GARCH-UGH	**3**	**2**	**2**	**2**	**2**
	(1.000, 0.997)	(0.538, 0.826)	(0.538, 0.826)	(0.538, 0.826)	(0.538, 0.826)
GARCH-EVT	6	5	5	6	7
	(0.128, 0.310)	(0.292, 0.569)	(0.292, 0.569)	(0.128, 0.310)	(0.049, 0.142)
0.995 Quantile					
Expected	15	15	15	15	15
UGH	25	27	27	34	45
	(0.018, 0.028)	(0.005, 0.010)	(0.005, 0.010)	(0.000, 0.000)	(0.000, 0.000)
GARCH-UGH	21	**18**	**15**	**14**	**12**
	(0.143, 0.295)	(0.452, 0.676)	(1.000, 0.927)	(0.793, 0.905)	(0.421, 0.689)
GARCH-EVT	**19**	19	20	20	20
	(0.320, 0.541)	(0.320, 0.541)	(0.219, 0.410)	(0.219, 0.410)	(0.219, 0.410)
0.99 Quantile					
Expected	30	30	30	30	30
UGH	47	56	55	59	67
	(0.004, 0.000)	(0.000, 0.000)	(0.000, 0.000)	(0.000, 0.000)	(0.000, 0.000)
GARCH-UGH	42	46	40	**38**	**34**
	(0.038, 0.064)	(0.006, 0.012)	(0.081, 0.127)	(0.159, 0.227)	(0.472, 0.523)
GARCH-EVT	**38**	**37**	**38**	**38**	36
	(0.159, 0.227)	(0.215, 0.292)	(0.159, 0.227)	(0.159, 0.227)	(0.286, 0.365)

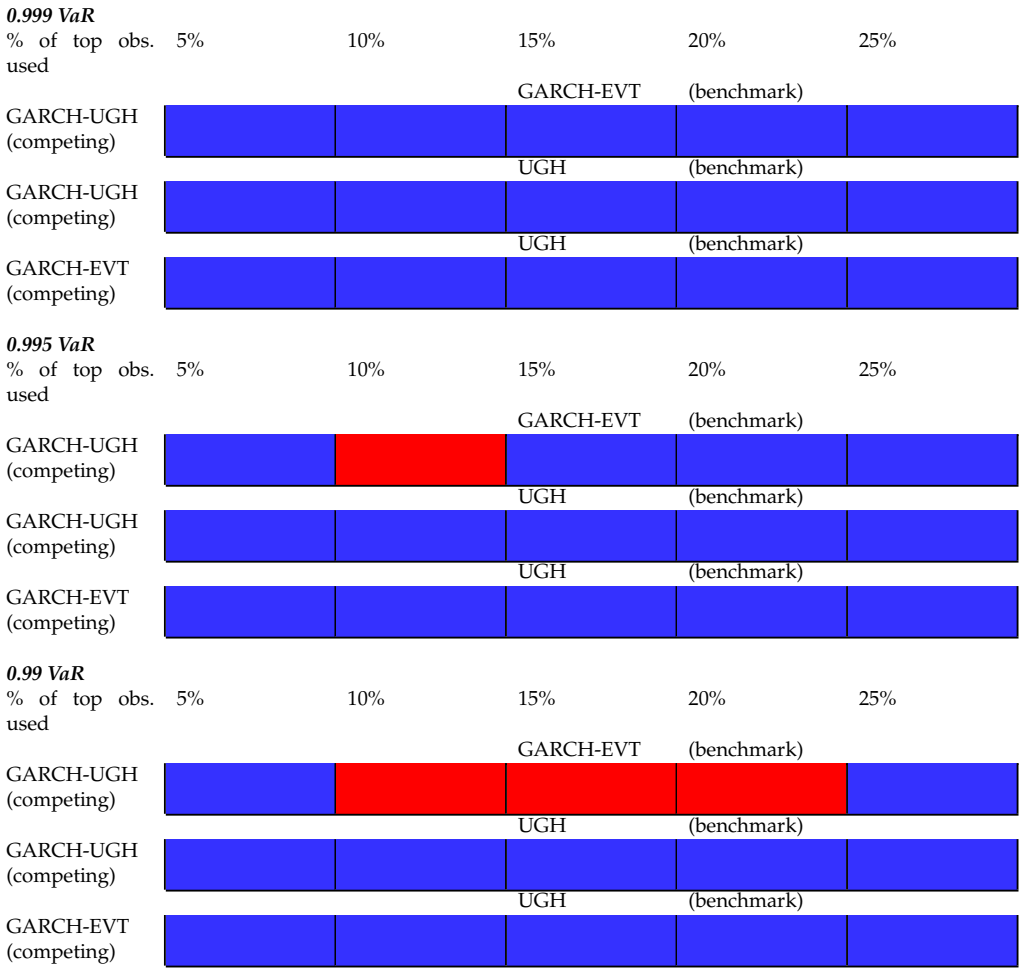
Notes: The closest numbers of VaR violations to theoretically expected ones are highlighted in bold. The *p*-values for the unconditional coverage test by Kupiec (1995) and conditional coverage test by Christoffersen (1998) are given in brackets in order.

threshold selections and four financial time series when $h = 0$ VaR scoring function of the form (3.16) is used. Recall that $h = 0$ and $h = 1$ in this context mean that 0-homogeneous and 1-homogeneous choices of VaR scoring functions, respectively, given in Section 3.5. The competing models are given in the vertical axis with the benchmark models along the horizontal axis. Using the t-statistic based on the DM test (3.11), we reject the hypothesis H_0^- at the test level 5% if $1 - \Phi(DM) \leq 0.05$ while the hypothesis H_0^+ is rejected if $\Phi(DM) \leq 0.05$. Under H_0^-, the comparative backtesting is passed for the competing model if the null hypothesis fails to be rejected. On the other hand, under H_0^+ the backtesting for the competing model is passed if the null hypothesis is rejected. The green zone corresponds to the case when H_0^- is not rejected and H_0^+ is rejected, which suggests that the competing model is considered as better than the benchmark model. The yellow zone is when only one of the backtestings under H_0^- and H_0^+ is passed and we cannot conclude which model performs the best. The red zone corresponds to the case when both backtestings fail to be passed, indicating a problem with the competing model.

As with the results of traditional VaR backtestings, it is illustrated that our proposed GARCH-UGH approach appears to be best overall. In 50 out of 60 cases, the GARCH-UGH approach is considered as better than GARCH-EVT approach based on the realized scores of VaR. We can also observe the similar trend as the results of the traditional backtestings that is, GARCH-UGH generally performs better than GARCH-EVT except when top 5% and 10% of observations are used due to the bias-variance trade-off. When GARCH-EVT is considered as better than GARCH-UGH, it indeed has the closest number of VaR violations to the theoretically expected numbers, which is consistent with the results of traditional backtestings again.

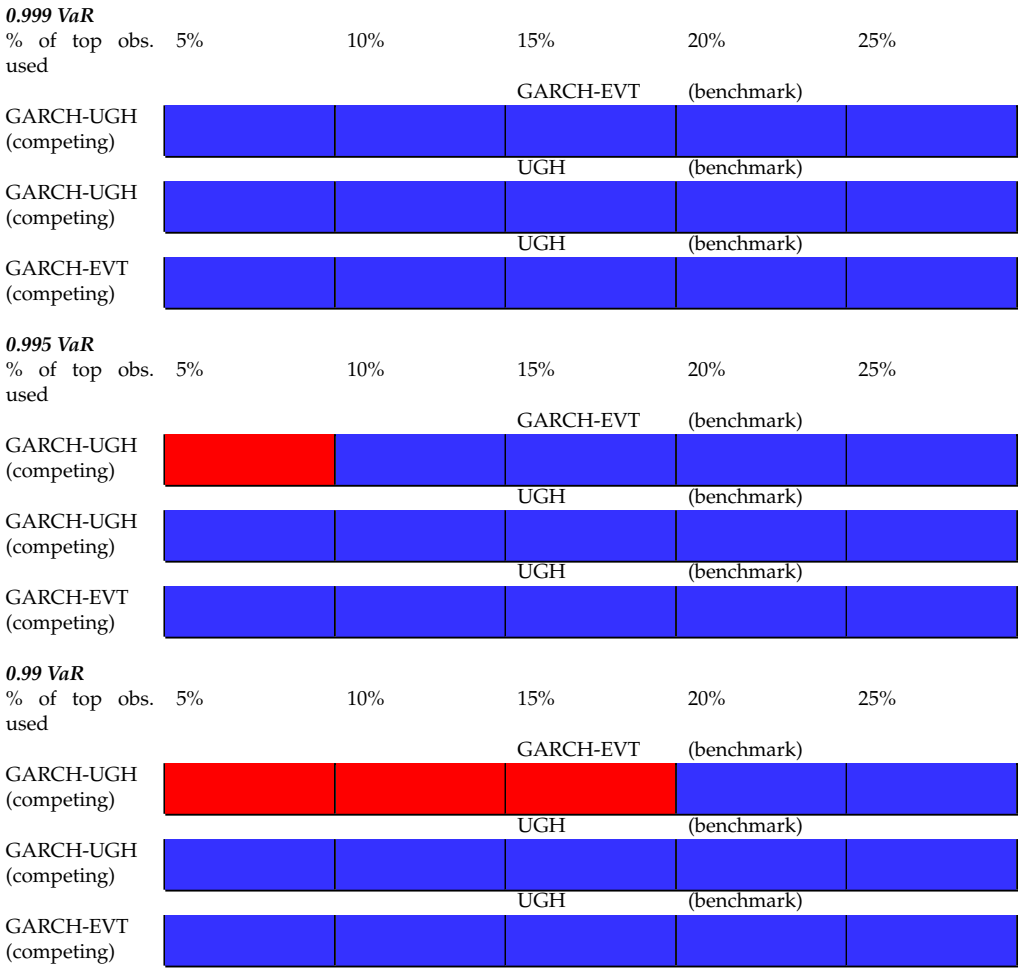
Tables 3.15-3.18 show the traffic light matrices for the $h = 1$ VaR scoring function given as the form (3.15). In 44 out of 60 cases, the GARCH-UGH approach is considered as better than the GARCH-EVT approach with 3 cases of no decisions taken. Comparative backtestings with two scoring functions and traditional backtestings result in a good agreement with the GARCH-UGH approach being the best estimator of VaR, while the unfiltered UGH being the worst estimator, i.e., failing the comparative backtestings against all the other methods. In the case of VaR, $h = 1$ scoring function has the worse discrimination ability than the $h = 0$ one especially

TABLE 3.11: Comparative backtesting: out-of-sample evaluations (traffic light matrices) of one-step ahead conditional VaR estimates by EVT-type methods from 8 December 1997 to 9 November 2009 at different quantile levels for the negative log-returns of DJ index by means of the Diebold-Mariano test using $h = 0$ VaR scoring function.



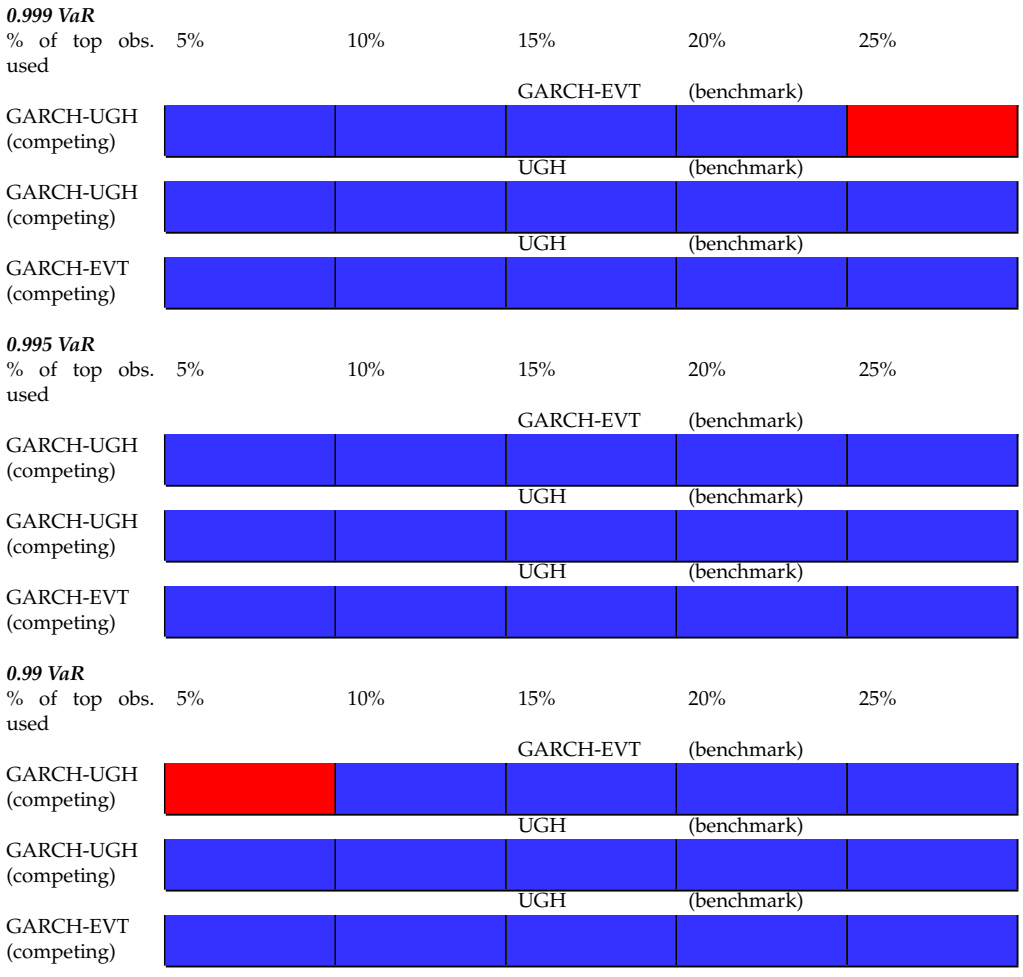
Notes: When the competing model in the row is swapped with the benchmark model in the column, the results are reversed.

TABLE 3.12: Comparative backtesting: out-of-sample evaluations (traffic light matrices) of one-step ahead conditional VaR estimates by EVT-type methods from 13 August 1997 to 16 July 2009 at different quantile levels for the negative log-returns of NASDAQ index by means of the Diebold-Mariano test using $h = 0$ VaR scoring function.



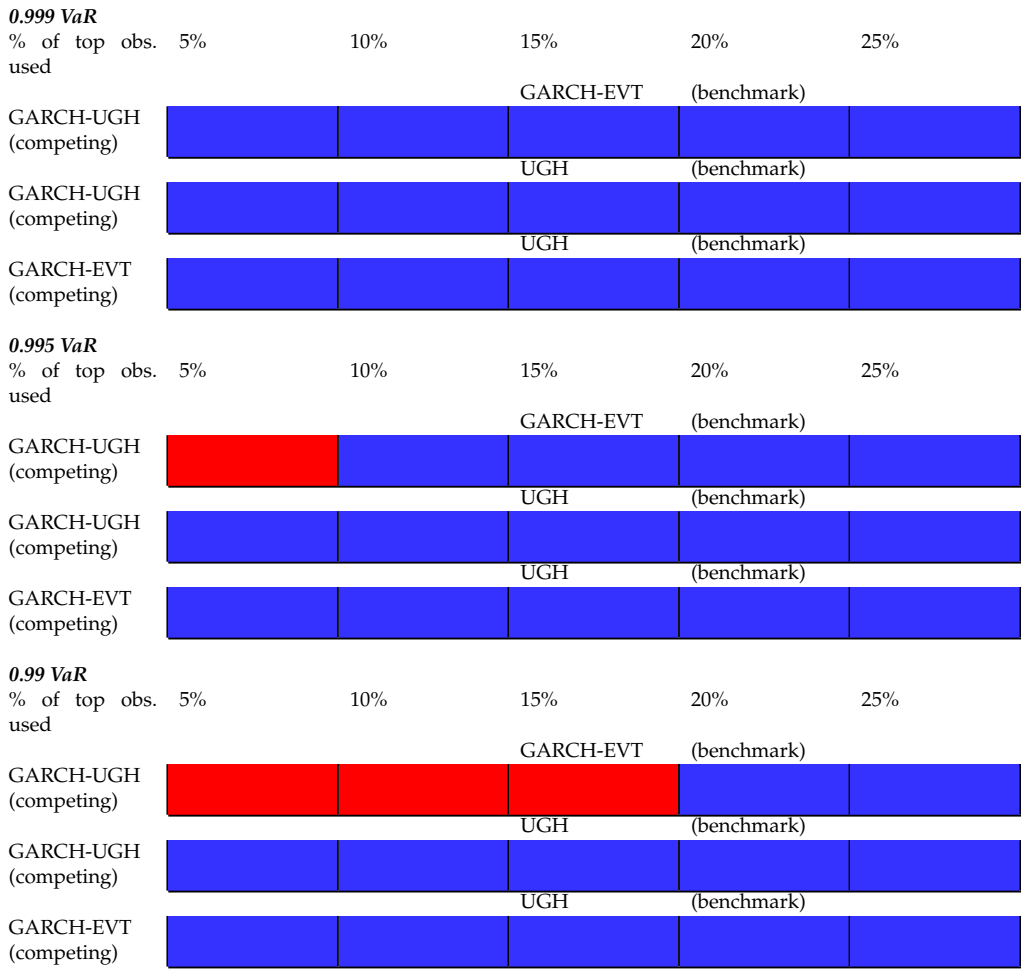
Notes: When the competing model in the row is swapped with the benchmark model in the column, the results are reversed.

TABLE 3.13: Comparative backtesting: out-of-sample evaluations (traffic light matrices) of one-step ahead conditional VaR estimates by EVT-type methods from 29 May 1997 to 12 August 2009 at different quantile levels for the negative log-returns of NIKKEI index by means of the Diebold-Mariano test using $h = 0$ VaR scoring function.



Notes: When the competing model in the row is swapped with the benchmark model in the column, the results are reversed.

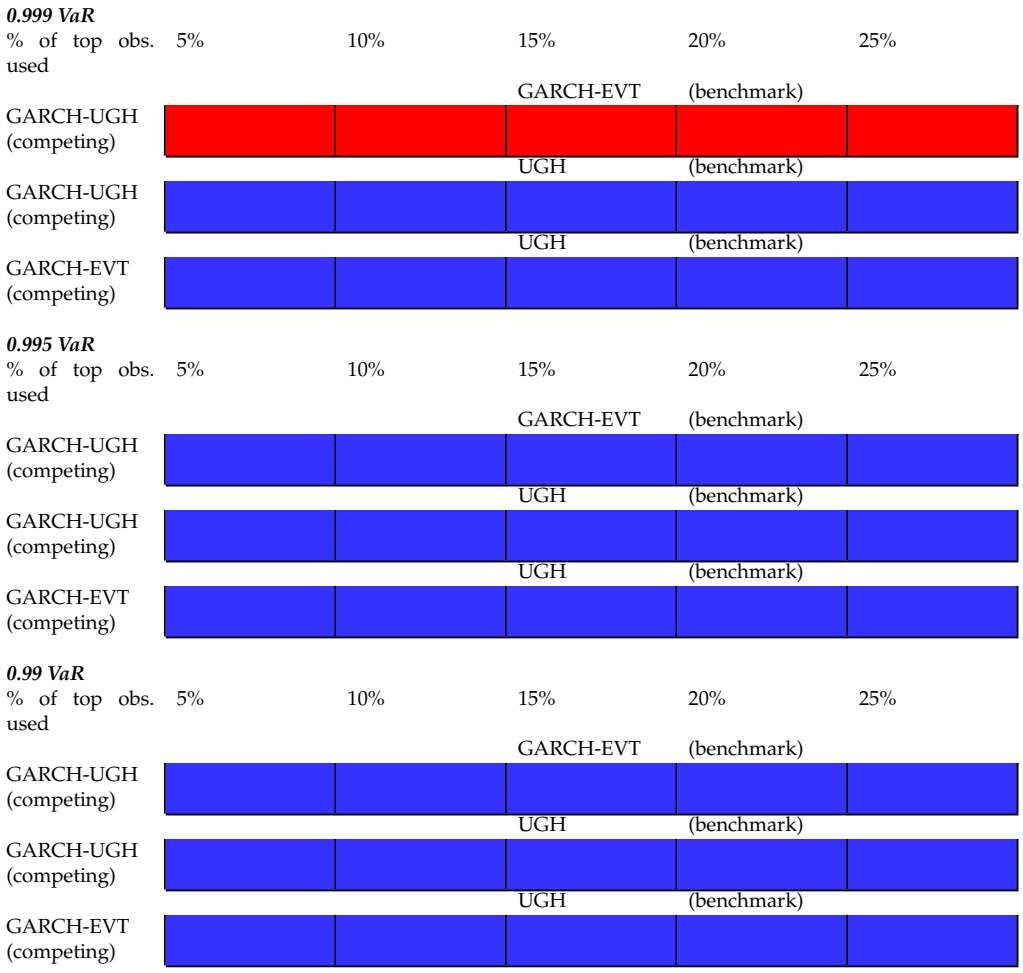
TABLE 3.14: Comparative backtesting: out-of-sample evaluations (traffic light matrices) of one-step ahead conditional VaR estimates by EVT-type methods from 28 September 2002 to 14 December 2010 at different quantile levels for the negative log-returns of JPY/GBP exchange rate by means of the Diebold-Mariano test using $h = 0$ VaR scoring function.



Notes: When the competing model in the row is swapped with the benchmark model in the column, the results are reversed.

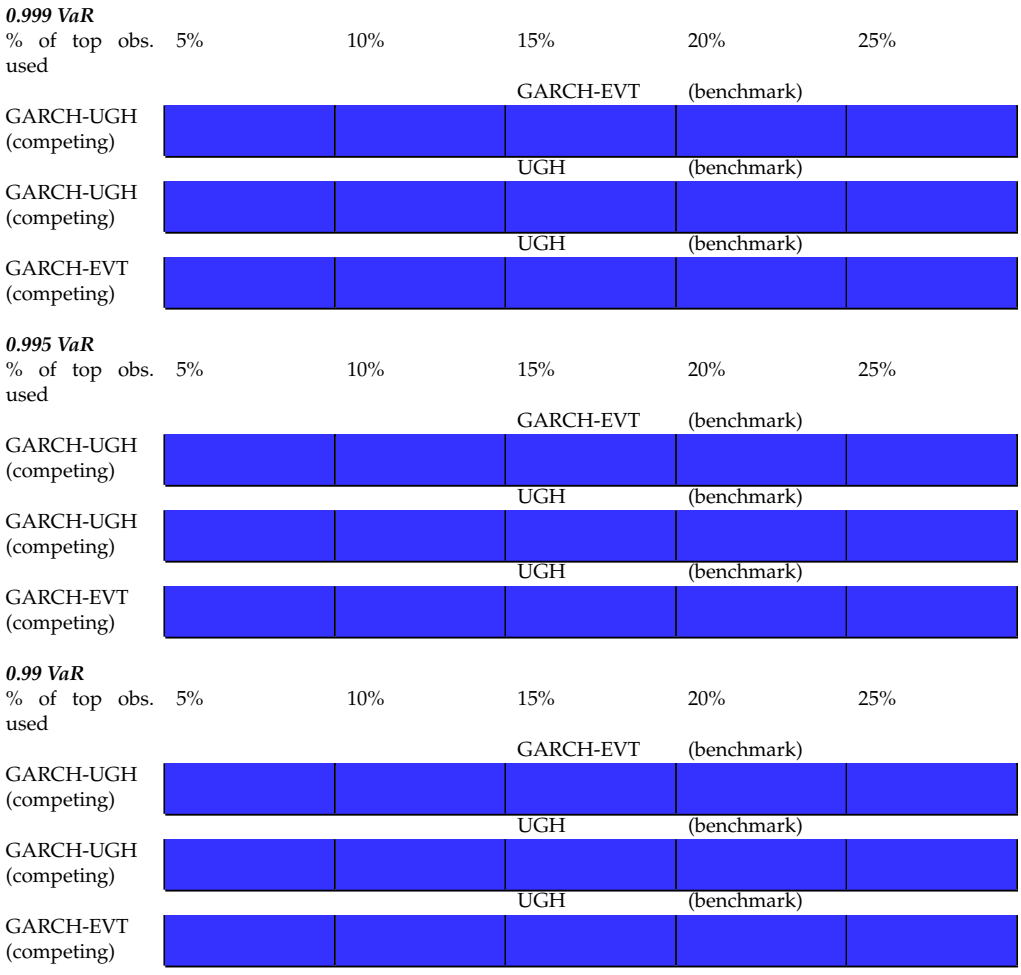
at level $\tau = 0.999$. However, it does not mean that the form $h = 0$ is the optimal scoring function for the VaR because there exists no theoretical support to use in practice. Comparative VaR backtestings rank the VaR estimation methods based on the realized VaR scores. They hence yield definitive answers to the cases when the estimation methods are all accepted or all rejected in the traditional VaR backtestings, especially when GARCH-UGH and GARCH-EVT approaches have the same number of VaR violations and are indistinguishable.

TABLE 3.15: Comparative backtesting: out-of-sample evaluations (traffic light matrices) of one-step ahead conditional VaR estimates by EVT-type methods from 8 December 1997 to 9 November 2009 at different quantile levels for the negative log-returns of DJ index by means of the Diebold-Mariano test using $h = 1$ VaR scoring function.



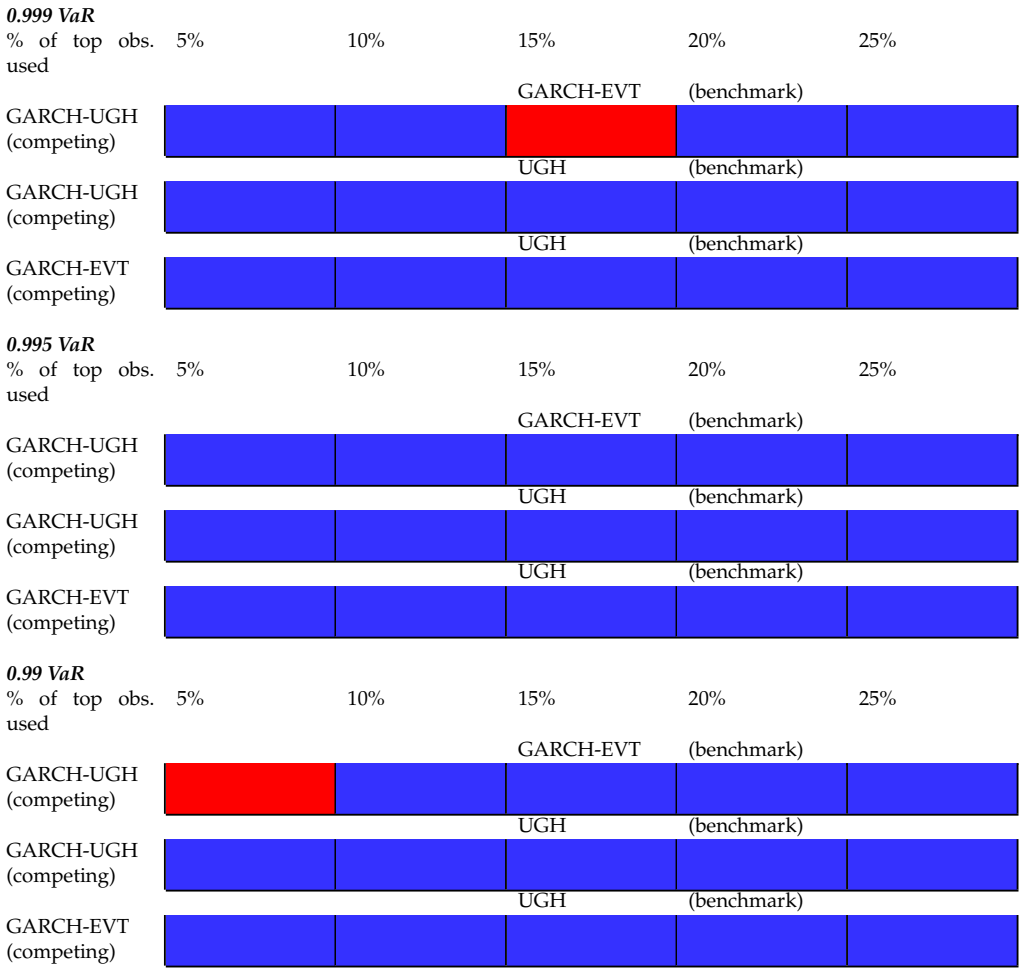
Notes: When the competing model in the row is swapped with the benchmark model in the column, the results are reversed.

TABLE 3.16: Comparative backtesting: out-of-sample evaluations (traffic light matrices) of one-step ahead conditional VaR estimates by EVT-type methods from 13 August 1997 to 16 July 2009 at different quantile levels for the negative log-returns of NASDAQ index by means of the Diebold-Mariano test using $h = 1$ VaR scoring function.



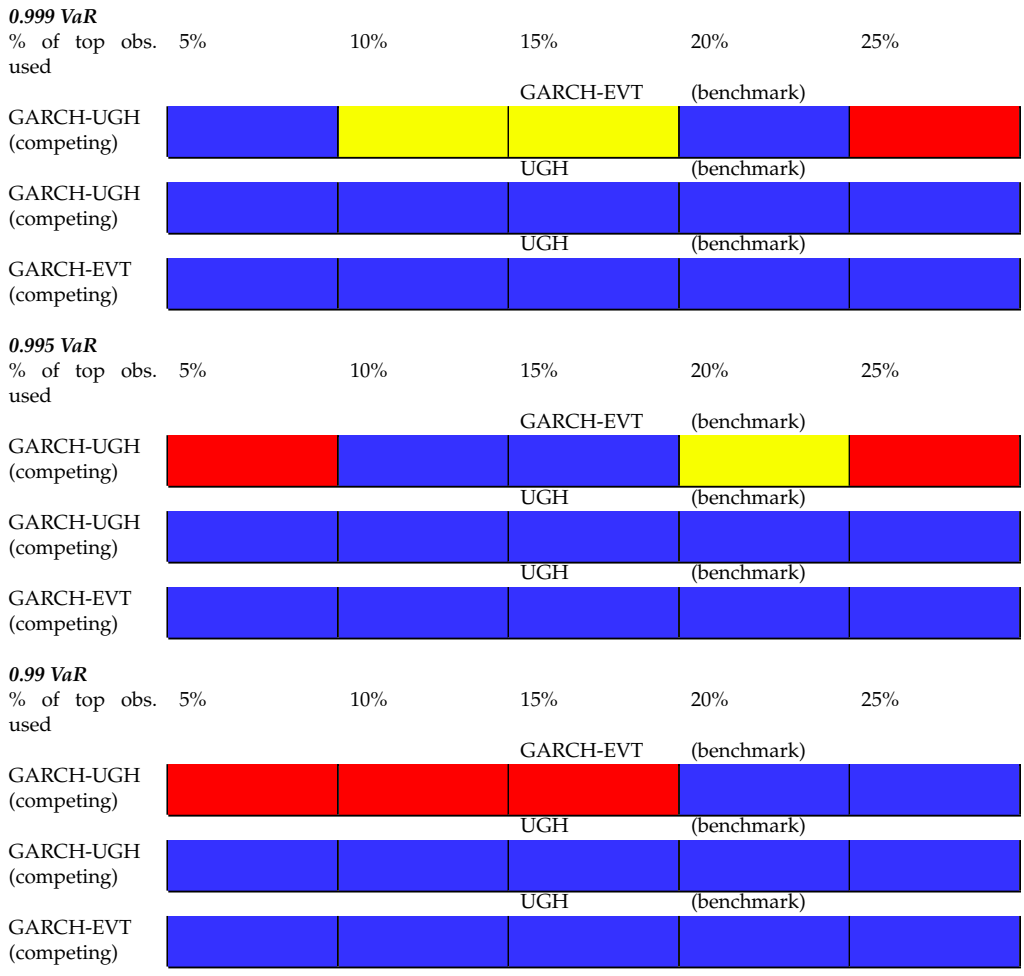
Notes: When the competing model in the row is swapped with the benchmark model in the column, the results are reversed.

TABLE 3.17: Comparative backtesting: out-of-sample evaluations (traffic light matrices) of one-step ahead conditional VaR estimates by EVT-type methods from 29 May 1997 to 12 August 2009 at different quantile levels for the negative log-returns of NIKKEI index by means of the Diebold-Mariano test using $h = 1$ VaR scoring function.



Notes: When the competing model in the row is swapped with the benchmark model in the column, the results are reversed.

TABLE 3.18: Comparative backtesting: out-of-sample evaluations (traffic light matrices) of one-step ahead conditional VaR estimates by EVT-type methods from 28 September 2002 to 14 December 2010 at different quantile levels for the negative log-returns of JPY/GBP exchange rate by means of the Diebold-Mariano test using $h = 1$ VaR scoring function.



Notes: When the competing model in the row is swapped with the benchmark model in the column, the results are reversed.

3.6.3.2 Comparison with basic estimation methods

Traditional VaR backtesting

Table 3.19 also supports the use of the GARCH-UGH approach for the estimation of dynamic extreme VaR because it outperforms the basic HS, GARCH-N and GARCH-t estimation methods. In 12 out of 12 cases our GARCH-UGH approach (with optimal sample fraction according to Tables 3.7-3.10) is closest to the mark. It also never fails either of the Kupiec and Christoffersen tests, while HS fails 8 and 7 times, GARCH-N fails 10 and 10 times, and GARCH-t fails 3 and 2 times out of 12 cases, respectively.

Comparative VaR backtesting

Tables 3.20-3.23 and 3.24-3.27 display the traffic light matrices of comparative VaR backtesting (see Section 3.5) for six estimation methods given in Section 3.3, three quantile levels and four financial time series when $h = 0$ (3.16) and $h = 1$ (3.15) VaR scoring functions are used, respectively. The optimal sample fraction for 3 EVT-type methods is selected based on the performance in the out-of-sample traditional VaR backtestings (see Tables 3.7-3.10).

As with the results of traditional VaR backtestings, it is illustrated that our proposed GARCH-UGH approach appears to be best overall, outperforming other five estimation methods. The two scoring functions result in a good agreement with GARCH-UGH approach being the better estimator in 11 out of 12 cases when compared to the basic HS, GARCH-N and GARCH-t approaches. It also suggests that there was no difference in the discrimination ability of both chosen VaR scoring functions having 2 cases of no decisions made in each function. HS, GARCH-N and UGH approaches generally perform worse than the other three approaches as they consider neither heavy-tail nor volatility.

We conclude from our empirical analysis that the proposed GARCH-UGH approach provides better one-step ahead dynamic extreme VaR estimates for financial time series than the benchmark conventional GARCH-EVT approach of McNeil and Frey (2000) and the other basic estimation approaches we have tested based on historical simulation or traditional fully parametric models. This can be seen from both

TABLE 3.19: Traditional backtesting: out-of-sample evaluations of one-step ahead conditional VaR estimates by basic estimation methods at different quantile levels for the negative log-returns of DJ, NASDAQ, NIKKEI indices and JPY/GBP exchange rate (time period given in Section 3.6.3) by means of the number of VaR violations, unconditional and conditional coverage tests.

	DJ	NASDAQ	NIKKEI	JPY/GBP
Testing window	3000	3000	3000	3000
Estimation window	1000	1000	1000	1000
0.999 Quantile				
Expected	3	3	3	3
HS	4	5	7	6
	(0.583, 0.855)	(0.292, 0.569)	(0.049, 0.142)	(0.128, 0.310)
GARCH-N	19	11	11	10
	(0.000, 0.000)	(0.000, 0.002)	(0.000, 0.002)	(0.001, 0.006)
GARCH-*t*	**3**	7	5	1
	(1.000, 0.997)	(0.049, 0.142)	(0.292, 0.569)	(0.179, 0.406)
GARCH-UGH	**3**	**3**	**3**	**3**
	(1.000, 0.997)	(1.000, 0.997)	(1.000, 0.997)	(1.000, 0.997)
0.995 Quantile				
Expected	15	15	15	15
HS	36	39	24	21
	(0.000, 0.000)	(0.000, 0.000)	(0.032, 0.042)	(0.143, 0.114)
GARCH-N	34	22	29	29
	(0.000, 0.000)	(0.090, 0.086)	(0.001, 0.004)	(0.001, 0.004)
GARCH-*t*	17	16	13	1
	(0.612, 0.798)	(0.798, 0.888)	(0.596, 0.821)	(0.000, 0.000)
GARCH-UGH	**14**	**15**	**15**	**15**
	(0.793, 0.905)	(1.000, 0.927)	(1.000, 0.927)	(1.000, 0.927)
0.99 Quantile				
Expected	30	30	30	30
HS	57	68	44	44
	(0.000, 0.000)	(0.000, 0.000)	(0.016, 0.022)	(0.016, 0.022)
GARCH-N	56	38	44	45
	(0.000, 0.000)	(0.159, 0.297)	(0.016, 0.029)	(0.010, 0.019)
GARCH-*t*	26	25	20	2
	(0.453, 0.601)	(0.345, 0.287)	(0.051, 0.130)	(0.000, 0.000)
GARCH-UGH	**31**	**30**	**30**	**34**
	(0.855, 0.711)	(1.000, 0.594)	(1.000, 0.738)	(0.472, 0.523)

Notes: The closest number of VaR violations to the theoretically expected number is highlighted in bold. The number of VaR violations for GARCH-UGH is when the optimal sample fraction is selected according to Tables 3.7-3.10. The p-values for the unconditional coverage test by Kupiec (1995) and conditional coverage test by Christoffersen (1998) are given in brackets in order.

TABLE 3.20: Comparative backtesting: out-of-sample evaluations (traffic light matrices) of one-step ahead conditional VaR estimates from 8 December 1997 to 9 November 2009 at different quantile levels for the negative log-returns of DJ index by means of the Diebold-Mariano test using $h = 0$ VaR scoring function.

0.999 VaR

	HS	GARCH-N	GARCH-*t*	UGH	GARCH-EVT	GARCH-UGH
HS						
GARCH-N						
GARCH-*t*						
UGH (25%)						
GARCH-EVT (5%)						
GARCH-UGH (15%)						

0.995 VaR

	HS	GARCH-N	GARCH-*t*	UGH	GARCH-EVT	GARCH-UGH
HS						
GARCH-N						
GARCH-*t*						
UGH (25%)						
GARCH-EVT (25%)						
GARCH-UGH (25%)						

0.99 VaR

	HS	GARCH-N	GARCH-*t*	UGH	GARCH-EVT	GARCH-UGH
HS						
GARCH-N						
GARCH-*t*						
UGH (25%)						
GARCH-EVT (15%)						
GARCH-UGH (20%)						

Notes: The optimal sample fraction is selected for 3 EVT-type methods based on the performance in the traditional out-of-sample backtestings (see Table 3.7).

TABLE 3.21: Comparative backtesting: out-of-sample evaluations (traffic light matrices) of one-step ahead conditional VaR estimates from 13 August 1997 to 16 July 2009 at different quantile levels for the negative log-returns of NASDAQ index by means of the Diebold-Mariano test using $h = 0$ VaR scoring function.

0.999 VaR

	HS	GARCH-N	GARCH-*t*	UGH	GARCH-EVT	GARCH-UGH
HS						
GARCH-N						
GARCH-*t*						
UGH (25%)						
GARCH-EVT (25%)						
GARCH-UGH (25%)						

0.995 VaR

	HS	GARCH-N	GARCH-*t*	UGH	GARCH-EVT	GARCH-UGH
HS						
GARCH-N						
GARCH-*t*						
UGH (15%)						
GARCH-EVT (10%)						
GARCH-UGH (15%)						

0.99 VaR

	HS	GARCH-N	GARCH-*t*	UGH	GARCH-EVT	GARCH-UGH
HS						
GARCH-N						
GARCH-*t*						
UGH (25%)						
GARCH-EVT (5%)						
GARCH-UGH (20%)						

Notes: The optimal sample fraction is selected for 3 EVT-type methods based on the performance in the traditional out-of-sample backtestings (see Table 3.8).

TABLE 3.22: Comparative backtesting: out-of-sample evaluations (traffic light matrices) of one-step ahead conditional VaR estimates from 29 May 1997 to 12 August 2009 at different quantile levels for the negative log-returns of NIKKEI index by means of the Diebold-Mariano test using $h = 0$ VaR scoring function.

0.999 VaR

	HS	GARCH-N	GARCH-*t*	UGH	GARCH-EVT	GARCH-UGH
HS						
GARCH-N						
GARCH-*t*						
UGH (20%)						
GARCH-EVT (10%)						
GARCH-UGH (10%)						

0.995 VaR

	HS	GARCH-N	GARCH-*t*	UGH	GARCH-EVT	GARCH-UGH
HS						
GARCH-N						
GARCH-*t*						
UGH (25%)						
GARCH-EVT (10%)						
GARCH-UGH (10%)						

0.99 VaR

	HS	GARCH-N	GARCH-*t*	UGH	GARCH-EVT	GARCH-UGH
HS						
GARCH-N						
GARCH-*t*						
UGH (20%)						
GARCH-EVT (10%)						
GARCH-UGH (20%)						

Notes: The optimal sample fraction is selected for 3 EVT-type methods based on the performance in the traditional out-of-sample backtestings (see Table 3.9).

TABLE 3.23: Comparative backtesting: out-of-sample evaluations (traffic light matrices) of one-step ahead conditional VaR estimates from 28 September 2002 to 14 December 2010 at different quantile levels for the negative log-returns of JPY/GBP exchange rate by means of the Diebold-Mariano test using $h = 0$ VaR scoring function.

0.999 VaR

	HS	GARCH-N	GARCH-*t*	UGH	GARCH-EVT	GARCH-UGH
HS						
GARCH-N						
GARCH-*t*						
UGH (25%)						
GARCH-EVT (15%)						
GARCH-UGH (5%)						

0.995 VaR

	HS	GARCH-N	GARCH-*t*	UGH	GARCH-EVT	GARCH-UGH
HS						
GARCH-N						
GARCH-*t*						
UGH (5%)						
GARCH-EVT (5%)						
GARCH-UGH (15%)						

0.99 VaR

	HS	GARCH-N	GARCH-*t*	UGH	GARCH-EVT	GARCH-UGH
HS						
GARCH-N						
GARCH-*t*						
UGH (5%)						
GARCH-EVT (25%)						
GARCH-UGH (25%)						

Notes: The optimal sample fraction is selected for 3 EVT-type methods based on the performance in the traditional out-of-sample backtestings (see Table 3.10).

TABLE 3.24: Comparative backtesting: out-of-sample evaluations (traffic light matrices) of one-step ahead conditional VaR estimates from 8 December 1997 to 9 November 2009 at different quantile levels for the negative log-returns of DJ index by means of the Diebold-Mariano test using $h = 1$ VaR scoring function.

0.999 VaR

	HS	GARCH-N	GARCH-*t*	UGH	GARCH-EVT	GARCH-UGH
HS						
GARCH-N						
GARCH-*t*						
UGH (25%)						
GARCH-EVT (5%)						
GARCH-UGH (15%)						

0.995 VaR

	HS	GARCH-N	GARCH-*t*	UGH	GARCH-EVT	GARCH-UGH
HS						
GARCH-N						
GARCH-*t*						
UGH (25%)						
GARCH-EVT (25%)						
GARCH-UGH (25%)						

0.99 VaR

	HS	GARCH-N	GARCH-*t*	UGH	GARCH-EVT	GARCH-UGH
HS						
GARCH-N						
GARCH-*t*						
UGH (25%)						
GARCH-EVT (15%)						
GARCH-UGH (20%)						

Notes: The optimal sample fraction is selected for 3 EVT-type methods based on the performance in the traditional out-of-sample backtestings (see Table 3.7).

TABLE 3.25: Comparative backtesting: out-of-sample evaluations (traffic light matrices) of one-step ahead conditional VaR estimates from 13 August 1997 to 16 July 2009 at different quantile levels for the negative log-returns of NASDAQ index by means of the Diebold-Mariano test using $h = 1$ VaR scoring function.

0.999 VaR

	HS	GARCH-N	GARCH-*t*	UGH	GARCH-EVT	GARCH-UGH
HS						
GARCH-N						
GARCH-*t*						
UGH (25%)						
GARCH-EVT (25%)						
GARCH-UGH (25%)						

0.995 VaR

	HS	GARCH-N	GARCH-*t*	UGH	GARCH-EVT	GARCH-UGH
HS						
GARCH-N						
GARCH-*t*						
UGH (15%)						
GARCH-EVT (10%)						
GARCH-UGH (15%)						

0.99 VaR

	HS	GARCH-N	GARCH-*t*	UGH	GARCH-EVT	GARCH-UGH
HS						
GARCH-N						
GARCH-*t*						
UGH (25%)						
GARCH-EVT (5%)						
GARCH-UGH (20%)						

Notes: The optimal sample fraction is selected for 3 EVT-type methods based on the performance in the traditional out-of-sample backtestings (see Table 3.8).

TABLE 3.26: Comparative backtesting: out-of-sample evaluations (traffic light matrices) of one-step ahead conditional VaR estimates from 29 May 1997 to 12 August 2009 at different quantile levels for the negative log-returns of NIKKEI index by means of the Diebold-Mariano test using $h = 1$ VaR scoring function.

0.999 VaR

	HS	GARCH-N	GARCH-*t*	UGH	GARCH-EVT	GARCH-UGH
HS						
GARCH-N						
GARCH-*t*						
UGH (20%)						
GARCH-EVT (10%)						
GARCH-UGH (10%)						

0.995 VaR

	HS	GARCH-N	GARCH-*t*	UGH	GARCH-EVT	GARCH-UGH
HS						
GARCH-N						
GARCH-*t*						
UGH (25%)						
GARCH-EVT (10%)						
GARCH-UGH (10%)						

0.99 VaR

	HS	GARCH-N	GARCH-*t*	UGH	GARCH-EVT	GARCH-UGH
HS						
GARCH-N						
GARCH-*t*						
UGH (20%)						
GARCH-EVT (10%)						
GARCH-UGH (20%)						

Notes: The optimal sample fraction is selected for 3 EVT-type methods based on the performance in the traditional out-of-sample backtestings (see Table 3.9).

TABLE 3.27: Comparative backtesting: out-of-sample evaluations (traffic light matrices) of one-step ahead conditional VaR estimates from 28 September 2002 to 14 December 2010 at different quantile levels for the negative log-returns of JPY/GBP exchange rate by means of the Diebold-Mariano test using $h = 1$ VaR scoring function.

0.999 VaR

	HS	GARCH-N	GARCH-*t*	UGH	GARCH-EVT	GARCH-UGH
HS						
GARCH-N						
GARCH-*t*						
UGH (25%)						
GARCH-EVT (15%)						
GARCH-UGH (5%)						

0.995 VaR

	HS	GARCH-N	GARCH-*t*	UGH	GARCH-EVT	GARCH-UGH
HS						
GARCH-N						
GARCH-*t*						
UGH (5%)						
GARCH-EVT (5%)						
GARCH-UGH (15%)						

0.99 VaR

	HS	GARCH-N	GARCH-*t*	UGH	GARCH-EVT	GARCH-UGH
HS						
GARCH-N						
GARCH-*t*						
UGH (5%)						
GARCH-EVT (25%)						
GARCH-UGH (25%)						

Notes: The optimal sample fraction is selected for 3 EVT-type methods based on the performance in the traditional out-of-sample backtestings (see Table 3.10).

the in-sample and the out-of-sample estimations at several quantile levels τ, including the very high $\tau = 0.999$ corresponding to a 99.9% VaR, and a large range of sample fractions k, due to the effect of the bias correction. Let us also point out that the GARCH-UGH method is carried out using an automatic recipe for the estimation of the extreme value index and extreme quantile, making it computationally cheap.

3.6.3.3 Constructing confidence interval of GARCH-UGH estimates

The corresponding plots of out-of-sample backtesting are shown in Figures 3.4-3.7 with the corresponding 95% asymptotic Gaussian confidence intervals corresponding to the GARCH-UGH estimation method in Figure 3.8: the confidence interval is given by

$$\left[\hat{\mu}_{t+1} + \hat{\sigma}_{t+1}\hat{q}_{1-p}(Z)\left(1 \pm \frac{1.96}{\sqrt{k}/\log(\frac{k}{np})} \times \sqrt{\frac{\hat{\gamma}^2_{k,k_\rho}}{\hat{\rho}^2_{k_\rho}}(\hat{\rho}^2_{k_\rho} + (1-\hat{\rho}_{k_\rho})^2)}\right)\right].$$

It is clearly seen that the GARCH-UGH and GARCH-EVT estimates have the same dynamics, with the bias correction shifting the estimate upwards or downwards depending on the rolling estimation window.

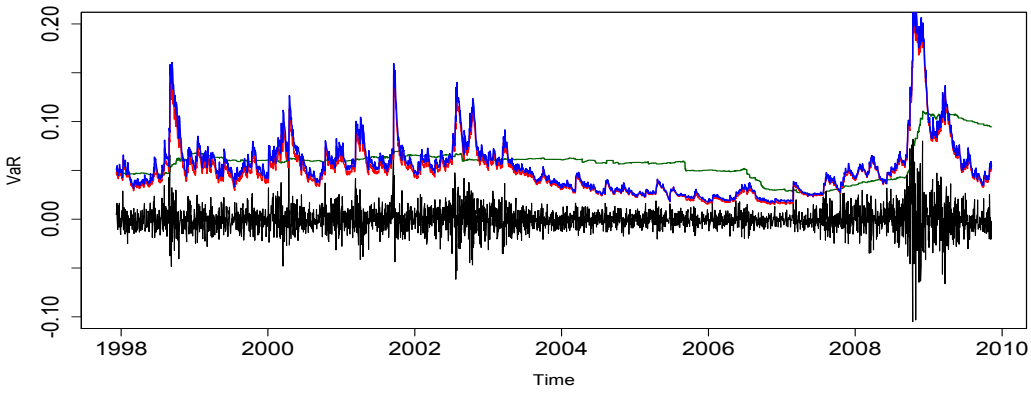


FIGURE 3.4: Out-of-sample backtesting of the DJ index from 8 December 1997 to 9 November 2009, and 99.9%-VaR estimates calculated using rolling estimation windows made of 1000 observations, with k corresponding to the top 15% observations from this window. GARCH-UGH (blue line), GARCH-EVT (red line) and UGH (dark green line) estimates are superimposed on the negative log-returns (black line).

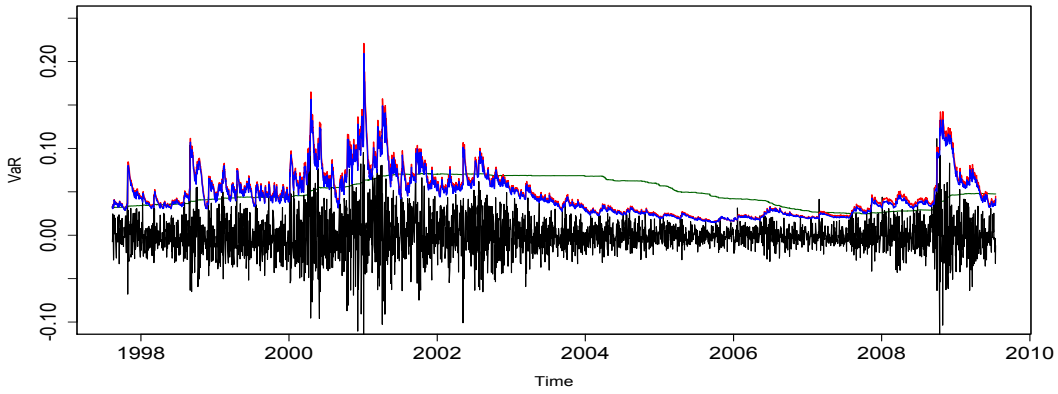


FIGURE 3.5: Out-of-sample backtesting of the NASDAQ index from 13 August 1997 to 16 July 2009, and 99.9%-VaR estimates calculated using rolling estimation windows made of 1000 observations, with k corresponding to the top 20% of observations from this window. GARCH-UGH (blue line), GARCH-EVT (red line) and UGH (dark green line) estimates are superimposed on the negative log-returns (black line).

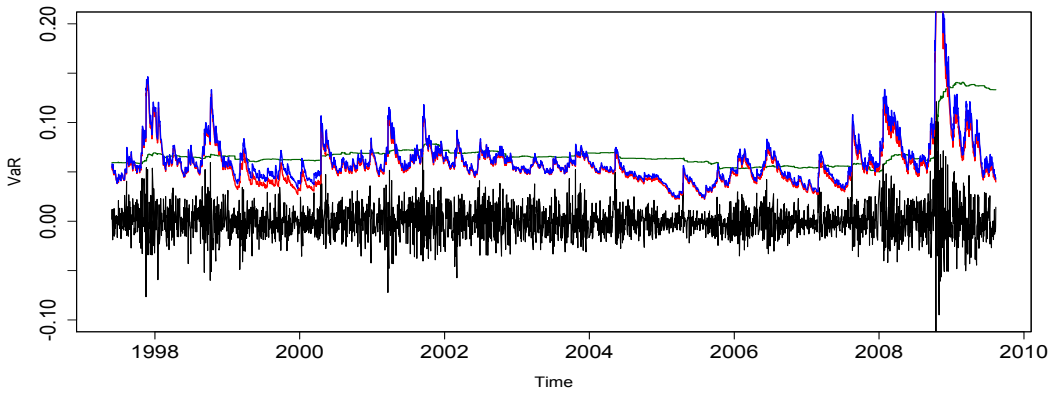


FIGURE 3.6: Out-of-sample backtesting of the NIKKEI index from 29 May 1997 to 12 August 2009, and 99.9%-VaR estimates calculated using rolling estimation windows made of 1000 observations, with k corresponding to the top 10% of observations from this window. GARCH-UGH (blue line), GARCH-EVT (red line) and UGH (dark green line) estimates are superimposed on the negative log-returns (black line).

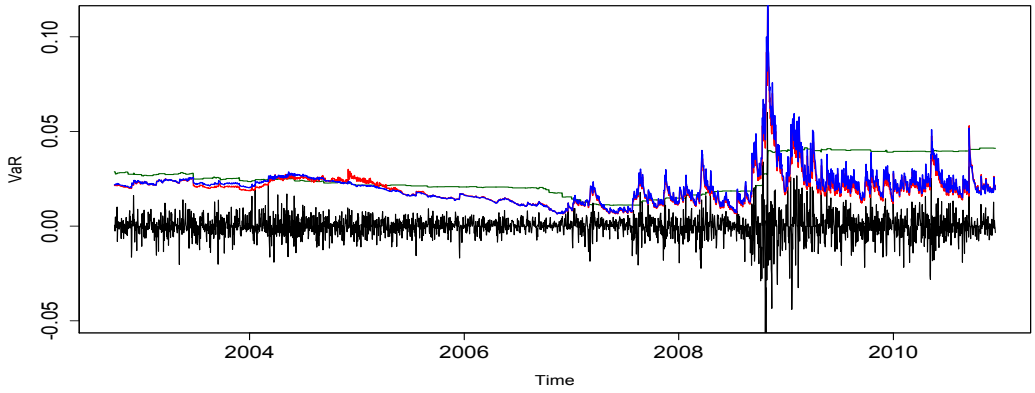


FIGURE 3.7: Out-of-sample backtesting of the JPY/GBP exchange rate from 28 September 2002 to 14 December 2010, and 99.9%-VaR estimates calculated using rolling estimation windows made of 1000 observations, with k corresponding to the top 10% of observations from this window. GARCH-UGH (blue line), GARCH-EVT (red line) and UGH (dark green line) estimates are superimposed on the negative log-returns (black line).

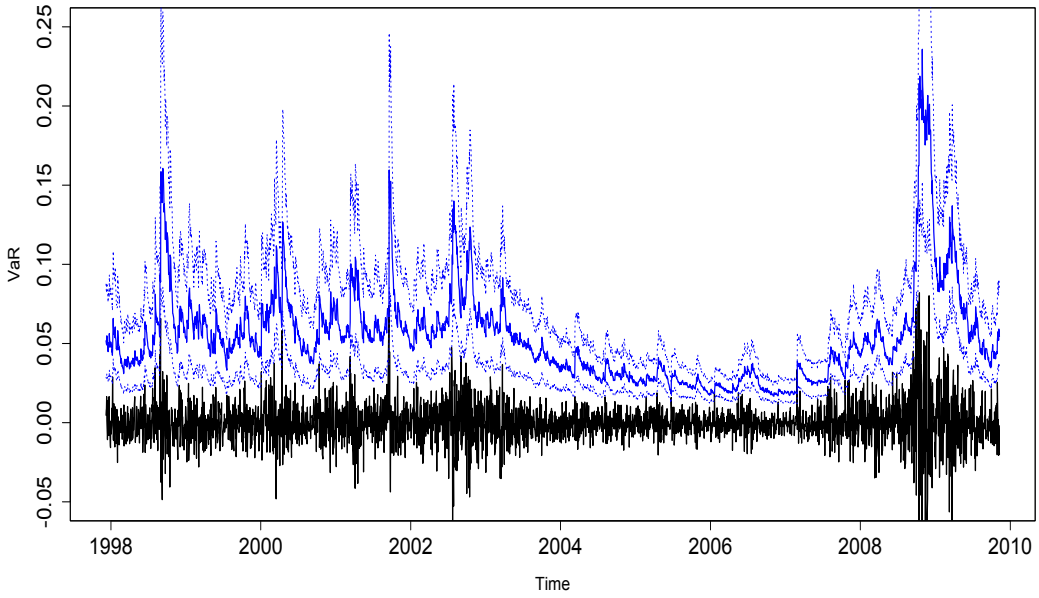


FIGURE 3.8: Out-of-sample backtesting of the DJ index from 8 December 1997 to 9 November 2009, and 99.9%-VaR estimates calculated using rolling estimation windows made of 1000 observations, with k corresponding to the top 15% observations from this window. GARCH-UGH (blue solid line) estimates are superimposed on the negative log-returns (black line) with the 95% asymptotic Gaussian confidence intervals (blue dashed line).

Chapter 4

Dynamic extreme Expected Shortfall estimation by GARCH-UGH

4.1 Introduction

Recall from Chapter 1 and 3 that recently the BCBS announced a change in the risk measure used for capital requirements in internal market risk models, moving from the VaR to the ES despite VaR's universality and conceptual simplicity. In practice there was a lively debate of which risk measure either VaR or ES would be best in regulatory framework over the last or two decades. In this section, we revise the pros and cons of both VaR and ES that were part of the debate, and describe the problems and aims of this chapter.

The following list discusses the pros and cons of both VaR and ES (not exhaustive):

- Tail risk information: VaR only measures the frequency of observations below or above the predictor and not their magnitude, i.e., severity of tail losses. This means that, while it is known that $100(1-\tau)\%$ of losses will be higher than the VaR q_τ at level τ, the VaR alone cannot give any further information about the size of these large losses. It is revealed for example during the turbulence of 2007-2008 crisis. The ES on the hand does give information of tail loss by definition as it is the mean of potential extreme losses exceeding the VaR.

- Coherence: VaR is not a coherent risk measure in the sense of Artzner et al. (1999), because it is not subadditive in general, meaning that it does not abide by the intuitive diversification principle stating that a portfolio built on several financial assets carries less risk than a portfolio solely consisting of one of these assets. Contrary to the VaR, ES is a coherent risk measure satisfying all axioms including monotonicity, homogeneity, translation invariance and subadditivity.

- Robustness: VaR is shown to be more robust than ES. We say that a risk measure is robust if it can accommodate model misspecification and is robust statistically regarding the changes in data. He et al. (2022) describe the statistical robustness of VaR and ES using four tools that are influence functions, asymptotic breakdown points, finite sample breakdown points and Hampel robustness (see also Emmer et al. 2015 for the detailed discussion).

- Elicitability: elicitability is a helpful decision-theoretic framework for the determination of optimal point forecasts, which can be used to compare the performance of different estimation methods of VaR and ES in the comparative backtesting (see Section 3.5 and 4.5). While the VaR (quantile) is elicitable, Gneiting (2011) points out that ES is not elicitable. This means that there exists no scoring function $S_{\text{ES}}(e, y)$ such that the ES estimate e of the true ES y can be obtained as the e that minimizes $S_{\text{ES}}(e, y)$. Given this definition, it is quite obvious that ES is not elicitable because there is no concrete realized data to be compared to the estimates of ES.

- Backtestability: the method of VaR backtesting is conceptually simple since it is based on the number of VaR violations, explained in Section 3.4. On the other hand, one of the major drawback of ES is its difficulty to be backtested. Researches from the academic sector and the practical sector are still struggling to find an optimal backtesting method that is both mathematically consistent and practically implementable. Recent theoretical progress in the research of ES has made the backtesting ES a viable task; see next Section 4.2 for a very recent literature review of backtesting procedures of ES. This difficulty of ES is mainly due to the fact that ES is highly model dependent and particularly

sensitive to the extreme tails as the estimation of ES relies on the estimation of VaR by definition. Clearly, VaR is not model dependent and easy to evaluate.

With regard to estimation of ES, there has not been sufficient investigation to establish the superiority of a certain estimator relative to the others in the literature, discussed in Section 4.2. In addition, no particular type of ES model is prescribed in the framework of Basel Committee on Banking Supervision (2019). In Section 4.3.1, we propose a novel approach of dynamic extreme ES estimation, which is based on our proposed GARCH-UGH approach from Chapter 3 and the use of asymptotic equivalence between VaR (quantile) and ES. Regarding the backtesting of ES, Basel Committee on Banking Supervision (2019) still demands financial institutions to use traditional VaR backtesting for ES. At the same time, we can expect that upcoming regulations will require them to backtest ES without using VaR backtesting method. We also tackle an urgent problem of which ES backtesting methods can be used in the practice (see Section 4.4 and 4.5).

4.2 Comprehensive study of ES

In this section, we review the ES from the perspectives of estimation and backtesting methods.

Estimation methods

There have been a number of literature that discuss the use of EVT for estimating (one-step and multi-step ahead) unconditional and conditional ES. We start from the two-step frameworks for the estimation of ES. McNeil and Frey (2000) use GARCH-EVT approach (see Section 4.3.2) to estimate 1-step ahead conditional ES and suggest to use a heavy-tailed distribution, preferably using EVT instead of normal distribution to model the standardized residuals after filtering in GARCH-EVT framework. Righi and Ceretta (2015) evaluate the several methods of unconditional, conditional and quantile/expectile regression-based models for 1-step ahead VaR and ES estimations. They find that the poor VaR estimates will result in the bad ES predictions. Bee et al. (2016) propose the two-step realized EVT approach, where financial returns are prewhitened with a high-frequency volatility model instead of GARCH model to

estimate 1-step and 10-step ahead VaR and ES. According to their results, GARCH-type filters perform slightly better than the high-frequency based filters although the realized EVT approach seems preferable at the longer time horizons when estimating multi-step ahead conditional ES. Novales and Garcia-Jorcano (2019) combine the semiparametric filtered historical simulation and EVT, i.e., POT approach, to estimate 10-step ahead conditional ES, which corrects the overestimation of risk by EVT-based models. Their result suggests that conditional EVT-based models to produce more accurate 1-step and 10-step ahead ES estimates than non-EVT based models.

We also look at the cases when EVT methods are not used. So and Wong (2012) estimate multi-step ahead ES under GARCH models that is computationally feasible to use in practice, compare to any Monte Carlo method that requires heavy computational effort to make the method widely applicable. Their estimation method combines the exact estimation of conditional kurtosis and GARCH models, which is based on the idea from Wong and So (2003). Lönnbark (2016) studies four different approaches including the method based on a skewed-t distribution to compare the estimations of multi-step ahead VaR and ES. Degiannakis and Potamia (2017) exploit the inter-day and intra-day volatility models to construct two-step GARCH-skewed t (see Section 4.3.2) and realized skewed t frameworks for the estimations of multi-step ahead conditional VaR and ES. Their models provide accurate estimations of VaR and ES in the case of 97.5% confidence level that is set up by Basel Committee on Banking Supervision (2019) but not in the case of the higher 99% confidence level.

Lastly we check the estimation of ES by means of the expectile. Taylor (2008) introduces the expectile-based unconditional VaR and ES that are only estimated at an intermediate level, i.e, not $q_\tau(Z)$ ($\tau \uparrow 1$), although financial institutions are typically interested in the extreme region. Daouia et al. (2018) and Daouia et al. (2020) propose a novel extreme expectile-based unconditional VaR and ES. Their method relies on the popular Hill estimator (3.6), Weissman quantile estimator (3.5), heavy-tailed property (3.4) and on the asymptotic equivalence between quantiles and expectiles (4.3) to transform quantile-based VaR and ES into expectile-based ones. Daouia et al.

(2021) introduce the estimator of EVI (tail index) that is based on weighted combinations of top order statistics and asymmetric least squares estimates. This resulting estimator called expectHill is used to estimate unconditional quantile-based and expectile-based ES. See Section 1.2.5 for the pros and cons of using the expectiles instead of quantiles.

Backtesting methods

In contrast to the estimations of unconditional and conditional ES where most of the existing models including the ones we referred and proposed for the VaR in Chapter 3 can easily be adapted to the ES, such adaptions are not straight-forward for backtesting ES estimates (Emmer et al. 2015). The main difficulty in backtesting ES, which is already discussed in Chapter 1 and Section 4.1, is its nonelicitability (Gneiting 2011; Fissler and Ziegel 2016). More specifically, there is no analogue to the hit sequence of VaR violations that lies at the heart of almost all traditional VaR backtestings apart from the comparative version. However, there are continuously growing literature regarding the traditional and comparative backtesting methods of ES, driven by the recent transition from VaR to ES in the Basel framework Basel Committee on Banking Supervision (2019). From the practical point of view, sudden appearance of multiple ES backtesting methods may not be that blessing because it is often difficult to find a middle ground between a reliable test of ES and the ease of the implementation in practice. A very recent comprehensive study of ES backtesting methods is given in Novales and Garcia-Jorcano (2019) and Deng and Qiu (2021), who check the performances of tests in terms of stability over different models, sensitivity to the sample sizes and computational burden.

We classify traditional ES backtesting methods into 5 groups by the inputs required for the tests as follows: using the whole or tail distribution of the financial returns, using multiple quantile levels of VaR, using ES, VaR and volatility (or shortfall deviation), using the pair of VaR and ES and using only ES. We also review the comparative ES backtesting, which is similar to the comparative VaR backtesting (see Section 3.5). As in every statistical method, every different ES backtesting methods which will be presented have their strengths and weaknesses.

Whole or tail distribution - For this group, they require the whole or tail distribution of the returns or equivalently the cumulative violation process $\int_0^\tau \mathbb{1}\{x_t >$

$\hat{q}_{s,t}\}ds$. Wong (2008) introduces the saddlepoint approximation test, which allows for detecting the deficiency of a model based on just one or two VaR violations. It is typically not model-free as it relies on a normal distribution and incurs a higher level of analytical costs for the researchers. Acerbi and Szekely (2014) introduce three model-free nonparametric tests (Z_1, Z_2, Z_3) that exploit the law of large numbers for the VaR violation process. As they are model-free they depend neither on the form nor on the parameters of the parent distribution. However, they have to satisfy assumptions of continuity of the distribution function, the probability density function of financial returns and also the independence. The test Z_1 is testing ES after VaR which is based on the idea that if VaR has been tested already we can separately test the magnitude of the realized VaR violations against the ES estimates. While Z_1 is insensitive to an excessive number of VaR violations as it is an average taken over these violations, the test Z_2 jointly evaluates frequency and magnitude of them. Acerbi and Szekely (2014) show that Z_2 is most powerful test among three tests and believe that Z_1 with the Basel Committee's Traffic Light coverage test in Section 3.4.3 or Z_2 alone represent valid ES backtesting method for Basel regulation. As far as we are aware, the test Z_3 is not used in the existing literature as it is less natural than other two. Costanzino and Curran (2015) develop a ES backtesting method, which exploits the fact that ES is an average of a continuum of VaR levels, and this method tests if the whole tail of the distribution beyond the VaR has been estimated correctly. It is similar to the unconditional coverage test of VaR by Kupiec (1995) in Section 3.4.1. Du and Escanciano (2015) propose ES backtesting method based on the cumulative violation process, which is the natural analogue of the conditional coverage test of VaR by Christoffersen (1998) in Section 3.4.2, extending the idea of Costanzino and Curran (2015). Löser et al. (2018) introduce a closely related test to Costanzino and Curran (2015) and Du and Escanciano (2015) in which the cumulative violation process is treated as a series of Bernoulli distributions and uniformly distributed r.v.s.

Multiple quantile levels - For this group, the estimates of ES are backtested indirectly by simultaneously backtesting a number of VaR estimates at different quantile levels instead of working on ES directly. This is natural due to the definition of ES (see Section 1.2.4). Emmer et al. (2015) propose a ES backtesting method based on a

simple linear quantile approximation. The ES estimate is obtained as the average of multiple VaRs at different quantile levels, which is given as follows:

$$\begin{aligned} e_\tau &= \frac{1}{\tau}\int_0^\tau q_s ds \\ &\approx \frac{1}{4}[q_\tau + q_{0.75\tau+0.25} + q_{0.5\tau+0.5} + q_{0.25\tau+0.75}]. \end{aligned} \tag{4.1}$$

where $\text{ES}_\tau = e_\tau$ and $\text{VaR}_\tau = q_\tau$. The Kupiec test (Kupiec 1995) is applied to all cases ($q_\tau, q_{0.75\tau+0.25}, q_{0.5\tau+0.5}$ and $q_{0.25\tau+0.75}$) and the ES estimate can be considered reliable if all VaR estimates successfully pass the test. Kratz et al. (2018) generalize the idea of (4.1) by considering quantile levels $\tau_1, \tau_2, ..., \tau_N$ of

$$\tau_j = \tau + \frac{j-1}{N}(1-\tau), \quad j = 1, ..., N$$

where τ may take the value of 0.975 corresponding to the Basel rules for banks. Their test called multinomial backtest could easily be performed as a regular routine as it is carried out in the same way as the Kupiec test on the financial data. Note that Basel Committee on Banking Supervision (2019) suggests to use a variant of ES backtesting method by Emmer et al. (2015) expressed by the Equation (4.1) based on two quantile levels at 0.975 and 0.99.

Triplet of ES, VaR and volatility - The ES backtesting methods that require a triplet of ES, VaR and volatility are quiet often used in the literature. One of the first and most frequently used ES backtesting method is the exceedance residual (ER) test by McNeil and Frey (2000).This test is based on the size of the discrepancy between the returns x_t and ES estimates $e_{\tau,t}$ in the event of VaR violation. In other words, they are interested in the ES-specified exceedance residuals exceeding the VaR estimates, defined as

$$er_t = (x_t - \hat{e}_{\tau,t})\mathbb{1}\{x_t > \hat{q}_{\tau,t}\}.$$

This er_t is a martingale difference sequence given that estimates $\hat{e}_{\tau,t}$ and $\hat{q}_{\tau,t}$ are true conditional on the information about the return process available up to time $t-1$, $\mathscr{F}_{t-1}$. When exceedance residuals are standardized by a given volatility estimates $er_t/\hat{\sigma}_t$, we call it standardized ER. This ES backtesting method relies on the definition of the martingale difference sequence and tests whether the expected value of the

standardized ER is zero, i.e., $E[er_t/\hat{\sigma}_t] = 0$. See Section 4.4.2 for the details of this test. Righi and Ceretta (2013) extend the ER test of McNeil and Frey (2000) using the shortfall deviation that is the dispersion only for the VaR violations, i.e., the square root of the truncated variance for VaR_τ, instead of the volatility estimates for the full sample. Nolde and Ziegel (2017) propose the ES backtesting based on the concept of conditional calibration (CC), which tests whether the expected value of the identification functions of the respective functional, i.e. ES, is zero. When CC test requires volatility estimates, it is said to be general CC test. See Section 4.4.3 for the details of this test.

Pair of VaR and ES - Examples include the non-standardized ER test of McNeil and Frey (2000) and the CC test of Nolde and Ziegel (2017) that are called simple (or raw) ER test and simple CC test: see again Section 4.4.2 and 4.4.3 for the details, respectively.

ES only - The estimates of ES are backtested directly in contrast to the above 4 methods, which use the auxiliary quantities, i.e., not exactly ES risk measures but rather some values including ES, instead of the ES itself. Bayer and Dimitriadis (2020b) introduce the first direct ES backtesting method based on their strict definition of backtesting stating that only ES estimates are allowed as an input variable besides the realized returns for backtesting ES. They introduce the expected shortfall regression (ESR) tests, which extend the classical method of Mincer and Zarnowitz (1969) to ES-specific versions. See Section 4.4.4 for the details of this test.

Comparative - Unlike traditional ES backtesting, comparative ES backtesting is based on the concept of joint elicitability of VaR and ES (Section 4.5.1) and the Diebold-Mariano (DM) test of Diebold and Mariano (1995), proposed by Fissler et al. (2015) and Nolde and Ziegel (2017). Comparative VaR backtesting is already introduced in Section 3.5 and ES-specific version will be introduced in Section 4.5. Acerbi and Szekely (2017) introduce a backtesting method called the ridge test based on a transformed version of the scoring function of VaR, which can be used for both model comparison and validation purposes.

In our empirical analysis of four financial time series (DJ, NASDAQ, NIKKEI and JPY/GBP) for backtesting and estimating ES (Section 4.6), we use simple and standardized ER tests of McNeil and Frey (2000), general and simple CC tests of

Nolde and Ziegel (2017), and thee types (strict, auxiliary and intercept) of ESR tests of Bayer and Dimitriadis (2020b).

4.3 Expected Shortfall (ES) estimation methods

No particular type of ES model is prescribed in the framework of Basel Committee on Banking Supervision (2019) although the BCBS announced a change in the risk measure used for capital requirements in internal market risk models, moving from VaR to ES. With regard to estimation of ES, there has not been sufficient investigation to establish the superiority of a certain estimator relative to the others in the literature, discussed in Section 4.2. The drawbacks of traditional approaches given in Section 1.3 and the need to accurately forecast the extreme market risks has motivated the development of an alternative method of ES estimation. We propose a novel approach of dynamic extreme ES estimation, which is based on our proposed GARCH-UGH approach (Kaibuchi et al. 2022) from Chapter 3 and the use of asymptotic equivalence between VaR (quantile) and ES.

4.3.1 GARCH-UGH method

The setting for the GARCH-UGH method for dynamic (extreme) ES is same as the dynamic extreme VaR in Section 3.2. Recall that we assume the dynamics of negative daily log-return X_t are governed by

$$X_t = \mu_t + \sigma_t Z_t,$$

where $\mu_t \in \mathbb{R}$ and $\sigma_t > 0$ denote the (conditional) mean and standard deviation, and the innovations Z_t form a strictly stationary white noise process, that is, they are i.i.d. with zero mean, unit variance and common marginal distribution function F_Z. Here we focus on the one-step ahead expected shortfall (ES), that is, the estimation of the conditional (extreme) ES of X_{t+1} given $\mathscr{F}_t$, whose order τ tends to 1 as the available sample size n goes to infinity. Similar to the VaR, the one-step ahead conditional ES of X_{t+1} can be written as

$$e_\tau(X_{t+1} \mid \mathscr{F}_t) = \mu_{t+1} + \sigma_{t+1} e_\tau(Z), \tag{4.2}$$

where $e_\tau(Z) = \mathrm{E}[Z \mid Z > q_\tau(Z)]$ is the τth unconditional ES of the marginal distribution of the innovations Z_t (see Section 1.2.4) and $q_\tau(Z)$ is the τth unconditional quantile (see Section 1.2.3).

In calculating this estimate (4.2), there are four main difficulties. First, one has to estimate μ_{t+1} and σ_{t+1}, which supposes that an appropriate model and estimation method have to be chosen. Second, the innovations Z_t are unobserved, which means that the estimation of $e_\tau(Z)$ has to be based on residuals following the estimation of μ_{t+1} and σ_{t+1}. Third and fourth difficulty are specific to our context. Third, we wish here to estimate a dynamic extreme conditional ES, that is, $e_\tau(X_{t+1} \mid \mathscr{F}_t)$ with τ very close to 1. In such contexts, it is well-known that traditional nonparametric estimators become inconsistent and adapted extrapolation methodologies have to be employed. Lastly, the resulting estimator of the τth unconditional ES of residuals $\hat{e}_\tau(Z)$ heavily depends on the estimated τth unconditional quantile $q_\tau(Z)$ by definition of ES and the EVI $\hat{\gamma}$ due to the use of asymptotic equivalence between VaR and ES (see Equation 4.4 later). In this sense, $\hat{e}_\tau(Z)$ may inherit the vexing defects of both $q_\tau(Z)$ and $\hat{\gamma}$.

In this framework for the ES estimation, we first follow the GARCH-UGH approach of Kaibuchi et al. (2022) where we estimate the mean and standard deviation in a GARCH-type model (Section 3.2.2) and use bias-corrected Weissman quantile estimator (3.10) to the filtered residuals for the estimation of unconditional $q_\tau(Z)$ ($\tau \uparrow 1$) (Section 3.2.3). We then rely on the asymptotic equivalence between quantile (VaR) and ES to estimate unconditional $e_\tau(Z)$ of residuals. The asymptotic equivalence between expectile and expectile-based ES, which is $epe_\tau(Z) = \mathrm{E}[Z \mid Z > ep_\tau(Z)]$ where ep_τ is the τth expectile (see Section 1.2.5), is given in Proposition 4 of Daouia et al. (2020). The asymptotic equivalence between quantile and expectile has also been found in Bellini and Di Bernardino (2015) and Daouia et al. (2018) as

$$\frac{ep_\tau}{q_\tau} \sim (\gamma^{-1} - 1)^{-\gamma}, \quad \tau \to 1, \tag{4.3}$$

where γ is the EVI given in Section 2.1.3. Therefore, the asymptotic equivalence of $\frac{epe_\tau}{ep_\tau}$ is nothing but one between quantile and ES expressed as $\frac{e_\tau}{q_\tau}$ as the factor (4.3) cancels out. Following Daouia et al. (2020), we now assume that $\mathrm{E}[Z_-] < \infty$ where

$Z_- = \min(Z, 0)$ denotes the negative part of Z. We further assume that first-order condition (3.4) holds with $0 < \gamma < 1$, implying that Z has a Pareto-type distribution. The assumption of $\mathrm{E}[Z_-] < \infty$ and $0 < \gamma < 1$ ensures that the first moment of Z exists, and thus ES of Z are well-defined. Then, as $\tau \to 1$,

$$\frac{e_\tau}{q_\tau} \sim (1-\gamma)^{-1}. \tag{4.4}$$

This relationship is also given in Novales and Garcia-Jorcano (2019). By further considering the second-order condition (3.7), Daouia et al. (2020) establish a precise control of the remainder term in the asymptotic equivalence of $\frac{epe_\tau}{ep_\tau}$, which is naturally extended to (4.4):

$$\begin{aligned}\frac{e_\tau}{q_\tau} = \frac{1}{1-\gamma}\Bigg(1 - \frac{\gamma^2(\gamma^{-1}-1)^\gamma}{q_\tau}(\mathrm{E}[Z] + o(1)) \\ + \frac{1-\gamma}{(1-\gamma-\rho)^2}(\gamma^{-1}-1)^{-\rho}A((1-\tau)^{-1})(1+o(1))\Bigg)\end{aligned} \tag{4.5}$$

where $\rho \leq 0$ is the second-order parameter and A is a positive or negative function converging to 0 at infinity, such that $|A|$ is regularly varying with index ρ (see Section 3.2.3).

Given estimates $\hat{\mu}_{t+1}$, $\hat{\sigma}_{t+1}$ and $\hat{q}_\tau(Z)$ of these quantities with the asymptotic equivalence $\frac{e_\tau}{q_\tau}$ (4.4), an estimate of $e_\tau(X_{t+1} \mid \mathscr{F}_t)$ is then

$$\hat{e}_\tau(X_{t+1} \mid \mathscr{F}_t) = \hat{\mu}_{t+1} + \hat{\sigma}_{t+1}\hat{q}_\tau(Z)(1-\hat{\gamma}_{k,k_\rho})^{-1}, \tag{4.6}$$

where $\hat{\gamma}_{k,k_\rho}$ is the bias-corrected Hill estimator of de Haan et al. (2016).

4.3.2 Other methods

In this section, we will explain briefly other five methods for ES estimation, which are used to compare with our proposed GARCH-UGH approach in the empirical analysis in Section 4.6. Their VaR estimation methods are already described in Section 3.3.

For Historical Simulation (HS) method, the τth unconditional ES is obtained by taking the mean of all returns X_t equal to or above the estimated VaR, which is

simply the empirical quantile of X_t at the desired quantile level.

For parametric GARCH-N and GARCH-t methods, these method use the same filtering step as explained in Section 3.2.3 and assume that the standardized residuals are normally and Student-t distributed, respectively. Since we only consider continuous distributions, the τth unconditional ES is estimated via the numerical integration following the definition of ES given as:

$$e_\tau(Z) = \mathrm{E}[Z \mid Z > q_\tau(Z)] = \frac{1}{1-\tau}\int_\tau^1 q_s(Z)ds.$$

where the q_τ is the τth quantile $\hat{F}_\tau^{-1}(Z)$ estimated in Section 3.3.

For UGH method, we use the UGH step (Section 3.2.3) directly to the series X_t without filtering and apply the asymptotic equivalence between VaR and ES (4.4) to estimate the τth unconditional ES.

For GARCH-EVT method, we consider the distribution function of excesses $Z - q_\tau(Z)$ over a fixed higher threshold $q_\tau(Z)$ under an approximately i.i.d. condition after filtering step. Note that we considered the excesses $Z - u$ over a high threshold u for the VaR estimation in Section 2.2.3 and 3.3. It is given in McNeil and Frey (2000) that excesses $Z - q_\tau(Z)$ also have a GPD distribution with the same EVI γ but a different scaling parameter β. For a quantile level τ, the estimator of unconditional ES of Z is given by

$$\hat{e}_\tau(Z) = \mathrm{E}[Z \mid Z > \hat{q}_\tau(Z)] = \hat{q}_\tau(Z)\left(\frac{1}{1-\hat{\beta}} + \frac{\hat{\beta} - \hat{\gamma}\hat{Z}_{n-k,n}}{(1-\hat{\gamma})\hat{q}_\tau(Z)}\right).$$

Then, the one-step ahead conditional ES based on GARCH-EVT is obtained by substituting $\hat{e}_\tau(Z)$ above into the Equation (4.2).

4.4 Traditional ES backtesting

We have discussed in Section 1.4 that the traditional backtesting is viewed as a model verification test. These tests are concerned with assessing some optimality property of a set of risk measure estimates and not suited to compare different estimation

methods for risk measures. They perform a statistical tests for the null hypothesis:

$$H_0 : \text{The risk measure estimates are correct.}$$

If H_0 is not rejected, then the risk measure estimates are deemed to be adequate. Generally, the traditional backtestings with the hypothesis H_0 are not relevant to elicitability of the risk measure and are not aimed at model comparison and ranking. Here our focus is to backtest a risk measure ES directly and indirectly by means of the exceedance residual (ER) test of McNeil and Frey (2000), the conditional calibration (CC) test of Nolde and Ziegel (2017) and the expected shortfall regression (ESR) test of Bayer and Dimitriadis (2020b), which are introduced in Section 4.2 briefly.

4.4.1 Problems of backtesting ES

We would like to briefly highlight two problems of backtesting ES before we look at the traditional ES backtesting methods in detail. As mentioned previously, one of the major drawback of ES is its difficulty to be backtested.

Firstly, the theoretical ES cannot be compared with the observed returns unlike the VaR (see Section 3.4) as one would not only look at the number of VaR violations $I_t = \mathbb{1}\{x_t > \hat{q}_\tau(X_t \mid \mathscr{F}_{t-1})\}$ but also their sizes. Secondly, most of the ES backtesting methods exist in the literature including the ones we use in the empirical analysis test the auxiliary quantities of ES rather than only ES itself unlike VaR (see Section 4.2). These tests require further inputs such as the VaR, the volatility and whole or tail distributions. Strictly speaking, these tests are not backtesting ES if we follow the strict definition of backtesting from Bayer and Dimitriadis (2020b). A rejection of the null hypothesis of these tests does not necessarily imply that the ES estimates are incorrect as the estimates of other inputs might be incorrect.

4.4.2 Exceedance residual test

Most frequently used traditional ES backtesting is the exceedance residual (ER) test proposed by McNeil and Frey (2000). This test is based on the size of the discrepancy between returns x_t and ES estimates $e_{\tau,t}$ in the event of VaR violations. In other

words, we are interested in the ES-specified exceedance residuals over the VaR estimates, defined as

$$er_t = (x_t - \hat{e}_{\tau,t})\mathbb{1}\{x_t > \hat{q}_{\tau,t}\}. \tag{4.7}$$

This quantity er_t is a martingale difference sequence given that estimates $\hat{e}_{\tau,t}$ and $\hat{q}_{\tau,t}$ are true conditional on the information about the return process available up to time $t-1$, $\mathscr{F}_{t-1}$.

In McNeil and Frey (2000), er_t, which is called the simple ER, is standardized by a given volatility estimates $\hat{\sigma}_t$, i.e., the volatility obtained in the GARCH step, to form the standardized ER $er_t/\hat{\sigma}_t$. Thus, we have two tests that are either based on simple (or raw) or standardized ER to carry out. Moreover, Righi and Ceretta (2013) extend the standardized ER test using the shortfall deviation (SD) that is the dispersion only for the VaR violations, i.e., the square root of the truncated variance for some quantile VaR_τ, instead of the volatility estimates for the full sample. Mathematically, it is defined as

$$\text{SD}_\tau = [\text{Var}(X|X > q_\tau(X))]^{1/2} = \left(\frac{1}{1-\tau}\int_\tau^1 (q_s(X) - e_s(X))^2 ds\right)^{1/2}.$$

Note that we do not use the SD-standardized version of the ER test in the empirical analysis given in Section 4.6 as it did not show the significance difference from the conventional standardized ER test in the pre-analysis.

Both simple and standardized ER test rely on the definition of the martingale difference sequence and test whether the expected value of the ER is zero, i.e., $\mu = E[er_t] = 0$ or $\mu = E[er_t/\hat{\sigma}_t] = 0$, as these residuals should behave like an i.i.d. sample with mean zero. McNeil and Frey (2000) use the estimate

$$\hat{\mu} = \frac{\sum_{t=1}^N er_t}{\sum_{t=1}^N \mathbb{1}\{x_t > \hat{q}_{\tau,t}\}} \tag{4.8}$$

and test the hypothesis of mean zero using a bootstrap test that makes no assumption about the underlying distribution of the residuals. In our analysis, we test μ against the one-sided alternative of $\mu < 0$, which indicates that ES is systematically underestimated and leads to insufficient provision of capital, and also against the two-sided alternative that $\mu \neq 0$ where the estimated ES is either systematically

underestimated or overestimated. In summary, two hypotheses are as follows:

$$H_0^{1s} : \mu \geq 0 \quad \text{against} \quad H_1^{1s} : \mu < 0, \quad \text{and}$$
$$H_0^{2s} : \mu = 0 \quad \text{against} \quad H_1^{2s} : \mu \neq 0.$$

It is clear from the Equations (4.7) and (4.8) that

$$\hat{\mu} = \frac{\sum_{t=1}^{N} x_t \mathbb{1}\{x_t > \hat{q}_{\tau,t}\}}{\sum_{t=1}^{N} \mathbb{1}\{x_t > \hat{q}_{\tau,t}\}} - \frac{\sum_{t=1}^{N} \hat{e}_{\tau,t} \mathbb{1}\{x_t > \hat{q}_{\tau,t}\}}{\sum_{t=1}^{N} \mathbb{1}\{x_t > \hat{q}_{\tau,t}\}}.$$

This means that the simple and standardized ER tests compare the empirical mean of realized returns x_t truncated at estimated VaR $\hat{q}_{\tau,t}$ to the estimated mean ES $\hat{e}_{\tau,t}$ when there are VaR violations. These tests thus rejects the null hypothesis whenever the discrepancy between the VaR and the ES is incorrect while simultaneous misspecifications of both estimates of VaR and ES cannot be detected. Note that the ES backtesting methods of Acerbi and Szekely (2014) are proposed in the same spirit as the ER test.

4.4.3 Conditional calibration test

Nolde and Ziegel (2017) propose the traditional ES backtesting method based on the concept of conditional calibration (CC), which tests whether the expected value of the identification functions of the respective functional, in our case ES, is zero.

Let $\Theta = (\eta_1, ..., \eta_d)$ be a vector of $d \geq 1$ risk measures η that are mappings from some collection of probability distributions $\mathcal{P}$ to real numbers. Definition 2 of Nolde and Ziegel (2017) states that the vector of risk measures Θ is identifiable with respect to $\mathcal{P}$ if there exists a identification function $I : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^d$ such that

$$\mathrm{E}[I(r, X)] = 0 \Leftrightarrow r = \Theta(X) \quad \forall X, \tag{4.9}$$

The identifiability is the close concept to the elicitability (see Section 1.2.5 and 4.5.1). In fact, when $d = 1$ identifiability implies elicitability under some additional assumptions (Acerbi and Szekely 2017). In most common cases, identifiability and

elicitability occur jointly when the Equation (4.9) represents the first-order stationarity condition of a scoring function

$$S(r,x) = \int^r I(y,x)dy$$

whose expected values of the score $\mathrm{E}[S(r,x)]$ has a global minimum in the prediction of a risk measure $r = \Theta(X)$. Since the risk measure can be obtained by minimizing the expected value of the score, a scoring function $S(r,x)$ is strictly consistent for r. The vector of risk measures Θ are thus elicitable because they admit a strictly consistent scoring function $S(r,x)$.

Although the identification functions are not uniquely defined, we use the identification function for the pair $(\mathrm{VaR}_\tau, \mathrm{ES}_\tau)$ at level $\tau \in (0,1)$ for the CC test given in Nolde and Ziegel (2017) as

$$I(q,e,X) = \begin{pmatrix} 1-\tau-\mathbb{1}\{X>q\} \\ q-e-\dfrac{1}{1-\tau}\mathbb{1}\{X>q\}(q-X), \end{pmatrix}$$

whose expected value is zero if and only if q and e equal the true VaR and ES of the X, respectively. We say that the sequence of estimations $\{q_t, e_t\}_{t=1,\dots,N}$ is conditionally calibrated for Θ if

$$\mathrm{E}[I(q_t, e_t, X_t) \mid \mathscr{F}_{t-1}] = 0$$

almost surely, for all $t \in \mathbb{N}$. Therefore, the conditional calibration (CC) test for a pair of estimates of the τth VaR $\hat{q}_t$ and the τth ES $\hat{e}_t$ is based on the one-sided and two-sided alternative hypotheses as follows:

$$H_0^{1s} : \mathrm{E}[I(\hat{q}_t,\hat{e}_t,X_t) \mid \mathscr{F}_{t-1}] \geq 0 \quad \text{against} \quad H_1^{1s} : \mathrm{E}[I(\hat{q}_t,\hat{e}_t,X_t) \mid \mathscr{F}_{t-1}] < 0, \quad \text{and}$$
$$H_0^{2s} : \mathrm{E}[I(\hat{q}_t,\hat{e}_t,X_t) \mid \mathscr{F}_{t-1}] = 0 \quad \text{against} \quad H_1^{2s} : \mathrm{E}[I(\hat{q}_t,\hat{e}_t,X_t) \mid \mathscr{F}_{t-1}] \neq 0,$$

almost surely, for all $t = 1, \dots, N$. We are particularly concerned with testing against the one-sided alternative of $\mathrm{E}[I(\hat{q}_t,\hat{e}_t,X_t) \mid \mathscr{F}_{t-1}] < 0$ that the issued ES estimates are systematically underestimated. From the practitioners' point of view, overestimation of ES_τ is not a problem as holding more capital than minimally required should always be allowed.

For this test, the requirement $\mathrm{E}[I(\hat{q}_t, \hat{e}_t, X_t) \mid \mathscr{F}_{t-1}] = 0$, almost surely, is equivalent to $\mathrm{E}[h_t^T I(\hat{q}_t, \hat{e}_t, X_t)] = 0$ for all $\mathscr{F}_{t-1}$ measurable $\mathbb{R}^2$ $(d = 2)$-valued functions h_t. Following Nolde and Ziegel (2017), we use a $\mathscr{F}_{t-1}$-measurable sequence $\{\mathbf{h}_t\}_{t=1,\dots,N}$ of $m \times 2$ $(d = 2)$-matrices $\mathbf{h}_t$, which are called test functions, to use the Wald-type test statistic:

$$T_{CC} = N\left(\frac{1}{N}\sum_{t=1}^{N} \mathbf{h}_t I(\hat{q}_t, \hat{e}_t, X_t)\right)^T \hat{\Omega}_N^{-1}\left(\frac{1}{N}\sum_{t=1}^{N} \mathbf{h}_t I(\hat{q}_t, \hat{e}_t, X_t)\right),$$

where

$$\hat{\Omega}_N = \frac{1}{N}\sum_{t=1}^{N} (\mathbf{h}_t I(\hat{q}_t, \hat{e}_t, X_t))(\mathbf{h}_t I(\hat{q}_t, \hat{e}_t, X_t))^T$$

is a consistent estimator of the covariance of the m-dimensional vector $\mathbf{h}_t I(\hat{q}_t, \hat{e}_t, X_t)$. Under the null hypothesis, the test statistic T_{CC} is asymptotically χ^2 distributed with m degrees of freedom.

We have two versions of the CC test similarly to the ER test of McNeil and Frey (2000) that are: the simple CC test, which uses a pair of (VaR, ES) only, and the general CC test, which requires the additional volatility estimates. For simple CC test, Nolde and Ziegel (2017) propose to use the $(m \times m)$ identity matrix for the test function $\mathbf{h}_t$ in both one-sided and two-sided hypotheses. For the general CC test, they and also Bayer and Dimitriadis (2020b) propose to use

$$\mathbf{h}_t = \begin{pmatrix} 1 & |\hat{q}_t| & 0 & 0 \\ 0 & 0 & 1 & \hat{\sigma}_t^{-1}, \end{pmatrix}$$

for the one-sided hypothesis and

$$\mathbf{h}_t = \hat{\sigma}_t \left(\frac{\hat{e}_t - \hat{q}_t}{\tau}, 1\right)$$

for the two-sided hypothesis, where $\hat{\sigma}_t$ is the estimate of the volatility, i.e., the volatility obtained in the GARCH step. Note that the CC test for a pair of (VaR, ES) is related to the ER test of McNeil and Frey (2000) in the case of backtesting ES. Moreover, the CC test for the VaR is closely related to the VaR backtesting method based on the number of VaR violations when we choose $\mathbf{h}_t = 1$ and an appropriate identification function of VaR.

4.4.4 Expected Shortfall regression test

Bayer and Dimitriadis (2020b) introduce the first direct ES backtesting method, which solely tests the ES in the sense that it only requires ES estimates as input variables rather than the auxiliary quantities based on for example, the VaR, the volatility and whole or tail distributions (see Section 4.2). This is motivated by their strict definition of backtesting in the following (already given briefly in Section 1.4). A backtest for the sequence of estimates $\{\hat{\eta}_t\}_{t=1,\dots,N}$ for the d-dimensional risk measure η relative to the sequence of realized returns $\{X_t\}_{t=1,\dots,N}$ is a function

$$f : \mathbb{R}^N \times \mathbb{R}^{N\times d} \to \{0,1\},$$

mapping the series of estimates of risk measures and returns onto the perspective test decision. It basically states that a backtesting for specific risk measure is only allowed to require estimates of this risk measure as input variables besides the realized returns. Strictly speaking, tests based on the auxiliary quantities are not backtesting ES. A rejection of the null hypothesis of these tests does not necessarily imply that the ES estimates are incorrect as the estimates of other inputs, e.g. the VaR and the volatility, might be incorrect.

They propose three types of the expected shortfall regression (ESR) tests for backtesting ES, extending the classical regression-based testing idea of Mincer and Zarnowitz (1969) to ES-specific versions. They regress the realized returns $\{X_t\}_{t=1,\dots,N}$ on the ES estimates $\{\hat{e}_t\}_{t=1,\dots,N}$ and intercept term by using a regression equation for the functional ES given as

$$X_t = \omega_1 + \omega_2 \hat{e}_t + \mathbf{u}_t^e, \tag{4.10}$$

where $\mathbf{u}_t^e$ is the error term and $\mathrm{ES}_\tau(\mathbf{u}_t^e \mid \mathscr{F}_{t-1}) = 0$ almost surely. Since the estimates $\hat{e}_t$ are calculated given $\mathscr{F}_{t-1}$, they estimate a regression model that models the τth conditional ES as linear function

$$\mathrm{ES}_\tau(X_t \mid \mathscr{F}_{t-1}) = \omega_1 + \omega_2 \hat{e}_t,$$

where X_t is the response variable and $\hat{e}_t$ is the explanatory variable including an intercept term. The hypothesis of this test is

$$H_0 : (\omega_1, \omega_2) = (0, 1) \quad \text{against} \quad H_1 : (\omega_1, \omega_2) \neq (0, 1).$$

For correctly specified ES estimates, the intercept ω_1 and slope ω_2 parameters equal zero and one, respectively, which is tested by using the Wald-type test statistics. However, we cannot estimate such a regression model, i.e., estimating the parameters ω, by maximum or generalized method of moment estimations using only the ES as a input variable because the ES is not elicitable (Gneiting 2011) as explained in Section 1.2.2 and 4.5.1, and there are no consistent score (or loss) and identification functions (Section 4.4.3), which could be used as objective functions for estimation methods. Bayer and Dimitriadis (2020b) use the ES regression equation (4.10) and also a feasible alternative by specifying an auxiliary VaR (quantile) regression equation given as

$$X_t = \beta_1 + \beta_2 \hat{q}_t + \mathbf{u}_t^q, \tag{4.11}$$

where $\mathbf{u}_t^q$ is the error term and $\text{VaR}_\tau(\mathbf{u}_t^q \mid \mathscr{F}_{t-1}) = 0$ almost surely. They jointly estimate the regression parameters (β, ω) by employing a scoring function of a pair (VaR, ES), which is based on joint elicitability (see Section 4.5.1), from Fissler and Ziegel (2016) for objective functions of maximum or generalized method of moment estimations.

The specification of the VaR regression equation in the joint regression model using Equations (4.10) and (4.11) allows for different ESR tests. First test is the auxiliary ESR test where the VaR estimates $\hat{q}_t$ are used as the explanatory variable in the Equation (4.11) but only the ES intercept and slope parameters ω are tested. The main drawback of the auxiliary ESR test is that it indeed requires both estimates of VaR and ES as input variables. Hence, it is formally a test of a pair (VaR, ES) like the ER test of McNeil and Frey (2000) and the CC test of Nolde and Ziegel (2017).

Second test is the strict ESR test where only ES estimates $\hat{e}_t$ are used as the explanatory variable in both Equations (4.10) and (4.11) and only the ES intercept and slope parameters ω are tested. The problem of this test is the potential model misspecification in the VaR regression equation as we use ES estimates $\hat{e}_t$ are used as the

explanatory variable.

Third test is the intercept ESR test where the slope parameter ω_2 is set to be 1 in the ES regression equation (4.10) and only the ES intercept parameter ω_1 is tested. The restricted regression equation after regressing the ES-specified exceedance errors $X_t - \hat{e}_t$ only on an intercept term is given as

$$X_t - \hat{e}_t = \beta_1 + \beta_2 \hat{e}_t + \mathbf{u}_t^q \quad \text{and} \quad X_t - \hat{e}_t = \omega_1 + \mathbf{u}_t^e.$$

Above equation allows us to define the one-sided and two-sided alternative hypotheses as follows:

$$H_0^{1s} : \omega_1 \geq 0 \quad \text{against} \quad H_1^{1s} : \omega_1 < 0, \quad \text{and}$$
$$H_0^{2s} : \omega_1 = 0 \quad \text{against} \quad H_1^{2s} : \omega_1 \neq 0.$$

The merit of using this test is that it solely requires the ES estimates as input variable like the strict ESR test. It also clarifies whether ES estimates are underestimated or overestimated unlike other two tests, which are generally unclear how the intercept ω_1 and slope ω_2 parameters influence the ES estimates. As mentioned previously, we are concerned with checking the underestimation of ES because the practitioners only have to prevent and penalize the underestimation of the financial risks.

Details of the Wald-type test statistics of three tests are provided in Bayer and Dimitriadis (2020b). Under the null hypothesis, the test statistic of the auxiliary and strict ESR tests is asymptotically χ^2 distributed with 2 degrees of freedom, while the intercept ESR is asymptotically χ^2 distributed with 1 degrees of freedom.

4.5 Comparative ES backtesting

Evaluating a sequence of risk measure estimates using a certain method is different from comparing estimation methods as we already discussed in Section 1.4. Recall that the comparative backtesting is better suited for model comparison on the basis of forecasting accuracy while traditional backtesting explained in Section 4.4 is viewed as a model verification. In practice (already illustrated in Section 3.6 and will be in Section 4.6), there are cases when traditional ES backtesting methods do

not yield definitive answers because the estimation methods are all accepted or all rejected. Moreover, traditional ES backtesting methods sometimes give a different conclusion in the decision of the estimation methods, i.e., one test rejects the null hypothesis of the underestimation of ES estimates while the other accept, which was not observed in the case of VaR backtesting. Using several methods of traditional ES backtesting is not sufficient and thus the use of comparative backtesting is desperately required especially for the ES. The comparative backtesting method for ES is based on the Diebold-Mariano test of Diebold and Mariano (1995), which is similar to the VaR version, but follows the idea of joint elicitability of VaR and ES (Fissler and Ziegel 2016), and uses the scoring function of a pair (VaR, ES) instead of the scoring function of VaR given in Section 3.5.

4.5.1 Elicitability and joint elicitability

In Section 1.2.2 and 4.1 we have mentioned elicitability briefly. Here we will explain the notion of the elicitability and the mathematical definitions of strictly consistent scoring function, elicitability and joint elicitability, which are the idea used in Section 3.5 as well. While we assessed some optimality property of a set of ES estimates in Section 4.4, we now want to compare the different ES estimates $\hat{e}_{\tau,t}$ provided by competing estimation procedures.

In this type of situation, we consider a scoring function S depending on both the estimate (i.e. forecast) and the realization. For example, a scoring function could be

$$
\begin{aligned}
S(r,x) &= (r-x)^2, \quad \text{(squared error)} \\
S(r,x) &= |r-x|, \quad \text{(absolute error)} \\
S(r,x) &= |(r-x)/x|, \quad \text{(absolute percentage error)} \\
S(r,x) &= |(r-x)/r)|, \quad \text{(relative error)}
\end{aligned}
$$

where r is the point estimate and x is the verifying observation. Each of the above scoring function S measures the distance between r and x. We generally take such scoring functions to be negatively oriented, that is, the smaller the better. Thus, the value r minimizing $S(r,x)$ is considered as the more accurate with respect to the observed value x. We say that a certain risk measure η is elicitable if we can find

at least one scoring function S such that $\eta(X)$ with X a random variable minimizes the expected value of the scoring function $\overline{S}(r) = \mathrm{E}[S(r, X)]$. In other words, a risk measure η is elicitable if there exists a scoring function S such that $r = \eta(X)$ leads to a minimum score $\overline{S}(r)$.

The formal definition of scoring function so that a risk measure is elicitable is now given. Let $\Theta = (\eta_1, ..., \eta_d)$ be a vector of $d \geq 1$ risk measures η that are mappings from some collection of probability distributions $\mathcal{P}$ to real numbers $\mathbb{R}$. The risk of X is given as $\Theta(X)$. Then, a scoring function $S : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$ is strictly consistent for Θ with respect to $\mathcal{P}$ if

$$\mathrm{E}[S(\Theta(X), X)] < \mathrm{E}[S(r, X)] \tag{4.12}$$

for all $r = (r_1, ..., r_d) \neq \Theta(X) = (\eta_1(X), ..., \eta_d(X))$ (Nolde and Ziegel 2017). The scoring function S is consistent if equality is included in the Equation (4.12). The vector of risk measures Θ is elicitable if there exists a scoring function that is strictly consistent for it, i.e., we have

$$\Theta(X) = \arg\min_{r \in \mathbb{R}} \mathrm{E}[S(r, X)]. \tag{4.13}$$

We look at a simple example before we present the cases for ES. We consider the mean functional, i.e., $\eta(X) = \mathrm{E}[X]$, which is elicited by the square error $S(r, x) = (r - x)^2$ (Gneiting 2011). As mentioned previously, this means that the value $\eta(X)$ minimizes the expected score

$$\overline{S}(r) = \mathrm{E}[S(r, X)] = \mathrm{E}[(r - X)^2], \quad \text{and hence}$$

$$\overline{S}(\eta(X)) = \mathrm{E}[(\mathrm{E}[X] - X)^2] < \overline{S}(r)$$

for any value of $r \neq \eta(X)$. We can see that the mean minimizes the mean squared error. However, it is not true for any scoring function S because for example, $\eta(X) = \mathrm{E}[X]$ does not minimize the expected score of the absolute percentage error $S(r, x) = |(r - x)/x|$.

Thus, the elicitability of a risk measure η provides the helpful framework for the determination of optimal point estimates by ranking the estimated values of $\eta(X)$ based on the expected score $\overline{S}$. In order to backtest a specific risk measure such as VaR and ES via the comparative backtesting method, we first need specific scoring functions that are strictly consistent for a certain risk measure, i.e., the estimates of that risk measure minimizes the expected score. In the financial literature, the existence of a scoring function for a certain risk measure gives a natural way to compare the accuracy of two different estimation models, i.e., to test the comparative hypothesis, which states one estimation model is better than another, by means of the Diebold-Mariano test (Diebold and Mariano 1995) on the difference of two realized scores.

While the VaR (quantile) is elicitable as it is the minimizers of the expected value of an appropriate piecewise linear score (see for example Bellini and Di Bernardino 2015 and Nolde and Ziegel 2017) given in Section 3.5, Gneiting (2011) points out that ES is not elicitable. This means that there exists no scoring function $S_{\text{ES}}(e, x)$ such that the ES estimate e of the true ES x can be obtained as the e that minimizes $S_{\text{ES}}(e, x)$. It is quite obvious that ES is not elicitable because there is no concrete realized data to be compared to the estimates of ES. Therefore, ES alone cannot be backtested comparatively to decide whether one estimation method is better than the other given only the realized returns.

However, it turns out that ES is jointly elicitable with VaR, thanks to the pioneering work done by Fissler et al. (2015) and Fissler and Ziegel (2016). Following the Equation (4.13), the joint elicitability of VaR and ES is defined as the existence of a strictly consistent scoring function (4.12) that solves

$$(\text{VaR}_\tau(X), \text{ES}_\tau(X)) = \arg\min_{(q,e)\in\mathbb{R}^2} \text{E}[S_{\text{VaR,ES}}(q, e, X)] \tag{4.14}$$

where VaR and ES are the unique minimizers of the expected score and the possible choices of $S_{\text{VaR,ES}}$ are given in Section 4.5.2. Note that the joint elicitability is also called as the coelicitability (He et al. 2022) and conditional elicitability (Emmer et al. 2015).

4.5.2 Scoring functions for the pair (VaR, ES)

It is given in Nolde and Ziegel (2017) that the possible choices of the scoring function $S_{\text{VaR,ES}}$ in the Equation (4.14) are of the form

$$\begin{aligned} S_{\text{VaR,ES}}(q,e,x) = \mathbb{1}\{x > q\}(-G_1(q) + G_1(x) - G_2(e)(q-x)) \\ + (1-\tau)(G_1(q) - G_2(e)(e-q) + \mathcal{G}_2(e)), \end{aligned} \tag{4.15}$$

where G_1 is increasing, $\mathcal{G}_2$ is strictly increasing and concave and $\mathcal{G}_2' = G_2$; see Corollary 5.5 of Fissler and Ziegel (2016) for a more detailed discussion. All scoring functions of the form (4.15) are strictly consistent for $(\text{VaR}_\tau, \text{ES}_\tau)$ with $\tau \in (0,1)$. The authors also mention that the above scoring functions remain strictly consistent as long as we set $G_1 = 0$ and use $\mathcal{G}_2$ that is strictly increasing and concave.

For the comparative ES backtesting using the DM test in our empirical analysis in Section 4.6, we use two particular scoring functions $S_{\text{VaR,ES}}$ introduced in Nolde and Ziegel (2017). First one is to choose $G_1(x) = 0$ and $\mathcal{G}_2 = x^{1/2}, x > 0$ in the general form (4.15), which leads to the $(1/2)$-homogeneous choice

$$S_{\text{VaR,ES}}(q,e,x) = \mathbb{1}\{x > q\}\frac{x-q}{2\sqrt{e}} + (1-\tau)\frac{q+e}{2\sqrt{e}}. \tag{4.16}$$

Second one is to choose $G_1(x) = 0$ and $\mathcal{G}_2 = \log x, x > 0$ in the general form (4.15), which leads to the 0-homogeneous choice given as

$$S_{\text{VaR,ES}}(q,e,x) = \mathbb{1}\{x > q\}\frac{x-q}{e} + (1-\tau)\left(\frac{q}{e} - 1 + \log(e)\right). \tag{4.17}$$

Concerning the power of Diebold-Mariano test we used in the comparative backtesting, this 0-homogeneous choice has been shown to yield more powerful test than the h-homogeneous choice with $h > 0$ in the volatility forecasting literature (see Nolde and Ziegel 2017 and Deng and Qiu 2021).

We will also present several other choices of strictly consistent scoring functions for a pair of VaR and ES available in the literature, which suppose that the financial returns x are the profits. In this case, a general form of the scoring function $S_{\text{VaR,ES}}$

(Fissler et al. 2015; Fissler and Ziegel 2016) is given by

$$\begin{aligned} S_{\text{VaR,ES}}(q,e,x) = &(\mathbb{1}\{x \le q\} - \tau)(G_1(q) - G_1(x)) \\ &+ \frac{1}{\tau} G_2(e)\mathbb{1}\{x \le q\}(q-x) + G_2(e)(e-q) - \mathcal{G}_2(e), \end{aligned} \tag{4.18}$$

where G_1 is increasing, $\mathcal{G}_2$ is strictly increasing and convex and $\mathcal{G}_2' = G_2$. The fact that ES is not elicitable can be clearly seen from the structure of $S_{\text{VaR,ES}}$ (4.18) as the first factor only depends on the VaR_τ whereas the second factor depends on both VaR_τ and ES_τ. Possible choices for functions G_1 and G_2 are $G_1(q) = q$ and $G_2(e) = \exp(e)$, $G_1(q) = q$ and $G_2(e) = \exp(e)/(1+\exp(e))$ from Fissler et al. (2015), $G_1(q) = 0$ and $G_2(e) = -1/e$ from Patton et al. (2019), and $G_1(q) = 0$ and $G_2(e) = \exp(e)$ from Deng and Qiu (2021). Moreover, Acerbi and Szekely (2014) proposed a 2-homogeneous scoring function for the pair (VaR_τ, ES_τ) under the additional assumption given as $\text{ES}_\tau(X) < W\text{VaR}_\tau(X)$ with a parameter $W > 0$. As we have already discussed in Section 3.5, strictly consistent scoring functions for the pair (VaR_τ, ES_τ) are not unique and hence we are unaware of which functions should be used in regulatory settings and in practice.

4.5.3 Diebold-Mariano test and traffic light approach

The framework of comparative backtesting for ES based on the Diebold-Mariano test of Diebold and Mariano (1995) is same as the test for the VaR given in Section 3.5 but uses the scoring functions of the pair (VaR, ES) introduced in Section 4.5.2 instead. Recall that the comparative backtesting is better suited for model comparison on the basis of forecasting accuracy while traditional backtesting is viewed as a model verification. In practice (already illustrated in Section 3.6.3), there are cases when traditional backtesting methods do not yield definitive answers because the estimation methods are all accepted or all rejected. The comparative backtestings enable to conduct direct comparisons of estimation methods when traditional backtestings are not working efficiently.

In order to compare the estimation performances of two models, say competing and benchmark models, and decide which one is better, we use the comparative version of the traffic light approach for the ES proposed by Fissler and Ziegel (2016) and

Nolde and Ziegel (2017), which is based on the Diebold-Mariano test (Diebold and Mariano 1995). In this comparative ES backtesting, we again consider the following two hypotheses:

H_0^- : The competing model predicts at least as well as the benchmark model,

H_0^+ : The competing model predicts at most as well as the benchmark model.

The null hypothesis H_0^- is an analogue of H_0 of traditional backtesting but adapted to a comparative setting. The other hypothesis H_0^+ is more conservative in the sense that a backtest is passed if we can reject H_0^+. From a regulatory perspective, when financial institutions propose a new internal model, they will need strong evidence to throw away the old one in favour of a new model.

For a sequence of ES estimates, $\hat{e}_{\tau,1}, \hat{e}_{\tau,2}, ..., \hat{e}_{\tau,N}$, and corresponding sequence of VaR estimates, $\hat{q}_{\tau,1}, \hat{q}_{\tau,2}, ..., \hat{q}_{\tau,N}$, and realized returns $x_1, x_2, ..., x_N$, the realized ES scores $S^1_{\text{VaR,ES}}(\hat{q}_{\tau,N}, \hat{e}_{\tau,N}, x_N)$ of the form (4.15) given in Section 4.5.2 are calculated for a competing model. Similarly, $S^2_{\text{VaR,ES}}(\hat{q}_{\tau,N}, \hat{e}_{\tau,N}, x_N)$ is formed for a benchmark model. The comparative ES backtesting treats $S_{\text{VaR,ES}}$ as a loss function and forms the t-statistic based on the DM test as follows:

$$DM = \frac{\sqrt{N}\bar{d}}{\hat{\sigma}_N}, \quad \bar{d} = \frac{1}{N}\sum_{t=1}^{N}(S^1_{\text{VaR,ES}}(\hat{q}_{\tau,N}, \hat{e}_{\tau,N}, x_N) - S^2_{\text{VaR,ES}}(\hat{q}_{\tau,N}, \hat{e}_{\tau,N}, x_N)), \quad (4.19)$$

where $\bar{d}$ is the sample mean of the loss differential of the pair (VaR, ES) estimates between the competing model (Model 1) and the benchmark model (Model 2), and $\hat{\sigma}_N$ is a suitable estimate of the asymptotic standard deviation of $\bar{d}$ already given in Section 3.5. Under proper mixing conditions, the test statistic is asymptotically standard normal $N(0,1)$; see Diebold and Mariano (1995) and Holzmann and Eulert (2014).

We finally recall the decisions taken in the comparative ES backtesting based on the DM test under the null hypotheses H_0^- and H_0^+. Under H_0^-, the comparative backtesting is passed for the competing model (Model 1) if the null hypothesis fails to be rejected. The competing model is then considered as better model than the benchmark (Model 2) in this specific situation and it simply means that this null hypothesis cannot be falsified. On the other hand, under H_0^+ the backtesting for the

competing model is passed if the null hypothesis is rejected. In terms of visualization of the results, the green zone corresponds to the case when H_0^- is not rejected and H_0^+ is rejected, which suggests that the competing model is considered as better than the benchmark model. The yellow zone is when only one of the backtestings under H_0^- and H_0^+ is passed and we cannot conclude which model performs the best. The red zone corresponds to the case when both backtestings fail to be passed, indicating a problem with the competing model.

4.6 Out-of-sample dynamic extreme ES estimation and backtesting

The purpose of this section is to examine the finite-sample performance of our proposed GARCH-UGH approach with other commonly used approaches because there has not been sufficient investigation to rank certain estimators of ES in the literature. We also tackle an urgent problem of which ES backtesting methods could be used in practice and provide some guidance to researchers and financial institutions with respect to quantitative risk management.

We consider again historical daily negative log-returns of three financial indices and an exchange rate, all made of $n = 4000$ observations (already introduced in Chapter 3):

- The Dow Jones Industrial Average (DJ) from 23 December 1993 to 9 November 2009;
- The Nasdaq Stock Market Index (NASDAQ) from 30 August 1993 to 16 July 2009;
- The Nikkei 225 (NIKKEI) from 14 May 1993 to 12 August 2009;
- The Japanese Yen-British Pound exchange rate (JPY/GBP) from 2 January 2000 to 14 December 2010.

The graphs of these financial time series and descriptive statistics with basic statistical tests are represented in Figure 3.1 and Table 3.1, respectively.

We compare our GARCH-UGH approach (see Section 4.3.1) with HS, GARCH-N, GARCH-t, the bias-reduced UGH and GARCH-EVT (see Section 4.3.2). A comparison with the basic estimation methods (HS, GARCH-N and GARCH-t) indicates the importance of extreme value methods in the estimation of the dynamic extreme ES. Moreover, a comparison with the UGH method (without filtering) allows us to see how effective filtering is, and a comparison with the GARCH-EVT method (not featuring bias reduction) will illustrate the benefit of bias reduction at the extreme value step after filtering.

We present out-of-sample evaluations of one-step ahead conditional ES estimates at different τ levels and choices of k by means of traditional and comparative ES backtesting methods discussed in Section 4.4 and 4.5, respectively. We compare our GARCH-UGH approach with other EVT-type methods and with basic estimation methods in Section 4.6.1 and 4.6.2, respectively. Note that the R package `esback` (Bayer and Dimitriadis, 2020a) has been used for traditional backtestings given in Section 4.4 and `sandwich` (Zeileis et al., 2022) for the estimation of the asymptotic standard deviation of the Equation 4.19 in the comparative backtesting given in Section 4.5.3 and also in Section 3.5.

4.6.1 Comparison with EVT-type methods

In order to carry out this out-of-sample backtest, we adopt a rolling window estimation approach. Specifically, we first fix a testing window W_T in each case, which corresponds to the periods of time considered in our in-sample evaluation of VaR given in Section 3.6.2 (8 December 1997 to 9 November 2009 for the Dow Jones, 13 August 1997 to 16 July 2009 for the NASDAQ, 29 May 1997 to 12 August 2009 for the Nikkei, 28 September 2002 to 14 December 2010 for the JPY/GBP exchange rate). At each time t in this testing window W_T, we use a window of length W_E of prior information in order to predict the conditional ES on time $t+1$ (with parameter estimates updated when the estimation window changes), which is then evaluated differently in each test of ES. Regarding the use of the GARCH-UGH method specifically, we retain the implementation suggested by the results of in-sample VaR backtesting (see Section 3.6.2). In other words, on a given data set and for a given value of k, if we observed during in-sample VaR backtesting that the choice $\hat{\rho}_{k_\rho} = -1$ performed better,

then we retain this choice for out-of-sample ES estimation; otherwise, we estimate ρ as indicated in Section 3.2.3 (see Tables A.1-A.4 in Appendix A).

Traditional ES backtesting

Tables 4.1-4.8 gather the numerical results of traditional ES backtestings for the comparison between the GARCH-UGH, GARCH-EVT and UGH methods. In the tables, ESR refers to the ES regression tests of Bayer and Dimitriadis (2020b), CC to the conditional calibration tests of Nolde and Ziegel (2017) and ER to the exceedance residuals tests of McNeil and Frey (2000) (details are given in Section 4.4).

Unlike the results obtained in the traditional VaR backtestings, it can be seen that our proposed GARCH-UGH is not dominant approach and in general all three EVT-type methods are performing similarly. Based on the strict, auxiliary and intercept ESR tests, the GARCH-UGH approach fails 17, 17 and 0, whereas the GARCH-EVT fails 0 in all tests and UGH fails 15, 18 and 12 times out of 60 cases, respectively. With respect to the general and simple CC tests, the GARCH-UGH and GARCH-EVT approaches fails 8 and 0, and 2 and 0, respectively, while the unfiltered UGH method fails 17 times out of 60 cases. Moreover, in the commonly used standardized and simple ER tests GARCH-UGH and GARCH-EVT fail 9 and 3, and 8 and 18, respectively, while the UGH method fails 3 times out of 60 cases.

Unlike the cases in the dynamic extreme VaR estimation, GARCH-UGH typically performs worse than or equal to other approaches, especially in DJ index and JPY/GBP exchange rate based on the ESR tests. In general, GARCH-UGH tends to overestimate the one-step ahead conditional ES. This is because the resulting estimator of τth conditional ES (4.6) heavily depends on the estimates of unconditional VaR $\hat{q}_\tau(Z)$ (3.10) by the definition of ES and the EVI $\hat{\gamma}$ (3.9) by the use of asymptotic equivalence between VaR and quantile (4.4). In that sense the GARCH-UGH approach may inherit the vexing defects of $\hat{q}_\tau(Z)$, $\hat{\gamma}$ and even the estimator of second-order parameter $\hat{\rho}$ discussed in Section 3.2.3. On the other hand, GARCH-EVT seems to produce reasonable ES estimates as it neither overestimates nor underestimates according to the numerical results except the ER tests.

Regarding the methods of traditional ES backtesting, it has been observed that the two types of ER tests perform differently in reference to the number of failures

for GARCH-UGH and GARCH-EVT approaches, disagreeing with the empirical application of Bayer and Dimitriadis (2020b). It is thought that the use of volatility obtained in the GARCH step instead of the shortfall deviation, i.e., the dispersion truncated for some VaR, may have had the effect to this issue. The numerical results of Tables also illustrate that the backtesting decisions based on the general CC test are more conservative in contrast to the corresponding simple CC test, which agrees with the simulation study of Nolde and Ziegel (2017).

In summary, traditional ES backtestings do not yield definitive answers because GARCH-UGH, GARCH-EVT and even unfiltered UGH approaches are either all accepted or all rejected in most cases, and they sometimes give a contradicted decision, i.e., one test rejects the null hypothesis of the underestimation of ES while the others accept. We thus check the performance of estimators via the comparative ES backtesting.

Comparative ES backtesting

Tables 4.9-4.12 display the traffic light matrices of comparative ES backtesting given in Section 4.5 for three EVT-type methods, three quantile levels, five different threshold selections and four financial time series when $h = \frac{1}{2}$ (VaR, ES) scoring function of the form (4.16) is used. The competing models are given in the vertical axis with the benchmark models along the horizontal axis. Using the t-statistic based on the DM test (4.19), we reject the hypothesis H_0^- at the test level 5% if $1 - \Phi(DM) \leq 0.05$ while the hypothesis H_0^+ is rejected if $\Phi(DM) \leq 0.05$. Under H_0^-, the comparative backtesting is passed for the competing model if the null hypothesis fails to be rejected. On the other hand, under H_0^+ the backtesting for the competing model is passed if the null hypothesis is rejected. The green zone corresponds to the case when H_0^- is not rejected and H_0^+ is rejected, which suggests that the competing model is considered as better than the benchmark model. The yellow zone is when only one of the backtestings under H_0^- and H_0^+ is passed and we cannot conclude which model performs the best. The red zone corresponds to the case when both backtestings fail to be passed, indicating a problem with the competing model.

In contradiction to the results of traditional ES backtestings, it is illustrated that our proposed GARCH-UGH approach appears to be best overall. In 46 out of 60

TABLE 4.1: Traditional backtesting: out-of-sample evaluations of one-step ahead conditional ES estimates by EVT-type methods from 8 December 1997 to 9 November 2009 at different quantile levels for the negative log-returns of DJ index.

Testing window	3000							
Estimation window	1000							
0.999 ES	% of top obs. used	Str. ESR	Aux. ESR	Int. ESR	Gen. CC	Sim. CC	Std. ER	Sim. ER
GARCH-UGH	5%	0.030	0.019	0.962	1.000	1.000	0.366	0.621
				(0.075)	(0.627)	(0.157)	(0.613)	(1.000)
	10%	0.002	0.001	0.982	1.000	1.000	1.000	0.743
				(0.036)	(0.751)	(0.029)	(0.743)	(0.849)
	15%	0.030	0.019	0.962	1.000	1.000	0.753	0.753
				(0.075)	(0.805)	(0.000)	(1.000)	(0.487)
	20%	0.002	0.001	0.982	1.000	1.000	0.743	0.743
				(0.036)	(0.873)	(0.017)	(1.000)	(1.000)
	25%	0.030	0.019	0.962	1.000	1.000	0.753	0.753
				(0.075)	(0.934)	(0.001)	(1.000)	(0.613)
GARCH-EVT	5%	0.499	0.554	0.540	1.000	1.000	0.240	0.366
				(0.920)	(0.499)	(0.649)	(0.487)	(1.000)
	10%	0.503	0.548	0.508	1.000	1.000	0.075	0.222
				(0.985)	(0.372)	(0.847)	(0.873)	(0.449)
	15%	0.499	0.554	0.540	1.000	1.000	0.231	0.296
				(0.920)	(0.384)	(0.852)	(0.457)	(0.548)
	20%	0.503	0.548	0.508	0.926	0.891	0.091	0.168
				(0.985)	(0.285)	(0.859)	(0.236)	(0.279)
	25%	0.499	0.554	0.540	0.932	0.723	0.089	0.113
				(0.920)	(0.242)	(0.779)	(0.258)	(0.175)
UGH	5%	0.671	0.789	0.477	-	0.374	-	0.805
				(0.955)	-	(0.049)	-	(0.377)
	10%	0.684	0.751	0.797	-	1.000	-	0.961
				(0.405)	-	(0.021)	-	(0.039)
	15%	0.671	0.789	0.477	-	1.000	-	0.969
				(0.955)	-	(0.000)	-	(0.031)
	20%	0.684	0.751	0.797	-	1.000	-	0.982
				(0.405)	-	(0.000)	-	(0.018)
	25%	0.671	0.789	0.477	-	0.719	-	1.000
				(0.955)	-	(0.000)	-	(0.000)
0.995 ES								
GARCH-UGH	5%	0.475	0.467	0.654	1.000	1.000	0.655	0.813
				(0.691)	(0.882)	(0.276)	(0.918)	(0.401)
	10%	0.286	0.322	0.811	1.000	1.000	0.750	0.886
				(0.379)	(0.426)	(0.051)	(0.652)	(0.201)
	15%	0.081	0.104	0.888	1.000	1.000	0.762	0.930
				(0.223)	(0.179)	(0.003)	(0.534)	(0.096)
	20%	0.030	0.019	0.962	1.000	1.000	0.814	0.952
				(0.075)	(0.078)	(0.000)	(0.356)	(0.060)
	25%	0.002	0.001	0.982	1.000	1.000	0.820	0.981
				(0.036)	(0.051)	(0.000)	(0.307)	(0.022)
GARCH-EVT	5%	0.438	0.439	0.651	1.000	1.000	0.638	0.812
				(0.698)	(0.925)	(0.251)	(0.951)	(0.411)
	10%	0.497	0.486	0.592	1.000	1.000	0.490	0.668
				(0.816)	(0.738)	(0.510)	(0.778)	(0.782)
	15%	0.451	0.491	0.589	1.000	1.000	0.432	0.604
				(0.821)	(0.655)	(0.535)	(0.708)	(0.897)
	20%	0.499	0.554	0.540	1.000	1.000	0.286	0.477
				(0.920)	(0.483)	(0.805)	(0.492)	(0.823)
	25%	0.503	0.548	0.508	1.000	1.000	0.245	0.409
				(0.985)	(0.438)	(0.872)	(0.409)	(0.711)
UGH	5%	0.093	0.102	0.045	-	0.025	-	0.691
				(0.090)	-	(0.000)	-	(0.634)
	10%	0.167	0.181	0.085	-	0.056	-	0.892
				(0.169)	-	(0.000)	-	(0.197)
	15%	0.378	0.444	0.255	-	0.225	-	0.996
				(0.509)	-	(0.000)	-	(0.006)
	20%	0.671	0.789	0.477	-	0.612	-	1.000
				(0.955)	-	(0.000)	-	(0.000)
	25%	0.684	0.751	0.797	-	1.000	-	0.999
				(0.405)	-	(0.001)	-	(0.001)

Notes: ESR refers to the ES regression tests of Bayer and Dimitriadis (2020b), CC to the conditional calibration tests of Nolde and Ziegel (2017) and ER to the exceedance residuals tests of McNeil and Frey (2000). The p-values for the one-sided test are given with the two-sided test in brackets.

TABLE 4.2: (Cont.) Traditional backtesting: out-of-sample evaluations of one-step ahead conditional ES estimates by EVT-type methods from 8 December 1997 to 9 November 2009 at different quantile levels for the negative log-returns of DJ index.

Testing window	3000							
Estimation window	1000							
0.99 ES	% of top obs. used	Str. ESR	Aux. ESR	Int. ESR	Gen. CC	Sim. CC	Std. ER	Sim. ER
GARCH-UGH	5%	0.443	0.454	0.600	1.000	1.000	0.467	0.559
				(0.799)	(0.453)	(0.809)	(0.807)	(0.992)
	10%	0.396	0.433	0.699	1.000	1.000	0.636	0.678
				(0.602)	(0.909)	(0.214)	(0.925)	(0.795)
	15%	0.254	0.383	0.832	1.000	1.000	0.711	0.787
				(0.336)	(0.606)	(0.096)	(0.726)	(0.491)
	20%	0.100	0.154	0.929	1.000	1.000	0.814	0.868
				(0.141)	(0.168)	(0.007)	(0.388)	(0.220)
	25%	0.006	0.029	0.989	1.000	1.000	0.801	0.891
				(0.021)	(0.069)	(0.000)	(0.402)	(0.131)
GARCH-EVT	5%	0.450	0.458	0.629	1.000	1.000	0.408	0.463
				(0.742)	(0.710)	(0.615)	(0.717)	(0.839)
	10%	0.446	0.435	0.622	1.000	0.883	0.152	0.242
				(0.755)	(0.390)	(0.951)	(0.294)	(0.453)
	15%	0.423	0.441	0.611	1.000	0.900	0.152	0.229
				(0.779)	(0.377)	(0.941)	(0.313)	(0.439)
	20%	0.506	0.426	0.606	0.978	0.837	0.058	0.130
				(0.788)	(0.253)	(0.917)	(0.146)	(0.263)
	25%	0.483	0.447	0.601	0.780	0.791	0.038	0.080
				(0.799)	(0.207)	(0.758)	(0.111)	(0.184)
UGH	5%	0.004	0.008	0.003	-	0.003	-	0.082
				(0.007)	-	(0.000)	-	(0.182)
	10%	0.027	0.024	0.022	-	0.006	-	0.368
				(0.045)	-	(0.000)	-	(0.705)
	15%	0.061	0.082	0.046	-	0.022	-	0.805
				(0.092)	-	(0.000)	-	(0.403)
	20%	0.197	0.201	0.133	-	0.066	-	0.948
				(0.265)	-	(0.000)	-	(0.079)
	25%	0.615	0.606	0.392	-	0.230	-	0.995
				(0.783)	-	(0.000)	-	(0.005)

Notes: ESR refers to the ES regression tests of Bayer and Dimitriadis (2020b), CC to the conditional calibration tests of Nolde and Ziegel (2017) and ER to the exceedance residuals tests of McNeil and Frey (2000). The p-values for the one-sided test are given with the two-sided test in brackets.

TABLE 4.3: Traditional backtesting: out-of-sample evaluations of one-step ahead conditional ES estimates by EVT-type methods from 13 August 1997 to 16 July 2009 at different quantile levels for the negative log-returns of NASDAQ index.

Testing window	3000							
Estimation window	1000							
0.999 ES	% of top obs. used	Str. ESR	Aux. ESR	Int. ESR	Gen. CC	Sim. CC	Std. ER	Sim. ER
GARCH-UGH	5%	0.307	0.304	0.492	1.000	0.415	0.867	0.155
				(0.984)	(0.222)	(0.470)	(0.444)	(0.580)
	10%	0.573	0.470	0.670	1.000	0.722	0.242	0.208
				(0.607)	(0.267)	(0.654)	(0.803)	(0.920)
	15%	0.307	0.304	0.492	1.000	1.000	0.181	0.893
				(0.984)	(0.350)	(0.085)	(0.376)	(0.222)
	20%	0.573	0.470	0.697	1.000	1.000	0.642	0.943
				(0.607)	(0.118)	(0.000)	(1.000)	(0.363)
	25%	0.307	0.304	0.492	0.000	1.000	0.244	0.756
				(0.984)	(0.126)	(0.000)	(1.000)	(0.244)
GARCH-EVT	5%	0.220	0.223	0.356	0.620	0.223	0.903	0.097
				(0.712)	(0.083)	(0.295)	(0.122)	(0.143)
	10%	0.236	0.237	0.322	0.225	0.162	0.059	0.065
				(0.645)	(0.059)	(0.258)	(0.069)	(0.082)
	15%	0.220	0.223	0.356	0.398	0.143	0.946	0.066
				(0.712)	(0.051)	(0.240)	(0.038)	(0.071)
	20%	0.236	0.237	0.322	0.168	0.121	0.049	0.051
				(0.645)	(0.047)	(0.215)	(0.052)	(0.053)
	25%	0.220	0.233	0.356	0.298	0.107	0.946	0.052
				(0.712)	(0.041)	(0.197)	(0.056)	(0.054)
UGH	5%	0.201	0.157	0.296	-	0.248	-	0.318
				(0.592)	-	(0.084)	-	(0.587)
	10%	0.030	0.014	0.870	-	0.570	-	0.615
				(0.261)	-	(0.183)	-	(0.878)
	15%	0.201	0.157	0.296	-	1.000	-	0.771
				(0.592)	-	(0.445)	-	(0.479)
	20%	0.030	0.014	0.870	-	1.000	-	0.876
				(0.261)	-	(0.059)	-	(0.228)
	25%	0.201	0.157	0.296	-	1.000	-	1.000
				(0.592)	-	(0.000)	-	(0.000)
0.995 ES								
GARCH-UGH	5%	0.226	0.232	0.408	0.619	0.646	0.104	0.142
				(0.816)	(0.243)	(0.530)	(0.203)	(0.275)
	10%	0.307	0.304	0.492	1.000	0.994	0.913	0.077
				(0.984)	(0.233)	(0.878)	(0.179)	(0.170)
	15%	0.381	0.365	0.527	0.697	0.750	0.082	0.056
				(0.946)	(0.188)	(0.916)	(0.154)	(0.123)
	20%	0.459	0.412	0.606	1.000	0.898	0.761	0.273
				(0.788)	(0.460)	(0.961)	(0.468)	(0.507)
	25%	0.573	0.470	0.697	0.503	0.867	0.203	0.213
				(0.607)	(0.428)	(0.847)	(0.413)	(0.414)
GARCH-EVT	5%	0.224	0.228	0.398	0.552	0.596	0.013	0.007
				(0.795)	(0.066)	(0.635)	(0.030)	(0.022)
	10%	0.220	0.233	0.356	0.625	0.488	0.988	0.006
				(0.712)	(0.038)	(0.236)	(0.016)	(0.012)
	15%	0.231	0.222	0.346	0.243	0.434	0.009	0.005
				(0.693)	(0.029)	(0.086)	(0.011)	(0.006)
	20%	0.221	0.236	0.331	0.212	0.380	0.008	0.005
				(0.661)	(0.027)	(0.063)	(0.009)	(0.006)
	25%	0.236	0.237	0.322	0.199	0.346	0.008	0.005
				(0.645)	(0.027)	(0.055)	(0.008)	(0.006)
UGH	5%	0.133	0.193	0.163	-	0.020	-	0.648
				(0.325)	-	(0.001)	-	(0.775)
	10%	0.201	0.157	0.296	-	0.070	-	0.805
				(0.592)	-	(0.001)	-	(0.440)
	15%	0.122	0.153	0.438	-	0.285	-	0.906
				(0.876)	-	(0.004)	-	(0.136)
	20%	0.091	0.073	0.660	-	1.000	-	0.961
				(0.680)	-	(0.003)	-	(0.044)
	25%	0.030	0.014	0.870	-	1.000	-	0.958
				(0.261)	-	(0.000)	-	(0.042)

Notes: ESR refers to the ES regression tests of Bayer and Dimitriadis (2020b), CC to the conditional calibration tests of Nolde and Ziegel (2017) and ER to the exceedance residuals tests of McNeil and Frey (2000). The p-values for the one-sided test are given with the two-sided test in brackets.

TABLE 4.4: (Cont.) Traditional backtesting: out-of-sample evaluations of one-step ahead conditional ES estimates by EVT-type methods from 13 August 1997 to 16 July 2009 at different quantile levels for the negative log-returns of NASDAQ index.

Testing window	3000							
Estimation window	1000							
0.99 ES	% of top obs. used	Str. ESR	Aux. ESR	Int. ESR	Gen. CC	Sim. CC	Std. ER	Sim. ER
GARCH-UGH	5%	0.113	0.104	0.461	0.717	0.841	0.991	0.012
				(0.921)	(0.098)	(0.770)	(0.031)	(0.038)
	10%	0.186	0.173	0.527	0.504	1.000	0.098	0.085
				(0.947)	(0.248)	(0.664)	(0.205)	(0.179)
	15%	0.364	0.329	0.672	0.368	1.000	0.704	0.293
				(0.657)	(0.565)	(0.740)	(0.556)	(0.534)
	20%	0.510	0.507	0.825	0.281	1.000	0.651	0.610
				(0.350)	(0.756)	(0.266)	(0.770)	(0.845)
	25%	0.127	0.446	0.971	0.043	0.473	0.205	0.797
				(0.058)	(0.353)	(0.020)	(0.439)	(0.426)
GARCH-EVT	5%	0.174	0.133	0.510	0.357	0.857	0.982	0.023
				(0.979)	(0.101)	(0.915)	(0.052)	(0.066)
	10%	0.176	0.163	0.503	0.119	0.528	0.016	0.016
				(0.995)	(0.067)	(0.632)	(0.028)	(0.040)
	15%	0.181	0.184	0.505	0.123	0.528	0.992	0.012
				(0.990)	(0.075)	(0.632)	(0.032)	(0.038)
	20%	0.200	0.202	0.519	0.026	0.328	0.001	0.008
				(0.963)	(0.040)	(0.108)	(0.008)	(0.022)
	25%	0.200	0.234	0.530	0.014	0.215	0.997	0.024
				(0.939)	(0.040)	(0.051)	(0.012)	(0.006)
UGH	5%	0.003	0.000	0.027	-	0.001	-	0.512
				(0.054)	-	(0.000)	-	(0.949)
	10%	0.004	0.001	0.065	-	0.004	-	0.748
				(0.130)	-	(0.000)	-	(0.578)
	15%	0.012	0.004	0.159	-	0.018	-	0.918
				(0.319)	-	(0.000)	-	(0.134)
	20%	0.028	0.012	0.353	-	0.114	-	0.971
				(0.706)	-	(0.000)	-	(0.033)
	25%	0.035	0.011	0.631	-	0.500	-	0.984
				(0.738)	-	(0.000)	-	(0.016)

Notes: ESR refers to the ES regression tests of Bayer and Dimitriadis (2020b), CC to the conditional calibration tests of Nolde and Ziegel (2017) and ER to the exceedance residuals tests of McNeil and Frey (2000). The p-values for the one-sided test are given with the two-sided test in brackets.

TABLE 4.5: Traditional backtesting: out-of-sample evaluations of one-step ahead conditional ES estimates by EVT-type methods from 29 May 1997 to 12 August 2009 at different quantile levels for the negative log-returns of NIKKEI index.

Testing window	3000							
Estimation window	1000							
0.999 ES	% of top obs. used	Str. ESR	Aux. ESR	Int. ESR	Gen. CC	Sim. CC	Std. ER	Sim. ER
GARCH-UGH	5%	0.575	0.754	0.761	1.000	1.000	0.291	0.709
				(0.478)	(0.815)	(0.807)	(1.000)	(1.000)
	10%	0.463	0.582	0.883	1.000	0.978	0.375	0.625
				(0.234)	(0.772)	(0.853)	(0.865)	(1.000)
	15%	0.575	0.754	0.761	0.225	0.719	0.000	1.000
				(0.478)	(0.379)	(0.012)	(0.000)	(0.000)
	20%	0.463	0.582	0.883	0.365	0.719	0.000	0.000
				(0.234)	(0.829)	(0.169)	(0.000)	(0.000)
	25%	0.575	0.754	0.761	0.000	0.068	0.000	0.000
				(0.478)	(0.317)	(0.000)	(0.000)	(0.000)
GARCH-EVT	5%	0.593	0.587	0.436	1.000	1.000	0.981	0.651
				(0.873)	(0.602)	(0.661)	(0.078)	(0.948)
	10%	0.593	0.596	0.440	1.000	0.929	0.265	0.315
				(0.881)	(0.431)	(0.858)	(0.819)	(0.559)
	15%	0.593	0.587	0.436	1.000	0.724	0.344	0.630
				(0.873)	(0.558)	(0.472)	(0.893)	(0.965)
	20%	0.593	0.596	0.440	1.000	0.564	0.721	0.436
				(0.881)	(0.394)	(0.462)	(0.439)	(0.663)
	25%	0.593	0.587	0.436	1.000	0.568	0.574	0.526
				(0.873)	(0.425)	(0.466)	(0.631)	(0.803)
UGH	5%	0.901	0.878	0.441	-	0.466	-	0.707
				(0.881)	-	(0.307)	-	(0.672)
	10%	0.307	0.658	0.761	-	0.804	-	0.694
				(0.478)	-	(0.384)	-	(0.573)
	15%	0.901	0.878	0.441	-	1.000	-	0.951
				(0.881)	-	(0.438)	-	(0.080)
	20%	0.307	0.658	0.761	-	1.000	-	0.742
				(0.478)	-	(0.002)	-	(0.258)
	25%	0.901	0.878	0.441	-	1.000	-	1.000
				(0.881)	-	(0.000)	-	(0.000)
0.995 ES								
GARCH-UGH	5%	0.672	0.677	0.511	1.000	0.750	0.753	0.262
				(0.977)	(0.462)	(0.981)	(0.454)	(0.480)
	10%	0.906	0.855	0.655	1.000	1.000	0.401	0.620
				(0.690)	(0.899)	(0.760)	(0.918)	(0.865)
	15%	0.740	0.797	0.677	1.000	1.000	0.435	0.578
				(0.645)	(0.994)	(0.793)	(0.999)	(0.958)
	20%	0.575	0.754	0.761	1.000	1.000	0.318	0.675
				(0.478)	(0.658)	(0.557)	(0.742)	(0.743)
	25%	0.463	0.582	0.883	0.536	0.578	0.322	0.657
				(0.234)	(0.580)	(0.180)	(0.840)	(0.857)
GARCH-EVT	5%	0.662	0.724	0.491	1.000	0.615	0.953	0.050
				(0.982)	(0.209)	(0.517)	(0.122)	(0.122)
	10%	0.667	0.605	0.507	1.000	0.592	0.866	0.118
				(0.986)	(0.314)	(0.752)	(0.278)	(0.251)
	15%	0.609	0.584	0.461	0.874	0.472	0.972	0.031
				(0.922)	(0.167)	(0.360)	(0.084)	(0.081)
	20%	0.593	0.587	0.436	0.719	0.426	0.982	0.024
				(0.873)	(0.109)	(0.122)	(0.029)	(0.041)
	25%	0.593	0.596	0.440	0.726	0.426	0.981	0.029
				(0.881)	(0.113)	(0.128)	(0.049)	(0.065)
UGH	5%	0.142	0.243	0.024	-	0.043	-	0.782
				(0.047)	-	(0.005)	-	(0.442)
	10%	0.298	0.278	0.192	-	0.092	-	0.859
				(0.384)	-	(0.005)	-	(0.263)
	15%	0.619	0.499	0.159	-	0.241	-	0.896
				(0.318)	-	(0.009)	-	(0.173)
	20%	0.901	0.878	0.441	-	0.634	-	0.944
				(0.881)	-	(0.012)	-	(0.071)
	25%	0.307	0.658	0.761	-	1.000	-	0.967
				(0.478)	-	(0.054)	-	(0.034)

Notes: ESR refers to the ES regression tests of Bayer and Dimitriadis (2020b), CC to the conditional calibration tests of Nolde and Ziegel (2017) and ER to the exceedance residuals tests of McNeil and Frey (2000). The p-values for the one-sided test are given with the two-sided test in brackets.

TABLE 4.6: (Cont.) Traditional backtesting: out-of-sample evaluations of one-step ahead conditional ES estimates by EVT-type methods from 29 May 1997 to 12 August 2009 at different quantile levels for the negative log-returns of NIKKEI index.

Testing window	3000							
Estimation window	1000							
0.99 ES	% of top obs. used	Str. ESR	Aux. ESR	Int. ESR	Gen. CC	Sim. CC	Std. ER	Sim. ER
GARCH-UGH	5%	0.674	0.620	0.493	1.000	1.000	0.617	0.457
				(0.986)	(0.679)	(0.829)	(0.683)	(0.807)
	10%	0.612	0.659	0.697	1.000	1.000	0.305	0.726
				(0.606)	(0.658)	(0.506)	(0.694)	(0.598)
	15%	0.380	0.384	0.795	1.000	1.000	0.138	0.897
				(0.409)	(0.129)	(0.075)	(0.263)	(0.188)
	20%	0.203	0.163	0.939	0.246	1.000	0.100	0.942
				(0.123)	(0.018)	(0.002)	(0.134)	(0.075)
	25%	0.771	0.489	0.763	1.000	1.000	0.126	0.939
				(0.475)	(0.066)	(0.036)	(0.213)	(1.000)
GARCH-EVT	5%	0.721	0.523	0.543	1.000	0.958	0.580	0.476
				(0.913)	(0.758)	(0.860)	(0.760)	(0.879)
	10%	0.753	0.669	0.582	1.000	0.841	0.699	0.331
				(0.835)	(0.548)	(0.982)	(0.546)	(0.610)
	15%	0.766	0.685	0.567	0.628	0.800	0.878	0.148
				(0.867)	(0.332)	(0.784)	(0.260)	(0.301)
	20%	0.766	0.570	0.572	0.681	0.795	0.871	0.172
				(0.856)	(0.332)	(0.787)	(0.257)	(0.333)
	25%	0.757	0.607	0.586	0.465	0.646	0.913	0.140
				(0.829)	(0.282)	(0.638)	(0.208)	(0.291)
UGH	5%	0.030	0.072	0.005	-	0.009	-	0.143
				(0.009)	-	(0.019)	-	(0.297)
	10%	0.044	0.054	0.008	-	0.019	-	0.278
				(0.016)	-	(0.026)	-	(0.529)
	15%	0.143	0.164	0.028	-	0.049	-	0.478
				(0.056)	-	(0.048)	-	(0.878)
	20%	0.400	0.382	0.092	-	0.140	-	0.732
				(0.185)	-	(0.079)	-	(0.582)
	25%	0.921	0.855	0.404	-	0.521	-	0.926
				(0.808)	-	(0.077)	-	(0.108)

Notes: ESR refers to the ES regression tests of Bayer and Dimitriadis (2020b), CC to the conditional calibration tests of Nolde and Ziegel (2017) and ER to the exceedance residuals tests of McNeil and Frey (2000). The p-values for the one-sided test are given with the two-sided test in brackets.

TABLE 4.7: Traditional backtesting: out-of-sample evaluations of one-step ahead conditional ES estimates by EVT-type methods from 28 September 2002 to 14 December 2010 at different quantile levels for the negative log-returns of JPY/GBP exchange rate.

Testing window	3000							
Estimation window	1000							
0.999 ES	% of top obs. used	Str. ESR	Aux. ESR	Int. ESR	Gen. CC	Sim. CC	Std. ER	Sim. ER
GARCH-UGH	5%	0.004	0.002	0.942	1.000	0.752	0.780	0.220
				(0.116)	(0.452)	(1.000)	(0.477)	(0.477)
	10%	0.000	0.000	0.998	0.510	0.719	0.000	1.000
				(0.003)	(0.776)	(0.572)	(0.000)	(0.000)
	15%	0.004	0.002	0.942	0.188	0.719	0.000	1.000
				(0.116)	(0.332)	(0.128)	(0.000)	(0.000)
	20%	0.000	0.000	0.998	0.036	0.719	0.000	1.000
				(0.003)	(0.202)	(0.011)	(0.000)	(0.000)
	25%	0.004	0.002	0.942	0.001	0.719	0.000	1.000
				(0.116)	(0.187)	(0.000)	(0.000)	(0.000)
GARCH-EVT	5%	0.905	0.506	0.290	1.000	0.688	0.950	0.050
				(0.580)	(0.213)	(0.471)	(0.112)	(0.112)
	10%	0.627	0.355	0.271	1.000	0.627	0.576	0.662
				(0.541)	(0.325)	(0.625)	(0.494)	(0.732)
	15%	0.905	0.506	0.290	1.000	0.678	1.000	0.815
				(0.580)	(0.336)	(0.640)	(0.070)	(0.885)
	20%	0.627	0.355	0.271	1.000	0.517	0.838	0.161
				(0.541)	(0.348)	(0.449)	(0.296)	(0.311)
	25%	0.905	0.506	0.290	1.000	0.482	0.769	0.243
				(0.580)	(0.391)	(0.309)	(0.408)	(0.438)
UGH	5%	0.060	0.028	0.145	-	0.388	-	0.039
				(0.290)	-	(0.135)	-	(0.105)
	10%	0.068	0.013	0.097	-	0.697	-	0.471
				(0.195)	-	(0.316)	-	(0.754)
	15%	0.060	0.028	0.145	-	0.946	-	0.612
				(0.290)	-	(0.450)	-	(0.939)
	20%	0.068	0.013	0.097	-	0.297	-	0.143
				(0.195)	-	(0.020)	-	(0.295)
	25%	0.060	0.028	0.145	-	0.136	-	0.005
				(0.290)	-	(0.005)	-	(0.015)
0.995 ES								
GARCH-UGH	5%	0.654	0.627	0.474	1.000	1.000	0.292	0.687
				(0.948)	(0.678)	(0.263)	(0.780)	(0.893)
	10%	0.389	0.344	0.551	1.000	1.000	0.258	0.677
				(0.889)	(0.203)	(0.190)	(0.538)	(0.689)
	15%	0.045	0.052	0.864	0.099	1.000	0.159	0.787
				(0.273)	(0.057)	(0.009)	(0.268)	(0.373)
	20%	0.004	0.002	0.942	0.001	1.000	0.117	0.897
				(0.116)	(0.013)	(0.000)	(0.146)	(0.137)
	25%	0.000	0.000	0.998	0.000	0.578	0.042	0.965
				(0.003)	(0.007)	(0.000)	(0.043)	(0.037)
GARCH-EVT	5%	0.675	0.767	0.279	1.000	0.940	0.943	0.034
				(0.557)	(0.240)	(0.652)	(0.151)	(0.118)
	10%	0.770	0.505	0.301	1.000	0.591	0.928	0.044
				(0.602)	(0.267)	(0.612)	(0.190)	(0.139)
	15%	0.783	0.564	0.315	1.000	0.688	0.773	0.246
				(0.629)	(0.467)	(0.531)	(0.427)	(0.453)
	20%	0.905	0.506	0.290	1.000	0.496	0.920	0.090
				(0.580)	(0.280)	(0.496)	(0.179)	(0.215)
	25%	0.627	0.355	0.271	1.000	0.443	0.955	0.045
				(0.541)	(0.236)	(0.447)	(0.125)	(0.136)
UGH	5%	0.262	0.781	0.145	-	0.157	-	0.139
				(0.290)	-	(0.116)	-	(0.295)
	10%	0.955	0.762	0.308	-	0.188	-	0.350
				(0.616)	-	(0.067)	-	(0.586)
	15%	0.399	0.721	0.240	-	0.304	-	0.708
				(0.481)	-	(0.168)	-	(0.654)
	20%	0.060	0.028	0.145	-	0.077	-	0.316
				(0.290)	-	(0.003)	-	(0.578)
	25%	0.068	0.013	0.097	-	0.017	-	0.093
				(0.195)	-	(0.000)	-	(0.207)

Notes: ESR refers to the ES regression tests of Bayer and Dimitriadis (2020b), CC to the conditional calibration tests of Nolde and Ziegel (2017) and ER to the exceedance residuals tests of McNeil and Frey (2000). The p-values for the one-sided test are given with the two-sided test in brackets.

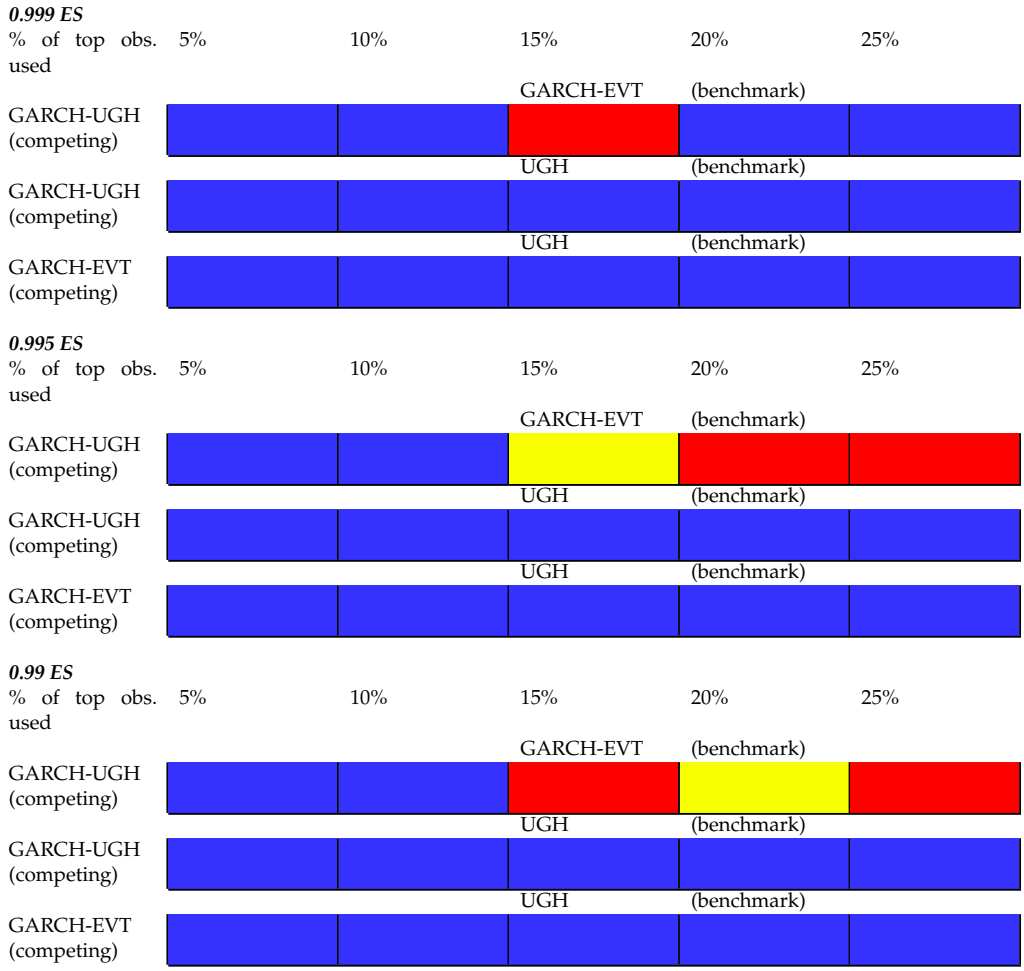
TABLE 4.8: (Cont.) Traditional backtesting: out-of-sample evaluations of one-step ahead conditional ES estimates by EVT-type methods from 28 September 2002 to 14 December 2010 at different quantile levels for the negative log-returns of JPY/GBP exchange rate.

Testing window	3000							
Estimation window	1000							
0.99 ES	% of top obs. used	Str. ESR	Aux. ESR	Int. ESR	Gen. CC	Sim. CC	Std. ER	Sim. ER
GARCH-UGH	5%	0.632	0.651	0.350	1.000	0.937	0.349	0.661
				(0.700)	(0.733)	(0.119)	(0.805)	(0.765)
	10%	0.609	0.587	0.557	1.000	1.000	0.126	0.890
				(0.885)	(0.038)	(0.008)	(0.204)	(0.173)
	15%	0.319	0.222	0.814	0.989	1.000	0.099	0.928
				(0.372)	(0.006)	(0.002)	(0.116)	(0.086)
	20%	0.066	0.040	0.944	0.102	1.000	0.068	0.971
				(0.112)	(0.000)	(0.000)	(0.071)	(0.033)
	25%	0.001	0.000	0.991	0.000	1.000	0.022	0.996
				(0.018)	(0.000)	(0.000)	(0.022)	(0.004)
GARCH-EVT	5%	0.714	0.689	0.282	1.000	0.927	0.640	0.362
				(0.564)	(0.648)	(0.393)	(0.648)	(0.649)
	10%	0.612	0.539	0.255	1.000	0.568	0.861	0.255
				(0.509)	(0.352)	(0.503)	(0.286)	(0.472)
	15%	0.619	0.620	0.263	1.000	0.651	0.758	0.325
				(0.526)	(0.465)	(0.426)	(0.455)	(0.599)
	20%	0.633	0.588	0.237	1.000	0.535	0.888	0.134
				(0.475)	(0.316)	(0.495)	(0.226)	(0.259)
	25%	0.695	0.986	0.246	1.000	0.509	0.929	0.082
				(0.492)	(0.266)	(0.552)	(0.167)	(0.184)
UGH	5%	0.320	0.306	0.083	-	0.048	-	0.155
				(0.166)	-	(0.031)	-	(0.314)
	10%	0.037	0.135	0.041	-	0.048	-	0.460
				(0.083)	-	(0.002)	-	(0.778)
	15%	0.148	0.450	0.066	-	0.061	-	0.776
				(0.132)	-	(0.003)	-	(0.482)
	20%	0.007	0.013	0.032	-	0.014	-	0.155
				(0.065)	-	(0.001)	-	(0.311)
	25%	0.036	0.163	0.021	-	0.001	-	0.005
				(0.041)	-	(0.000)	-	(0.028)

Notes: ESR refers to the ES regression tests of Bayer and Dimitriadis (2020b), CC to the conditional calibration tests of Nolde and Ziegel (2017) and ER to the exceedance residuals tests of McNeil and Frey (2000). The p-values for the one-sided test are given with the two-sided test in brackets.

cases, the GARCH-UGH approach is considered as better than GARCH-EVT approach with 2 cases of no definitive answers based on the realized scores of (VaR, ES). We can also observe in comparative backtesting that GARCH-UGH is regarded as the better estimator than the GARCH-EVT even if it fails one of the traditional ES backtesting, i.e., ESR tests.

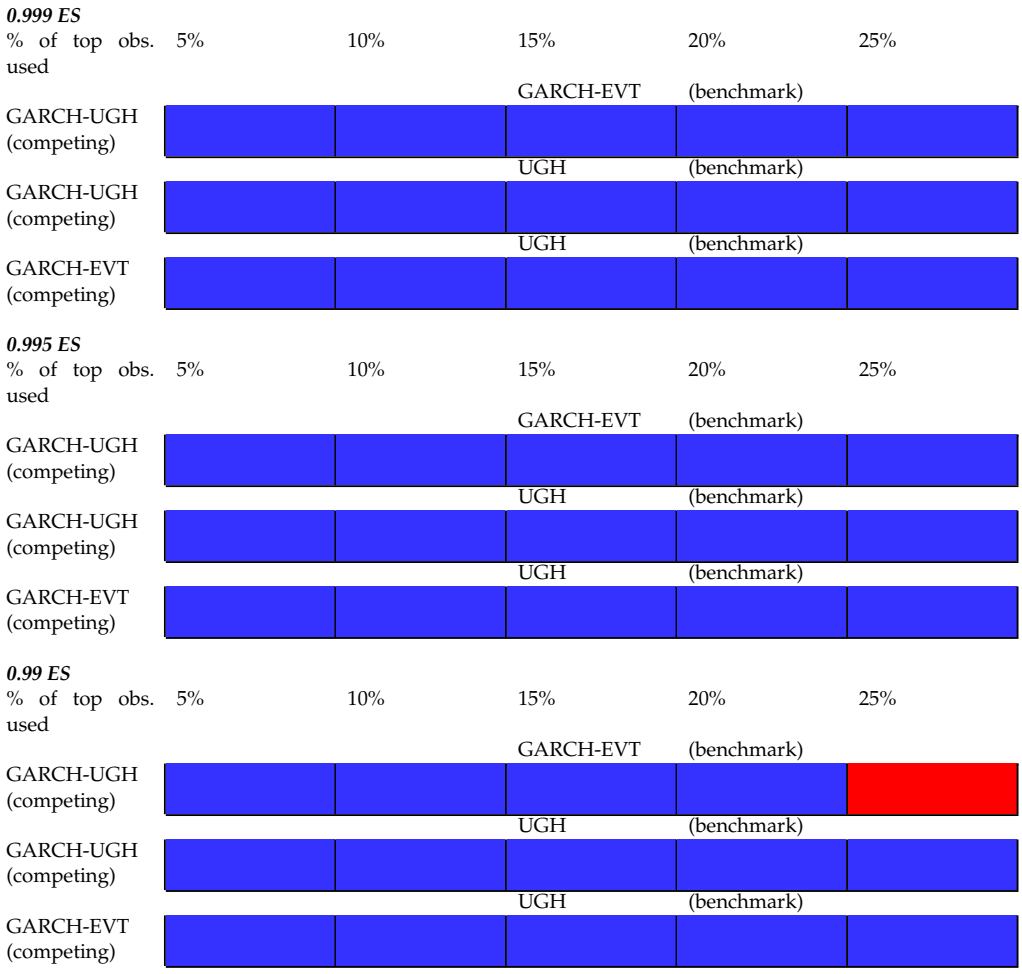
TABLE 4.9: Comparative backtesting: out-of-sample evaluations (traffic light matrices) of one-step ahead conditional ES estimates by EVT-type methods from 8 December 1997 to 9 November 2009 at different quantile levels for the negative log-returns of DJ index by means of the Diebold-Mariano test using $h = \frac{1}{2}$ (VaR, ES) scoring function.



Notes: When the competing model in the row is swapped with the benchmark model in the column, the results are reversed.

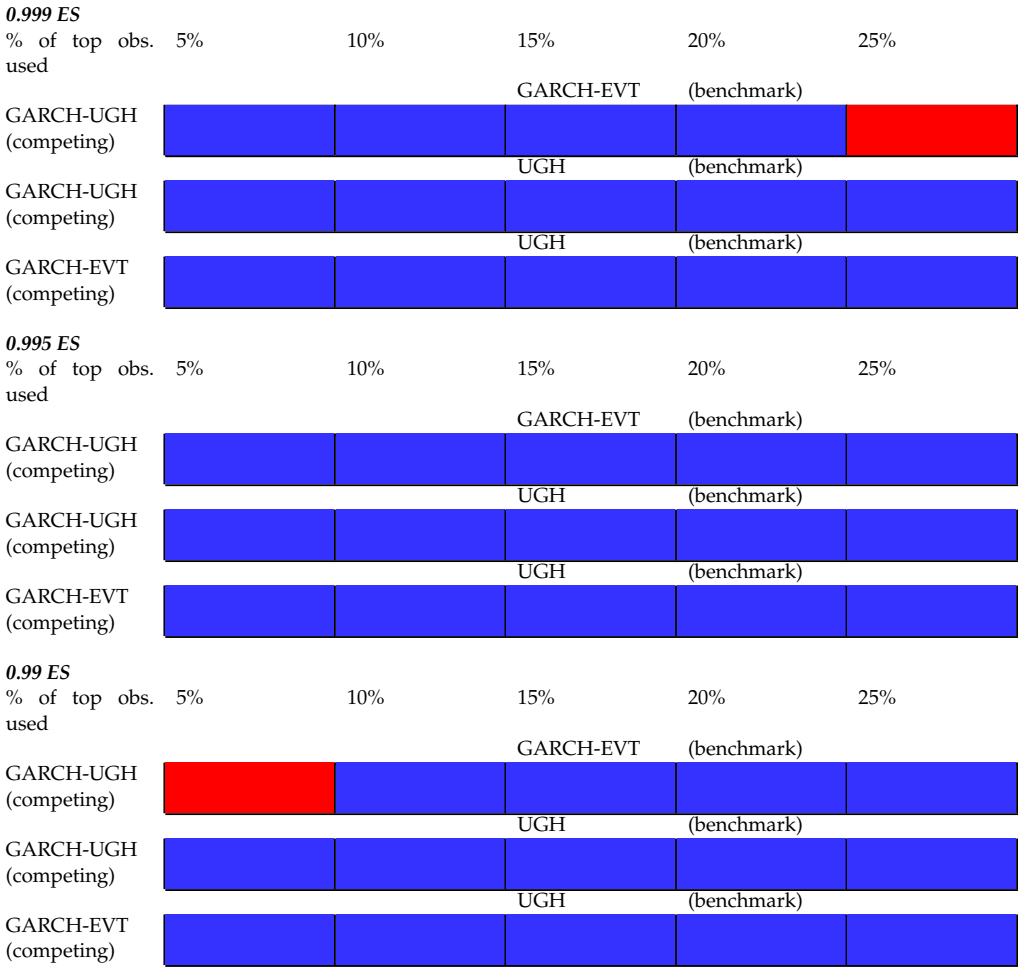
Tables 4.13-4.16 show the traffic light matrices for the $h = 0$ (VaR, ES) scoring function given as the form (4.17). In 47 out of 60 cases, the GARCH-UGH approach is considered as better than the GARCH-EVT approach with 1 case of no decision

TABLE 4.10: Comparative backtesting: out-of-sample evaluations (traffic light matrices) of one-step ahead conditional ES estimates by EVT-type methods from 13 August 1997 to 16 July 2009 at different quantile levels for the negative log-returns of NASDAQ index by means of the Diebold-Mariano test using $h = \frac{1}{2}$ (VaR, ES) scoring function.



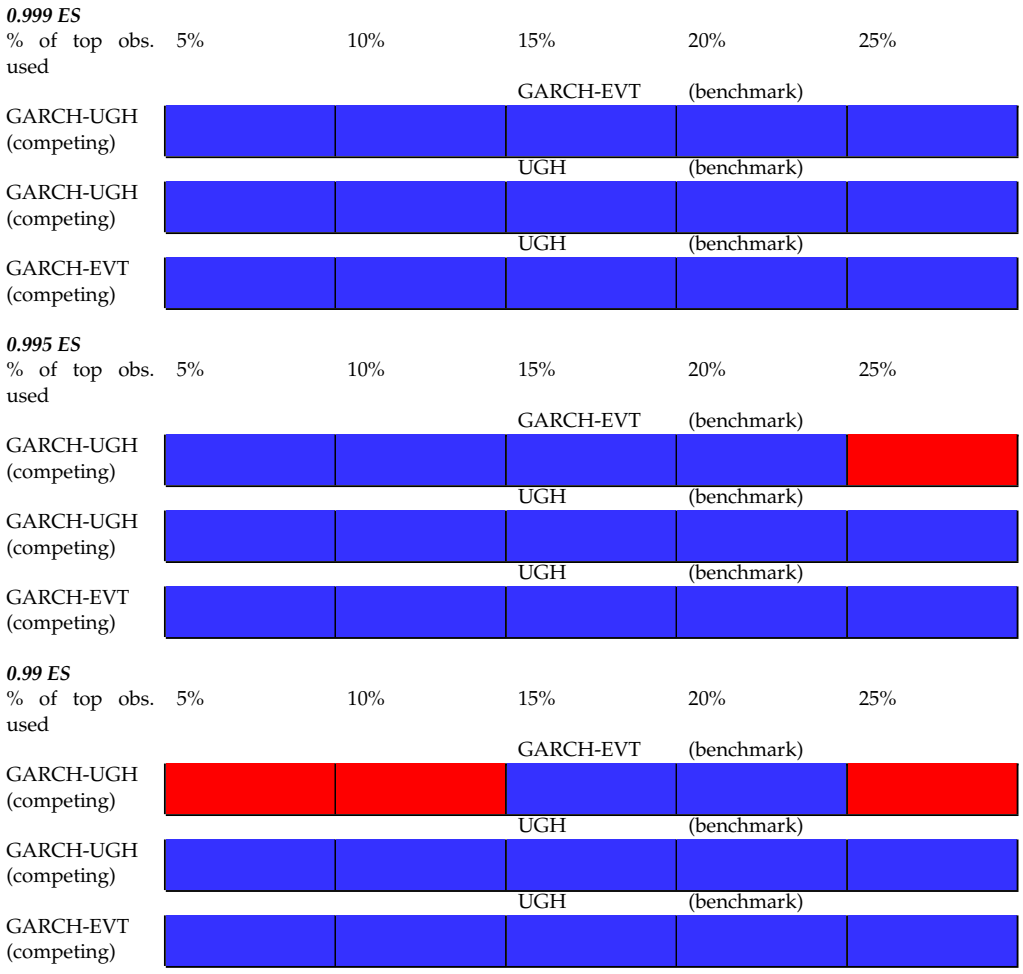
Notes: When the competing model in the row is swapped with the benchmark model in the column, the results are reversed.

TABLE 4.11: Comparative backtesting: out-of-sample evaluations (traffic light matrices) of one-step ahead conditional ES estimates by EVT-type methods from 29 May 1997 to 12 August 2009 at different quantile levels for the negative log-returns of NIKKEI index by means of the Diebold-Mariano test using $h = \frac{1}{2}$ (VaR, ES) scoring function.



Notes: When the competing model in the row is swapped with the benchmark model in the column, the results are reversed.

TABLE 4.12: Comparative backtesting: out-of-sample evaluations (traffic light matrices) of one-step ahead conditional ES estimates by EVT-type methods from 28 September 2002 to 14 December 2010 at different quantile levels for the negative log-returns of JPY/GBP exchange rate by means of the Diebold-Mariano test using $h = \frac{1}{2}$ (VaR, ES) scoring function.



Notes: When the competing model in the row is swapped with the benchmark model in the column, the results are reversed.

taken. Comparative backtestings with two scoring functions result in a good agreement with the GARCH-UGH approach being the best estimator of ES, while the unfiltered UGH being the worst estimator, i.e., failing the comparative backtestings against other EVT-type methods. Regarding the discrimination ability of two chosen scoring functions, $h = \frac{1}{2}$ and $h = 0$ (VaR, ES) scoring functions yield 2 and 1 times of no definitive answers out of 60 cases (see also Section 4.6.2 for the similar trend). Although it is a subtle difference between two scoring functions, the 0-homogenous (VaR, ES) scoring function allows us to study heavier-tailed processes than the $\frac{1}{2}$-homogenous one, which is also found in Nolde and Ziegel (2017).

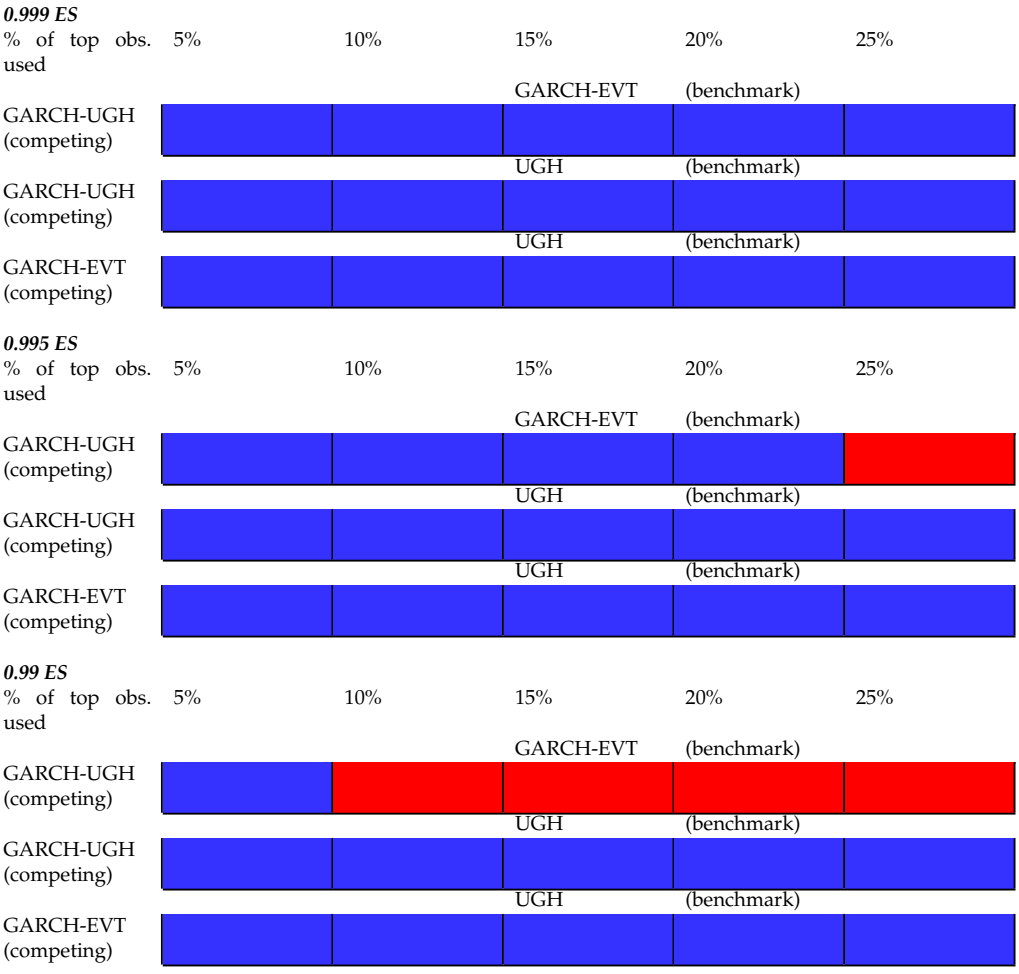
They indeed yield definitive answers to the cases when the estimation methods are all accepted or all rejected, although they are not always in line with the results of traditional ES backtestings. However, results of comparative ES backtestings with two scoring functions are more or less same as the results of traditional and comparative VaR backtestings based on the number of VaR violations. Righi and Ceretta (2015) remark that an ES model should be precise when VaR violations occur so the requirement of a model is that it obtains the correct number of VaR violations compared to the expected ones. It is thus suggested that better VaR estimates tend to produce better ES estimates.

4.6.2 Comparison with basic estimation methods

Traditional ES backtesting

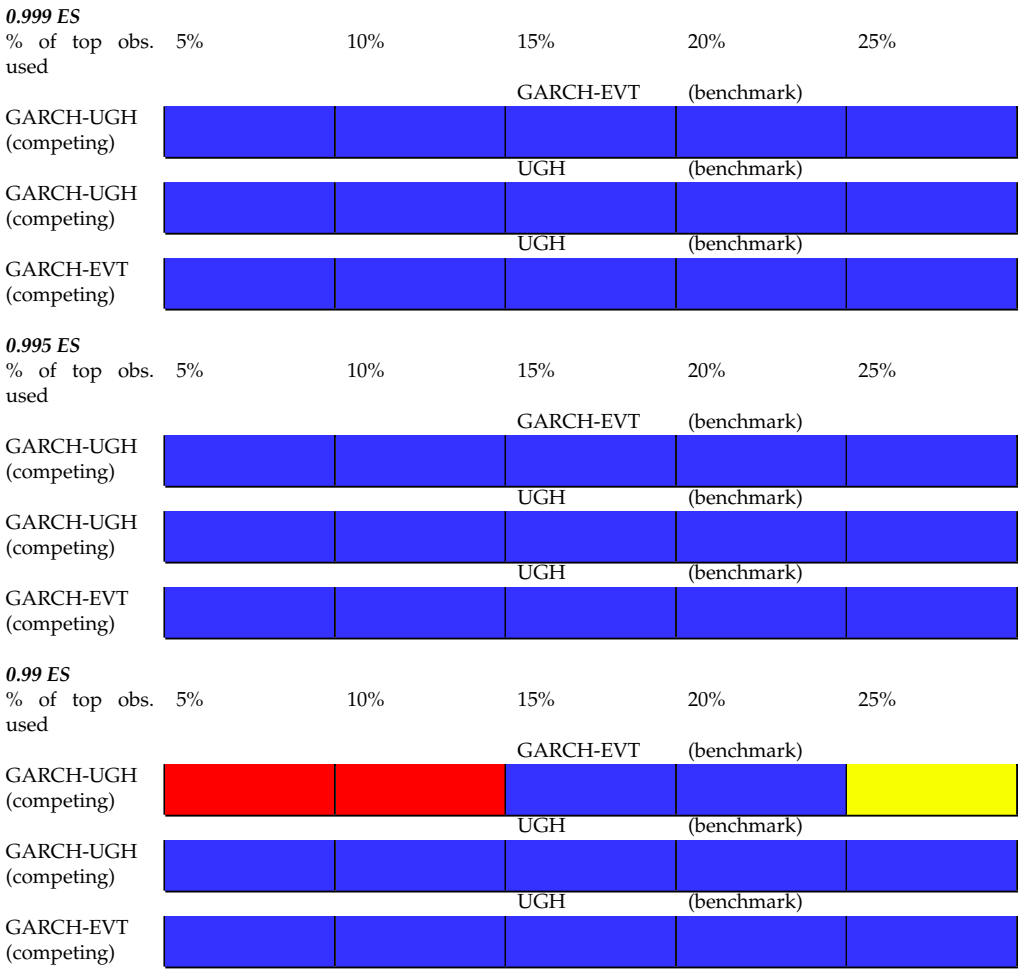
Tables 4.17-4.20 show the numerical results of traditional ES backtestings for the comparison between the HS, GARCH-N and GARCH-t methods. Note that in our case the sample fraction for GARCH-UGH approach is regarded as optimal when the certain threshold ranging between 5% and 25% has the closet number of VaR violations to the theoretically expected ones, pass both the Kupiec and Christoffersen tests (see Tables 3.7-3.10) and pass all traditional ES backtestings (see Tables 4.1-4.8). Tables support the use of GARCH-UGH approach for the estimation of dynamic (extreme) ES because it outperforms basic estimation methods, while the GARCH-N method performing the worst. Our GARCH-UGH approach with optimal sample fraction never fails the traditional ES backtestings except the strict and auxiliary ESR tests; in both tests it fails 2 times out of 12 cases. Focusing on the commonly used

TABLE 4.13: Comparative backtesting: out-of-sample evaluations (traffic light matrices) of one-step ahead conditional ES estimates by EVT-type methods from 8 December 1997 to 9 November 2009 at different quantile levels for the negative log-returns of DJ index by means of the Diebold-Mariano test using $h = 0$ (VaR, ES) scoring function.



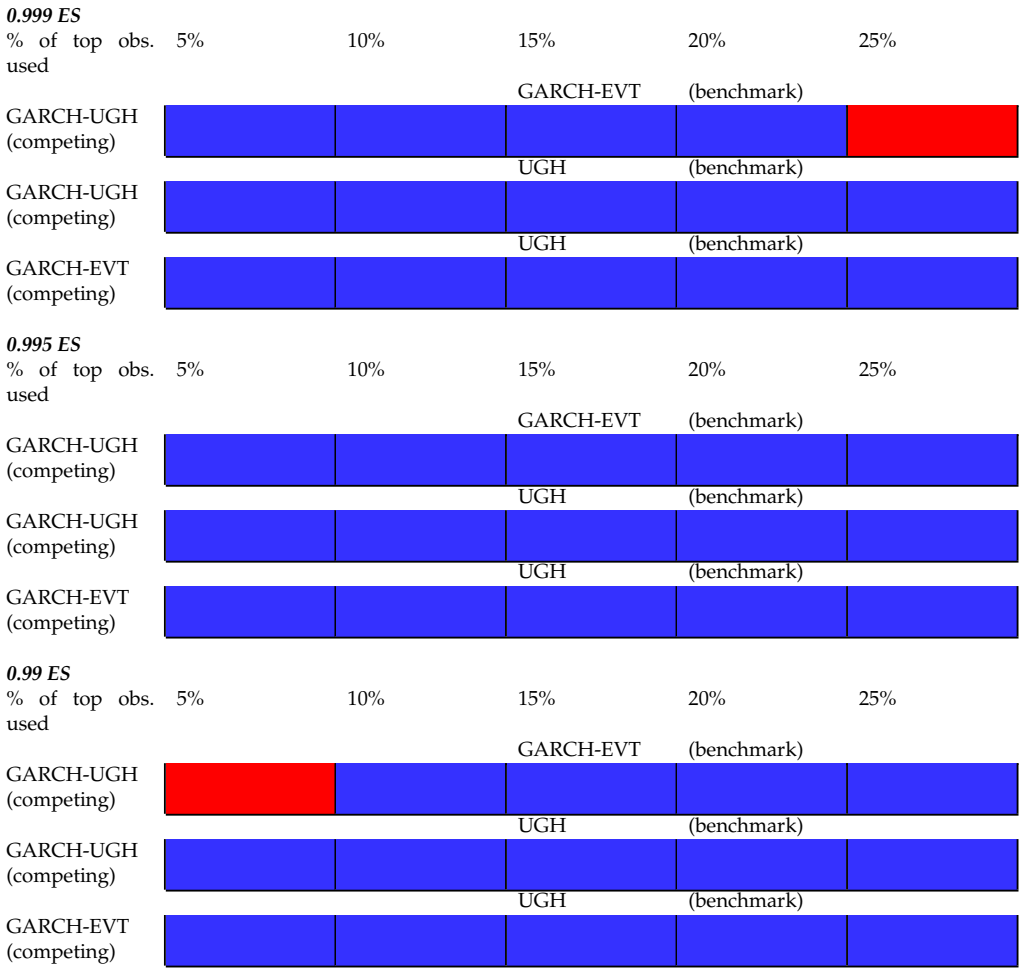
Notes: When the competing model in the row is swapped with the benchmark model in the column, the results are reversed.

TABLE 4.14: Comparative backtesting: out-of-sample evaluations (traffic light matrices) of one-step ahead conditional ES estimates by EVT-type methods from 13 August 1997 to 16 July 2009 at different quantile levels for the negative log-returns of NASDAQ index by means of the Diebold-Mariano test using $h = 0$ (VaR, ES) scoring function.



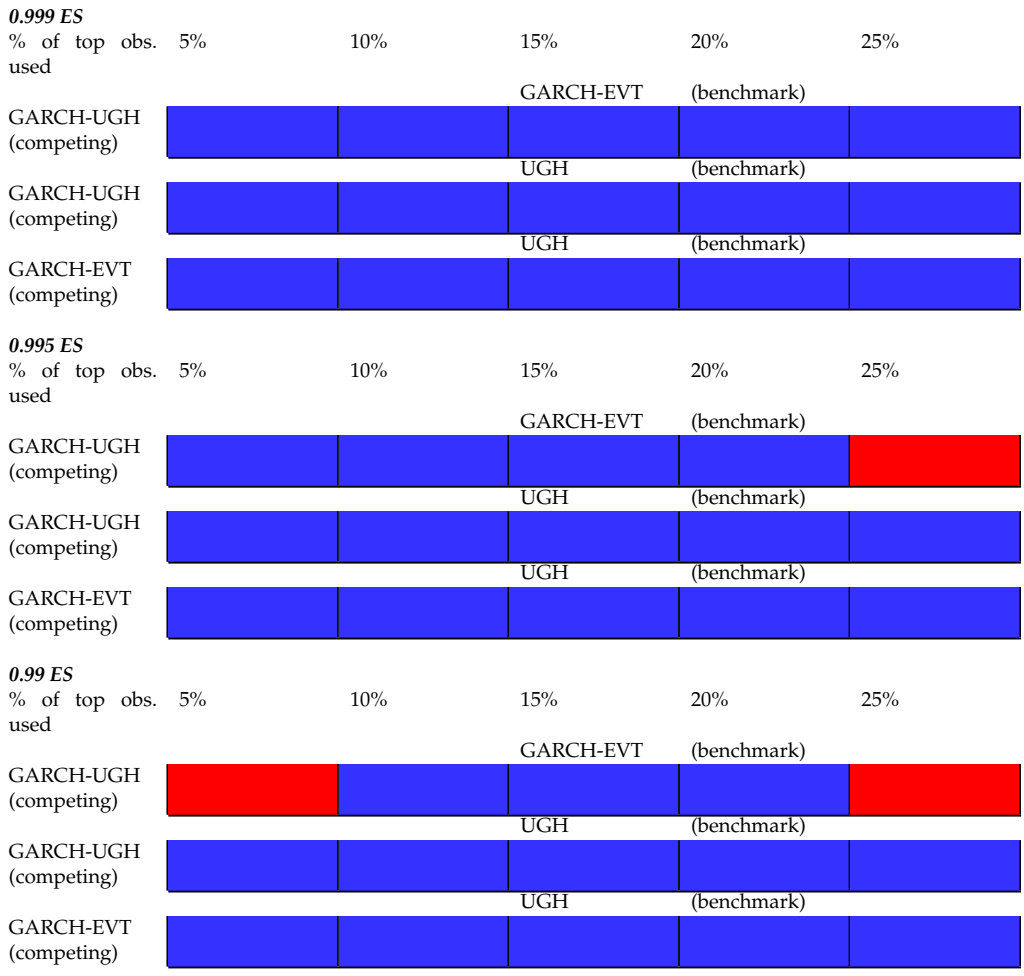
Notes: When the competing model in the row is swapped with the benchmark model in the column, the results are reversed.

TABLE 4.15: Comparative backtesting: out-of-sample evaluations (traffic light matrices) of one-step ahead conditional ES estimates by EVT-type methods from 29 May 1997 to 12 August 2009 at different quantile levels for the negative log-returns of NIKKEI index by means of the Diebold-Mariano test using $h = 0$ (VaR, ES) scoring function.



Notes: When the competing model in the row is swapped with the benchmark model in the column, the results are reversed.

TABLE 4.16: Comparative backtesting: out-of-sample evaluations (traffic light matrices) of one-step ahead conditional ES estimates by EVT-type methods from 28 September 2002 to 14 December 2010 at different quantile levels for the negative log-returns of JPY/GBP exchange rate by means of the Diebold-Mariano test using $h = 0$ (VaR, ES) scoring function.



Notes: When the competing model in the row is swapped with the benchmark model in the column, the results are reversed.

simple CC and simple ER tests, HS fails 6 and 5 times, GARCH-N fails 9 and 11 times, and GARCH-t fails 4 and 5 times out of 12 cases, respectively.

TABLE 4.17: Traditional backtesting: out-of-sample evaluations of one-step ahead conditional ES estimates by basic estimation methods from 8 December 1997 to 9 November 2009 at different quantile levels for the negative log-returns of DJ index.

Testing window	3000						
Estimation window	1000						
0.999 ES	Str. ESR	Aux. ESR	Int. ESR	Gen. CC	Sim. CC	Std. ER	Sim. ER
HS	0.049	0.012	0.012	-	0.166	-	0.096
			(0.024)	-	(0.123)	-	(0.096)
GARCH-N	0.001	0.001	0.001	0.017	0.009	0.001	0.001
			(0.002)	(0.038)	(0.001)	(0.002)	(0.003)
GARCH-t	0.334	0.452	0.688	0.902	0.750	0.000	0.244
			(0.594)	(0.213)	(0.760)	(0.000)	(0.244)
GARCH-UGH	0.030	0.019	0.962	1.000	1.000	0.753	0.753
(25%)			(0.075)	(0.934)	(0.001)	(1.000)	(0.613)
0.995 ES							
HS	0.102	0.061	0.034	-	0.029	-	0.219
			(0.067)	-	(0.002)	-	(0.429)
GARCH-N	0.001	0.001	0.001	0.003	0.002	0.000	0.000
			(0.002)	(0.004)	(0.002)	(0.001)	(0.001)
GARCH-t	0.534	0.499	0.583	1.000	1.000	0.402	0.540
			(0.834)	(0.634)	(0.743)	(0.665)	(0.973)
GARCH-UGH	0.286	0.322	0.811	1.000	1.000	0.750	0.886
(10%)			(0.379)	(0.426)	(0.051)	(0.652)	(0.201)
0.99 ES							
HS	0.006	0.010	0.004	-	0.004	-	0.021
			(0.007)	-	(0.001)	-	(0.063)
GARCH-N	0.001	0.001	0.001	0.001	0.001	0.000	0.000
			(0.002)	(0.002)	(0.001)	(0.000)	(0.000)
GARCH-t	0.362	0.367	0.704	0.675	0.646	0.103	0.219
			(0.593)	(0.290)	(0.708)	(0.227)	(0.426)
GARCH-UGH	0.100	0.154	0.929	1.000	1.000	0.814	0.868
(20%)			(0.141)	(0.168)	(0.007)	(0.388)	(0.220)

Notes: ESR refers to the ES regression tests of Bayer and Dimitriadis (2020b), CC to the conditional calibration tests of Nolde and Ziegel (2017) and ER to the exceedance residuals tests of McNeil and Frey (2000). The p-values for the one-sided test are given with the two-sided test in brackets. The optimal sample fraction for GARCH-UGH is selected based on the performance in traditional out-of-sample backtestings of VaR and ES (see Tables 4.1, 4.2, 4.9 and 4.13).

Comparative ES backtesting

Tables 4.21-4.24 and 4.25-4.28 display the traffic light matrices of comparative ES backtesting (see Section 4.5) for GARCH-UGH with optimal sample fraction and basic estimation methods given in Section 4.3.2, three quantile levels and four financial time series when $h = \frac{1}{2}$ (4.16) and $h = 0$ (4.17) (VaR, ES) scoring functions are used, respectively. The optimal sample fraction is selected based on the performance in the out-of-sample traditional VaR and ES backtestings as explained previously.

TABLE 4.18: Traditional backtesting: out-of-sample evaluations of one-step ahead conditional ES estimates by basic estimation methods from 13 August 1997 to 16 July 2009 at different quantile levels for the negative log-returns of NASDAQ index.

Testing window	3000						
Estimation window	1000						
0.999 ES	Str. ESR	Aux. ESR	Int. ESR	Gen. CC	Sim. CC	Std. ER	Sim. ER
HS	0.082	0.110	0.112	-	0.118	-	0.035
			(0.234)	-	(0.155)	-	(0.035)
GARCH-N	0.039	0.052	0.144	0.110	0.039	0.994	0.005
			(0.287)	(0.025)	(0.044)	(0.006)	(0.005)
GARCH-*t*	0.232	0.204	0.234	0.382	0.138	0.941	0.076
			(0.474)	(0.052)	(0.236)	(0.066)	(0.094)
GARCH-UGH	0.573	0.470	0.697	1.000	1.000	0.642	0.943
(20%)			(0.607)	(0.118)	(0.000)	(1.000)	(0.363)
0.995 ES							
HS	0.155	0.126	0.120	-	0.009	-	0.421
			(0.240)	-	(0.001)	-	(0.768)
GARCH-N	0.085	0.080	0.076	0.042	0.117	0.000	1.000
			(0.151)	(0.089)	(0.012)	(0.001)	(0.000)
GARCH-*t*	0.240	0.239	0.232	0.327	0.909	0.012	0.991
			(0.463)	(0.333)	(0.055)	(0.024)	(0.016)
GARCH-UGH	0.459	0.412	0.606	1.000	0.898	0.761	0.273
(20%)			(0.788)	(0.460)	(0.961)	(0.468)	(0.507)
0.99 ES							
HS	0.005	0.000	0.026	-	0.001	-	0.150
			(0.053)	-	(0.000)	-	(0.313)
GARCH-N	0.033	0.032	0.133	0.191	0.069	1.000	0.000
			(0.266)	(0.016)	(0.135)	(0.002)	(0.005)
GARCH-*t*	0.244	0.250	0.373	0.175	0.367	0.994	0.006
			(0.746)	(0.037)	(0.124)	(0.004)	(0.016)
GARCH-UGH	0.510	0.507	0.825	0.281	1.000	0.651	0.610
(20%)			(0.350)	(0.756)	(0.266)	(0.770)	(0.845)

Notes: ESR refers to the ES regression tests of Bayer and Dimitriadis (2020b), CC to the conditional calibration tests of Nolde and Ziegel (2017) and ER to the exceedance residuals tests of McNeil and Frey (2000). The p-values for the one-sided test are given with the two-sided test in brackets. The optimal sample fraction for GARCH-UGH is selected based on the performance in traditional out-of-sample backtestings of VaR and ES (see Tables 4.3, 4.4, 4.10 and 4.14).

TABLE 4.19: Traditional backtesting: out-of-sample evaluations of one-step ahead conditional ES estimates by basic estimation methods from 29 May 1997 to 12 August 2009 at different quantile levels for the negative log-returns of NIKKEI index.

Testing window	3000						
Estimation window	1000						
0.999 ES	Str. ESR	Aux. ESR	Int. ESR	Gen. CC	Sim. CC	Std. ER	Sim. ER
HS	0.036	0.053	0.047	-	0.114	-	0.025
			(0.097)	-	(0.118)	-	(0.025)
GARCH-N	0.028	0.029	0.017	0.083	0.030	0.984	0.021
			(0.034)	(0.033)	(0.042)	(0.017)	(0.022)
GARCH-*t*	0.603	0.411	0.875	1.000	1.000	0.436	0.631
			(0.232)	(0.880)	(0.491)	(0.895)	(0.743)
GARCH-UGH	0.463	0.582	0.883	1.000	0.978	0.375	0.625
(10%)			(0.224)	(0.772)	(0.853)	(0.865)	(1.000)
0.995 ES							
HS	0.103	0.130	0.037	-	0.042	-	0.048
			(0.073)	-	(0.077)	-	(0.117)
GARCH-N	0.025	0.024	0.012	0.052	0.019	0.999	0.008
			(0.024)	(0.028)	(0.019)	(0.005)	(0.018)
GARCH-*t*	0.342	0.350	0.705	0.992	0.867	0.803	0.268
			(0.590)	(0.415)	(0.856)	(0.401)	(0.573)
GARCH-UGH	0.906	0.855	0.655	1.000	1.000	0.401	0.620
(10%)			(0.690)	(0.899)	(0.760)	(0.918)	(0.865)
0.99 ES							
HS	0.036	0.021	0.006	-	0.016	-	0.065
			(0.012)	-	(0.034)	-	(0.143)
GARCH-N	0.026	0.027	0.021	0.047	0.017	1.000	0.000
			(0.043)	(0.014)	(0.033)	(0.002)	(0.003)
GARCH-*t*	0.545	0.392	0.852	0.002	0.037	0.963	0.029
			(0.296)	(0.157)	(0.063)	(0.096)	(0.075)
GARCH-UGH	0.612	0.659	0.697	1.000	1.000	0.305	0.726
(10%)			(0.606)	(0.658)	(0.506)	(0.694)	(0.598)

Notes: ESR refers to the ES regression tests of Bayer and Dimitriadis (2020b), CC to the conditional calibration tests of Nolde and Ziegel (2017) and ER to the exceedance residuals tests of McNeil and Frey (2000). The p-values for the one-sided test are given with the two-sided test in brackets. The optimal sample fraction for GARCH-UGH is selected based on the performance in traditional out-of-sample backtestings of VaR and ES (see Tables 4.5, 4.6, 4.11 and 4.15).

TABLE 4.20: Traditional backtesting: out-of-sample evaluations of one-step ahead conditional ES estimates by basic estimation methods from 28 September 2002 to 14 December 2010 at different quantile levels for the negative log-returns of JPY/GBP exchange rate.

Testing window	3000						
Estimation window	1000						
0.999 ES	Str. ESR	Aux. ESR	Int. ESR	Gen. CC	Sim. CC	Std. ER	Sim. ER
HS	0.080	0.198	0.098	-	0.148	-	0.022
			(0.198)	-	(0.252)	-	(0.022)
GARCH-N	0.262	0.431	0.315	0.378	0.136	0.972	0.032
			(0.629)	(0.139)	(0.079)	(0.054)	(0.068)
GARCH-*t*	0.000	0.000	0.819	0.000	0.000	0.000	0.000
			(0.299)	(0.380)	(0.000)	(0.000)	(0.000)
GARCH-UGH	0.004	0.002	0.942	1.000	0.752	0.780	0.220
(5%)			(0.116)	(0.452)	(1.000)	(0.477)	(0.477)
0.995 ES							
HS	0.173	0.774	0.196	-	0.198	-	0.223
			(0.393)	-	(0.283)	-	(0.422)
GARCH-N	0.071	0.059	0.022	0.094	0.034	0.998	0.004
			(0.044)	(0.062)	(0.023)	(0.007)	(0.020)
GARCH-*t*	0.000	0.000	0.830	0.000	0.000	0.000	0.000
			(0.341)	(0.317)	(0.000)	(0.000)	(0.000)
GARCH-UGH	0.389	0.344	0.551	1.000	1.000	0.258	0.677
(10%)			(0.889)	(0.203)	(0.190)	(0.538)	(0.689)
0.99 ES							
HS	0.367	0.728	0.169	-	0.163	-	0.483
			(0.338)	-	(0.094)	-	(0.846)
GARCH-N	0.055	0.115	0.016	0.062	0.022	1.000	0.001
			(0.032)	(0.029)	(0.034)	(0.003)	(0.006)
GARCH-*t*	0.000	0.000	0.998	0.000	0.000	1.000	0.000
			(0.044)	(0.943)	(0.000)	(0.000)	(0.000)
GARCH-UGH	0.319	0.222	0.814	0.989	1.000	0.099	0.928
(15%)			(0.372)	(0.006)	(0.002)	(0.116)	(0.086)

Notes: ESR refers to the ES regression tests of Bayer and Dimitriadis (2020b), CC to the conditional calibration tests of Nolde and Ziegel (2017) and ER to the exceedance residuals tests of McNeil and Frey (2000). The p-values for the one-sided test are given with the two-sided test in brackets. The optimal sample fraction for GARCH-UGH is selected based on the performance in traditional out-of-sample backtestings of VaR and ES (see Tables 4.7, 4.8, 4.12 and 4.16).

As with the results of traditional VaR backtestings, it is illustrated that our proposed GARCH-UGH approach appears to be best overall, outperforming three basic estimation methods. The two scoring functions result in a good agreement with GARCH-UGH approach being the better estimator in 10 out of 12 cases when compared to the HS, GARCH-N and GARCH-t approaches. HS, GARCH-N and UGH approaches generally perform worse than the GARCH-UGH approach as they consider neither heavy-tail nor volatility.

TABLE 4.21: Comparative backtesting: out-of-sample evaluations (traffic light matrices) of one-step ahead conditional VaR estimates by basic estimation methods from 8 December 1997 to 9 November 2009 at different quantile levels for the negative log-returns of DJ index by means of the Diebold-Mariano test using $h = \frac{1}{2}$ (VaR, ES) scoring function.

0.999 ES

	HS	GARCH-N	GARCH-t	GARCH-UGH
HS				
GARCH-N				
GARCH-t				
GARCH-UGH (25%)				

0.995 ES

	HS	GARCH-N	GARCH-t	GARCH-UGH
HS				
GARCH-N				
GARCH-t				
GARCH-UGH (10%)				

0.99 ES

	HS	GARCH-N	GARCH-t	GARCH-UGH
HS				
GARCH-N				
GARCH-t				
GARCH-UGH (20%)				

Notes: The optimal sample fraction for GARCH-UGH is selected based on the performance in traditional out-of-sample backtestings of VaR and ES (see Tables 4.1, 4.2, 4.9 and 4.13).

TABLE 4.22: Comparative backtesting: out-of-sample evaluations of one-step ahead conditional VaR estimates by basic estimation methods from 13 August 1997 to 16 July 2009 at different quantile levels for the negative log-returns of NASDAQ index by means of the Diebold-Mariano test using $h = \frac{1}{2}$ (VaR, ES) scoring function.

0.999 ES

	HS	GARCH-N	GARCH-*t*	GARCH-UGH
HS				
GARCH-N				
GARCH-*t*				
GARCH-UGH (20%)				

0.995 ES

	HS	GARCH-N	GARCH-*t*	GARCH-UGH
HS				
GARCH-N				
GARCH-*t*				
GARCH-UGH (20%)				

0.99 ES

	HS	GARCH-N	GARCH-*t*	GARCH-UGH
HS				
GARCH-N				
GARCH-*t*				
GARCH-UGH (20%)				

Notes: The optimal sample fraction for GARCH-UGH is selected based on the performance in traditional out-of-sample backtestings of VaR and ES (see Tables 4.3, 4.4, 4.10 and 4.14).

TABLE 4.23: Comparative backtesting: out-of-sample evaluations of one-step ahead conditional VaR estimates by basic estimation methods from 29 May 1997 to 12 August 2009 at different quantile levels for the negative log-returns of NIKKEI index by means of the Diebold-Mariano test using $h = \frac{1}{2}$ (VaR, ES) scoring function.

0.999 ES

	HS	GARCH-N	GARCH-*t*	GARCH-UGH
HS				
GARCH-N				
GARCH-*t*				
GARCH-UGH (10%)				

0.995 ES

	HS	GARCH-N	GARCH-*t*	GARCH-UGH
HS				
GARCH-N				
GARCH-*t*				
GARCH-UGH (10%)				

0.99 ES

	HS	GARCH-N	GARCH-*t*	GARCH-UGH
HS				
GARCH-N				
GARCH-*t*				
GARCH-UGH (10%)				

Notes: The optimal sample fraction for GARCH-UGH is selected based on the performance in traditional out-of-sample backtestings of VaR and ES (see Tables 4.5, 4.6, 4.11 and 4.15).

TABLE 4.24: Comparative backtesting: out-of-sample evaluations (traffic light matrices) of one-step ahead conditional VaR estimates by basic estimation methods from 28 September 2002 to 14 December 2010 at different quantile levels for the negative log-returns of JPY/GBP exchange rate by means of the Diebold-Mariano test using $h = \frac{1}{2}$ (VaR, ES) scoring function.

0.999 ES

	HS	GARCH-N	GARCH-*t*	GARCH-UGH
HS				
GARCH-N				
GARCH-*t*				
GARCH-UGH (5%)				

0.995 ES

	HS	GARCH-N	GARCH-*t*	GARCH-UGH
HS				
GARCH-N				
GARCH-*t*				
GARCH-UGH (10%)				

0.99 ES

	HS	GARCH-N	GARCH-*t*	GARCH-UGH
HS				
GARCH-N				
GARCH-*t*				
GARCH-UGH (15%)				

Notes: The optimal sample fraction for GARCH-UGH is selected based on the performance in traditional out-of-sample backtestings of VaR and ES (see Tables 4.7, 4.8, 4.12 and 4.16).

TABLE 4.25: Comparative backtesting: out-of-sample evaluations (traffic light matrices) of one-step ahead conditional VaR estimates by basic estimation methods from 8 December 1997 to 9 November 2009 at different quantile levels for the negative log-returns of DJ index by means of the Diebold-Mariano test using $h = 0$ (VaR, ES) scoring function.

0.999 ES

	HS	GARCH-N	GARCH-*t*	GARCH-UGH
HS				
GARCH-N				
GARCH-*t*				
GARCH-UGH (25%)				

0.995 ES

	HS	GARCH-N	GARCH-*t*	GARCH-UGH
HS				
GARCH-N				
GARCH-*t*				
GARCH-UGH (10%)				

0.99 ES

	HS	GARCH-N	GARCH-*t*	GARCH-UGH
HS				
GARCH-N				
GARCH-*t*				
GARCH-UGH (20%)				

Notes: The optimal sample fraction for GARCH-UGH is selected based on the performance in traditional out-of-sample backtestings of VaR and ES (see Tables 4.1, 4.2, 4.9 and 4.13).

TABLE 4.26: Comparative backtesting: out-of-sample evaluations (traffic light matrices) of one-step ahead conditional VaR estimates by basic estimation methods from 13 August 1997 to 16 July 2009 at different quantile levels for the negative log-returns of NASDAQ index by means of the Diebold-Mariano test using $h = 0$ (VaR, ES) scoring function.

0.999 ES

	HS	GARCH-N	GARCH-*t*	GARCH-UGH
HS				
GARCH-N				
GARCH-*t*				
GARCH-UGH (20%)				

0.995 ES

	HS	GARCH-N	GARCH-*t*	GARCH-UGH
HS				
GARCH-N				
GARCH-*t*				
GARCH-UGH (20%)				

0.99 ES

	HS	GARCH-N	GARCH-*t*	GARCH-UGH
HS				
GARCH-N				
GARCH-*t*				
GARCH-UGH (20%)				

Notes: The optimal sample fraction for GARCH-UGH is selected based on the performance in traditional out-of-sample backtestings of VaR and ES (see Tables 4.3, 4.4, 4.10 and 4.14).

TABLE 4.27: Comparative backtesting: out-of-sample evaluations (traffic light matrices) of one-step ahead conditional VaR estimates by basic estimation methods from 29 May 1997 to 12 August 2009 at different quantile levels for the negative log-returns of NIKKEI index by means of the Diebold-Mariano test using $h = 0$ (VaR, ES) scoring function.

0.999 ES

	HS	GARCH-N	GARCH-*t*	GARCH-UGH
HS				
GARCH-N				
GARCH-*t*				
GARCH-UGH (10%)				

0.995 ES

	HS	GARCH-N	GARCH-*t*	GARCH-UGH
HS				
GARCH-N				
GARCH-*t*				
GARCH-UGH (10%)				

0.99 ES

	HS	GARCH-N	GARCH-*t*	GARCH-UGH
HS				
GARCH-N				
GARCH-*t*				
GARCH-UGH (10%)				

Notes: The optimal sample fraction for GARCH-UGH is selected based on the performance in traditional out-of-sample backtestings of VaR and ES (see Tables 4.5, 4.6, 4.11 and 4.15).

TABLE 4.28: Comparative backtesting: out-of-sample evaluations (traffic light matrices) of one-step ahead conditional VaR estimates by basic estimation methods from 28 September 2002 to 14 December 2010 at different quantile levels for the negative log-returns of JPY/GBP exchange rate by means of the Diebold-Mariano test using $h = 0$ (VaR, ES) scoring function.

0.999 ES

	HS	GARCH-N	GARCH-*t*	GARCH-UGH
HS				
GARCH-N				
GARCH-*t*				
GARCH-UGH (5%)				

0.995 ES

	HS	GARCH-N	GARCH-*t*	GARCH-UGH
HS				
GARCH-N				
GARCH-*t*				
GARCH-UGH (10%)				

0.99 ES

	HS	GARCH-N	GARCH-*t*	GARCH-UGH
HS				
GARCH-N				
GARCH-*t*				
GARCH-UGH (15%)				

Notes: The optimal sample fraction for GARCH-UGH is selected based on the performance in traditional out-of-sample backtestings of VaR and ES (see Tables 4.7, 4.8, 4.12 and 4.16).

Chapter 5

Conclusions

5.1 Value-at-Risk (VaR)

In Chapter 3 we introduce an extension of the two-step GARCH-EVT approach from McNeil and Frey (2000) for dynamic extreme VaR estimation, based on a semiparametric bias-reduced extreme quantile estimator from de Haan et al. (2016). This differs from the other papers published in the econometric literature by introducing a finite-sample improvement at the extreme value step, rather than using a more complicated filter than the AR(1)-GARCH(1,1) filter. Estimation of VaR is still needed in practice although Basel Committee on Banking Supervision (2019) changed the risk measure for capital requirements in the internal market risk model from VaR to ES because sensible estimation of ES is based on correctly specified VaR estimates by the definition of ES.

We conclude from our empirical analysis that the proposed GARCH-UGH approach provides better one-step ahead dynamic extreme VaR estimates for financial time series than the benchmark conventional GARCH-EVT approach of McNeil and Frey (2000) and the other basic estimation approaches we have tested based on historical simulation or traditional fully parametric models. This can be seen from both the in-sample and the out-of-sample estimations at several quantile levels τ, including the very high $\tau = 0.999$ corresponding to a 99.9% VaR, and a large range of sample fractions k, due to the effect of the bias correction, from the perspective of in-sample and out-of-sample traditional VaR backtestings. The numerical results of comparative VaR backtesting, which is based on the Diebold-Mariano test, also support the use of the GARCH-UGH approach by yielding definitive answers to the

cases when GARCH-UGH and GARCH-EVT approaches are either all accepted or all rejected in the traditional VaR backtestings. Our finite-sample results also illustrate that the GARCH-UGH method leads to one-step ahead extreme conditional VaR estimates that are less sensitive to the choice of sample fraction, and hence mitigates the difficulty in selecting the optimal number of observations for the estimations. Finally, the computational cost of GARCH-UGH is lower than that of conventional GARCH-EVT: the extreme value step in the GARCH-UGH method is semiparametric with an automatic and fast recipe for the estimations of the one-step ahead extreme conditional VaRs, while the competing GARCH-EVT method is based on a parametric fit of the Generalized Pareto Distribution (GPD) to the residuals using Maximum Likelihood Estimation.

5.2 Expected Shortfall (ES)

In Chapter 4 we extend the GARCH-UGH approach to the dynamic extreme ES estimation by means of the asymptotic equivalence between quantile (VaR) and ES. This is motivated by the fact that there has not been sufficient investigation to establish the superiority of a certain estimator of ES relative to the others in the literature and no particular type of ES model is prescribed in the framework of the BCBS.

We conclude from our empirical analysis that the proposed GARCH-UGH approach provides better one-step ahead dynamic extreme ES estimates for financial time series than other EVT-type and basic estimation methods according to the numerical results of comparative ES backtesting based on the joint elicitability of VaR and ES with the support of traditional ES backtestings that are not always in line with the comparative results. Our results of both traditional VaR and ES backtestings also show that poor VaR estimates induce the imperfect ES estimates because incorrect number of VaR violations can lead to low p-values in the traditional ES backtestings.

In contrast to estimation of dynamic extreme ES where most of the existing models including the ones we referred and proposed for the VaR estimation can easily be adapted to the ES, such adaptions are not straight-forward for backtesting ES estimates. Regarding the backtesting of ES, Basel Committee on Banking Supervision

(2019) still demands financial institutions to use traditional VaR backtesting for ES. At the same time, we can expect that upcoming regulations will require them to backtest ES without using VaR backtesting method. We also tackle an urgent problem of which ES backtesting methods can be used in the practice. Based on the strict definition of backtesting for specific risk measure, i.e., backtesting for ES is only allowed to use estimates of ES and realized returns as input variables, a rejection of the traditional ES tests does not necessarily imply that the ES is misspecified, but the estimates for the input components are misspecified. On the other hand, it is true that the ES is strongly related to the VaR through its definition and joint elicitability, and thus reasonable to backtest both quantities jointly while checking the different aspects of mathematical properties as explained in Section 4.4. With respect to the comparative ES backtesting based on the Diebold-Mariano test, it is very helpful for practitioners to select better performing methods among competing alternatives as traditional ES backtestings do not often yield definitive answers because estimation methods are either all accepted or all rejected and they sometimes give a contradicted decision, i.e., one test rejects the null hypothesis of the underestimation of ES while the others accept. Moreover, the Diebold-Mariano test using joint elicitability of VaR and ES with two hypotheses has strong discrimination ability in terms of guarding models that are not truly better than the benchmark. From a regulatory perspective, it is conservative in the sense than when a new competing model is proposed, it will need strong evidence to overthrow the old benchmark model in favour of new competing model. The drawback of comparative ES backtesting is that as consistent scoring functions of the pair (VaR, ES) are not unique, at present there exists no particular optimal scoring function with any theoretical guarantee and using the wrong scoring function for a specific risk measure may result in biased results.

5.3 Overall conclusion

This thesis considered about quantitative risk management using Extreme Value Theory (EVT). We specifically focused on the use of EVT to study extreme financial market risk, which is the risk of losses arising from movements in market prices,

from a quantitative point of view. The use of EVT in the estimation of risk measures such as VaR and ES was natural and effective because the occurrence of extreme financial market events are increasing in the post 1980s, and commonly used methods, for example, historical simulation and normal distribution, tend to underestimate the risk measures severely.

Although Basel Committee on Banking Supervision (2019) changed the risk measure for capital requirements in the internal market risk model from VaR to ES, estimation of VaR is still needed in practice. This is because financial institutions now face the paradox of using ES for computing their market risk capital requirements and using VaR for backtesting ES. More specifically, the sensible estimation of ES is based on correctly specified VaR estimates by the definition of ES. For this reason, both estimation and backtesting of VaR are still important even now. In addition, we believe that it is necessary to incorporate dynamic changes in the market to reflect the most updated risk level. We therefore proposed a new two-step bias-reduced estimation methodology for the estimation of one-step ahead dynamic extreme VaR, called GARCH-UGH (Unbiased Gomes-de Haan), whereby financial returns are first filtered using an AR-GARCH model, and then a bias-reduced estimator of extreme quantiles is applied to the standardized residuals in Chapter 3. Our results indicate that the GARCH-UGH estimates of the dynamic extreme VaR are more accurate than those obtained either by historical simulation, conventional AR-GARCH filtering with Gaussian or Student-t innovations, or AR-GARCH filtering with standard extreme value estimates, both from the perspective of traditional and comparative VaR backtestings.

With regard to the estimation of ES, there has not been sufficient investigation to establish the superiority of a certain estimator relative to the others in the literature and no particular type of ES model is prescribed in the framework of Basel Committee on Banking Supervision (2019). We thus proposed a novel approach of dynamic extreme ES estimation, which is based on our proposed GARCH-UGH approach and the use of asymptotic equivalence between VaR (quantile) and ES, in Chapter 4. Our results show that the GARCH-UGH approach produces more accurate ES estimates than those obtained by basic estimation methods, both from the perspective of traditional and comparative ES backtestings. When compared to other EVT-type

methods, comparative backtestings with chosen two scoring functions result in a good agreement with the GARCH-UGH approach being the best estimator of ES, while traditional backtestings are not always in line with the superiority of our proposed approach.

Regarding the backtesting of ES, Basel Committee on Banking Supervision (2019) still demands financial institutions to use traditional VaR backtesting for ES. At the same time, we can expect that upcoming regulations will require them to backtest ES without using VaR backtesting methods. We also tackled an urgent problem of which ES backtesting methods can be used in the practice. With reference to the strict definition of backtesting given by Bayer and Dimitriadis (2020b), we understand that a backtesting for specific risk measure should only require its estimates and the realized returns as input variables. In contrast to the VaR, fulfilling this definition for ES is very difficult task because ES is strongly related to the VaR through its definition and joint elicitability. As in every statistical method, each of different ES backtesting methods, which is presented in Chapter 4, has its strengths and weaknesses. We thus strongly suggest adopting a two-stage backtesting framework, i.e., the use of both traditional and comparative backtestings for risk measures that will enhance the regulatory framework for financial institutions by providing the correct incentives for accuracy of risk measure estimates. More precisely, the comparative backtesting methods can be used by financial institutions internally to select better performing methods among competing alternatives.

Use of traditional backtesting only like in the regulatory framework of BCBS might not be sufficient because passing a traditional backtesting, i.e., the null hypothesis of the risk measurement procedure is correct is not rejected, does not mean that one can be sure that the null hypothesis is correct (Nolde and Ziegel 2017). In practice as shown in our empirical analysis, there are cases when traditional backtesting methods do not yield definitive answers because competitive estimation methods are all accepted or all rejected. The comparative backtestings enable us to conduct direct comparisons of estimation methods when traditional backtestings do not work efficiently. In the case of VaR, they have more power than the traditional ones based on assessing the number of VaR violations since the scoring functions consider not only the fact that the VaR violations are occurred but also the

information on the size of the violation. Supplementing with comparative backtestings is essential, and hence can adequately quantify the risks even though they still have some drawbacks to consider for the practical use, e.g., there exists no optimal scoring function with any theoretical guarantee. We think that the major challenge of Basel Committee on Banking Supervision (2019) in the implementation of the ES as a risk measure for market risk is the unavailability of simple tools for its evaluation as explained. We also believe that the findings of the estimation and backtesting of risk measures for tail risks in volatile financial market given in Chapter 3 and 4 would be useful for developing regulatory framework of BCBS and monetary policies aimed at mitigating tail risks.

5.4 Future studies

We highlight five possible directions for further investigations. The first one is that one could replace the AR(1)-GARCH(1,1) filter by a more sophisticated filter. Which filter should be used is not obvious: one could think about replacing the AR(1) part by an ARMA(p, q) part, or the GARCH(1,1) part by a GARCH(p, q) part (or a more complicated asymmetric version), or both (see Section 3.1). This may make it possible to even better account or the volatility dynamics, whose accurate estimation and prediction are key.

The second one is the extension of our GARCH-UGH approach to the estimation of the multiple-step ahead conditional extreme VaR and ES. In many cases, a long-term analysis for the goal of making long-term investment decisions is just as relevant and important as the short-term risk management. This is because certain regulations such as those advocated by BCBS require the estimation of the 10-day ahead VaR at the 99% confidence level, rather than merely the one-step ahead VaR. In Basel II (Basel Committee on Banking Supervision 2009), financial institutions were asked to derive 10-day VaR by a simple square-root-of-time scaling of 1-day VaR in calculating market risk. Similarly, Basel III (Basel Committee on Banking Supervision 2019) proposed to use the the square-root-of-time scaling to calculate multi-day ES from 1-day ES. The difficulty in estimating multi-day VaR and ES is obtaining enough homogeneous data on multi-day returns over non-overlapping

periods. Using overlapping returns would misrepresent the tail behavior of the return distributions leading to significant error in extreme VaR and ES estimates. This motivated the use of scaling law such as square-root-of-time and applying EVI instead of square root (Dacorogna et al. 2001) although it is known that it may lead to severe biases (Christoffersen 1998; Danielsson and Zigrand 2006; Wang et al. 2010; Lönnbark 2016; Novales and Garcia-Jorcano 2019).

Danielsson and de Vries (1997), McNeil and Frey (2000) and also Novales and Garcia-Jorcano (2019) tackle this challenging problem of the multiple-step ahead estimation using a bootstrap methodology, but bootstrapping with heavy tails is known to be very difficult to calibrate, especially in the extreme value setup we considered here. The development of an adaptation of the GARCH-UGH method to the multiple-step ahead setup for extreme VaR and ES estimations, which stays computationally manageable and accurate is well beyond the scope of the current paper. Moreover, it is difficult to construct the formal backtesting methods for multi-step VaR based on counting the number of VaR violations and even harder for ES because we are using overlapping multi-step returns.

The third one is the enhancement of our GARCH-UGH approach for the estimation of dynamic extreme ES. While our approach is the second-order approximation assuming the second-order condition (3.7) on the tail quantile function, we used the first-order asymptotic equivalence between VaR and ES (4.4) to approximate ES from VaR. Instead one could instead use the second-order asymptotic equivalence based on the Equation 4.5, which precisely controls the remainder term neglected in the first-order approximation.

The fourth one is the consideration of scoring function specific to the extreme order instead of the intermediate order versions used in the comparative VaR and ES backtestings (see Equations 3.14 and 4.15). It is interesting to have such scoring functions, which intensely react to the tail risks, as we compared the performance of risk measures based on the estimation of extreme quantiles for EVT-type methods.

The five and final perspective is the estimation of alternative dynamic risk measures as a way of solving the drawbacks of VaR and ES that we highlighted in Sections 1.2.2 and 4.1. One candidate will be the expectile risk measure (see Newey and Powell 1987 for the original definition of expectiles in a regression context given

in Section 1.2.5), which takes into account both the frequency of extreme observations and their magnitude, and is also shown to be a coherent and elicitable risk measure in Ziegel (2016) unlike VaR and ES. Moreover, sample expectiles produce a class of smooth curves as functions of the probability level τ unlike sample quantiles (Daouia et al. 2018). The use of expectiles has recently received substantial attention from the perspective of risk management as an alternative tool for quantifying tail risk (see for example Daouia et al. 2018; Daouia et al. 2020), but the case of dynamic estimation of extreme expectiles in a financial time series context has not been considered yet. The development of a GARCH-UGH-based method for the estimation of dynamic extreme expectiles based on the use of the asymptotic equivalence between quantiles and expectiles (4.3) will thus be an interesting complement to the present thesis. Of course, there exists no universally preferred risk measure: the expectile only has an implicit formulation in general, and is more difficult to interpret than the VaR and ES.

Appendix A

In-sample estimation of second-order parameters

TABLE A.1: In-sample evaluations of one-step ahead conditional VaR estimates from 8 December 1997 to 9 November 2009 at different quantile levels for the negative log-returns of DJ index by means of the number of VaR violations, unconditional and conditional coverage tests.

Testing window	3000				
% of top obs. used	5%	10%	15%	20%	25%
DJ:					
0.999 Quantile					
Expected	3	3	3	3	3
GARCH-UGH	2	2	2	1	1
(Gomes)	(0.538, 0.826)	(0.538, 0.826)	(0.538, 0.826)	(0.179, 0.406)	(0.179, 0.406)
GARCH-UGH	2	2	2	**2**	**2**
($\hat{\rho}_{k_\rho} = -1$)	(0.538, 0.826)	(0.538, 0.826)	(0.538, 0.826)	(0.538, 0.826)	(0.538,0.826)
0.995 Quantile					
Expected	15	15	15	15	15
GARCH-UGH	19	14	14	15	15
(Gomes)	(0.320, 0.541)	(0.793, 0.905)	(0.793, 0.905)	(1.000, 0.927)	(1.000, 0.927)
GARCH-UGH	**15**	21	20	20	23
($\hat{\rho}_{k_\rho} = -1$)	(1.000, 0.927)	(0.143, 0.295)	(0.218, 0.410)	(0.218, 0.410)	(0.055, 0.133)
0.99 Quantile					
Expected	30	30	30	30	30
GARCH-UGH	27	28	29	31	33
(Gomes)	(0.576, 0.669)	(0.711, 0.717)	(0.854, 0.741)	(0.855, 0.711)	(0.588, 0.598)
GARCH-UGH	34	35	35	36	41
($\hat{\rho}_{k_\rho} = -1$)	(0.472, 0.523)	(0.371, 0.444)	(0.371, 0.444)	(0.286, 0.365)	(0.056, 0.091)

Notes: The VaR violations when $\hat{\rho}_{k_\rho} = -1$ performed better, i.e., had the closest number of violations to theoretically expected ones, are highlighted in bold. If we observe that the choice $\hat{\rho}_{k_\rho} = -1$ perform better, then we retain this choice for out-of-sample estimation. The p-values for the unconditional coverage test by Kupiec (1995) and conditional coverage test by Christoffersen (1998) are given in brackets in order.

TABLE A.2: In-sample evaluations of one-step ahead conditional VaR estimates from 13 August 1997 to 16 July 2009 at different quantile levels for the negative log-returns of NASDAQ index by means of the number of VaR violations, unconditional and conditional coverage tests.

Testing window	3000				
% of top obs. used	5%	10%	15%	20%	25%
NASDAQ:					
0.999 Quantile					
Expected	3	3	3	3	3
GARCH-UGH	4	4	4	4	2
(Gomes)	(0.583, 0.855)	(0.583, 0.855)	(0.583, 0.855)	(0.583, 0.855)	(0.538, 0.826)
GARCH-UGH	6	6	5	5	5
($\hat{\rho}_{k_\rho} = -1$)	(0.128, 0.370)	(0.128, 0.370)	(0.292, 0.569)	(0.292, 0.569)	(0.292, 0.569)
0.995 Quantile					
Expected	15	15	15	15	15
GARCH-UGH	14	14	12	11	11
(Gomes)	(0.793, 0.905)	(0.793, 0.905)	(0.421, 0.689)	(0.277, 0.532)	(0.277, 0.532)
GARCH-UGH	20	20	**14**	**14**	**13**
($\hat{\rho}_{k_\rho} = -1$)	(0.218, 0.410)	(0.218, 0.410)	(0.793, 0.905)	(0.793, 0.905)	(0.596, 0.821)
0.99 Quantile					
Expected	30	30	30	30	30
GARCH-UGH	23	23	23	25	25
(Gomes)	(0.180, 0.341)	(0.180, 0.341)	(0.180, 0.341)	(0.345, 0.519)	(0.345, 0.519)
GARCH-UGH	41	41	41	36	36
($\hat{\rho}_{k_\rho} = -1$)	(0.056, 0.091)	(0.056, 0.091)	(0.056, 0.091)	(0.286, 0.427)	(0286, 0.427)

Notes: The VaR violations when $\hat{\rho}_{k_\rho} = -1$ performed better, i.e., had the closest number of violations to theoretically expected ones, are highlighted in bold. If we observe that the choice $\hat{\rho}_{k_\rho} = -1$ perform better, then we retain this choice for out-of-sample estimation. The *p*-values for the unconditional coverage test by Kupiec (1995) and conditional coverage test by Christoffersen (1998) are given in brackets in order.

TABLE A.3: In-sample evaluations of one-step ahead conditional VaR estimates from 29 May 1997 to 12 August 2009 at different quantile levels for the negative log-returns of NIKKEI index by means of the number of VaR violations, unconditional and conditional coverage tests.

Testing window	3000				
% of top obs. used	5%	10%	15%	20%	25%
NIKKEI:					
0.999 Quantile					
Expected	3	3	3	3	3
GARCH-UGH	4	1	4	1	1
(Gomes)	(0.583, 0.885)	(0.179, 0.406)	(0.583, 0.885)	(0.179, 0.406)	(0.179, 0.406)
GARCH-UGH	4	**2**	2	**4**	1
($\hat{\rho}_{k_\rho} = -1$)	(0.583, 0.855)	(0.538, 0.826)	(0.538, 0.826)	(0.583, 0.855)	(0.179, 0.406)
0.995 Quantile					
Expected	15	15	15	15	15
GARCH-UGH	13	13	12	11	9
(Gomes)	(0.596, 0.821)	(0.596, 0.821)	(0.421, 0.689)	(0.277, 0.532)	(0.093, 0.238)
GARCH-UGH	13	13	**13**	**13**	**12**
($\hat{\rho}_{k_\rho} = -1$)	(0.596, 0.821)	(0.596, 0.821)	(0.596, 0.821)	(0.596, 0.821)	(0.421, 0.689)
0.99 Quantile					
Expected	30	30	30	30	30
GARCH-UGH	26	25	26	31	23
(Gomes)	(0.453, 0.601)	(0.345, 0.519)	(0.453, 0.601)	(0.855, 0.711)	(0.180, 0.341)
GARCH-EVT	26	25	34	34	**28**
($\hat{\rho}_{k_\rho} = -1$)	(0.453, 0.601)	(0.345, 0.519)	(0.472, 0.523)	(0.472, 0.523)	(0.711, 0.666)

Notes: The VaR violations when $\hat{\rho}_{k_\rho} = -1$ performed better, i.e., had the closest number of violations to theoretically expected ones, are highlighted in bold. If we observe that the choice $\hat{\rho}_{k_\rho} = -1$ perform better, then we retain this choice for out-of-sample estimation. The p-values for the unconditional coverage test by Kupiec (1995) and conditional coverage test by Christoffersen (1998) are given in brackets in order.

TABLE A.4: In-sample evaluations of one-step ahead conditional VaR estimates from 28 September 2002 to 14 December 2010 at different quantile levels for the negative log-returns of JPY/GBP exchange rate by means of the number of VaR violations, unconditional and conditional coverage tests.

Testing window	3000				
% of top obs. used	5%	10%	15%	20%	25%
JPY/GBP:					
0.999 Quantile					
Expected	3	3	3	3	3
GARCH-UGH	3	2	1	1	1
(Gomes)	(1.000, 0.997)	(0.538, 0.826)	(0.179, 0.406)	(0.179, 0.406)	(0.179, 0.406)
GARCH-UGH	3	2	**3**	**2**	**2**
($\hat{\rho}_{k_\rho} = -1$)	(1.000, 0.997)	(0.538, 0.826)	(1.000, 0.997)	(0.538, 0.826)	(0.538, 0.826)
0.995 Quantile					
Expected	15	15	15	15	15
GARCH-UGH	16	14	14	14	16
(Gomes)	(0.798, 0.888)	(0.793, 0.905)	(0.793, 0.905)	(0.793, 0.905)	(0.798, 0.888)
GARCH-UGH	21	20	18	19	20
($\hat{\rho}_{k_\rho} = -1$)	(0.143, 0.295)	(0.218, 0.410)	(0.452, 0.676)	(0.320, 0.541)	(0.218, 0.410)
0.99 Quantile					
Expected	30	30	30	30	30
GARCH-UGH	31	32	31	29	22
(Gomes)	(0.855, 0.612)	(0.717, 0.609)	(0.855, 0.612)	(0.854, 0.556)	(0.123, 0.259)
GARCH-UGH	42	46	45	47	50
($\hat{\rho}_{k_\rho} = -1$)	(0.038, 0.064)	(0.006, 0.012)	(0.010, 0.019)	(0.004, 0.007)	(0.001, 0.002)

Notes: The VaR violations when $\hat{\rho}_{k_\rho} = -1$ performed better, i.e., had the closest number of violations to theoretically expected ones, are highlighted in bold. If we observe that the choice $\hat{\rho}_{k_\rho} = -1$ perform better, then we retain this choice for out-of-sample estimation. The p-values for the unconditional coverage test by Kupiec (1995) and conditional coverage test by Christoffersen (1998) are given in brackets in order.

Appendix B

Supplementary simulations

B.1 Simulation setup

Our aim in this section is to compare our GARCH-UGH approach for dynamic extreme VaR estimation with other five approaches discussed in Section 3.3 when the GARCH model is misspecified and the tail of the innovations is not explicitly heavy, i.e., assuming normal innovations. The former case is within expectations in the applications but the latter case is beyond the scope of the assumption because assuming heavy-tail is ubiquitous in financial risk management.

To complete this comparison, we consider two data generating processes for simulating the observations used in our simulation study. The GARCH model we used is defined as

$$X_t = \sigma_t Z_t$$

where $\sigma_t > 0$ denote the (conditional) standard deviation and the innovations Z_t is a strict white noise process, that is, they are i.i.d. with zero mean, unit variance and common marginal distribution function F_Z. We assume that for each t, σ_t are measurable with respect to the $\sigma-$algebra $\mathscr{F}_{t-1}$ representing the information about the return process available up to time $t-1$. The sequence X_t follows a GARCH(p,q) process if, for each t,

$$\sigma_t^2 = \alpha_0 + \sum_{j=1}^{p} \alpha_j X_{t-j}^2 + \sum_{l=1}^{q} \beta_l \sigma_{t-l}^2$$

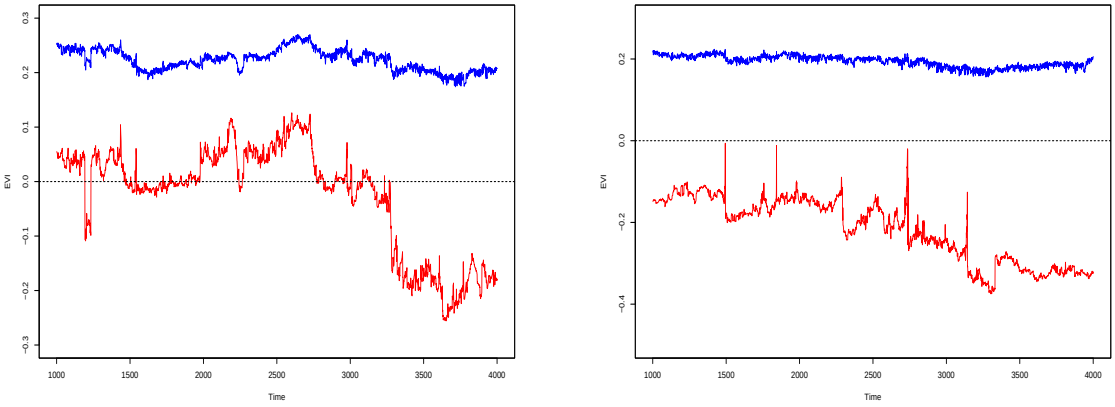
where $\alpha_j, \beta_l > 0$. In our simulation study, we specifically consider the GARCH(1,2) models where all parameters are equal to the estimates when GARCH(1,2) model is fitted to NIKKEI index from 20 February 2008 to 22 October 2021. They are as

follows

- Model 1: GARCH(1,2) model with Student t innovations with 6.775757 degrees of freedom and $\hat{\alpha}_0 = 0.000005$, $\hat{\alpha}_1 = 0.115598$, $\hat{\beta}_1 = 0.863749$ and $\hat{\beta}_2 = 0.000009$.
- Model 2: GARCH(1,2) model with normal innovations and $\hat{\alpha}_0 = 0.000007$, $\hat{\alpha}_1 = 0.131618$, $\hat{\beta}_1 = 0.832505$ and $\hat{\beta}_2 = 0.004905$.

Figure B.1 illustrates the out-of-sample estimation of the Extreme Value Index (EVI) by GARCH-UGH and GARCH-EVT approaches using the top 10% and 15% of observations from rolling estimation window made of 1000 observations for models 1 and 2, respectively. For model 1, the distribution of Z_t after filtering with misspecified AR(1)-GARCH(1,1) model is heavy-tail because the innovations is set as being Student t. The estimates of EVI by GARCH-UGH are stable between 0.2 and 0.3 whereas GARCH-EVT yields fluctuating estimates ranging from 0.1 to -0.3. Note that the heavy-tailed feature of the GARCH models does not depend on whether the innovations follow a heavy-tailed distribution (de Haan et al. 2016). Nonetheless, we used t innovations because empirical studies support using heavy-tailed innovations for modeling financial time series (see for example McNeil and Frey 2000). For model 2, it is not clear whether the tail is explicitly heavy or not by assuming normal innovations. It can be seen from Figure B.1b that estimated value of EVI is decreased but it does not guarantee that the distribution of Z_t possesses non-heavy tail. While GARCH-UGH continues to produce EVI estimates as 0.2, GARCH-EVT now yields $\hat{\gamma} < 0$, which means that it is short-tailed. We use normal distribution as the distribution of innovations to consider the unexpected scenario for financial risk management.

In the next section, we present the results of out-of-sample evaluations of one-step ahead conditional VaR estimates at different τ levels and choices of k by means of traditional and comparative backtestings. The procedure of out-of-sample estimation and backtesting is given in Section 3.6.3.



(a) GARCH(1,2) process with t innovations

(b) GARCH(1,2) process with normal innovations

FIGURE B.1: Out-of-sample estimation of extreme value index (EVI) by GARCH-UGH (blue line) and GARCH-EVT (red line) approaches using the top 10% and 15% of observations from rolling estimation windows made of 1000 observations for the data generated from GARCH(1,2) process with t and normal innovations, respectively.

B.2 Simulation results

Tables B.1 and B.2 gather the numerical results for the comparison between the GARCH-UGH, GARCH-EVT and UGH approaches from the perspective of traditional VaR backtestings. The corresponding plots of out-of-sample backtesting are shown in Figures B.2-B.5. For model 1 with t innovations, it can be seen that our GARCH-UGH approach appears to be best overall. In 11 out of 15 cases, the GARCH-UGH approach yields the closest number of VaR violations to the theoretically expected numbers, while the GARCH-EVT and the unfiltered UGH methods yield the closest ones 6 and 4 times, respectively. All three EVT-type methods never fail both Kupiec and Christoffersen tests. Surprisingly, for model 2 with normal innovations it is shown again that the GARCH-UGH approach still appears to be best overall by means of the number of VaR violations. In 12 out of 15 cases, GARCH-UGH approach yields the closest number of VaR violations, while the UGH methods fares worst. On no occasion GARCH-UGH and GARCH-EVT approaches fail the Christoffersen test, while UGH approach fails 7 times out of 15 cases. Based on the results of traditional VaR backtestings, GARCH-UGH generally performs better than other EVT-type approaches even if the filtering GARCH model is misspecified and the innovations Z_t are normally distributed.

TABLE B.1: Traditional backtesting: out-of-sample evaluations of one-step ahead conditional VaR estimates by EVT-type methods at different quantile levels for the data generated from GARCH(1,2) process with t innovations by means of the number of VaR violations, unconditional and conditional coverage tests.

Testing window	3000				
Estimation window	1000				
% of top obs. used	5%	10%	15%	20%	25%
0.999 Quantile					
Expected	3	3	3	3	3
UGH	5	2	1	2	1
	(0.292, 0.569)	(0.538, 0.826)	(0.179, 0.406)	(0.538, 0.826)	(0.179, 0.406)
GARCH-UGH	**4**	**3**	**3**	**3**	**3**
	(0.583, 0.855)	(1.000, 0.997)	(1.000, 0.997)	(1.000, 0.997)	(1.000, 0.997)
GARCH-EVT	**4**	4	4	4	4
	(0.583, 0.885)	(0.583, 0.885)	(0.583, 0.885)	(0.583, 0.885)	(0.583, 0.855)
0.995 Quantile					
Expected	15	15	15	15	15
UGH	18	**15**	12	10	9
	(0.567, 0.190)	(1.000, 0.927)	(0.421, 0.689)	(0.168, 0.374)	(0.093, 0.238)
GARCH-UGH	**15**	**15**	14	14	**15**
	(1.000, 0.927)	(1.000, 0.927)	(0.793, 0.905)	(0.793, 0.905)	(1.000, 0.927)
GARCH-EVT	**15**	**15**	**15**	**15**	**15**
	(1.000, 0.927)	(1.000, 0.927)	(1.000, 0.927)	(1.000, 0.927)	(1.000, 0.927)
0.99 Quantile					
Expected	30	30	30	30	30
UGH	31	31	**31**	**31**	**27**
	(0.855, 0.612)	(0.855, 0.612)	(0.855, 0.612)	(0.855, 0.612)	(0.576, 0.433)
GARCH-UGH	**30**	**30**	28	32	**33**
	(1.000, 0.738)	(1.000, 0.738)	(0.711, 0.717)	(0.717, 0.663)	(0.588, 0.598)
GARCH-EVT	27	26	27	24	24
	(0.576, 0.669)	(0.453, 0.601)	(0.576, 0.669)	(0.254, 0.430)	(0.254, 0.430)

Notes: The closest numbers of VaR violations to theoretically expected ones are highlighted in bold. The p-values for the unconditional coverage test by Kupiec (1995) and conditional coverage test by Christoffersen (1998) are given in brackets in order.

TABLE B.2: Traditional backtesting: out-of-sample evaluations of one-step ahead conditional VaR estimates by EVT-type methods at different quantile levels for the data generated from GARCH(1,2) process with normal innovations by means of the number of VaR violations, unconditional and conditional coverage tests.

Testing window	3000				
Estimation window	1000				
% of top obs. used	5%	10%	15%	20%	25%
0.999 Quantile					
Expected	3	3	3	3	3
UGH	4	1	1	1	1
	(0.583, 0.855)	(0.179, 0.406)	(0.179, 0.406)	(0.179, 0.406)	(0.179, 0.406)
GARCH-UGH	**3**	**3**	**3**	**3**	**3**
	(1.000, 0.997)	(1.000, 0.997)	(1.000, 0.997)	(1.000, 0.997)	(1.000, 0.997)
GARCH-EVT	**3**	**3**	**3**	4	4
	(1.000, 0.997)	(1.000, 0.997)	(1.000, 0.997)	(0.583, 0.885)	(0.583, 0.855)
0.995 Quantile					
Expected	15	15	15	15	15
UGH	26	23	19	**16**	**12**
	(0.010, 0.000)	(0.055, 0.007)	(0.320, 0.170)	(0.798, 0.888)	(0.421, 0.689)
GARCH-UGH	17	**14**	**14**	9	9
	(0.612, 0.798)	(0.793, 0.905)	(0.793, 0.905)	(0.093, 0.238)	(0.093, 0.238)
GARCH-EVT	**15**	13	13	13	**12**
	(1.000, 0.927)	(0.596, 0.821)	(0.596, 0.821)	(0.596, 0.821)	(0.421, 0.689)
0.99 Quantile					
Expected	30	30	30	30	30
UGH	35	35	37	35	35
	(0.371, 0.000)	(0.371, 0.000)	(0.215, 0.000)	(0.371, 0.000)	(0.371, 0.000)
GARCH-UGH	**33**	**33**	**32**	**29**	**28**
	(0.588, 0.598)	(0.588, 0.598)	(0.717, 0.663)	(0.854, 0.741)	(0.710, 0.717)
GARCH-EVT	**27**	25	24	24	23
	(0.576, 0.669)	(0.345, 0.519)	(0.254, 0.430)	(0.254, 0.430)	(0.180, 0.341)

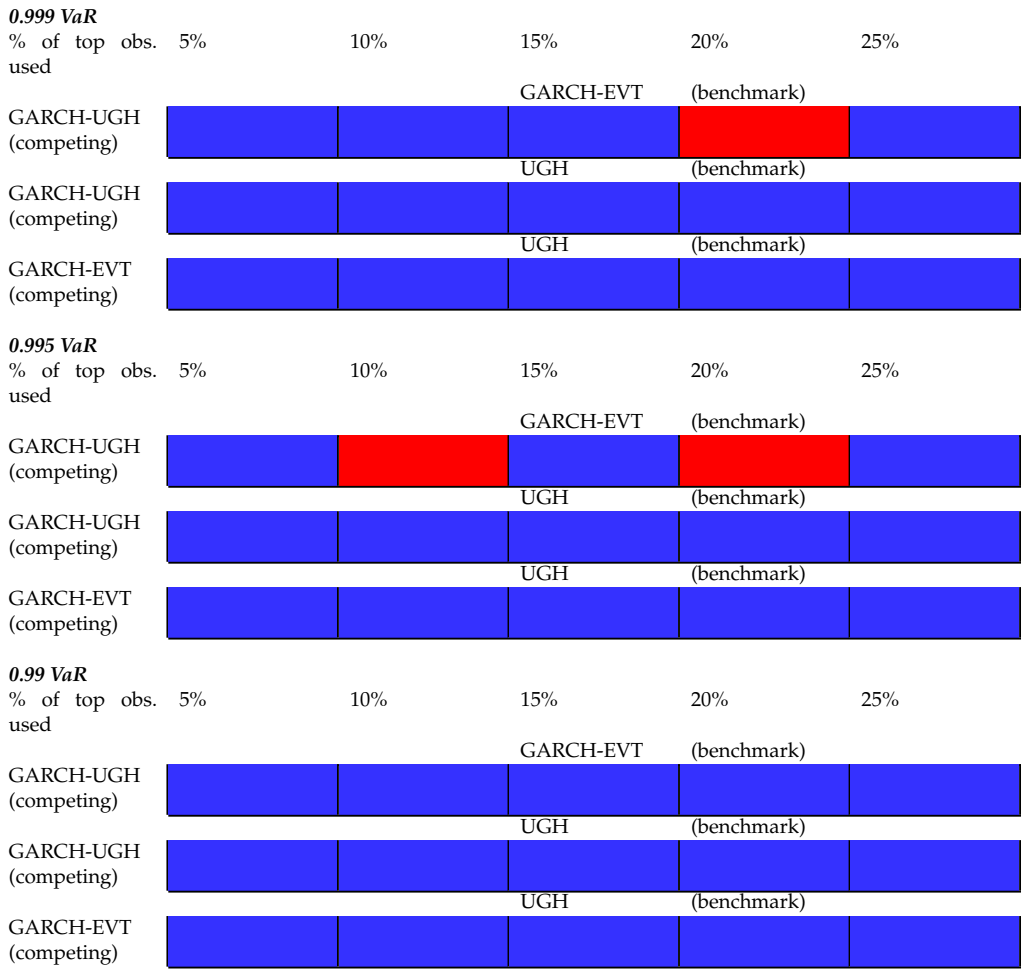
Notes: The closest numbers of VaR violations to theoretically expected ones are highlighted in bold. The p-values for the unconditional coverage test by Kupiec (1995) and conditional coverage test by Christoffersen (1998) are given in brackets in order.

Tables B.3 and B.4 display the traffic light matrices of comparative VaR backtesting given in Section 3.5 for three EVT-type methods and for the data generated from GARCH(1,2) process with t innovations when $h = 0$ (3.16) and $h = 1$ (3.15) VaR scoring functions are used, respectively. Tables B.5 and B.6 display the cases when the innovations are normally distributed. The competing models are given in the vertical axis with the benchmark models along the horizontal axis. Using the t-statistic based on the DM test (3.11), we reject the hypothesis H_0^- at the test level 5% if $1 - \Phi(DM) \leq 0.05$ while the hypothesis H_0^+ is rejected if $\Phi(DM) \leq 0.05$. Under H_0^-, the comparative backtesting is passed for the competing model if the null hypothesis fails to be rejected. On the other hand, under H_0^+ the backtesting for the competing model is passed if the null hypothesis is rejected. The green zone corresponds to the case when H_0^- is not rejected and H_0^+ is rejected, which suggests that the competing model is considered as better than the benchmark model. The yellow zone is when only one of the backtestings under H_0^- and H_0^+ is passed and we cannot conclude which model performs the best. The red zone corresponds to the case when both backtestings fail to be passed, indicating a problem with the competing model.

It is illustrated that our proposed GARCH-UGH approach appears to be best overall for GARCH(1,2) model with t innovations. In 22 out of 30 cases, the GARCH-UGH approach is considered as better than GARCH-EVT approach based on the realized scores of VaR. Comparative backtestings with two scoring functions and traditional backtestings result in a good agreement with the GARCH-UGH approach being the best estimator of VaR, while the unfiltered UGH being the worst estimator, i.e., failing the comparative backtestings against all the other methods. On the other hand, when innovations are normally distributed, GARCH-EVT approach is considered as better than GARCH-UGH approach in 19 out of 30 cases, which contradicts with the results of traditional VaR backtestings. Figure B.2 suggests that the GARCH-UGH approach overestimates the VaR during the low-volatile period although it has the closest number of VaR violations to the theoretically expected numbers. Comparative VaR backtestings rank the VaR estimation methods based on the realized VaR scores. They hence yield definitive answers to the cases when

the estimation methods are all accepted or all rejected in the traditional VaR backtestings, especially when GARCH-UGH and GARCH-EVT approaches have the same number of VaR violations and are indistinguishable.

TABLE B.3: Comparative backtesting: out-of-sample evaluations (traffic light matrices) of one-step ahead conditional VaR estimates by EVT-type methods at different quantile levels for the data generated from GARCH(1,2) process with t innovations by means of the Diebold-Mariano test using $h = 0$ VaR scoring function.

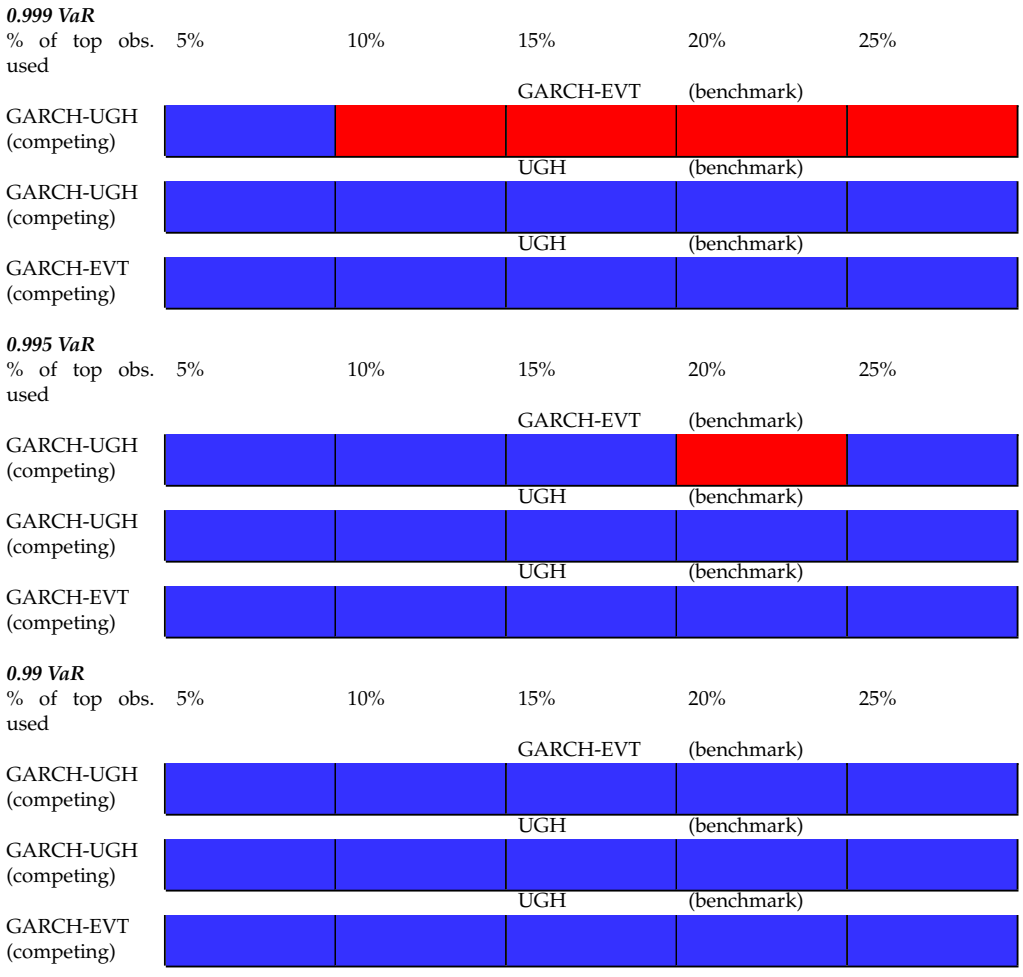


Notes: When the competing model in the row is swapped with the benchmark model in the column, the results are reversed.

Comparison with basic estimation methods

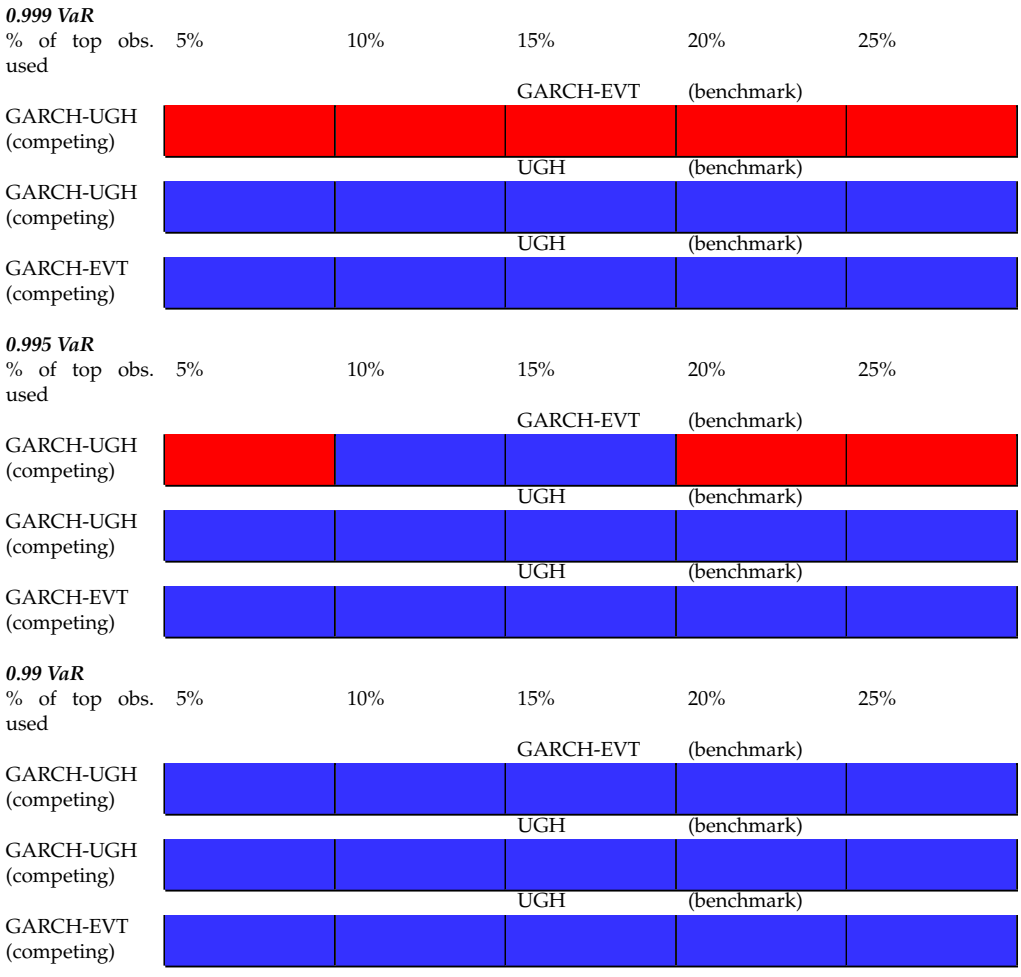
Table B.7 also supports the use of the GARCH-UGH approach for the estimation of dynamic extreme VaR even supposing the GARCH model is misspecified and the innovations Z_t are normally distributed because it outperforms the basic HS, GARCH-N and GARCH-t estimation methods. In 5 out of 6 cases our GARCH-UGH approach (with optimal sample fraction according to Tables B.1-B.2) is closest

TABLE B.4: Comparative backtesting: out-of-sample evaluations (traffic light matrices) of one-step ahead conditional VaR estimates by EVT-type methods at different quantile levels for the data generated from GARCH(1,2) process with t innovations by means of the Diebold-Mariano test using $h = 1$ VaR scoring function.



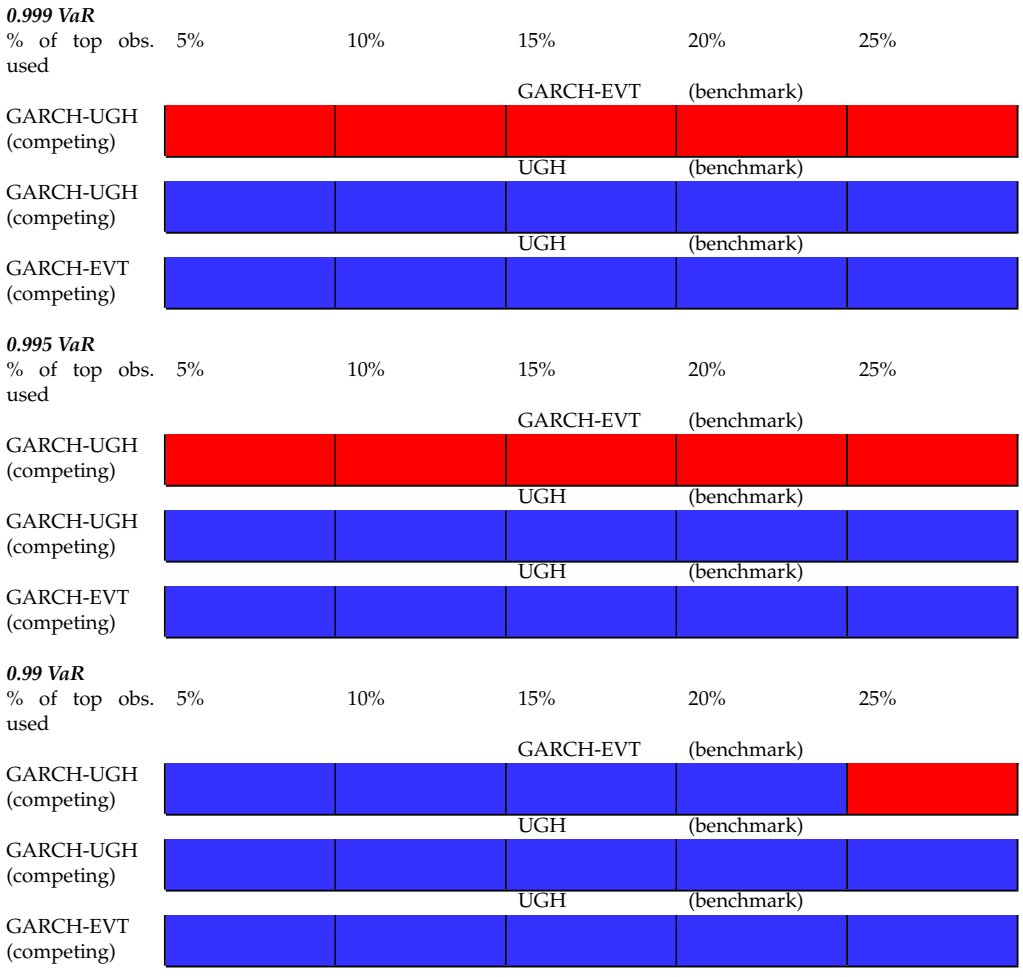
Notes: When the competing model in the row is swapped with the benchmark model in the column, the results are reversed.

TABLE B.5: Comparative backtesting: out-of-sample evaluations (traffic light matrices) of one-step ahead conditional VaR estimates by EVT-type methods at different quantile levels for the data generated from GARCH(1,2) process with normal innovations by means of the Diebold-Mariano test using $h = 0$ VaR scoring function.



Notes: When the competing model in the row is swapped with the benchmark model in the column, the results are reversed.

TABLE B.6: Comparative backtesting: out-of-sample evaluations (traffic light matrices) of one-step ahead conditional VaR estimates by EVT-type methods at different quantile levels for the data generated from GARCH(1,2) process with normal innovations by means of the Diebold-Mariano test using $h = 1$ VaR scoring function.



Notes: When the competing model in the row is swapped with the benchmark model in the column, the results are reversed.

to the mark. In fact, HS performs best in terms of the number of VaR violations for the remaining one case but fails the Chirstoffersen test. GARCH-UGH approach also never fails either of the Kupiec and Christoffersen tests, while HS fails 0 and 1 times, GARCH-N fails 1 and 1 times, and GARCH-*t* fails 1 and 1 times out of 6 cases, respectively.

Tables B.8 and B.9 display the traffic light matrices of comparative VaR backtesting (see Section 3.5) for six estimation methods given in Section 3.3, three quantile levels and the data generated from GARCH(1,2) process with t innovations when $h = 0$ (3.16) and $h = 1$ (3.15) VaR scoring functions are used, respectively. Tables B.10 and B.11 illustrate the cases for normal innovations. The optimal sample fraction for 3 EVT-type methods is selected based on the performance in the out-of-sample traditional VaR backtestings (see Tables B.1-B.2).

As with the results of traditional VaR backtestings, it is illustrated that our proposed GARCH-UGH approach appears to be best overall for the GARCH(1,2) model with t innovations even if the filtering model is misspecified. The two scoring functions result in a good agreement with GARCH-UGH approach being the best estimator when compared to the other five methods because it considers both the volatility change and the heavy-tail of the distribution of the residuals. Inversely, when the innovations are normally distributed, GARCH-UGH is not dominant approach anymore compared to the basic HS, GARCH-N and GARCH-*t* approaches. In this case, GARCH-EVT approach appears to be best overall because it can handle the cases when the tail of distribution of the residuals is not clearly heavy-tail. It fits the generalized Pareto distribution to the residuals for the estimation of EVI and the range of EVI is not limited to $\gamma > 0$.

TABLE B.7: Traditional backtesting: out-of-sample evaluations of one-step ahead conditional VaR estimates by basic estimation methods at different quantile levels for the data generated from GARCH(1,2) process with t and normal innovations by means of the number of VaR violations, unconditional and conditional coverage tests.

	GARCH(1,2) with t innovations	GARCH(1,2) with normal innovations
Testing window	3000	3000
Estimation window	1000	1000
0.999 Quantile		
Expected	3	3
HS	**3**	1
	(1.000, 0.997)	(0.179, 0.406)
GARCH-N	12	2
	(0.000, 0.000)	(0.538, 0.826)
GARCH-t	1	2
	(0.179, 0.406)	(0.538, 0.826)
GARCH-UGH	**3**	**3**
(10%)	(1.000, 0.997)	(1.000, 0.997)
0.995 Quantile		
Expected	15	15
HS	14	21
	(0.793, 0.151)	(0.143, 0.114)
GARCH-N	23	10
	(0.055, 0.133)	(0.168, 0.374)
GARCH-t	9	10
	(0.093, 0.238)	(0.168, 0.374)
GARCH-UGH	**15**	**14**
(5%)	(1.000, 0.927)	(0.793, 0.905)
0.99 Quantile		
Expected	30	30
HS	29	**30**
	(0.854, 0.556)	(1.000, 0.001)
GARCH-N	40	21
	(0.081, 0.127)	(0.081, 0.188)
GARCH-t	13	21
	(0.000, 0.002)	(0.081, 0.188)
GARCH-UGH	**30**	29
(10%)	(1.000, 0.738)	(0.854, 0.741)

Notes: The closest number of VaR violations to the theoretically expected number is highlighted in bold. The number of VaR violations for GARCH-UGH is when the optimal sample fraction is selected according to Tables B.1 and B.2. The p-values for the unconditional coverage test by Kupiec (1995) and conditional coverage test by Christoffersen (1998) are given in brackets in order.

TABLE B.8: Comparative backtesting: out-of-sample evaluations (traffic light matrices) of one-step ahead conditional VaR estimates at different quantile levels for the data generated from GARCH(1,2) process with t innovations by means of the Diebold-Mariano test using $h = 0$ VaR scoring function.

0.999 VaR

	HS	GARCH-N	GARCH-t	UGH	GARCH-EVT	GARCH-UGH
HS						
GARCH-N						
GARCH-t						
UGH (5%)						
GARCH-EVT (10%)						
GARCH-UGH (10%)						

0.995 VaR

	HS	GARCH-N	GARCH-t	UGH	GARCH-EVT	GARCH-UGH
HS						
GARCH-N						
GARCH-t						
UGH (10%)						
GARCH-EVT (10%)						
GARCH-UGH (5%)						

0.99 VaR

	HS	GARCH-N	GARCH-t	UGH	GARCH-EVT	GARCH-UGH
HS						
GARCH-N						
GARCH-t						
UGH (15%)						
GARCH-EVT (15%)						
GARCH-UGH (10%)						

Notes: The optimal sample fraction is selected for 3 EVT-type methods based on the performance in the traditional out-of-sample backtestings (see Table B.1).

TABLE B.9: Comparative backtesting: out-of-sample evaluations (traffic light matrices) of one-step ahead conditional VaR estimates at different quantile levels for the data generated from GARCH(1,2) process with t innovations by means of the Diebold-Mariano test using $h = 1$ VaR scoring function.

0.999 VaR

	HS	GARCH-N	GARCH-t	UGH	GARCH-EVT	GARCH-UGH
HS						
GARCH-N						
GARCH-t						
UGH (5%)						
GARCH-EVT (10%)						
GARCH-UGH (10%)						

0.995 VaR

	HS	GARCH-N	GARCH-t	UGH	GARCH-EVT	GARCH-UGH
HS						
GARCH-N						
GARCH-t						
UGH (10%)						
GARCH-EVT (10%)						
GARCH-UGH (5%)						

0.99 VaR

	HS	GARCH-N	GARCH-t	UGH	GARCH-EVT	GARCH-UGH
HS						
GARCH-N						
GARCH-t						
UGH (15%)						
GARCH-EVT (15%)						
GARCH-UGH (10%)						

Notes: The optimal sample fraction is selected for 3 EVT-type methods based on the performance in the traditional out-of-sample backtestings (see Table B.1).

TABLE B.10: Comparative backtesting: out-of-sample evaluations (traffic light matrices) of one-step ahead conditional VaR estimates at different quantile levels for the data generated from GARCH(1,2) process with normal innovations by means of the Diebold-Mariano test using $h = 0$ VaR scoring function.

0.999 VaR

	HS	GARCH-N	GARCH-*t*	UGH	GARCH-EVT	GARCH-UGH
HS						
GARCH-N						
GARCH-*t*						
UGH (5%)						
GARCH-EVT (5%)						
GARCH-UGH (5%)						

0.995 VaR

	HS	GARCH-N	GARCH-*t*	UGH	GARCH-EVT	GARCH-UGH
HS						
GARCH-N						
GARCH-*t*						
UGH (20%)						
GARCH-EVT (5%)						
GARCH-UGH (10%)						

0.99 VaR

	HS	GARCH-N	GARCH-*t*	UGH	GARCH-EVT	GARCH-UGH
HS						
GARCH-N						
GARCH-*t*						
UGH (20%)						
GARCH-EVT (5%)						
GARCH-UGH (20%)						

Notes: The optimal sample fraction is selected for 3 EVT-type methods based on the performance in the traditional out-of-sample backtestings (see Table B.2).

TABLE B.11: Comparative backtesting: out-of-sample evaluations (traffic light matrices) of one-step ahead conditional VaR estimates at different quantile levels for the data generated from GARCH(1,2) process with normal innovations by means of the Diebold-Mariano test using $h = 1$ VaR scoring function.

0.999 VaR

	HS	GARCH-N	GARCH-*t*	UGH	GARCH-EVT	GARCH-UGH
HS						
GARCH-N						
GARCH-*t*						
UGH (5%)						
GARCH-EVT (5%)						
GARCH-UGH (5%)						

0.995 VaR

	HS	GARCH-N	GARCH-*t*	UGH	GARCH-EVT	GARCH-UGH
HS						
GARCH-N						
GARCH-*t*						
UGH (20%)						
GARCH-EVT (5%)						
GARCH-UGH (10%)						

0.99 VaR

	HS	GARCH-N	GARCH-*t*	UGH	GARCH-EVT	GARCH-UGH
HS						
GARCH-N						
GARCH-*t*						
UGH (20%)						
GARCH-EVT (5%)						
GARCH-UGH (20%)						

Notes: The optimal sample fraction is selected for 3 EVT-type methods based on the performance in the traditional out-of-sample backtestings (see Table B.2).

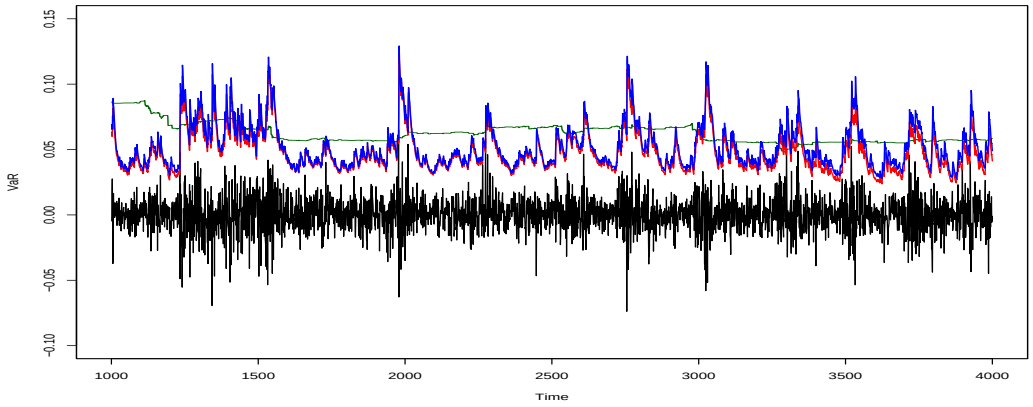


FIGURE B.2: Out-of-sample backtesting of the data generated from GARCH(1,2) process with t innovations and 99.9%-VaR estimates calculated using rolling estimation windows made of 1000 observations, with k corresponding to the top 10% observations from this window. GARCH-UGH (blue line), GARCH-EVT (red line) and UGH (dark green line) estimates are superimposed on the simulated process (black line).

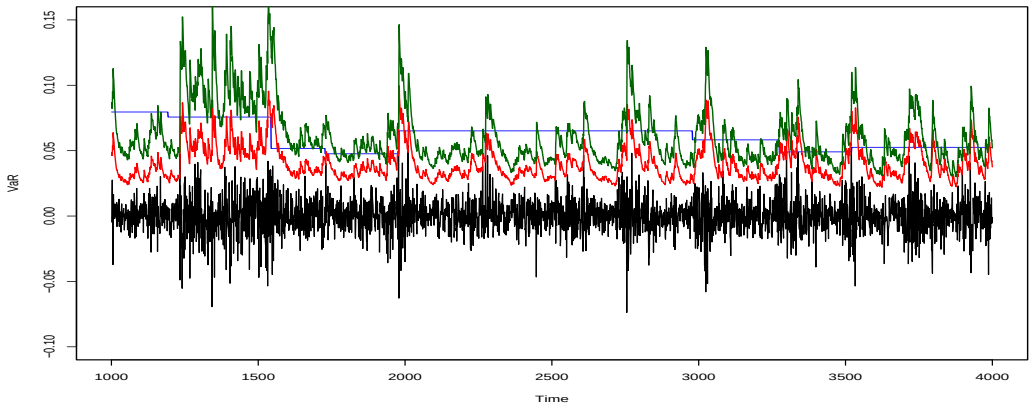


FIGURE B.3: Out-of-sample backtesting of the data generated from GARCH(1,2) process with t innovations and 99.9%-VaR estimates by HS (blue line), GARCH-N (red line) and GARCH-t (dark green line) calculated using rolling estimation windows made of 1000 observations, which are superimposed on the simulated process (black line).

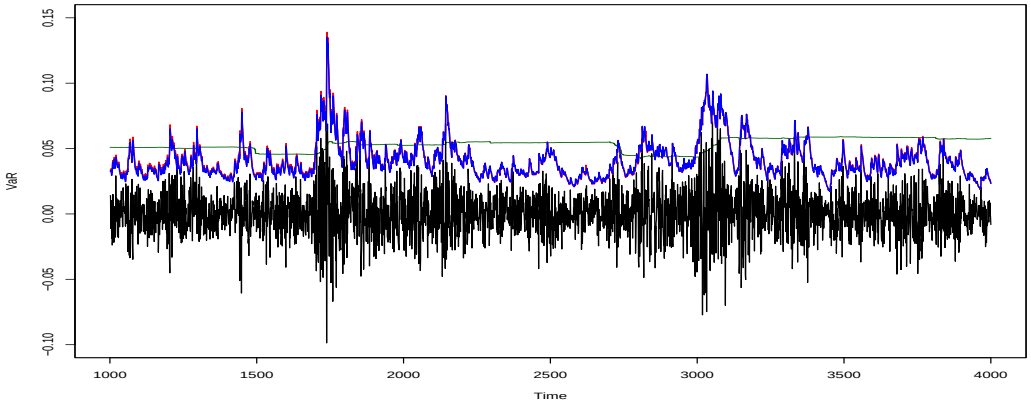


FIGURE B.4: Out-of-sample backtesting of the data generated from GARCH(1,2) process with normal innovations and 99.5%-VaR estimates calculated using rolling estimation windows made of 1000 observations, with k corresponding to the top 10% observations from this window. GARCH-UGH (blue line), GARCH-EVT (red line) and UGH (dark green line) estimates are superimposed on the simulated process (black line).

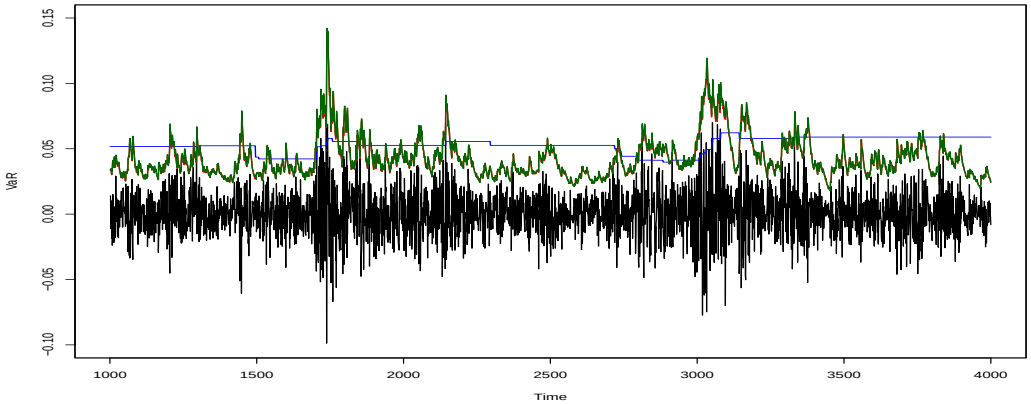


FIGURE B.5: Out-of-sample backtesting of the data generated from GARCH(1,2) process with normal innovations and 99.5%-VaR estimates by HS (blue line), GARCH-N (red line) and GARCH-t (dark green line) calculated using rolling estimation windows made of 1000 observations, which are superimposed on the simulated process (black line).

Bibliography

Abuzayed, B. et al. (2021). "Systemic risk spillover across global and country stock markets during the COVID-19 pandemic". In: *Economic Analysis and Policy* 71, pp. 180–197.

Acerbi, C. and B. Szekely (2014). *Backtesting Expected Shortfall*. https://www.msci.com/documents/10199/22aa9922-f874-4060-b77a-0f0e267a489b. Accessed: 2022-3-14.

— (2017). *General properties of backtestable statistics*. Available at SSRN: https://ssrn.com/abstract=2905109 or https://dx.doi.org/10.2139/ssrn.2905109.

Andrews, D.W.K. (1991). "Heteroskedasticity and autocorrelation consistent covariance matrix estimation". In: *Econometrica* 59.3, pp. 817–858.

Artzner, P. et al. (1999). "Coherent Measures of Risk". In: *Mathematical Finance* 9, pp. 203–228.

Barone-Adesi, G., K. Giannopoulos, and L. Vosper (2002). "Backtesting Derivative Portfolios with Filtered Historical Simulation (FHS)". In: *European Financial Management* 8.1, pp. 31–58.

Basel Committee on Banking Supervision (2009). *Revisions to the Basel II market risk framework*. https://www.bis.org/publ/bcbs148.pdf. Accessed: 2020-10-14.

— (2013). *Fundamental review of the trading book: A revised market risk framework*. https://www.bis.org/publ/bcbs265.pdf. Accessed: 2020-10-14.

— (2019). *Minimum captial requirements for market risk*. https://www.bis.org/bcbs/publ/d457.pdf. Accessed: 2022-3-28.

Bayer, S. and T. Dimitriadis (2020a). *esback: Expected Shortfall Backtesting*. R package version 0.3.0. URL: https://CRAN.R-project.org/package=esback.

— (2020b). "Regression-Based Expected Shortfall Backtesting". In: *Journal of Financial Econometrics*. Available at https://doi.org/10.1093/jjfinec/nbaa013.

Bee, M., D. J. Dupuis, and L. Trapin (2016). "Realizing the extremes: Estimation of tail-risk measures from a high-frequency perspective". In: *Journal of Empirical Finance* 36, pp. 86–99.

Beirlant, J. et al. (2004). *Statistics of Extremes: Theory and Applications*. Wiley.

Bellini, F. and E. Di Bernardino (2015). "Risk management with expectiles". In: *The European Journal of Finance* 23.6, pp. 487–506.

Bellini, F., I. Negri, and M. Pyatkova (2019). "Backtesting VaR and expectiles with realized scores". In: *Statistical Methods and Applications* 28, pp. 119–142.

Bollerslev, T. (1986). "Generalized autoregressive conditional heteroskedasticity". In: *Journal of Econometrics* 31, pp. 307–327.

Byström, H. N. E. (2004). "Managing extreme risks in tranquil and volatile markets using conditional extreme value theory". In: *International Review of Financial Analysis* 13, pp. 133–152.

Caeiro, F. and M.I. Gomes (2010). "An asympototically unbiased moment estimator of a negative extreme value index". In: *Discussiones Mathematicae Probability and Statistics* 30.1, pp. 5–19.

Cai, J.-J., L. de Haan, and C. Zhou (2013). "Bias correction in extreme value statistics with index around zero". In: *Extremes* 16, pp. 173–201.

Chakraborty, G., G.R. Chandrashekhar, and G. Balasubramanian (2021). "Measurement of extreme market risk: Insights from a comprehensive literature review". In: *Cogent Economics and Finance* 9.1.

Chavez-Demoulin, V. and A. Guillou (2018). "Extreme quantile estimator for β-mixing time series and applications". In: *Insurance: Mathematics and Economics* 83, pp. 59–74.

Chavez-Demoulin, V., A. C. Davison, and A. J. McNeil (2005). "Estimating Value-at-Risk: A point process approach". In: *Quantitative Finance* 5, pp. 227–234.

Christoffersen, P. F. (1998). "Evaluating interval forecasts". In: *International Economic Review* 39.4, pp. 841–862.

Coles, S. (2001). *An introduction to statistical modelling of extreme values*. Springer.

Cont, R. (2001). "Empirical properties of asset returns: stylized facts and statistical issues". In: *Quantitative Finance* 1.2, pp. 223–236.

Cont, R., R. Deguest, and G. Scandolo (2010). "Robustness and sensitivity analysis of risk measurement procedures". In: *Quantitative Finance* 10.6, pp. 593–606.

Corsi, F. (2009). "A Simple Approximate Long-Memory Model of Realized Volatility". In: *Journal of Financial Econometrics* 7.2, pp. 174–196.

Costanzino, N. and M. Curran (2015). *Backtesting General Spectral Risk Measures with Application to Expected Shortfall*. Available at SSRN: https://ssrn.com/abstract=2514403 or https://dx.doi.org/10.2139/ssrn.2514403.

Dacorogna, M. et al. (2001). "Extremal Forex Returns in Extremely Large Data Sets". In: *Extremes* 4.2, pp. 105–127.

Danielsson, J. (2011). *Financial Risk Forecasting: The Theory and Practice of Forecasting Market Risk with Implementation in R and Matlab*. Wiley.

Danielsson, J. and C. de Vries (1997). *Value-at-Risk and Extreme Returns*. FMG-Discussion Paper NO 273, Financial Markets Group, London School of Economics.

Danielsson, J. and Y. Morimoto (2000). "Forecasting Extreme Financial Risk: A Critical Analysis of Practical Methods for the Japanese Market". In: *Discussion paper no. 2000-E-8, Institute for Monetary and Econometric Studies, Bank of Japan*.

Danielsson, J. and J-P. Zigrand (2006). "On time-scaling of risk and the square-root-of-time rule". In: *Journal of Banking and Finance* 30.10, pp. 2701–2713.

Daouia, A., S. Girard, and G. Stupfler (2018). "Estimation of tail risk based on extreme expectiles". In: *Journal of the Royal Statistical Society: Series B* 80, pp. 263–292.

— (2020). "Tail expectile process and risk assessment". In: *Bernoulli* 26.1, pp. 531–556.

— (2021). "ExpectHill estimation, extreme risk and heavy tails". In: *Journal of Econometrics* 221.1, pp. 97–117.

de Haan, L. and A. Ferreira (2006). *Extreme Value Theory An Introduction*. Springer Series in Operations Research and Financial Engineering. New York: Springer.

de Haan, L., C. Mercadier, and C. Zhou (2016). "Adapting extreme value statistics to financial time series: dealing with bias and serial dependence". In: *Finance and Stochastics* 20, pp. 321–354.

Degiannakis, S. and A. Potamia (2017). "Multiple-days-ahead value-at-risk and expected shortfall forecasting for stock indices, commodities and exchange rates:

Inter-days versus intra-day data". In: *International Review of Financial Analysis* 49, pp. 176–190.

Dekkers, A.L.M., J.H.J. Einmahl, and L. de Haan (1989). "A Moment Estimator for the Index of an Extreme-Value Distribution". In: *The Annals of Statistics* 17.4, pp. 1833–1855.

Deng, K. and J. Qiu (2021). "Backtesting expected shortfall and beyond". In: *Quantitative Finance* 21.7, pp. 1109–1125.

Diebold, F. X., T. Schuermann, and J. D. Stroughair (2000). "Pitfalls and Opportunities in the Use of Extreme Value Theory in Risk Management". In: *The Journal of Risk Finance* 1, pp. 30–35.

Diebold, F.X. and R.S. Mariano (1995). "Comparing Predictive Accuracy". In: *Journal of Business and Economic Statistics* 13.3, pp. 253–263.

Drees, H. (1995). "Refined Pickands Estimators of the Extreme Value Index". In: *The Annals of Statistics* 23.6, pp. 2059–2080.

— (2003). "Extreme quantile estimation for dependent data, with applications to finance". In: *Bernoulli* 9, pp. 617–657.

Du, Z. and J.C. Escanciano (2015). *Backtesting Expected Shortfall: Accounting for Tail Risk*. Available at SSRN: https://ssrn.com/abstract=2548544 or https://dx.doi.org/10.2139/ssrn.2548544.

Echaust, K. and M. Just (2020). "Value at Risk Estimation Using the GARCH-EVT Approach with Optimal Tail Selection". In: *Mathematics* 8, p. 114.

Embrechts, P., C. Klüppelberg, and T. Mikosch (1997). *Modelling Extremal Events for Insurance and Finance*. Springer.

Emmer, S., M. Kratz, and D. Tasche (2015). "What is the best risk measure in practice? A comparison of standard measures". In: *Journal of Risk* 18.2, pp. 31–60.

Engle, R.F. and S. Manganelli (2004). "CAViaR: Conditional Autoregressive Value at Risk by Regression Quantiles". In: *Journal of Business and Economic Statistics* 22.4, pp. 367–381.

Ergen, I. (2015). "Two-step methods in VaR prediction and the importance of fat tails". In: *Quantitative Finance* 15.6, pp. 1013–1030.

Fernandez, V. (2005). "Risk management under extreme events". In: *International Review of Financial Analysis* 14, pp. 113–148.

Fissler, T. and J.F. Ziegel (2016). "Higher order elicitability and Osband's principle". In: *The Annals of Statistics* 44.4, pp. 1680–1707.

— (2021). "On the elicitability of range value at risk". In: *Statistics and Risk Modeling* 38.1-2, pp. 25–46.

Fissler, T., J.F. Ziegel, and T. Gneiting (2015). "Expected Shortfall is jointly elicitable with Value at Risk - Implications for backtesting". In: *Risk*. https://doi.org/10.48550/arXiv.1507.00244.

Foss, S., D. Korshunov, and S. Zachary (2013). *An Introduction to Heavy-Tailed and Subexponential Distributions*. Springer.

Fraga Alves, M.I., M.I. Gomes, and L. de Haan (2003). "A New Class of Semi-Parametric Estimators of the Second Order Parameter". In: *PORTUGALIAE MATHEMATICA* 60.2, pp. 193–213.

Francq, C. and J.-M. Zakoïan (2004). "Maximum Likelihood Estimation of Pure GARCH and ARMA-GARCH Processes". In: *Bernoulli* 10.4, pp. 605–637.

— (2010). *GARCH Models: Structure, Statistical Inference and Financial Applications*. John Wiley & Sons.

Furió, D. and F. J. Climent (2013). "Extreme value theory versus traditional GARCH approaches applied to financial data: a comparative evaluation". In: *Quantitative Finance* 13.1, pp. 45–63.

Galanos, A. and T. Kley (2022). *rugarch: Univariate GARCH Models*. R package version 1.4-7. URL: https://CRAN.R-project.org/package=rugarch.

Gneiting, T. (2011). "Making and evaluating point forecasts". In: *Journal of the American Statistical Association* 106.494, pp. 746–762.

Goegebeur, Y., J. Beirlant, and T. de Wet (2010). "Kernel estimators for the second order parameter in extreme value statistics". In: *Journal of Statistical Planning and Inference* 140, pp. 2632–2652.

Gomes, M. I., M. J. Martins, and M. Neves (2000). "Alternatives to a Semi-Parametric Estimator of Parameters of Rare Events - The Jacknife Methodology". In: *Extremes* 3.3, pp. 207–229.

Gomes, M. I., D. Pestana, and F. Caeiro (2009). "A note on the asymptotic variance at optimal levels of a bias-corrected Hill estimator". In: *Statistics & Probability Letters* 79, pp. 295–303.

Gomes, M.I. and M.J. Martins (2002). “"Asymptotically Unbiased" Estimators of the Tail Index Based on External Estimation of the Second Order Parameter”. In: *Extremes* 5, pp. 5–31.

Gomes, M.I., L. de Haan, and L. Peng (2002). “Semi-parametric Estimation of the Second Order Parameter in Statistics of Extremes”. In: *Extremes* 5, pp. 387–414.

Hall, P. and A.H. Welsh (1985). “Adaptive Estimates of Parameters of Regular Variation”. In: *The Annals of Statistics* 13.1, pp. 331–341.

He, X.D., S. Kou, and X. Peng (2022). “Risk Measures: Robustness, Elicitability, and Backtesting”. In: *Annual Review of Statistics and Its Applications* 9, pp. 141–166.

Hendricks, D. (1996). “Evaluation of Value-at-Risk Models Using Historical Data”. In: *Economic Policy Review* 2.1, pp. 39–70.

Hill, B. M. (1975). “A simple general approach to inference about the tail of a distribution”. In: *The Annals of Statistics* 3, pp. 1163–1174.

Hill, J. B. (2015). “Tail index estimation for a filtered dependent time series”. In: *Statistica Sinica* 25.2, pp. 609–629.

Hofert, M. and K. Hornik (2016). *qrmdata: Data Sets for Quantitative Risk Management Practice*. R package version 2016-01-03-1. URL: https://CRAN.R-project.org/package=qrmdata.

Holzmann, H. and M. Eulert (2014). “The role of the information set for forecasting—with applications to risk management”. In: *The Annals of Applied Statistics* 8.1, pp. 595–621.

Hosking, J.R.M. and J.R. Wallis (1987). “Parameter and Quantile Estimation for the Generalized Pareto Distribution”. In: *Technometrics* 29.3, pp. 339–349.

Jalal, A. and M. Rockinger (2008). “Predicting tail-related risk measures: The consequences of using GARCH filters for non-GARCH data”. In: *Journal of Empirical Finance* 15, pp. 868–877.

Kaibuchi, H., Y. Kawasaki, and G. Stupfler (2022). “GARCH-UGH: a bias-reduced approach for dynamic extreme Value-at-Risk estimation in financial time series”. In: *Quantitative Finance* 22.7, pp. 1277–1294.

Karmakar, M. and S. Paul (2019). “Intraday portfolio risk management using VaR and CVaR: A CGARCH-EVT-Copula approach”. In: *International Journal of Forecasting* 35, pp. 699–709.

Koenker, R. and G.W. Bassett (1978). "Regression Quantiles". In: *Econometrica* 46.1, pp. 33–50.

Kou, S., X. Peng, and C.C. Heyde (2013). "External Risk Measures and Basel Accords". In: *Mathematics of Operations Research* 38.3, pp. 393–417.

Kratz, M., Y.H. Lok, and A.J. McNeil (2018). "Multinomial VaR Backtests: A simple implicit approach to backtesting expected shortfall". In: *Journal of Banking and Finance* 88, pp. 393–407.

Kupiec, P. (1995). "Techniques for verifying the accuracy of risk management models". In: *Journal of Derivatives* 3, pp. 73–84.

Li, M., C. Lai, and L. Xiao (2022). "Did COVID-19 increase equity market risk exposure? Evidence from China, the UK, and the US". In: *Applied Economics Letters* 29.6, pp. 567–571.

Lönnbark, C. (2016). "Approximation methods for multiple period Value at Risk and Expected Shortfall prediction". In: *Quantitative Finance* 16.6, pp. 947–968.

Löser, R., D. Wied, and D. Ziggel (2018). "New backtests for unconditional coverage of expected shortfall". In: *Journal of Risk* 21.4, pp. 1–21.

Mason, D.M. (1982). "Laws of Large Numbers for Sums of Extreme Values". In: *The Annals of Probability* 10.3, pp. 754–764.

McNeil, A. J. and R. Frey (2000). "Estimation of tail-related risk measures for heteroscedastic financial time series: an extreme value approach". In: *Journal of Empirical Finance* 7, pp. 271–300.

McNeil, A. J., R. Frey, and P. Embrechts (2015). *Quantitative Risk Management: Concepts, Techniques and Tools - Revised Edition*. Princeton University Press.

Mincer, J. and V. Zarnowitz (1969). *The Evaluation of Economic Forecasts*. URL : http : //www.nber.org/chapters/c1214. National Bureau of Economic Research.

Newey, W. K. and J. L. Powell (1987). "Asymmetric Least Squares Estimation and Testing". In: *Econometrica* 55.4, pp. 819–847.

Newey, W. K. and K.D. West (1994). "Automatic Lag Selection in Covariance Matrix Estimation". In: *Review of Economic Studies* 61, pp. 631–653.

Nolde, N. and J.F. Ziegel (2017). "Elicitability and backtesting: Perspectives for banking regulation". In: *The Annals of Applied Statistics* 11.4, pp. 1833–1874.

Novales, A. and L. Garcia-Jorcano (2019). "Backtesting extreme value theory models of expected shortfall". In: *Quantitative Finance* 19.5, pp. 799–825.

Orlowski, L.T. (2012). "Financial crisis and extreme market risks: Evidence from Europe". In: *Review of Financial Economics* 21.3, pp. 120–130.

Patton, A.J., J.F. Ziegel, and R. Chen (2019). "Dynamic semiparametric models for expected shortfall (and Value-at-Risk)". In: *Journal of Econometrics* 211.2, pp. 388–413.

Peng, L. (1998). "Asymptotically unbiased estimators for the extreme-value index". In: *Statistics and Probability Letters* 38.2, pp. 107–115.

Pickands, J. (1975). "Statistical Inference Using Extreme Order Statistics". In: *The Annals of Statistics* 3.1, pp. 119–131.

Reiss, R.-D. and M. Thomas (2007). *Statistical Analysis of Extreme Values*. Birkhäuser Verlag.

Resnick, S. (2007). *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling*. Springer.

Righi, M. and P.S. Ceretta (2013). "Individual and Flexible Expected Shortfall Backtesting". In: *Journal of Risk Model Validation* 7.3, pp. 3–20.

— (2015). "A comparison of Expected Shortfall estimation models". In: *Journal of Economics and Business* 78, pp. 14–47.

Scarrott, C. and A. MacDonald (2012). "A Review of Extreme Value Threshold Estimation and Uncertainty Quantification". In: *REVSTAT* 10.1, pp. 33–60.

Sharma, M. (2012). "Evaluation of Basel III revision of quantitative standards for implementation of internal models for market risk". In: *IIMB Management Review* 24.4, pp. 234–244.

So, M.K.P. and C-M. Wong (2012). "Estimation of multiple period expected shortfall and median shortfall for risk management". In: *Quantitative Finance* 12.5, pp. 739–754.

Taylor, J.W. (2008). "Estimating Value at Risk and Expected Shortfall Using Expectiles". In: *Journal of Financial Econometrics* 6.2, pp. 231–252.

Wang, J-N., J-H. Yeh, and N. Y-P. Cheng (2010). *How Accurate is the Square-Root-Of-Time Rule at Scaling Tail Risk: A Global Study*. Available at SSRN: https://ssrn.com/abstract=1671586 or http://dx.doi.org/10.2139/ssrn.1671586.

Weissman, I. (1978). "Estimation of parameters and large quantiles based on the k largest observations". In: *Journal of the American Statistical Association* 73, pp. 812–815.

Wong, C-M. and M.K.P. So (2003). "On conditional moments of GARCH models, with applications to multiple period value at risk estimation". In: *Statistica Sinica* 13.4, pp. 1015–1044.

Wong, W.K. (2008). "Backtesting trading risk of commerical banks using expected shortfall". In: *Journal of Banking and Finance* 32.7, pp. 1404–1415.

Yi, Y., X. Feng, and Z. Huang (2014). "Estimation of extreme value-at-risk: An EVT approach for quantile GARCH model". In: *Econometric Letters* 124, pp. 378–381.

Youssef, M., L. Belkacem, and K. Mokni (2015). "Value-at-Risk estimation of energy commodities: A long-memory GARCH-EVT approach". In: *Energy Economics* 51, pp. 99–110.

Zeileis, A. (2004). "Econometric Computing with HC and HAC Covariance Matrix Estimators". In: *Journal of Statistical Software* 11.10, pp. 1–17.

Zeileis, A. et al. (2022). *sandwich: Robust Covariance Matrix Estimators*. R package version 3.0-2. URL: https://CRAN.R-project.org/package=sandwich.

Zhang, D., M. Hu, and Q. Ji (2020). "Financial markets under the global pandemic of COVID-19". In: *Finance Research Letters* 36. https://doi.org/10.1016/j.frl.2020.101528.

Zhao, L-T. et al. (2019). "Oil price risk evaluation using a novel hybrid model based on time-varying long memory". In: *Energy Economics* 81, pp. 70–78.

Ziegel, J. F. (2016). "Coherence and elicitability". In: *Mathematical Finance* 26.4, pp. 901–918.