Charles University in Prague

Faculty of Social Sciences
Institute of Economic Studies

MASTER THESIS

Measuring Extremes: Empirical Application on European Markets

Author: **Bc. Durmus Ozturk**

Supervisor: **Mgr. Krenar Avdulaj**

Academic Year: **2014/2015**

Declaration of Authorship

The author hereby declares that he compiled this thesis independently, using only the listed resources and literature.

The author grants to Charles University permission to reproduce and to distribute copies of this thesis document in whole or in part.

Prague, January 5, 2015

Signature

Acknowledgments

I would like to express a special appreciation and thanks to my supervisor Mgr. Krenar Avdulaj for his aspiring guidance, encouragement, and constructive and friendly advises during the entire time of writing thesis. I would also like to express my gratitude also to my family, friends and colleagues for their immensely support throughout my studies.

Abstract

This study employs Extreme Value Theory and several univariate methods to compare their Value-at-Risk and Expected Shortfall predictive performance. We conduct several out-of-sample backtesting procedures, such as unconditional coverage, independence and conditional coverage tests. The dataset includes five different stock markets, PX50 (Prague, Czech Republic), BIST100 (Istanbul, Turkey), ATHEX (Athens, Greece), PSI20 (Lisbon, Portugal) and IBEX35 (Madrid, Spain). These markets have different financial histories and data span over twenty years. We analyze the global financial crisis period separately to inspect the performance of these methods during the high volatility period. Our results support the most common findings that Extreme Value Theory is one of the most appropriate risk measurement tools. In addition, we find that GARCH family of methods, after accounting for asymmetry and fat tail phenomena, can be equally useful and sometimes even better than Extreme Value Theory based method in terms of risk estimation.

Keywords Extreme Value Theory, Value-at-Risk, Expected Shortfall, Out-of-Sample Backtesting

Author's e-mail ozturkdurmus@windowslive.com
Supervisor's e-mail ies.avdulaj@gmail.com

Contents

List of Tables

List of Figures

Abbreviations

EVT	Extreme Value Theory
GPD	Generalized Pareto Distribution
GEV	Generalized Extreme Value Distribution
VaR	Value-at-Risk
ES	Expected Shortfall
MLE	Maximum Likelihood Estimation
QMLE	Quasi Maximum Likelihood Estimation
HS	Historical Simulation
FHS	Filtered Historical Simualtion
EWMA	Exponentially Weighted Moving Average
ARMA	Autoregressive Moving Average
ARCH	Autoregressive Conditional Heteroskedasticity
GARCH	Generalized Autoregressive Conditional Heteroskedasticity
EGARCH	Exponentially Generalized Autoregressive Conditional Heteroskedasticity
GJR-GARCH	Glosten-Jagannathan-Runkle Generalized Autoregressive Conditional Heteroskedasticity
POT	Peaks Over Threshold
BMM	Block Maxima-Minima Method
GFC	Global Financial Crisis
ACF	Autocorrelation Function
iid	Identically and Independently Distributed
CLT	Central Limit Theorem
DGPD	Dynamic Generalized Pareto Distribution

Master Thesis Proposal

Author	Bc. Durmus Ozturk
Supervisor	Mgr. Krenar Avdulaj
Proposed topic	Measuring Extremes: Empirical Application on European Markets

Topic characteristics Market risk is one of the most important phenomena in financial risk management and financial regulation. Worldwide financial instabilities and last global financial crisis again indicated the importance of assessing the rare and extreme events. That is, assessing the probabilities of these extreme and rare events is one of the crucial subjects in risk management since these events may have catastrophic financial consequences. Therefore, accuracy of risk estimates regarding stochastic market movements must be adequate in order to protect the capital and financial markets from such losses. Particularly, the overestimation of a risk may result in a capital holding that is excessive to cover losses, and oppositely the underestimation of a risk may result in a capital holding that fails to cover incurred losses.

Standard tool that has been established and widely used for downside risk assessment of a market portfolio is Value-at-Risk (VaR), defined as a quantile based market risk measure that gives a single number as an output. It is defined as the loss of a portfolio with respect to a given probability and time horizon. Because of its simplicity, it became standard risk measure after its release in early 90s by J.P Morgan with the RiskMetrics-Exponentially weighted moving average (EWMA) method. This followed its use for capital requirements as introduced in Basel 2 framework. However, VaR was widely criticized (e.g. not coherent - not satisfying subbadditivity property- Artzner et al. 1999), and other risk measures are hence established, such as Expected Shortfall (ES), Return Level, etc.

One of the most important and discussed issues in the accuracy of VaR and other quantile based risk measures is that this measure crucially depends on the models used to estimate them. Although many different methods and measures trying to explain these events are developed, most of them fail to provide accurate estimates of risk. Namely, they entail strong assumptions and face the problem of stylized facts of financial time series such as fat-tails, asymmetry, dependance, volatility clustering etc. Extreme Value Theory (EVT) exactly appears here to provide better solution to modelling extreme events as it does not take into account the whole distribution but rather tails of the distribution and thus provides better fit for the distributions of these extreme events. It also allows for the asymmetry in the tails of a distribution which is observed as the one of the stylized fact in the financial time series. In addition, the cooperation of EVT with many other parametric and nonparametric methods, which can be used to model stylized facts of the financial time series return distribution, gives a strong power to the usefulness of the theory dealing with risk assessment of extreme and rare events.

Global financial instability and last global financial crisis and its catastrophic effects faced by financial institutions yielded a motivation to conduct this study. First of all, this study will provide a deep analysis of the theoretical issues and consequences of the application of EVT-based models. Consequently, the main goal of this paper is to focus on the appropriateness of the EVT for measuring the probability of extreme events and assessing the risk defined in terms of VaR, ES. Different dimensions of the risk modelling will be used; their accuracy will be examined and relatively compared with chosen methods. Empirical analysis of the chosen methods will be done on the following European Stock Markets, BIST100 (Istanbul Stock Exchange), PX50 (Prague Stock Exchange), IBEX35 (Madrid Stock Exchange), ASE (Athens Stock Exchange), PSI20 (Lisbon-Portuguese Stock Exchange) for the period covering last global financial crisis. The accuracy of the models will also be tested in the financial crisis period. Particularly, Backtesting methodology will be applied in order to check relative performance of the conducted models.

Hypotheses

- Stylized facts such as non-normality, dependence, fat-tailness, asymmetry, etc. are present in the underlying market indices.
- Models trying to capture these stylized facts are better than those that

do not count for them in terms of their predictive performance for VaR and ES estimation.

- EVT-based method is superior to the non-EVT based ones in terms of their predictive performance for VaR and ES estimation.
- EVT-based method is significantly better than other methods, especially in the volatile periods, namely for predicting the extreme risks such as global financial crisis.

Methodology The main part of this study will be the methodology since it will encompass the conduction of different methods and assign the accuracy of their risk estimates. However, the focus will be on the EVT-based methods. By comparing the predictive performance of baseline models in terms of risk estimates, the study will assess the relative accuracy of these methods.

Firstly, the preliminary data analysis from chosen stock market indices will be done. Particularly, the log return of the indices will be tested by various statistical techniques to reveal the characteristics of the data and make a brief comparison among the chosen markets. Secondly, different methods will be used to analyze the data after preliminary analysis. Following methods will be conducted.

EVT-based methods: Peaks over Threshold (POT) method that uses Generalized Pareto distribution (GPD). In addition, 2 step approach established by McNeil and Frey (2000) will be applied. Namely, this approach uses AR-GARCH type methods to filter data, saves residuals from this procedure and fits them with GPD as in the POT method. ARCH-GARCH family of methods: These methods will be used within different dimensions such as symmetric, asymmetric, with normal distribution and student-t distribution etc. Traditional methods: These methods will be Normal method, RiskMetrics, Historical Simulation etc Some additional methods: semiparametric methods such as FHS, CaViaR

After modelling the data with aforementioned methods, the resulting risk estimates will be generated in terms of VaR and ES. Accordingly, risk estimates will be calculated not only for left tail, i.e long position, but also for right tail, i.e. position considering 95%, 99% and 99.5% confidence level.

Last part of methodology will be Backtesting procedure with the aim to check the relative predictive performance of the methods applied. In addition, volatile period will be evaluated separately as a sub-sample.

Note: The statistical and econometric estimation will be done in R software.

Outline

1. Introduction
2. Theoretical Background
3. Related Work
4. The Model
5. Empirical Verification
6. Conclusion

Core bibliography

1. McNeil, A. J. & R. Frey (2000): "Estimation of tail-related risk measures for heteroscedastic financial time series: an extreme value approach." *Journal of Empirical Finance* **7(3–4)**: pp. 271–300. Special issue on Risk Management
2. Kupiec, P. H. (1995): "Techniques for verifying the accuracy of risk measurement models." *Journal of Derivatives* **3(2)**: pp. 73–184.
3. Christoffersen, P. (1995): "Evaluating interval forecasts." *International Economic Review* **39**: pp. 841–862.
4. Christoffersen, P., J. Hahn, & A. Inoue (2001): *"Testing and comparing value-at-risk measures" Journal of Empirical Finance* **8(3)**: pp. 325–342. Washington, DC: FIAS.
5. Embrechts, P., S. I. Resnick, & G. Samorodnitsky (1999): "Extreme value theory as a risk management tool." *North American Actuarial Journal* **3(2)**: pp. 30–41.
6. Gencay, R., & F. Selcuk (2004): "Extreme value theory and value-at-risk: Relative performance in emerging markets." *International Journal of Forecasting* **20(2)**: pp. 287–303.

Author

Supervisor

Chapter 1

Introduction

The Global Financial Crisis of 2007-2008 is often considered to be the most severe one ever since 1930s and the Great Depression. It caused worldwide financial losses which were accompanied by the general industrial and economic slowdown. Moreover, the effects of crisis were felt due to the global financial network and contagion. Hence, the need to predict such events with accuracy and to allow individuals, firms and institutions to prevent or to be ready for such events becomes essential.

Market crash of 1987 increased attention to define relevant risk measures for the purpose of risk management. Consequently, Value-at-Risk (VaR) became an industry standard as a tool to measure market risk in early 1990s with the JP Morgan RiskMetrics. Moreover, VaR was established as the primary determinant of the required risk capital against the occurrence of the potential losses (on Banking Supervision & for International Settlements 2004). Since then, VaR is the most widely used market risk measure and it is simply defined as the worst expected loss for some time period at a given confidence level (McNeil *et al.* 2005). However, VaR is often criticized mainly for not being coherent risk measure, i.e. it is not subadditive (Artzner *et al.* 1999). Therefore, Artzner *et al.* (1999) proposed Expected Shortfall (ES) as an alternative risk measure, which is in fact coherent. Although VaR and ES seem to be sound risk measures, their accuracy crucially depends on the models used to estimate Danielsson & Vries (2000). With regard to this, variety of models are proposed in the literature that can be grouped under non-parametric, parametric and semi-parametric approaches (Gencay & Selcuk 2004).

In parametric approaches, one imposes a specific distribution assumption to model the data at hand, with a normal distribution being a common choice.

On the other side, non-parametric approaches with the most popular one being Historical Simulation (HS) that does not rely on any assumption about the underlying distribution and it directly estimates risk measures based on the empirical distribution. However, these two approaches often face some problems. For example, HS method gives equal importance to the each observation and thus, it may lead to inaccurate estimation of VaR and ES. Similarly, parametric approaches face the problem of model risk since we do not know the true underlying distribution. Finally, semi-parametric approaches combine the advantages and significantly reduce the problems of both non-parametric and parametric approaches. With regard to this, Extreme Value Theory (EVT) offers a semi-parametric approach to calculate risk measures as a combination of HS with parametric extreme value distributions (McNeil *et al.* 2005). Furthermore, as a well-developed theory in the field of probability, EVT concentrates on the behavior of extremes. Similarly, the fact that VaR and ES are also concerned with the extreme quantiles of the distribution, these concepts naturally match together.

The main issue in the VaR and ES application is to decide which methodology to implement. These methodologies are mainly based on the characteristics of the financial data at hand. Empirical literature, following the pioneering studies of Mandelbrot (1963) and Fama (1965), documented consistent findings about financial markets known as *stylized facts* such as fat-tails, asymmetry, dependence, volatility clustering, etc. These empirical findings can significantly alter the estimation of risk measures and therefore should be taken into account in order to get the accurate risk estimates. In line with this, McNeil & Frey (2000) developed a method based on EVT - so called 2 step procedure which is applied by fitting a dynamic ARMA-GARCH type model to the data in the first step, and in the second step Generalized Pareto Distribution (GPD) is fitted to the respective residuals obtained within the previous step. By doing this, one can account for the dependence and stochastic nature of the volatility. In addition, given the fact that EVT can account for fat-tails and asymmetry by focusing separately on each of the two tails of the distribution, makes 2 step procedure a viable approach for calculation of risk measures.

EVT approach to VaR computation was considered by many empirical studies concerned with the financial risk measurement and overall, the following conclusions emerges: first, EVT based risk estimates are at least as precise as estimates from other methodologies regardless of the analyzed confidence level; second, when one goes further in the tail, i.e. 99% and more, EVT based methods outperform other methods in terms of VaR, since it is a theory being concerned with extremes.

Hence, in line with the above conclusions, the main goal of this study is to focus on the appropriateness of the EVT for measuring the probability of extreme events and assessing the risk defined in terms of VaR and ES. The research questions are as follows: first, stylized facts such as non-normality, dependence, fat-tailness, asymmetry, etc. are present in the underlying market indices; second, models trying to capture these stylized facts are better than those that do not count for these in terms of their predictive performance for VaR and ES estimation; third, EVT-based method is superior to the non-EVT based ones in terms of their predictive performance for VaR and ES estimation; and the last, EVT-based method is significantly better than other methods, especially in the volatile periods; namely, for predicting the extreme risks such as global financial crisis.

Briefly, this study examines the predictive performance of EVT based methods by comparing it with the variety of methods. Namely, these methods are standard normal distribution, HS, unconditional EVT method based on GPD, RiskMetrics-EWMA method, two best selected conditional ARMA-GARCH family methods, Filtered Historical Simualtion (FHS) method and conditional EVT method based on GPD (2 step procedure). For the purpose of methods comparison, this study examines daily closing prices from five different European capitalization weighted stock market indices; PX50 (Prague, Czech Republic), BIST100 (Istanbul, Turkey), ATHEX (Athens, Greece), PSI20 (Lisbon, Portugal), IBEX35 (Madrid, Spain). The data span is the same for all markets covering the period from 1st March 1994 until 24th February 2014. This time span covers the period of many crises that had significant impact on the underlying markets such as; Asian financial crisis (1997), Turkish banking crisis (2001), Global financial crisis (2008-2010), and lately the European debt crisis (2012). The full sample size ranges from 4933 daily observations for PX50 to 5123 daily observations for PSI20. The respective comparison of methods for VaR is based on so called - backtesting procedure by employing unconditional coverage test developed by Kupiec (1995), as well as independence and conditional coverage tests developed by Christoffersen (1998). Similarly, ES estimates are backtested using the tests developed by McNeil & Frey (2000).

This study contributes to the empirical literature on financial risk estimation by providing the extensive and detailed description, subsequent application and detailed comparison of the most popular and most commonly used respective methods. Also, all of the relevant methods are applied to five different stock markets that have not been combined for these purposes before. In addition, as opposed to the largest number of respective studies that usually cover period of several years, our study uses a very long data span (approximately twenty years) that covers periods of several crises that affected the chosen markets. Such a large time period coverage allows us to compare the predictive performance of the chosen methods in a long span of out-of-sample observations in which the estimates and its actual counterparts are compared.

The study is organized as follows: second chapter presents the theoretical foundations of VaR, ES and univariate EVT, and outlines the empirical findings regarding VaR and ES under EVT framework. Third chapter presents the competing univariate methods considered in this study, as well as how these methods are implemented. In addition, it presents the backtesting methodology which we base our criterion for competing univariate methods. Fourth chapter presents the data and empirical results obtained by following the structure of methodology. And finally, chapter five concludes.

Chapter 2

Theoretical Foundations and Literature

This chapter presents the theoretical foundations and outlines the empirical findings about VaR and ES in the EVT framework. In the first section, we define the main properties of risk measures in terms of VaR and ES. Section two briefly presents observed empirical facts about financial time series. After that, we summarize theoretical findings of EVT, its relation with the central limit theorem and statistical methods developed on EVT[1]. Finally, we outline the main findings of EVT approaches to VaR and ES by analyzing the comparative literature and we state our research question.

2.1 Value at Risk and Expected Shortfall

The need for the appropriate risk management for financial institutions, especially after the market crash of 1987, increased attention to define relevant risk measures. Consequently, VaR became an industry standard as a tool to measure market risk in early 1990s with the JP Morgan RiskMetrics. Furthermore, it is established as the primary determinant of the required risk capital against the occurrence of the potential losses[2] (on Banking Supervision & for International Settlements 2004). Since then, VaR is the most widely used market risk measure and it is defined as the worst expected loss for some time period

[1] Note that we will only consider univariate EVT. We refer reader for detailed analysis of Extreme Value Theory to Embrechts *et al.* (1997) and McNeil *et al.* (2005).

[2] In regulatory norms, it is required to calculate VaR at time horizon equal to 10 days, i.e the period which financial institution liquidate its position and 99% confidence level. Then, the resulting VaR is used in the calculation of market risk capital (on Banking Supervision & for International Settlements 2004).

at a given confidence level (McNeil *et al.* 2005), or in probability terms, it is a quantile of the underlying distribution of returns. Following McNeil *et al.* (2005), VaR is formally defined as follows:

$$\begin{aligned} VaR^{\alpha} &= \inf\{x \in \mathbb{R} : P(X > x) \leq 1 - \alpha\} \\ &= \inf\{x \in \mathbb{R} : F_X(x) \geq \alpha\} \\ &= F_X^{-1}, \end{aligned} \tag{2.1}$$

where $\alpha \in (0, 1)$ and it represents the confidence level (or similarly probability), F_X is the cumulative distribution function of random variable X and F_X^{-1} is the generalized inverse[3] (or similarly the quantile function) of the distribution function F_X.

VaR, as a quantile based risk measure, focuses on the tail of the distribution and hence captures the extreme losses which are rare. It can be effectively applied regardless of the assumed specific distribution (Danielsson & Vries 2000). But the major advantage of VaR is that it gives a single number as an output and the respective interpretation is quite easy and straightforward, which makes it effective tool for the financial institutions.

Anyways, VaR is often criticized mainly for two reasons. First criticism stems from its definition since it is concentrated on the cut-off between the center and the tail of a distribution, thus completely disregarding the information beyond this cut-off, i.e. it does not give information about the losses exceeding the respective VaR. Second and the most problematic critics arise from its theoretical coherency. Artzner *et al.* (1999) presented a theoretical concept of coherent risk measures[4], and showed that VaR is not a coherent risk measure as it violates the axiom of subadditivity, i.e. it can happen that $VaR^{\alpha}(w_1X+w_2Y) > w_1VaR^{\alpha}(X)+w_2VaR^{\alpha}(Y)$, where $w_1, w_2 \in (0, 1)$. Hence, VaR contradicts with the diversification effects, in which we lower the level of risk[5].

Given these two flaws of VaR, Artzner *et al.* (1999) proposed ES as an

[3] Generalized inverse of the distribution function F_X is defined as $F_X^{-1} = \inf\{x \in \mathbb{R} : F_X(x) \geq \alpha\}$.

[4] If risk measure satisfies the axioms of monotonicity, positive homogeneity, subadditivity, and translation invariance, then it is called coherent (see Artzner *et al.* (1999) for detailed documentation).

[5] One can avoid this flaw by considering portfolios that are composed of the same set of underlying elliptically distributed risk factors (McNeil *et al.* 2005).

alternative risk measure which is in fact coherent[6]. ES simply defined as average of the VaR values that exceeds VaR^{α}, formally:

$$ES^{\alpha}(X) = \frac{1}{1-\alpha}\int_{\alpha}^{1} VaR_{\varphi}^{\alpha}(X)\mathrm{d}_{\varphi}, \qquad \alpha \in (0,1), \tag{2.2}$$

or assuming F_X is a continuous (integrable):

$$ES^{\alpha} := \frac{\mathrm{E}[XI_{\{X \geq VaR^{\alpha}\}}]}{1-\alpha} = \mathrm{E}\big[X|X \geq VaR^{\alpha}\big]. \tag{2.3}$$

Therefore, by definition, ES provides information beyond the VaR since it takes into account the size of the losses further in the tail of the distribution unlike the VaR. In addition, given the coherency of ES, makes it an effective candidate for an appropriate risk measure.

Although VaR and ES theoretically seem sound risk measures, their application in reality is challenging. That is, the accuracy of the estimation crucially depends on the models which differ due to their way of addressing the problem of the estimation procedure (Danielsson & Vries 2000). Variety of models are proposed in the literature and they can be grouped under three different categories[7](Gencay & Selcuk 2004):

- non-parametric methods, e.g. HS, Hill Estimator, etc;
- parametric methods: e.g. Gaussian, Riskmetrics, GARCH, etc;
- semi-parametric methods: e.g. EVT, FHS, CaViaR, etc.

One of the most famous method is non-parametric historical simulation which does not make any assumption about the underlying distribution of the time series and it directly estimates risk measures based on the empirical distribution. However, it may lead to inaccurate estimates of VaR and ES as a result of giving an equal importance to the each observation.

On the other side, parametric methods impose a distribution assumption to model the observations, i.e. model is fitted to the data and estimated parameters are used for the calculation of VaR and ES. For example, the simplest parametric approach to calculate VaR and ES is based on the normality assumption; given that the observations are normally distributed with mean μ and variance σ^2, VaR and ES are calculated as follows:

[6] Formal proof of ES being coherent can be found in McNeil *et al.* (2005)

[7] reader can find detailed analysis and implementation of the methods from each category in chapter methodology.

$$VaR^{\alpha} = \mu + \sigma \Phi^{-1}(\alpha), \tag{2.4}$$

$$ES^{\alpha} = \mu + \sigma \frac{\Psi\big(\Phi^{-1}(\alpha)\big)}{1-\alpha}, \tag{2.5}$$

where $\alpha \in (0,1)$ and represents a confidence level and Φ^{-1} represents inverse function (or quantile) of the standardized normal distribution while Ψ stands for the analogous density function.

The main problem of parametric methods is the explicit distribution assumption for the observations because we do not know the true underlying distribution, i.e. model risk. Hence the accuracy of the risk estimation can vary with the assumed distribution assumption, resulting either in underestimation or overestimation of the risk, unless we adequately model the data.

Furthermore, the semi-parametric methods combine the advantages of both non-parametric and parametric methods and avoid some of their difficulties. For example, an EVT method so called 2 step procedure can be given in this group, as it combines both non-parametric historical simulation and parametric generalized pareto distribution (McNeil & Frey 2000), i.e. it focuses only on the tail of the distribution and reduce the model risk that parametric methods face, similarly it aims to model asymptotic behavior of the tail and it avoids the main problems that non-parametric methods face[8].

2.2 Stylized Facts

The main issue in the application of VaR and ES is to decide which methodology to implement. For this reason, it is essential to understand the underlying assumptions of these methodologies together with their way of addressing the problem of estimation procedure. After this initial step, one can decide for the method which he thinks to be closer to his objectives. Hence, the selection of the methodology is a first and important step in order to calculate risk measures.

These methodologies are mainly based on the characteristics of the financial data at hand. Empirical literature, following the pioneering studies of Mandelbrot (1963) and Fama (1965), documented consistent findings about financial

[8] See Chapter 2 for the theoretical foundations of EVT and Subsection 3.2.4 for the implementation of 2 step procedure.

markets known as *stylized facts*. These empirical findings can significantly alter the estimation of risk measures and therefore should be taken into account in order to achieve accurate estimation of risk. These stylized facts can be summarized as follows:

fat-tails – the distributions of financial returns are leptokurtic, i.e. the tail of a distribution is heavier than the tail of a normal distribution, implying that the probability of extreme returns are higher than the one predicted under normality assumption.

asymmetry – the distributions of financial returns are typically negatively skewed, i.e. the left tail of the distribution is larger than the right tail, implying that the downward and upward asset price movements are not equal (also known as leverage effect).

dependence – financial time series exhibit autocorrelation, i.e. asset returns are significantly correlated, implying that the negative (positive) price changes tend to be followed by negative (positive) price changes.

time varying volatility – volatility (standard deviation) is in fact time dependent, i.e. it fluctuates over time, implying that the one can observe periods characterized with high volatility or low volatility at various points in time.

volatility clustering – squared returns are significantly correlated, i.e. factors of volatility have a tendency to cluster in time, implying that the large (small) price changes tend to be followed by large (small) price changes.

conditional fat-tails – the distribution of residuals is still leptokurtic, even after the correction of volatility clustering, e.g. using Generalized Autoregressive Conditional Heteroskedasticity (GARCH) family of models.

The methodologies for risk measurement are basically differing based on the accountability of these above mentioned stylized facts (capable to capture either some or all of them). Hence, it is crucial to understand the exact nature of the financial data at hand in order to accurately estimate risk.

2.3 Extreme Value Theory

The first step is the choice of the approach when working with EVT and those can be classified as parametric and non-parametric approach. The main difference between these two approaches is that the first one is based on fitting the parametric distribution that should reflect the underlying distribution of extremes, while the second one is based on the shape parameter (or tail index) estimation that should reflect behavior of the tail of the distribution.

These approaches are based on either estimating the shape parameter (or tail index) which characterizes the behavior of the tail of the distribution, i.e. non-parametric approach, or fitting a parametric distribution to which the extremes of the underlying distribution eventually obey, i.e. parametric approach.

The first applications of EVT to finance started with the latter approach in early 1990s, out of which the most commonly used once were based on Hill estimator. Soon after, the successful development of parametric approaches followed with two possible applications that differ according to the definition of *extreme value*. Namely, extreme value can be defined either as the maxima-minima of the iid random variables in a non-overlapping sub-samples (BMM) or as the iid random variables in a sample that exceeds a predefined threshold u (POT). Unlike the non-parametric approaches (Hill estimator), parametric approaches (POT) have shown to have better stability in respect to the choice of the threshold u (McNeil & Frey 2000). In addition, parametric methods allow for modelling any type of extreme value distribution, as opposed to the nonparametric which is used for modelling fat tailed distributions only.

Before we start to explain approaches to EVT, we will present main theoretical findings that EVT is based on.

Fisher-Tippett and Gnedenko Theorems EVT has sound theorems with the similarity of the well-known Central Limit Theorem (CLT) in probability theory. CLT states that sufficiently large number of Identically and Independently Distributed (iid) random variables converges to a given normal distribution (also known as a bell-shaped), regardless of the underlying distribution. Similarly, Fisher-Tippet and Gnedenko guarantee that the maximum-minimum of these sufficiently large iid random variables converges to one of three given distributions family. Thus, the key difference between these two theories is that the CLT takes into account iid random variables while EVT concentrates on the maxima-minima of these iid random varibles. In addition, unlike the CLT which considers a unique distribution, EVT provides three different types of asymptotic distributions.

In mathematical terms, first one considers n iid random variables $X_1, \cdots, X_n$, with a cumulative distribution function F. Then we can define the maximum[9] of these variables as $M_n = max\{X_1, \cdots, X_n\}$. Distribution function of collection M_n is defined as F^n and basically we are interested in the asymptotic distribution of M_n. Then, taking the limit will result in following degenerate[10] distribution function:

$$\lim_{n \to +\infty} F^n(x) = \begin{cases} 1, & \text{if } F(x) = 0 \\ 0, & \text{if } F(x) < 1. \end{cases}$$

Fisher & Tippett (1928) standardized the set of maximum random variables M_n (affine transformation) with location and scale parameters d_n and c_n, respectively, i.e. norming constants, and studied the asymptotic behavior of these standardized set $\frac{M_n - d_n}{c_n}$. Furthermore, with the assumption of the existence of such location d_n and scale c_n parameters in whole sequence, Gnedenko (1943) proved the convergence of these standardized set of maximum random variables to one out of three non-degenerate distributions family. These non-degenerate distributions are formulated as follows:

- *Fréchet* : $H_\xi(x) = \begin{cases} 1, & \text{if} \quad x \leq 0 \\ \exp\{-x^{-1/\xi}\}, & \text{if} \quad x > 0 \end{cases} \qquad \xi > 0;$
- *Gumbel* : $H_0(x) = \exp\{-\exp(-x)\}, \ x \in \mathbb{R};$
- *Weibull* : $H_\xi(x) = \begin{cases} \exp\{x^{1/\xi}\}, & \text{if} \quad x \leq 0 \\ 1, & \text{if} \quad x > 0 \end{cases} \qquad \xi > 0,$

where ξ is called a shape parameter that considers the tail behavior of the distribution and it is determined by the distribution function F in a unique way. These three types of distributions are called *extreme value distributions* and have following characteristics:

- Fréchet type is representing the asymptotic distribution for fat-tailed distributions, i.e. with power-decaying tails. Examples of this type are student-t, log-gamma and Pareto distributions, etc.
- Gumbel type is representing the asymptotic distribution for light-tailed distributions, i.e. exponentially decaying tails. Such as normal, log-normal, gamma, chi-squared distributions etc.

[9] Similarly, minimum can be defined as $min\{X_1, \cdots, X_n\} = -max\{-X_1, \cdots, -X_n\}$; however, we will document the following only based on the maxima.

[10] Degenerate distribution is the probability distribution of a random variable that is characterized with the single value, i.e. tossing a coin or rolling a dice. On the other side, non-degenerate distribution function F can be defined as, $\exists x | F(x) \in [0, 1]$.

- Weibull type is representing the asymptotic distribution for short-tailed distributions, i.e. with finite right endpoint. The beta distribution can be given as an example of this type.

In finance, particular importance is given to the Fréchet family of extreme value distribution because of the stylized fact that most financial series are fat-tailed, implying an asymptotic behavior of Fréchet type distribution.

Reciprocal of the shape parameter ξ is called as the *tail index* $\alpha := \frac{1}{\xi} > 0$ and it is directly related to the tail behavior of the distribution. For instance, one can verify that the degrees of freedom in the student-t distribution are equal to tail index α with the similar type.

2.3.1 Non-parametric Approaches to EVT

We will briefly present the non-parametric approaches to EVT, though detailed review of the non-parametric approaches to EVT can be found in Embrechts *et al.* (1999) and McNeil *et al.* (2005).

The first and the most used non-parametric approach that have been studied in the EVT literature is Hill estimator and it is introduced by Hill (1975). It dominates other non-parametric approaches in the case of estimation of the tail index α, i.e. if the asymptotic behavior of the distribution belongs to fat-tailed Fréchet family. If not, Hill estimator cannot be employed and one should consider some other type of non-parametric estimator. Assuming the positive sequence of random variables (to ensure logarithm), $X_{1,n} \geq \cdots \geq X_{n,n}$, Hill estimator $\hat{\alpha}$ can be formulated as follows:

$$\hat{\alpha}^{H}_{k,n} = \left[\frac{\sum_{i=1}^{k} (\ln X_{i,n} - \ln X_{k,n})}{k} \right]^{-1},$$

where k is the value determining the cut-off level between the observations that belong to the tail and center of the distribution. Thus, Hill estimator crucially depends on the cut-off level k which is the main concern of this method in empirical applications.

Next, the non-parametric estimator proposed by Pickands III (1975), is called Pickands estimator. It is employed to estimate the shape parameter ξ and it gives an estimate of shape parameter ξ regardless of the asymptotic behavior of the distribution, i.e. any of the three Fréchet, Gumbel, Weibull extreme value distributions. On the other side, high volatility of the Pickands estimator can be given as its drawback, as it was found in the simulation studies of Kearns & Pagan (1997).

Similarly, Dekkers & De Haan (1989) also introduced a non-parametric estimator - so called Dekkers-Einmahl-de Haan (DEdH) estimator, which modifies the Hill estimator and extends it to account for any kind of extreme value distribution, not only Fréchet type. Moreover, the Hill estimator derived from the small sample will yield a biased estimator for the tail index $\hat{\alpha}$ as a result of small number of extremes considered. Another modification of the Hill estimator that gives an unbiased estimator for the tail index $\hat{\alpha}$ in the respective case can be found in Huisman *et al.* (2001).

2.3.2 Parametric Approaches to EVT

Block Maxima-Minima Method and GEV

Jenkinson (1955) introduced a generalized framework for the three extreme value distributions that we introduced under Fisher–Tippett and Gnedenko theorems. It is called *Generalized Extreme Value Distribution (GEV)*[11] and it nests the Gnedenko's representation of three extreme value distributions. It is formulated as follows:

$$H_{\mu,\sigma,\xi}(x) = \begin{cases} \exp\big\{-(1+\xi\frac{x-\mu}{\sigma})^{-1/\xi}\big\}, & \xi \neq 0 \qquad \& \qquad 1+\xi\frac{x-\mu}{\sigma} > 0, \\ \exp\big\{-\exp(-\frac{x-\mu}{\sigma})\big\}, & \xi = 0, \end{cases} \tag{2.6}$$

where μ, σ and ξ represent location, scale and shape parameter, respectively. Notice that the shape parameter ξ distinguishes the type of the extreme value distribution where $\xi > 0$ delivers Fréchet type, $\xi = 0$ delivers Gumbel type and if $\xi < 0$ the GEV distribution is Weibull type.

The first parametric approach to EVT is *Block Maxima-Minima Method (BMM)* method that employs GEV distribution to a particular set of maxima-minima of iid random observations. Its application consists of following two steps:

- given a random sample of iid data $X_1, \cdots, X_N$ that are drawn from distribution F, first divide this random sample into m non-overlapping sub-samples that each consists of n observations, that is $m \times n = N$,

[11] it is also known as *Jenkinson–Von Mises representation* of extreme value distributions.

where $m > 0$ and $N > n$. And let M_i denote the maximum of the ith sub-sample, where $i = 1, \cdots, m$.

- Next, fit the GEV distribution to the sequence of maxima $M_1, \cdots, M_m$ in order to obtain estimates $\hat{\xi}, \hat{\mu}$ and $\hat{\sigma}$ of parameters, with the assumption that $F \in MDA(H_\xi)$[12], where $\xi, \mu, \sigma \in \mathbb{R}$ and $\sigma > 0$.

The respective estimation can be done by means of Maximum Likelihood Estimation (MLE) with the following constraints:

$$\sigma > 0 \qquad \& \qquad 1 + \xi \frac{M_i - \mu}{\sigma} > 0, \qquad \text{for } i = 1, \cdots, m.$$

A main strength of BMM method is its natural interpretation of the problem to create the maxima-minima sequence of the observations. These observations can be naturally structured in blocks. For example, daily observations can be divided into monthly or quarterly blocks. On the other side, main drawback of the BMM method is in the definition of extreme values since many non-extreme values in a single block could be considered as extreme, or oppositely, many extreme values in a single block could be omitted as a result of clustering.

Peaks Over Threshold Method and GPD

Another parametric approach to EVT is *Peaks Over Threshold (POT)* method and it is as natural as the BMM method. In this approach extremes are defined as the observations which exceed the set (high) threshold level u. Unlike the BMM method, POT method uses data more efficiently and allows for a more parsimonious modelling (McNeil *et al.* 2005; Këllezi & Gilli 2006). In addition, it is based on solid set of theoretical foundations similar to the BMM method as documented by Balkema & De Haan (1974) and Pickands III (1975).

In POT method, we are interested in the distribution of exceedances over the set threshold level u which is in fact conditional on observations exceeding u. This distribution of interest is called *excess distribution*, and by definition, the excess distribution over the threshold u takes the following form:

[12]

Definition 1 ($F \in MDA(H_\xi)$). *If there exists a sequence of norming constants such that the asymptotic distribution of the standardized maxima is of Fréchet (or Gumbel, or Weibull) type with shape parameter ξ, we will say that F is in the maximum domain of attraction of the Fréchet (Gumbel, or Weibull, respectively) distribution H_ξ (Rocco 2011).*

$$F_u(x) = P(X - u \le x | X > u) = \frac{F(x+u) - F(u)}{1 - F(u)}, \qquad 0 \le x < x_F,$$

where $x_F = \sup\{x \in \mathbb{R} : F(x) < 1\}$, and F is the cumulative distribution function of a random variable X.

Additionally, Balkema & De Haan (1974) showed that the asymptotic distribution of excess distribution F_u converges to a *GPD*. Therefore, GPD is formulated as follows:

$$G_{\xi,\beta}(x) = \begin{cases} 1 - \left(1 + \xi \frac{x}{\beta}\right)^{-1/\xi}, & \xi \neq 0, \\ 1 - e^{-\frac{x}{\beta}}, & \xi = 0, \end{cases} \quad \text{where} \begin{cases} x \ge 0, & \text{if } \xi \ge 0, \\ 0 \le x \le -\beta/\xi, & \text{if } \xi < 1, \end{cases} \tag{2.7}$$

where $\beta > 0$ and it represents the scale parameter whereas ξ is the shape parameter as in the case of GEV distribution.

The parametric POT method that employs GPD, similar to BMM method, is applied by the following two steps:

- given a random sample of iid data $X_1, \cdots, X_N$ that are drawn from distribution F, first determine a threshold level u and denote the data exceeding this threshold level with $\tilde{X}_j$, where $j = 1, \cdots, N_u$.
- Next fit the GPD to the sequence of exceedances $\tilde{X}_1, \cdots, \tilde{X}_{N_u}$ in order to obtain estimates of parameters $\hat{\xi}$ and $\hat{\beta}$, with the assumption that $F_u(y) = G_{\xi,\beta}(y)$, where $\xi, \beta \in \mathbb{R}$ and $\beta > 0$.

Again, the respective parameter estimation can be done by means of MLE with the following constraints[13]:

$$\sigma > 0 \qquad and \qquad 1 + \xi\frac{X_j - u}{\beta} > 0, \qquad for \qquad j = 1 \cdots N_u.$$

McNeil & Frey (2000), based on the simulation studies, discusses the efficiency of the non-parametric Hill estimator and parametric GPD under MLE.

[13] Majority of the empirical literature employs MLE for the estimation of parameters, i.e. maximizing the log-likelihood function with respect to shape and scale parameters ξ, β. However, Tolikas et al. (2007) uses probability weighted moments (PWM)and argues that the estimates are more efficient than the one under MLE procedure.

They argue that Hill estimator does not estimate the parameters efficiently and hence, nor the VaR. Moreover, they conclude that the GPD provides better estimates of VaR especially further in the tail, i.e. high quantiles.

2.3.3 Issues with the Application of EVT Approaches

All of the EVT approaches above face the common problem of threshold level selection, i.e. parameter estimates derived are sensitive to the choice of the threshold level[14]. For example, different number of blocks m in the BMM method or different threshold level u (k) in the POT method (Hill method) will result significantly different parameter estimates, hence different estimates of the risk. In addition, the determination of threshold u is difficult since it faces the variance-bias trade-off. In other words, if u is set high, only few data will belong to the tail as extreme, yielding huge and inefficient variance estimators (or similarly standard errors). On the other side, setting u low will result in biased parameter estimates, because of the many data considered as extreme, though they are not. Thus, the choice of the threshold is a main issue for the application of EVT approaches.

In literature, several procedures are adopted for the purpose of threshold selection. However, there is not any satisfactory procedure available. Many authors conventionally follow widely used suggestions, i.e. 5%-15% of the data considered as extreme. Moreover, graphical tools are commonly used for this issue; the most used one being Hill plot when dealing with Hill estimator[15]. It plots the estimates of the shape parameter ξ (or tail index α) with the varying level of the threshold value k.

Another widely used graphical tool for the purpose of threshold selection is *mean excess plots*, $(u, e_n(u))$ where $e_n(u)$ is called sample mean excess function and it is defined as follows:

$$e_n(u) = \frac{\sum_{i=k}^{n}(x_i^n - u)}{n-k+1}, \qquad k = min\{i|x_i^n > u\} \quad \& \quad x_1^n < u < x_n^n, \tag{2.8}$$

where, $n-k+1$ represents number of data that exceeds the threshold level $u > 0$. Therefore, the criterion for the selection is based on the behavior of the estimated sample mean excess function which should be approximately linear

[14] For example, Lux (2001) states that the different conclusions in the studies of German stock returns are due to the different choice of the threshold values.

[15] Analogously, Pickands plot can be conducted in the case of Pickands estimator

if the data over u is distributed according to GPD (Këllezi & Gilli 2006). In addition, the sign of the shape parameter ξ will be determined by the slope of this linear trend, i.e. upward linear relationship with a positive gradient above the threshold indicates a positive shape parameter $\xi > 0$ and hence, a heavy tail (Fréchet family).

These graphical tools are very useful but still far from being a statistical tool. Hence, McNeil & Frey (2000) adopted a procedure called *mean squared error minimization* using Monte Carlo simulation which simultaneously takes into account both the variance and bias. That is, minimization of mean squared error will eventually imply minimum variance and bias, hence optimum choice of threshold u, too. However, this procedure has some drawbacks. First, the simulation is done based on the known distribution F which contradicts with the EVT approach that can be drawn from an unknown distribution. Second, if the excess distribution fails to be iid, the variance will be even greater (Kearns & Pagan 1997).

Another crucial problem for the application of EVT approach to financial data is the unrealistic iid hypothesis. As stated in Section 2.2, the financial time series exhibits autocorrelation in both absolute and squared values, hence violating the respective hypothesis. Therefore, one should take into account the dependence before applying EVT. This can be pursued mainly in two ways. The first one is done by extending the EVT framework with additional hypothesis, i.e. strictly stationary time series. The second one is so called two-step procedure suggested by Diebold *et al.* (1998) and first implemented in the study of McNeil & Frey (2000)[16]. Briefly, procedure is applied by fitting a dynamic ARMA-GARCH type model to the data in the first-step, and in the second step GPD (or GEV)[17] is fitted to the respective residuals obtained within first step. Main intuition is based on the property that the standardized residuals extracted from the dynamic model will be approximately iid. Hence, one could account for the dependence structure of the time series using the standardized residuals instead of raw series for the application of EVT.

Another issue for the application of EVT is the frequency of data. Low frequency, i.e. monthly, data will not be significant for EVT approaches unless

[16] Note that this procedure is of main concern on this study. The application of same study can be also found in Kuester *et al.* (2006); Neftci (2000); Gencay & Selcuk (2004); Bali (2007); Andjelic *et al.* (2010); Ergen (2010); Andersen & Pedersen (2010). Multivariate setting can be found in Nystrom & Skoglund (2002); Avdulaj (2012)

[17] One can also apply non-parametric EVT approaches in the second step. However Kuester *et al.* (2006) argues that the distribution assumption for the innovations at the first step is indeed significantly effecting the results of non-parametric approaches in the second step.

the time range is large, as a result of extreme values considered which are, by definition, rare. On the other side, high frequency, i.e. minute-to-minute, data is found to improve the significance of the EVT approaches, since the accuracy of parameter estimates is high due to huge amount of data that characterize the behavior of the tail Lux (2001). Yet, in this case, one should consider seasonality component of high frequency data before applying EVT. However, vast majority of the empirical literature use daily frequency data. Since EVT approaches are concerned with the amount of data governing the tail and possible solution in order to make it viable is to either expand time span of the data or jointly model the both tails. Latter is sensible in the case of symmetry while former could be impractical since the very old data may deteriorate results instead of improve.

Given the stylized fact that the distribution of financial data is fat-tailed, EVT stands as the viable approach since it can account for the heaviness of the tail. In addition, one can model both tails of the distribution separately, thus also capturing the asymmetry.

2.4 Value at Risk and Expected Shortfall under EVT framework

EVT approaches, by its very nature, concentrates on the behavior of the extreme observations. Similarly, VaR and ES are concerned with the extreme quantiles of the distribution; hence, these concepts naturally match with each other. In addition, EVT makes it possible to estimate quantiles of distribution (even out of sample quantiles) and thus, one can calculate VaR and ES using EVT approaches.

Simply, VaR and ES estimates under GPD are calculated as follows[18]:

$$\widehat{VaR^{\alpha}(X)} = u + \frac{\hat{\beta}}{\hat{\xi}}\left[\Big(\frac{n(1-\alpha)}{k}\Big)^{-\hat{\xi}} - 1\right], \tag{2.9}$$

$$\widehat{ES^{\alpha}(X)} = \frac{1}{1-\alpha}\int_{\alpha}^{1} \widehat{VaR_{\varphi}(X)}\,\mathrm{d}\varphi = \frac{\widehat{VaR^{\alpha}(X)}}{1-\hat{\xi}} + \frac{\hat{\beta}-\hat{\xi}u}{1-\hat{\xi}}, \tag{2.10}$$

[18] Detailed derivation of VaR and ES under GPD can be found in McNeil *et al.* (2005). Similarly, derivation of VaR and ES under GEV and non-parametric Hill estimator can be found in Longin (2000) and Danielsson & Vries (1998), respectively.

where $\alpha \in (0, 1)$ and it represents the confidence level, n is the total number of observations, u is the threshold level and k is the number of extremes exceeding this threshold u. Moreover, $\hat{\xi}$ and $\hat{\beta}$ represent the estimated shape and scale parameters using MLE.

In the case these above formulations is applied to raw returns, i.e. without taking the dependence structure of time series into account, it is called an unconditional EVT approach to VaR and ES estimation. However, conditional estimates, i.e. estimates that conditional on past information, are always of interest because of the stylized fact that financial time series exhibit autocorrelation. Conditional methods can react quickly to the changes in the market by modelling the stochastic structure of the volatility, thus yielding more accurate risk estimates than unconditional methods, especially during the stress periods (McNeil & Frey 2000; Danielsson & Vries 2000). However, conditional methods show large variability over time, which makes unfavorable for the regulator purposes. For example, 2 step procedure that we explain the implementation in the following Chapter 3, can be given as a conditional EVT approach to VaR and ES, in which one can model the conditional mean and variance using ARMA-GARCH family models and then use the standardized residuals (extracted) in the calculation of VaR and ES using above equations (McNeil & Frey 2000).

2.5 Comparative Studies

EVT approach to VaR computation was considered by many empirical studies concerned with the financial risk measurement but the first and the seminal applications can be found in the studies of **??**.

Danielsson & Vries (2000), using data from several U.S. stocks, compare the accuracy of one step ahead VaR predictions from two different EVT methods and two conventional methods, namely J.P Morgan RiskMetrics and historical simulation. Authors state that the RiskMetrics method performs the best when confidence level is equal to 95%. They explain this outcome as in the 95% confidence we are sufficiently inside the sample. However, when going further in the tail, i.e. 99, 99.5%, RiskMetrics consistently underestimates the risk, with ever larger biases. In other words, we are not sufficiently in the tail at the 95% confidence in order EVT to be efficient. But at the 99 and 99.5% confidence, EVT is appropriately employed and outperforms other methods as a result of more accurate VaR prediction since we are concerned with extreme events.

Similarly, Pownall & Koedijk (1999) investigated the relative performance of EVT and RiskMetrics techniques in terms of VaR calculation during Asian crisis and found that the conditional EVT approach provides better VaR estimations than the RiskMetrics.

Moreover, Longin (2000) compared four classical VaR methods (i.e. normal distribution modelling, HS, GARCH and Exponentially Weighted Moving Average (EWMA)) with EVT on the S&P500 index period covering January 1962 to December 1993. This study was the first to consider both short and long positions (both tails) in the application of EVT approach to VaR and stated that the both left and right tails are important. The author concludes that EVT is better suited for the calculation of VaR, as it explicitly considers the event risk by focusing on extreme events.

2 step procedure which is firstly implemented in the study of McNeil & Frey (2000), using data from S&P500 index, DAX index, US dollar/British pound exchange rate, BMW share prices and price of gold. Authors employ several backtesting procedures to compare accuracy of VaR and ES estimates under the conditional EVT (2 step procedure), unconditional EVT (GPD), conditional normal (GARCH with normal innovations) and conditional t methods (GARCH with t innovations). Results show that the conditional EVT approach is superior to the rest of the methods in terms of VaR computation. Conditional EVT method gives better estimates than methods that do not account for heavy-tails and stochastic nature of the volatility. Moreover, they conclude that the fat-tailed distributions, preferably EVT, should be considered for modelling the innovations in the case of ES estimation employed.

In addition, Neftci (2000), by estimating with daily data from several exchange and interest rates (overall period is from January 1990 to October 1995), compares the accuracy of VaR in the EVT method with the traditional one based on the normal distribution. Author finds that the VaRs calculated under EVT method is 20%-30% greater than the standard approach. Based on the empirical out-of-sample forecasting, results remarkably addresses the EVT since VaRs calculated are more precise than the standard approach. Therefore, author concludes that the VaR estimations under EVT approach are more robust and accurate for capturing both the occurrence rate and the size of extreme events in financial markets. However, Tolikas et al. (2007) who studied the data from DAX index reached the conclusion that, when using sufficiently large amount of data, historical simulation can perform equally good as the EVT based methods.

Similarly, Gencay & Selcuk (2004) investigates the relative performance of six models, i.e. variance-covariance approach with both normal and student-t distribution, historical simulation, unconditional and conditional EVT approaches, for estimating one day ahead VaR considering both tails at different quantiles. Using daily data from nine emerging (stock) markets, they compare the relative performance of each model in a dynamic setting (backtesting). Their results show that especially at higher quantiles, i.e. 99% and more, conditional EVT method clearly dominates others in terms of VaR estimation. Hence, they conclude that the EVT is indispensable for risk management in emerging economies, particularly in VaR calculations.

citebali1, using daily data from Dow Jones Industrial Average (DJIA) equity index for the period May 1896 to December 2000, i.e. 28,758 observations, compares the out-of-sample forecasting performance of EVT based methods against the normal and skewed-t distribution. With the application of unconditional (frequency of violations) and conditional coverage (dependence of violations) tests in backtesting, they found that the extreme value distributions provide more precise VaR estimates rather than normal and skewed-t distributions.

Kuester *et al.* (2006), compares the out-of-sample VaR performance of several univariate unconditional and conditional methods, using daily return data from NASDAQ Composite Index period covering more than 30 years, i.e. 7681 daily observations. They find that most approaches show inadequate performance, although some of them are acceptable under the regulatory norms. Moreover, they find that the conditional two step procedure which combines GARCH family model with a GPD performs overall best, following conditional filtered historical simulation.

Many other comparison studies can be found[19], but overall, the following conclusions emerge:

- EVT based risk estimates are at least as precise as estimates from other methodologies regardless of the analyzed confidence level.

- When one goes further in the tail, i.e. 99% and more, EVT based methods outperform other methods in terms of VaR, since it is a theory being concerned with extremes.

Hence, in line with the above conclusions, the main goal of this study is to focus on the appropriateness of the EVT for measuring the probability of extreme

[19] Those are Andjelic *et al.* (2010),Assaf (2009),Cifter (2011),Djakovic *et al.* (2011),Furio & Climent (2013),Lee *et al.* (2006),Mutu *et al.* (2011),Nystrom & Skoglund (2002),Samanta & LeBaron (2005),Trzpiot & Majewska (2010),Zikovic & Aktan (2009),Andersen & Pedersen (2010), etc.

events and assessing the risk defined in terms of VaR and EVT. Briefly, this study examines the predictive performance of EVT based methods by relatively comparing with the variety of methods. Namely, standard normal distribution, historical simulation, unconditional EVT method based on GPD, RiskMetrics-EWMA method, two best selected conditional ARMA-GARCH family methods with normal and student-t distributed innovations, filtered historical simulation and conditional method based on GPD (2 step procedure). Respective comparison for VaR is based on so called backtesting procedure by employing the unconditional coverage test developed by Kupiec (1995), as well as independence and conditional coverage tests developed by Christoffersen (1998). Similarly, ES estimates are backtested using the tests developed by McNeil & Frey (2000). Implementation and detailed review of the above mentioned methods and quantitative tests are presented in the following chapter.

Chapter 3

Methodology

In this chapter we first briefly present the dynamic models considered in this study. Next, we document competing univariate methods and explain how they are implemented. Finally, we present the Backtesting methodology which we base our criterion for competing methods.

For the purposes of risk measuring the price changes are quite often used. This measure can be expressed in different ways such as absolute, relative or logarithmic form (price change). In the case when the price change relative to its initial value is used, one is then actually considering a price return. Following standard practice in financial literature and risk management, this study uses logarithmic price changes to measure the change in value of a portfolio which is also known as continuously-compounded return. The most common reason for using the returns instead of prices in practice is their more "favorable" statistical properties. In addition, the stock price series are typically integrated of order one, i.e. are non-stationary, whereas the logarithmic returns are already stationary and ergodic, which makes them more convenient to work with (Campbell *et al.* 1997). One of the reasons for choosing the logarithmic returns for this study is the fact that these returns are assumed to be normally distributed and serially independent under the Geometric Brownian Motion (GBM) for the asset price. Accordingly, for a long time, a big part of the available and often used methodologies regarding the risk management is based on the GBM assumption.

Therefore, Logarithmic return R_t at time t is calculated as follows:

$$R_t = \ln P_t - \ln P_{t-1} = \ln \left[\frac{P_t}{P_{t-1}} \right], \tag{3.1}$$

where P_t stands for the price of an underlying stock at time t. Note that t represents one business day in this study. For the purposes of simplicity, from now on-wards, logarithmic returns will be refereed just as returns. Along with this, note that this study is focused on equity securities only, and the methods description that follows can be as well applied to broader set of asset types.

3.1 Dynamic Models for Varying Volatility

This section will briefly present the univariate dynamic models that are used to account for autocorrelation, time varying volatility and asymmetric effects of positive and negative shocks. This stylized facts were already elaborated in Section 2.2 and shall be captured with the aim of adequate modeling.

Let's start with the following assumption; where individual return series processes $\{R_t\}_{t\in\mathbb{Z}}$ of the considered stocks are adapted to the filtration $\{\mathcal{F}_t\}_{t\in\mathbb{Z}}$ that represents the accrual information over time. The algebra $\mathcal{F}_t$ represents the available information at time t and typically this will be the information contained in past and present values of the time series itself $(R_s)_{s\geq t}$. The corresponding filtration is known as the natural filtration. In other words, the following Autoregressive Moving Average (ARMA) process can be thought of as putting particular structure on the conditional mean μ_t of the process. Autoregressive Conditional Heteroskedasticity (ARCH) and GARCH processes will later be seen to put structure on the conditional variance σ_t^2.
Next, assume that each of these individual return series follows a stationary ARMA process of order p and q (ARMA(p, q)) with the following form[1]:

$$R_t = \mu_t + \varepsilon_t, \tag{3.2}$$

where μ_t is the conditional mean and ε_t is the innovation with conditional mean of zero and conditional variance of σ_t^2, i.e. $\varepsilon_t \sim (0, \sigma_t^2)$, measurable with respect to $\mathcal{F}_{t-1}$.

The family of classical ARMA processes is widely used in many traditional applications of time series analysis. By allowing the conditioning of the process mean on past realizations, these models can capture the serial (linear) dependence structure of the time series. Those are covariance-stationary processes that are constructed based on white noise. Whether the process is strictly stationary or not, will depend on the exact nature of the driving white noise,

[1] Note that this is a constant mean process where orders p and q are equal to 0

also known as the process of innovations. If the innovations are iid or simply from a strictly stationary process, then ARMA will also be a strictly stationary process.

Thus, assume the following model:

$$R_t = \omega + \sum_{i=1}^{p} \phi_i R_{t-i} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \varepsilon_t, \qquad \varepsilon_t = \sigma_t z_t \tag{3.3}$$

where $\{z_t\}_{t\in\mathbb{Z}}$ is a strict white noise process with zero mean and variance equal to 1, i.e. $z_t \sim N(0,1)$.

Furthermore, three GARCH-type models are considered in this study, for the purpose of modelling the dynamics of the aforementioned strictly positive-valued process $\{\sigma_t\}_{t\in\mathbb{Z}}$. Those models are ARCH, Exponentially Generalized Autoregressive Conditional Heteroskedasticity (EGARCH) and Glosten-Jagannathan-Runkle Generalized Autoregressive Conditional Heteroskedasticity (GJR-GARCH) model. In order to choose the dynamic time series model and the respective orders for the empirical estimation, this study will base its evaluation on different information criteria, such as Akaike, Bayesian, Schwarz and Hannan-Quinn for the fitted models. In addition, the models' ability to fit and pre-whiten the respective return series will be tested by some diagnostic tests (See Section 4.3).

3.1.1 GARCH Model

The need to capture serial volatility dependence in financial return series, gave a rise to different models. Accordingly Engle (1982) introduced the ARCH) for modelling the conditional variance regarding the time varying volatility. The author suggested this variance to be modelled as a linear function of the squared past innovations and thus, the general form of the ARCH(q) model of order q can be specified as follows:

$$\sigma_t^2 = \omega + \sum_{j=1}^{q} \alpha_j \varepsilon_{t-j}^2, \tag{3.4}$$

where, $\omega > 0$ and $\alpha_j \geq 0,$ for $j = 1 \cdot 2 \cdots q$, in order to keep the conditional variance positive.

In order to adequately fit the equation above, a large number of squared lagged residuals shall be considered. In other words, the ARCH lag q will be large which makes the model more complicated and hence, less convenient to use. In line with this, Bollerslev (1986) proposed a more parsimonious specification, i.e. the model where the persistent volatility movements can be captured with no need to estimate the large number of coefficients as opposed to the ARCH family of models. Therefore, the respective model is known as GARCH MODEL which, in its general GARCH(p, q) form, can be represented as follows:

$$\sigma_t^2 = \omega + \sum_{i=1}^{p} \beta_i \sigma_{t-i}^2 + \sum_{j=1}^{q} \alpha_j \varepsilon_{t-j}^2, \tag{3.5}$$

where, $\omega > 0,\ \alpha_j \geq 0,$ for $j = 1 \cdot 2 \cdots q,$ and $\beta_i \geq 0,$ for $i = 1 \cdot 2 \cdots p.$

GARCH family models represent the generalized ARCH-type model version that extends the basic ARCH specification by allowing for the lag structure to be more flexible. Within such a structure, the squared conditional volatility σ_t^2 is defined as a linear function of both past innovations ε_{t-j}^2 and past squared conditional volatilities σ_{t-i}^2. According to the GARCH, the best forecast for the next day variance is a weighted average of all the constant, i.e. long term variance, today's variance estimate, and new information expressed by the most recent squared residuals (respective information was unavailable when today's prediction was made). The empirical studies evidence that the GARCH-type models are effective models for the future conditional variance prediction.

Even though these two family of models successfully account for both the autocorrelation and volatility persistence, the asymmetry that was empirically observed in financial time series remains an unsolved issue for the risk management. Symmetry of ARCH and GARCH family models indicate that positive and negative shocks (innovations) have the same effect with regards to volatility. In other words, neither the plus nor the minus shock sign will affect the conditional volatility and the only squared innovations will be included in the equation for conditional variance (Campbell *et al.* 1997).

The empirical literature also evidenced that the positive shocks to volatility have higher correlation with the negative shocks to returns compared to the positive shocks to the returns. This might be due to the fact that negative returns shocks are found to drive the volatility up, or this causality might go the opposite way. Nevertheless, this larger impact of negative shocks is not in line with the stylized facts. As a matter of fact, such leverage represents the debt to equity ratio, i.e. the price leverage increases as the price falls, is known as the leverage effect (Black 1976). In the next sections, it is introduced a two different GARCH-type models extensions in order consider the respective asymmetry.

3.1.2 EGARCH Model

In order to find the model that captures the asymmetries in the return-volatility relationship, Nelson (1991) suggested the EGARCH which, in its general EGARCH(p, q) form, can be represented as follows:

$$\begin{aligned} \ln(\sigma_t^2) = \omega &+ \sum_{i=1}^{p} \beta_i ln(\sigma_{t-i}^2) + \sum_{j=1}^{q} \gamma_j \frac{\varepsilon_{t-j}}{\sigma_{t-j}} \\ &+ \sum_{j=1}^{q} \alpha_j \Big[\frac{|\varepsilon_{t-j}|}{\sigma_{t-j}} - E\big(\frac{|\varepsilon_{t-j}|}{\sigma_{t-j}}\big)\Big], \end{aligned} \tag{3.6}$$

Within the EGARCH model context, there is no need for inequality constraints to ensure the conditional variance to be positive since the conditional variance is parameterized using natural logarithm as opposed to lagged conditional variances. Hence, the standardized shocks' expected value $E(|\varepsilon_{t-j}|/\sigma_{t-j})$, will depend on the assumed distribution of shocks.

In general, EGARCH models are advantageous over the GARCH models due to 3 main reasons. Firstly, EGARCH models capture the asymmetry in the tail of the distribution by allowing the different effects of volatility to come from the positive and negative shocks. That is, if one would assume the leverage effect to hold, then the asymmetry coefficient γ_j shall be negative, whereby larger effect on the future volatility shall be found from negative opposed to the same size positive shocks. Secondly, EGARCH models also allow for the greater volatility impact coming from the large shocks. Thirdly, EGARCH models capture the volatility persistence (clustering) by one coefficient only β_i as opposed to the GARCH models that capture with the β_i and α_j coefficients combined together.

3.1.3 GJR-GARCH Model

In line with Nelson (1991) who aimed to find the model that will unambiguously capture the asymmetries in the return-volatility relationships, Glosten *et al.* (1993) accordingly proposed a model that can capture the respective asymmetry by permitting the past negative and positive shocks to have different effects on the conditional variance. This model is known as GJR-GARCH and it differs from the EGARCH model in terms of capturing asymmetry. Namely, within the EGARCH model context, the coefficient γ_j that is applied to the innovation ε_{t-j} is used for measuring the asymmetry, whereas the GJR-GARCH model represents the asymmetry by the coefficient γ_j that enters to the model as the Boolean indicator. Hence, the general GJR-GARCH(p, q) form can be represented as follows:

$$\sigma_t^2 = \omega + \sum_{i=1}^{p} \beta_i \sigma_{t-i}^2 + \sum_{j=1}^{q} \alpha_j \varepsilon_{t-j}^2 + \sum_{j=1}^{q} \gamma_j \varepsilon_{t-j}^2 I_{\{\varepsilon_{t-j}>0\}}, \tag{3.7}$$

If one would assume the leverage effect to hold, the asymmetry coefficient γ_j shall be positive. Accordingly, in the above equation I represents the function of Boolean indication taking the value of zero in the case of positive shocks ($z_t \geq 0$) and the value of 1 in the case of negative ones ($z_t < 0$).[2]

3.1.4 Quasi Maximum Likelihood Estimation for Dynamic Models

In order to fit the GARCH-type models, one shall start off from a certain assumptions for the innovation distribution. In general, it is assumed that the innovation distribution of the return series process $\{R_t\}_{t\in\mathbb{Z}}$ is not conditionally normally distributed, but instead, model will be estimated by assuming normally-distributed innovations to fit the GARCH models. In such a way, it is appled a Quasi Maximum Likelihood Estimation (QMLE),also known as Pseudo Maximum Likelihood Estimation. QMLE is applied to the parameters as if the normality assumption of innovations was satisfied, but even if innovations were not normally distributed, it would give us a parameter and conditional volatility estimates. Speaking of that, Bollerslev & Wooldridge (1992) found the QMLE estimators to be asymptotically normally distributed and consistent, despite the possibility of inadequate normality assumption.[3]

3.2 Method Selection and Implementation

This part of the thesis outlines the selected univariate methods and explains how they are implemented for the risk estimation in terms of VaR and ES. These methods are known as univariate as they are dealing with only one risk factor; that is, logarithmic returns calculated from daily closing price series. For each

[2] Note that GJR-GARCH model is similar to the Threshold GARCH model, the only difference is that latter one models conditional variance, not the standard deviation (Zakoian 1994).

[3] Prove of asymptotically normal and consistent QMLE estimators in the context of general GARCH(p, q) model can be found in Kokoszka *et al.* (2003)

of the methods considered, one-day ahead VaR and ES estimates are extracted at the 95, 99 and 99.5% confidence levels. In addition, this study does not only consider the left tail that corresponds to the long position but also the right tail that corresponds to the short position on the respective underlying asset. However, for the sake of simplicity, the methods discussion that follows will be done only for the left tail[4].

This study method selection is mainly based on the stylized facts accountability. Moreover, their practical usage is also considered. Naturally, not all of the stylized facts are accounted within each method. However, such methods comparison allows to realize which of the stylized facts should be captured to get better risk estimation in terms of VaR and ES.

This study considers the following methods: HS, Gaussian Normal method (Normal), Static EVT method (GPD), EWMA, GARCH type methods (GARCH-n, GARCH-t), FHS method and Dynamic EVT method. The latter method is also known as Dynamic Generalized Pareto Distribution (DGPD) and it is the main method under the focus of this study. All in all, methods used in this study are however incomplete and thus, cannot be considered as the indicators of possible strategies.

3.2.1 Historical Simulation and Filtered Historical Simulation Methods

Historical Simulation

The first method considered is HS. Arguably, it is one of the most popular risk estimation method. HS is a simple non-parametric risk estimation method based on historical returns distribution[5]. Namely, the risk of the portfolio can be simply estimated by taking into consideration the past risk factor evolutions. Next, the non-parametric feature implies no distributional assumptions regarding the true empirical returns distribution, i.e. returns are not subject to any distributional constraint and hence, there is not any respective parameter derivation. This represents the main advantage of HS method. Furthermore, the past returns volatility is considered to be unconditional and constant over the sample period, which leads to very slow incorporation of changes in the market. Another crucial assumption of the HS method is that the past re-

[4] The same analysis for the right tail is done by simply changing the sign of the return series and then applying the same methods as described in the case of the left tail

[5] Historical simulation technique with its variations can be found in Mahoney (1995)

turns distribution is a good indication and representative of the likely future returns. This assumption from another side represents one of the HS method drawbacks, since past returns can be very poor tool for predicting the future extreme events. In this case, estimates based on past observations that might not repeat may lead to inaccurate future forecasts. HS method also gives the equal importance to the historical and recent observations and thus, does not capture the impact of the volatility increase in the recent data. For example, if the observed period is too large, the most recent observations will have the same impact as the most distant ones since they are weighted equally, Brooks *et al.* (2005).

The fact that there are not any distributional assumptions makes the HS method implementation quite simple and straightforward. Basically, the first step is to make the logarithmic return series in an ascending order. Then, the historical VaR will be the absolute value of the return in such ordered series that corresponds to the derived index value based on a certain confidence level. For example, if the modeled number of logarithmic returns is 1000 (n=1000) and confidence level is $\alpha = 95\%$, then the VaR estimate will be the fifty first sequence of the 1000 logarithmic returns made in ascending order. Similarly, the respective ES estimate will correspond to the average of the values that exceed the fifty first sequence, that is VaR value. In mathematical terms, Let R_t denote the returns at time t, n is the amount of historical returns to be modeled, given the ordered sequence $R_1 \cdots R_n$, where $R_1 < R_2 < \cdots < R_n$, one-day ahead VaR and ES are calculated as follows:

$$\widehat{VaR}_{t+1}^{\alpha} = \mid R_{[n(1-\alpha)+1]} \mid, \tag{3.8}$$

$$\widehat{ES}_{t+1}^{\alpha} = \frac{\sum_{i=1}^{[n(1-\alpha)+1]} \mid R_i \mid}{[n(1-\alpha)+1]}, \tag{3.9}$$

To conclude, HS method shall be used with caution, since by its nature it cannot provide any extreme event prediction that is more extreme than the minimum value in the considered sample, or that is worse than the sample extreme return. Moreover, HS method is quite sensitive to the choice of a sample size. Therefore, a sample size choice can significantly affect and alter the value predicted by HS method. Nevertheless, as it was argued by Mahoney (1995), this method was found by the comparison studies to be a good benchmark. That is why this study applies the HS method as a benchmark against the more sophisticated alternatives.

Filtered Historical Simulation

As it was already pointed out, HS method cannot account for the changing market dynamics, even though it is simple and straightforward. Accordingly, some researches attempted to offer some solution for such drawbacks and proposed different synthetic models that combine historical simulation approach with the parametric models. The most commonly used method out of those is a FHS that was introduced by the early studies of Hull & White (1998) and Barone-Adesi & Giannopoulos (1996); Barone-Adesi *et al.* (1998). This method is based on the combination of parametric risk factor modelling with a non-parametric innovations modelling and thus, it is considered to be a semi-parametric method. Also, by applying the dynamic time series approach, FHS accounts for the time-varying volatility and imposes no distributional assumptions regarding the innovations. Hence, FHS uses the empirical distribution of innovations that allows an asymmetric (skewed) innovations and heavy-tailed distribution to be incorporated naturally.

This study applies the FHS method, by firstly fitting the ARMA-GARCH type model to the returns. In such a way serial correlation is removed (ARMA) and respective volatility clustering is accounted (GARCH). For the subsequent parameter estimation we use here a QMLE that was already introduced in Subsection 3.1.4 and get the estimates of the conditional mean μ_{t+1} and conditional volatility σ_{t+1} . Then, We use these estimates to get the standardized residuals:

$$\hat{Z}_s = \frac{R_s - \hat{\mu}_s}{\hat{\sigma}_s}, \qquad \text{where}, s = t - n + 1 \cdots \cdot t, \tag{3.10}$$

Assuming these standardized residuals as observations coming from the innovation distribution F_z, the conditional 1-period VaR and ES are calculated as follows:

$$\widehat{VaR}^{\alpha}_{t+1} = \hat{\mu}_{t+1} + \hat{\sigma}_{t+1}\widehat{VaR}^{\alpha}(Z), \tag{3.11}$$

$$\widehat{ES}^{\alpha}_{t+1} = \hat{\mu}_{t+1} + \hat{\sigma}_{t+1}\widehat{ES}^{\alpha}(Z), \tag{3.12}$$

The $\widehat{VaR}^{\alpha}(Z)$ and $\widehat{ES}^{\alpha}(Z)$ stands for the estimated VaR and ES from the empirical distribution of standardized residuals Z, respectively. These were obtained by applying the estimation procedure described within the unconditional HS method description.

3.2.2 RiskMetrics-Exponentially Weighted Moving Average

When it comes to the well know methods for volatility forecasting, EWMA was a main VaR calculations tool in the beginning of 1990s, after it became popularized by the JP Morgan RiskMetrics department. EWMA is conditional volatility method and the biggest strengths of those methods is that many of the financial time series common stylized facts such as persistence, time-varying volatility, volatility clustering, etc. can be easily captured. Moreover, EWMA is popular due to its simplicity and significant advantages over the traditional methods for volatility forecasting that are known as a Simple Moving Average models (SMA). Namely, while the traditional methods were based on the fixed or equally weighted moving averages, EWMA naturally relies on the exponentially weighted moving averages for the forecasts of the normal distribution variances (volatilities) and covariances (correlations) in terms of risk modelling.

In general, the EWMA method captures the dynamic volatility features by using the historical observations. As it was already mentioned, its main advantage was assignment of different weights to the observations. In other words, the highest weight will be assigned to the most recent observations and accordingly the lowest weight will be assigned to the oldest observation implying that the respective weights quickly decline with going back in time. Such weights assignment allows the volatility to have immediate response to the market booms as in such way volatility can react to the market jumps of a bigger scale. Thus, the advantage of this method is twofold. Specifically, besides the faster volatility reaction to market shocks due to bigger weight assignment to the recent data compared to the past one, this method also enables the exponential volatility decline following the large shocks because the respective "shock" observation weight will fail.

Assume R_t denotes daily logarithmic returns and it is subject to filtration $\{\mathcal{F}_t\}_{t\in\mathbb{Z}}$ that represents the available information set up to time t. Conditional return process can be written as $R_t|\mathcal{F}_{t-1} \sim N(\mu_t, \sigma_t^2)$, where μ_t is the conditional mean and σ_t^2 is the conditional variance. According to the JP Morgan RiskMetrics (1996), the following expression could determine the model dynamics:[6]

$$R_t = \sigma_t \varepsilon_t, \qquad \varepsilon_t \sim N(0,1), \tag{3.13}$$

$$\sigma_t^2 = (1-\lambda)\sum_{t=1}^{n} \lambda^{t-1} R_t^2, \qquad 0 < \lambda < 1, \tag{3.14}$$

where n is the number of returns observations and λ represents the exponential (decay) factor ranging between 0 and 1, showing the volatility persistence. Accordingly $1-\lambda$ is a parameter showing the speed at which the volatility absorbs the market shocks. This parameter, in other words, determines the relative weights assigned to return observations, as well as the nominal data amount used in the volatility estimation. The most recent observation has $(1-\lambda)$ weight, the second most recent one has $(1-\lambda)\lambda$, etc. until the oldest observation that has $(1-\lambda)\lambda^{n-1}$ weight. Hence, we can conclude that the higher decay factor implies the longer past returns. If the decay factor equals to 1, the respective volatility will be solely explained by its past values.

Following the JP Morgan RiskMetrics (1996) again, assuming mean to be 0 and infinitive available data, volatility equation can be rephrased with a recursive substitution and then one day variance forecast can be expressed by EWMA estimator feature as follows:

$$\sigma_{t+1}^2 = \lambda\sigma_t^2 + (1-\lambda)R_t^2, \tag{3.15}$$

The two issues arise in terms of computation of the RiskMetrics-EWMA volatility estimates. Those are the distributional mean and a factor for exponential decay λ. Generally, one can find an optimal decay factors set for each covariance that will give the symmetric and positive definite covariance matrix. Accordingly, RiskMetrics popularized the method that chooses one optimal decay factor that will be used for the whole covariance matrix estimation. Also, based on JP Morgan RiskMetrics (1996), $\lambda = 0.94$ was found optimal in terms of one-day forecast, while $\lambda = 0.97$ was found optimal for forecasts of one month or 25 trading days. Hence, the large value of the exponential factor implies that the total variance would be merely affected by the current one.

[6] Note that this is a special type of GARCH model, known as Integrated GARCH (IGARCH) Model that imposes $\alpha = 1-\lambda$ and $\beta = \lambda$ and $\omega = 0$ coefficient constraints (Bollerslev 1986).

Then, the exponential factor close to unity increases the smoothness of time series (JP Morgan RiskMetrics 1996).

And finally, given the confidence level α, one-day VaR and ES estimates are calculated as follows:

$$\widehat{VaR}_{t+1}^{\alpha} = \hat{\sigma}_{t+1}\Phi^{-1}(\alpha), \tag{3.16}$$

$$\widehat{ES}_{t+1}^{\alpha} = \hat{\sigma}_{t+1}\frac{\Psi\big(\Phi^{-1}(\alpha)\big)}{1-\alpha}, \tag{3.17}$$

where, Φ^{-1} represents inverse function of the standardized normal distribution while Ψ stands for the analogous density function; next, $\hat{\sigma}_{t+1}$ represent the one-day ahead conditional variance forecasted by previous equation (Equation 3.15).

3.2.3 GARCH-n and GARCH-t Methods

As it was already mentioned, this study considers three dynamic GARCH-type models; namely, GARCH, EGARCH, and GJR-GARCH. In line with this, the specific dynamic model to be used in this study will be chosen in Section 4.3. Hence, in this section we will refer to the respective dynamic model to be used as the GARCH-type model.

GARCH-n and GARCH-t methods are two of the few conditional methods that are considered in this study. These methods are fitted with the dynamic ARMA-GARCH-type models using the Maximum Likelihood Estimator. In order to estimate the GARCH-n method, innovations are assumed to follow conditional normal distribution. However, empirical evidences suggested that standardized residuals have a tail that is often fatter compared to the normal distribution, i.e. leptokurtic. Hence, in order to account for this, GARCH model based on the assumption of t-distributed innovations will also be estimated, which is named as the GARCH-t method throughout this paper .

For the GARCH-n method, given the innovations that are assumed to be normally distributed and the confidence level α, the conditional one-day VaR and ES estimates can be calculated as follows:

$$\widehat{VaR}_{t+1}^{\alpha} = \hat{\mu}_{t+1} + \hat{\sigma}_{t+1}\Phi^{-1}(\alpha), \tag{3.18}$$

$$\widehat{ES}^{\alpha}_{t+1} = \hat{\mu}_{t+1} + \hat{\sigma}_{t+1}\frac{\Psi(\Phi^{-1}(\alpha))}{1-\alpha}, \tag{3.19}$$

where, Φ^{-1} represents inverse function of the standardized normal distribution while Ψ stands for the analogous density function; whereas, $\hat{\mu}_{t+1}$ and $\hat{\sigma}_{t+1}$ represent the GARCH estimates for the next period's mean and variance that are fitted using the normally-distributed innovations.

For the GARCH-t method, given the innovations that are assumed to be t-distributed and the confidence level α, the conditional 1-period VaR and ES estimates can be calculated as follows:

$$\widehat{VaR}^{\alpha}_{t+1} = \hat{\mu}_{t+1} + \hat{\sigma}_{t+1}\sqrt{\frac{(\hat{\nu}-2)}{\hat{\nu}}}t^{-1}_{\hat{\nu}}(\alpha), \tag{3.20}$$

$$\widehat{ES}^{\alpha}_{t+1} = \hat{\mu}_{t+1} + \hat{\sigma}_{t+1}\sqrt{\frac{(\hat{\nu}-2)}{\hat{\nu}}}\frac{g_{\hat{\nu}}(t^{-1}_{\hat{\nu}}(\alpha))}{1-\alpha}\Big(\frac{\hat{\nu}+(t^{-2}_{\hat{\nu}}(\alpha))^2}{\hat{\nu}-1}\Big), \tag{3.21}$$

where, t^{-1}_{ν} represents the inverse function of standardized t-distribution while g_{ν} stands for the analogous density function; whereas, $\hat{\mu}_{t+1}$ and $\hat{\sigma}_{t+1}$ represent the GARCH estimates for the next period's mean and variance that are fitted using the t-distributed innovations.

3.2.4 Dynamic GPD Approach–2 Step Procedure

DGPD is a conditional method, firstly proposed by McNeil & Frey (2000), that follows a two-step methodology for VaR and ES estimation. Namely, it is based on POT model under the EVT. In this method, the conditional return distribution (i.e. its central part) is modeled by non-parametric HS and the tails are modeled by parametric GPD. Hence this method is also known as semi-parametric approach for risk calculation. POT model is chosen mainly for two reasons. Firstly, the data is used more efficiently under this method as opposed to BMM model. Secondly, POT method also allows for VaR and ES estimation. As it was already mentioned, the underlying EVT assumption of iid series is often violated in typical return series as a result of conditional heteroskedasticity. Thus, in line with McNeil & Frey (2000), data series is filtered by using dynamic ARMA-GARCH type models in order to create an iid series. This procedure is known as pre-whitening process and it is applied before fitting GPD to the tails. This method is implemented in two steps (2 step procedure-2 step approach).

Within the first step, ARMA-GARCH type model fitted to the returns. For the subsequent parameter estimation QMLE is used to get the estimates of the conditional mean $\hat{\mu}_{t+1}$ and conditional standard deviation $\hat{\sigma}_{t+1}$. In such a way, estimates of standardized residuals are extracted as follows:

$$\hat{Z}_s = \frac{R_s - \hat{\mu}_s}{\hat{\sigma}_s}, \qquad \text{where}, s = t - n + 1 \cdot \cdots t, \tag{3.22}$$

these standardized residuals are assumed as observations coming from the iid innovation distribution F_z.

Within the Second step, GPD is fitted to the standardized residuals that are exceeding threshold u by using MLE. But firstly, threshold value u need to be chosen. As there is no generally accepted way of choosing the threshold value denoted by u(see Subsection 2.3.3), respective empirical studies proposed different tools for this issue. Namely, different graphical techniques were discussed by Embrechts *et al.* (1997) for this purpose, but this method would be impractical for this study as it would be necessary to test threshold value u in each step of the rolling estimation. Instead, this study follows McNeil & Frey (2000) by fixing the number of the threshold exceedances k and by equalizing the value u with the estimated empirical $q = k/n$ quantile from the standardized residuals distribution. Following McNeil & Frey (2000), number of threshold exceedances k are set to 100 out of 1000 in-sample observations. in regard to that, findings of McNeil *et al.* (2005) show that setting k to 100 when sample size equals 1000 observation shall yield good VaR and ES estimates[7]. Furthermore, the study of McNeil & Frey (2000) also confirms the GPD robustness respective to the choice of k. However, we will investigate the reasonability of the set exceedances $k = 100$ in our data, by studying the behavior of the estimated shape parameter and VaR estimations with varying exceedances and their corresponding threshold values.

After selecting the threshold value u and fitting GPD to the standardized residuals that are exceeding this threshold, conditional 1-period VaR and ES estimates can be calculated as follows:

$$\widehat{VaR}^{\alpha}_{t+1} = \hat{\mu}_{t+1} + \hat{\sigma}_{t+1}\widehat{VaR}^{\alpha}(Z), \tag{3.23}$$

[7] The findings of some similar studies, i.e. those of McNeil & Frey (2000) and Kuester *et al.* (2006), show that 80-150 exceedances interval is suitable for the t-distributed data with different degrees of freedom.

$$\widehat{ES}^{\alpha}_{t+1} = \hat{\mu}_{t+1} + \hat{\sigma}_{t+1}\widehat{ES}^{\alpha}(Z), \tag{3.24}$$

The $\widehat{VaR}^{\alpha}(Z)$ and $\widehat{ES}^{\alpha}(Z)$ stand for the estimated VaR and ES from the distribution of standardized residuals Z calculated using Equation 2.9 and Equation 2.10, respectively. Also note that we consider a method that only the second step of this procedure applied straight to the raw returns, i.e. not iid, in order to see the reasonability of the pre-whitening procedure, and we denote this method as (static) GPD.

3.3 Backtesting Methodology

In the preceding sections of this study, we presented different methods for one-day VaR and ES estimations at the time t based on the return distribution. When these estimation methods are repeatedly applied during some time period, i.e. testing window, one can monitor the methods performance and also compare how these perform relative to each other. Such a statistical monitoring technique is known as the backtesting method. Hence, this section will briefly present the backtesting methodology to compare the methods accuracy as well as their relative performance in each market. Moreover, given that the Basel Committee requires the banks to calculate the 10 day VaR for the purposes of market risk, the relative performance comparison of a different methods built for the ten-days holding period instead of a one-day might be relevant as well (?). Nevertheless, two main issues often arise when using the ten-day risk measures. Firstly, even though the risk measures are computed based on the assumption of a constant portfolio structure, the financial institutions mostly make daily portfolio adjustments. Secondly, according to the McNeil *et al.* (2005), such an overlapping ten-days return may result in dependence, which can make the statistical inference in the comparison even more complicated. But one could avoid this dependence issue if the backtesting would be based on non-overlapping periods (i.e. if, after each estimation, the estimation window would be moved ten-days forward). Nevertheless, this might worsen the statistical validity of respective tests as the observations number would be significantly reduced.

In order to backtest the methods on the historical returns series $R_1 \cdots R_m$, we will repeatedly apply the respective methods by using the so called rolling estimation window covering the n (daily) return observations (in-sample period). To be concrete, if we start from the first n series observations, the mentioned estimation window will be repeatedly moved one-day ahead resulting in $p = (m - n)$, one-day $\widehat{VaR}_t^{\alpha}$ and $\widehat{ES}_t^{\alpha}$ estimates. These estimates are further going to be backtested against the actual observed values R_{t+1} over the subsequent time period, i.e. testing window (out-of-sample period). Following McNeil & Frey (2000), we set the estimation window length to $n = 1000$ which leaves us with very long testing window span - around 4000. Whenever the observed realization is greater than the estimated ES (conditional or unconditional) $R_{t+1} > \widehat{VaR}_t^{\alpha}$, it is called a violation. This enables us to define a violation indicator and violation counter over the respective testing window; that is, $V = \sum_{j=1}^{p} I_{\{R_{t+j} > \widehat{VaR}_{t+j-1}^{\alpha}\}}$, where I represents the violation indicator and V stands for the violations counter over the testing window p. Hence, application of this rolling estimation technique will allow us to compare the resulting VaR and ES estimates against the actual return value and accordingly to compare the methods accuracy and their relative performance.

We will start the methods evaluation regarding their VaR estimation ability by comparing the actual (observed) number of VaR violations (i.e. exceedances) against the expected number of VaR violations given the different confidence levels used in this study; that is, 95, 99 and 99.5% . Regarding the ES estimates, the comparison of actual returns when the VaR violation occurred will be used for evaluating the methods performance. Here it would be important to note that while testing a particular hypothesis, two types of error might occur. The first one is so called type I error occurring when the correct model is rejected and the second one is called type II error occurring when the incorrect model is accepted. Within the risk management context, the type 2 error may be very costly (certainly more costly than the type 1 error) which yields a need to impose high threshold for accepting the risk model validity. In addition, the choice of appropriate confidence level regarding the VaR and ES estimations has some implications. Namely, larger confidence level for VaR will result in smaller violations number which will make the model more difficult to validate, i.e. accept. For example, if one would consider a 95% level, he would observe more violations as compared to the case of 99% level and hence should get a better test regarding the model accuracy. Nevertheless, this study takes into consideration this issue and solves it by considering a very long data span for the respective testing window. That is, we will use 4000 observations for the testing window, which leaves us with 200, 40 and 20 expected violations for 95, 99 and 99.5% confidence levels, respectively.

The subsequent sections will present the criterias used for the comparison of methods applied in this study as well as the different statistical tests for comparison of these criteria which are built on the violations frequency and time dynamics [8]

3.3.1 Methods for Value at Risk Backtesting

The methods valuation in terms of VaR will be done according to two criteria. Namely, these criteria refer to matching the actual VaR^{α} violations portions κ with the expected violations portions $\Pi = (1 - \alpha)$ and whether the VaR violations are scattered randomly over the time, all given the confidence level α.

The case when the actual violations portion is large in comparison to the expected violations portion $\kappa > \Pi$ usually indicates that the method is underestimating the risk and hence, this might result in financial distress due to lack of the risk capital. Oppositely, the case when the actual violations portion is small in comparison to the expected violations portion $\kappa < \Pi$ usually indicates that the method is overestimating the risk, which often results in unnecessarily large risk capital allocation. As it was argued by Gencay & Selcuk (2004), if we are not binding the policy purposes, the case of smaller actual violation portion against its expected value does not have to be better than the opposite case. As a result of this, from the risk managers point of view, aim is to get close to the equality $\kappa = \Pi$.

If we assume the return observations sample $R_1, \cdots R_m$ and the associated VaR estimates $\widehat{VaR}^{\alpha}_{n+1}, \cdots \widehat{VaR}^{\alpha}_{m}$, one might get the actual κ and expected violations Π, $\kappa = \Pi$ match; but in the case when the violations do not fall randomly in time but are grouped within the same short time period, the respective financial distress risk will be significantly higher than it would be in the former case. Such a group of violations might indicate the higher probability of having a new VaR violation given the respective violation occurrence, i.e. positive violations dependence. In the empirical literature, this was found to be an indicator for the model misspecification.

[8] These statistical tests are described in details in the studies of Kupiec (1995), Christoffersen (1998), McNeil & Frey (2000).

Unconditional Coverage Test for the frequency of VaR violations

In order to test whether there is a significant difference between the actual violations portions κ and expected violations portions Π given some method, we will use the Unconditional Coverage Test that was proposed by Kupiec (1995). This test is applied by testing the Null Hypothesis that looks as follows: $H_0 : \kappa = \Pi = 1 - \alpha$. Within this hypothesis, the binomial distribution is used with parameters p and $(1-\alpha)$ for calculating the probability to detect violations V in the respective testing window $p = m - n$. Accordingly, the Likelihood ratio statistics below shall be applied to test the statistical significance of the $\kappa = \Pi = 1 - \alpha$ equality.

$$LR_{uc} = -2\ln\left[\frac{(1-\Pi)^{p-V}\Pi^{V}}{(1-\hat{\kappa})^{p-V}\hat{\kappa}^{V}}\right] \sim \chi^2(1), \tag{3.25}$$

where, p represents the number of observations within the respective testing window, V stands for the number of violations, and $p-V$ stands for the number of non-violations. Next, the $\hat{\kappa} = V/p$ is used for the estimation of κ violations portion. In addition, this test is characterized by the asymptotic chi-squared distribution with the degree of freedom equal to one $\chi^2(1)$.

This test has two main drawbacks. First, as it was already recognized by the author himself, the statistical weakness of this test is reflected in the fact that the respective sample size corresponds to the current regulatory framework of 1 year. Second, there is a probability of this test falling to reject a model producing the grouped violations, as it considers the returns frequency only, without the time of their occurrence. Therefore, the backtesting methodology shall not be based on only the Unconditional Coverage Test (Campbell 2005). In addition, many other authors found that this test cannot be used in the case of small samples as it does not count for the volatility clustering phenomenon; hence, it shall not be used on its own for the VaR estimates accuracy.

Independence Test for the time dynamics of VaR violations

Even though the focus of the Unconditional Coverage test is based on the violations frequency solely, the respective theoreticians argue that these violations are still expected to be scattered randomly in time. Whereas, the adequate VaR methods should react to the volatilities and correlations that are changing such that violations are mutually independent, those less adequate ones usually yield a successive violations sequence (Finger 2005). Hence, as the catastrophic events are much more likely to happen when the large losses occur in rapid sequences as opposed to case of single violations happening sporadically, the detection of the violations clustering is the main expectation of the VaR users.

In regard to this, Christoffersen (1998) developed a so-called independence test for the time dynamics of violations that attempts to solve the respective issue by examining the VaR violations occurrence time; thus, taking into account the independence of violations. In other words, the respective Null Hypothesis of independence states that the probability of violation in the next day is not depended upon the occurrence of violation today. In order to test this Null Hypothesis, we will define the following set up that was already created by Christoffersen (1998). Let a i and j denote a conditions that will occur today and tomorrow, respectively, and also assume that these can take values of either 0 or 1 that stand for the non-violation and violation, respectively. Next, lets represent the observations by p_{ij} once when i and j occurred in time, successively. The resulting outcome can be hence presented with the help of 2×2 table below:

	$i = 0$	$i = 1$	
$j = 0$	p_{00}	p_{10}	$p_{00} + p_{10}$
$j = 1$	p_{01}	p_{11}	$p_{01} + p_{11}$
	$p_{00} + p_{01}$	$p_{10} + p_{11}$	p

Next, let κ_i denote the estimated probability to observe the violation tomorrow that depends on the condition i which is today. In such a way the violation proportions $\hat{\kappa}_0$ and $\hat{\kappa}_1$ may be estimated as follows:

$$\hat{\kappa}_0 = \frac{p_{01}}{p_{00} + p_{01}}, \qquad \hat{\kappa}_1 = \frac{p_{11}}{p_{10} + p_{11}}, \qquad \textit{and} \qquad \hat{\kappa} = \frac{p_{01} + p_{11}}{p_{00} + p_{01} + p_{10} + p_{11}},$$

The model used for VaR estimation will be accurate if the tomorrow's violation is not depended upon the occurrence of violation today. In mathematical terms, the respective null hypothesis takes the form $H_0 : \hat{\kappa}_0 = \hat{\kappa}_1$. Accordingly, the Likelihood ratio statistics below shall be applied to test the statistical significance of the $\hat{\kappa}_0 = \hat{\kappa}_1$ equality:

$$LR_{ind} = -2 \ln \left[\frac{(1 - \hat{\kappa})^{p_{00} + p_{10}} \hat{\kappa}^{p_{01} + p_{11}}}{(1 - \hat{\kappa}_0)^{p_{00}} \hat{\kappa}_0^{p_{01}} (1 - \hat{\kappa}_1)^{p_{10}} \hat{\kappa}_1^{p_{11}}} \right] \sim \chi^2(1), \tag{3.26}$$

Conditional Coverage Test for the frequency and time dynamics of VaR violations

Furthermore, Christoffersen (1998) proposed the joint test of independence and correct coverage that is known as the Conditional Coverage for the VaR violations frequency and time dynamics. By combining the mentioned two tests, i.e. correct violations portion and violations independence; this synthetic test allows us to test the both properties of the adequate VaR method. Thus, the Conditional Coverage test will sum the both former likelihood ratio tests and will look as follows:

$$LR_{cc} = LR_{uc} + LR_{ind} \sim \chi^2(2), \tag{3.27}$$

Since both LR_{ind} and LR_{uc} are characterized by the chi-squared distribution with the degrees of freedom equal to 1 $\chi^2(1)$, the conditional coverage will be also characterized by the asymptotically chi-squared distribution, but with the degrees of freedom equal to two $\chi^2(2)$.

This approach has a significant contribution that is reflected in its capability to test the separate hypothesis for violations independence and frequency together with the joint hypothesis of VaR method having the accurate independent violations frequency. The above depicted Christoffersen (1998)'s framework permits us to inspect whether the grouped violations, or inaccurate coverage, or both together, caused the test failure. Basically, this can be tested by the separate likelihood ratio tests statistics calculation that will, for the respective critical values, use the chi-squared distribution with degrees of freedom equal to one. As argued by Campbell (2005), there is a possibility of having the joint test passed without accordingly having the separate likelihood ratio tests both successfully passed. This implies that the respective tests should be individually tested no matter of the joint test results.

3.3.2 Methods for Expected Shortfall Backtesting

As it was already discussed in Section 2.1, ES stands for the convenient risk measure as it satisfies the property of subadditivity and thus being coherent. Therefore, we want to make the becktesting of our ES estimates. This can be done by checking for the difference between the tomorrow's return R_{t+1} and today's expected shortfall estimate ES_t that depends on the tomorrow's return R_{t+1} exceeding the today's VaR estimate $R_{t+1} > \widehat{VaR}_t^{\alpha}$. Hence, given the VaR violation occurrence, we can run the methods evaluation according to the ES estimate and observed return discrepancy. Formal test suggested by McNeil & Frey (2000) where they defined the exceedance residuals υ_t, i.e. the residuals calculated in the event of VaR violation, looks as follows:

$$\upsilon_{t+1} = \frac{R_{t+1} - ES_t^\alpha}{\sigma_{t+1}} = Z_{t+1} - E\big[Z|Z > Z^\alpha\big], \tag{3.28}$$

Given the model $R_t = \mu_t + \sigma_t Z_t$, exceedance residuals υ, conditional on VaR violations $R_{t+1} > VaR_t^\alpha$ (or equivalently $z_{t+1} > z^\alpha$), are iid, i.e. $\upsilon_t \sim N(0,1)$. It is clear that forming an empirical version of these exceedance residuals on the days when VaR violation occur $R_{t+1} > VaR_t^\alpha$ is possible by replacing the estimates of expected shortfall $\widehat{ES}_t^\alpha$ and the standard deviation $\hat{\sigma}_{t+1}$. Under the Null Hypothesis of correct ES estimates and correctly modeled parameters, $E\big[z|z > z^\alpha\big]$, these residuals shall have an iid properties, i.e. zero mean and unit variance. In order to test this hypothesis, bootstrapping technique is used as it makes no assumption regarding the distribution of exceedance residuals. Authors then perform a one sided t-test in contrast to the alternative hypothesis of residuals having mean bigger than zero. In other words, under this alternative hypothesis, an expected shortfall is assumed to be systematically underestimated as this is the expected failure direction.[9]

[9] This test can be conducted for both the plain and standardized residuals.

Chapter 4

Empirical Results

Financial crisis of 2007-2008 caused worldwide financial losses which was accompanied by the general industrial and economic slowdown. Morevoer, the effects of crisis can be still felt due to the global financial network and contagion. Hence, the need to predict such events with accuracy and to allow individuals, firms and institutions to prevent or to be ready for such events becomes essential. The motivation for this study was hence derived from the need to find the best method that will predict extreme events. In order to achieve our aim, we selected five European stock markets to analyze and we applied 8 methods in total to compare their accuracy, correctness and predictability.

This chapter presents the data and empirical results obtained following the structure of the Methodology. After introducing the selected data, we will first start by the preliminary analysis and descriptive statistics where we provide the general information regarding our market series. Next, we will choose the appropriate dynamic model and specification by comparing different models based on different information criterias and analyzing the fit achieved as well as the residuals. Next, the tail analysis will be conducted for EVT method in order to see how appropriate is the selected threshold values. And lastly, backtesting methodology will be applied in order to compare the performance of the selected methods in the underlying markets.

4.1 Data

Namely, this study examines daily closing prices from five different capitalization weighted stock market indices; PX50 (Prague, Czech Republic), BIST100 (Istanbul, Turkey), ATHEX (Athens, Greece), PSI20 (Lisbon, Portugal), IBEX35

(Madrid, Spain). Daily closing prices are taken from the Thomson Reuters Eikon website for each of the markets [1]. The data span is the same for all markets covering the period from 1st March 1994 until 24th February 2014. This time span covers the period of many crises that had significant impact on the underlying markets such as; Asian financial crisis (1997), Turkish banking crisis (2001), Global financial crisis (2008-2010), and lately the European debt crisis (2012). Even though the date's interval is similar for all the markets, number of observation differs. That is due to different work-off days in different countries as this study sample covers different markets. The full sample size ranges from 4933 daily observations for PX50 to 5123 daily observations for PSI20.

Table 4.1: Data

Markets	Observations	Start Date	End Date
PX50	4933	1994-03-01	2014-02-24
BIST100	4989	1994-03-01	2014-02-24
ATHEX	4986	1994-03-01	2014-02-24
PSI20	5123	1994-03-01	2014-02-24
IBEX35	5036	1994-03-01	2014-02-24

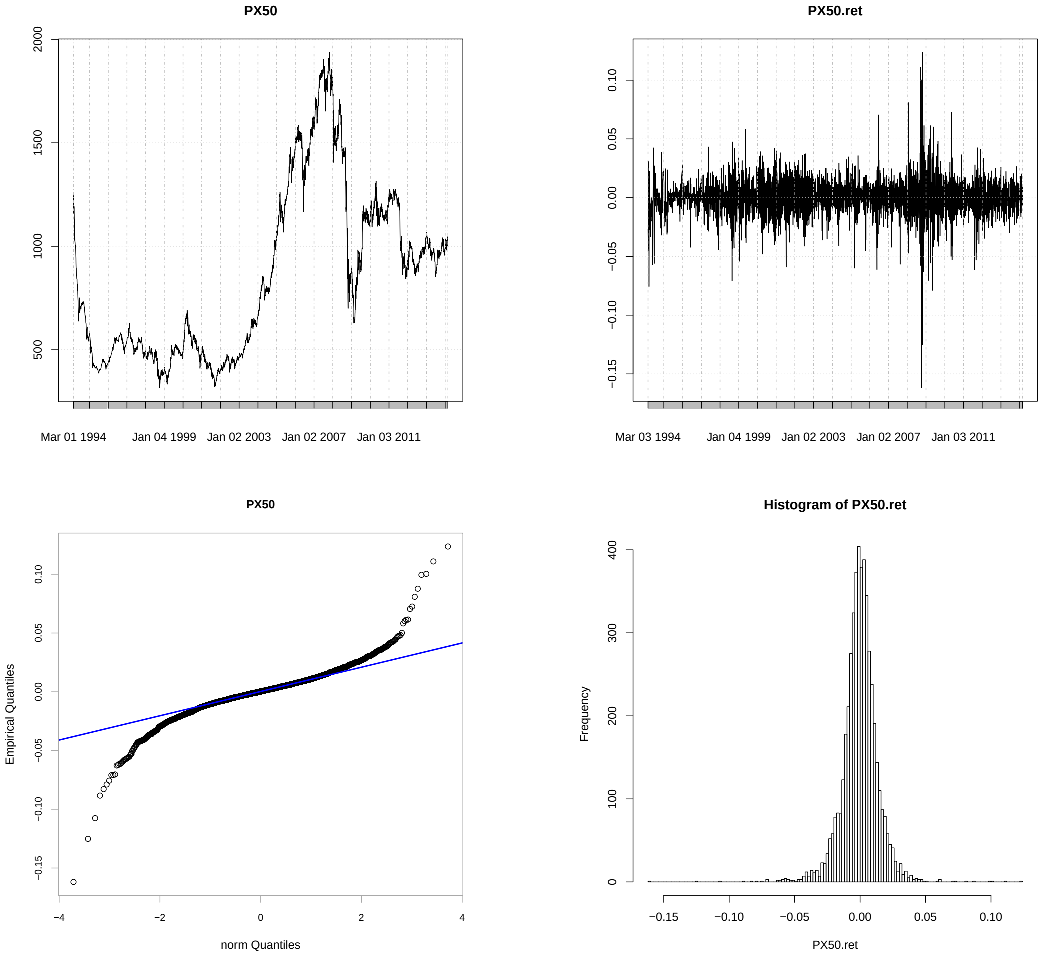
4.2 Preliminary Analysis and Descriptive Statistics

We start this section by presenting the main findings from preliminary analysis. Figure 4.1 shows time series plot of prices and logarithmic returns as well as histogram and QQ-plot for PX50. Same figures can be found in Appendix for other markets. Next, Table 4.2 shows the descriptive statistics of all the markets for both full-sample and in-sample period. In-sample period covers first 1000 observations from all the markets separately. Considering the full-sample; highest daily average return is in BIST100 (0.0012) while other markets have daily average returns around 0. This is possibly as a consequence of high inflation rates in Turkey compare to other markets over the considered period. Not surprisingly, BIST100 has the highest standard deviation (0.0258) among

[1] https://thomsonreuterseikon.com/

the markets. ATHEX stands out with the second highest standard deviation (0.0179). PX50, PSI20 and IBEX35 have almost similar standard deviation of the daily stock returns such as; 0.0142, 0.0115, 0.0148 respectively.

Figure 4.1: Preliminary Plots for PX50



Source: author's computations.

When we check the minimum and maximum returns in the full-sample, we can again see the lowest and highest daily return in BIST100 with minimum return -0.1998 and maximum return 0.1777. PX50 lies as the second market which has the lowest minimum daily return (-0.1619) while IBEX35 lies as the second market which has the highest maximum daily return (0.1348). Positive sample excess kurtosis values in all the markets indicate that the distribution of returns in all the markets are fat tailed. This can be approved by the QQ-plot in Figure 4.1. Empirical quantiles of the logarithmic returns are plotted against the normal quantiles in QQ-plot in order to see the variations in the tails of the distributions. The reversed S shape of the observations approves that the empirical quantiles of the returns tend to be larger than the quantiles of a normal distribution suggesting that the normality is a poor assumption for the underlying market's return distributions. Moreover, negative sample skewness in all the markets indicates the asymmetric tail behavior of the return distributions, meaning that the negative returns are more likely than the positive returns in these markets.

In-sample observations also shows similar characteristics with the full-sample observations. BIST100 stands out with the highest average (0.0031), maximum (0.1232), standard deviation (0.0279) and the lowest minimum (-0.1290) of the daily returns.

Table 4.2: Descriptive Statistics: refers to raw logarithmic returns; kurtosis calculated as excess kurtosis.

	Full-Sample				
	PX50	BIST100	ATHEX	PSI20	IBEX35
Observations	4932	4988	4985	5122	5035
Minimum	-0.1619	-0.1998	-0.1021	-0.1038	-0.0959
Median	0.0002	0.0014	0.0002	0.0001	0.0008
Mean	-0.0000	0.0012	0.0000	0.0001	0.0002
Maximum	0.1236	0.1777	0.1343	0.1020	0.1348
Variance	0.0002	0.0007	0.0003	0.0001	0.0002
Stdev	0.0142	0.0258	0.0179	0.0115	0.0148
Skewness	-0.4694	-0.0718	-0.0303	-0.3461	-0.0115
Kurtosis	11.1600	5.0618	3.6745	8.2266	4.8066
	In-Sample				
	PX50	BIST100	ATHEX	PSI20	IBEX35
Observations	1000	1000	1000	1000	1000
Minimum	-0.0757	-0.1290	-0.0774	-0.0706	-0.0536
Median	-0.0005	0.0034	0.0002	0.0007	0.0012
Mean	-0.0010	0.0031	0.0003	0.0009	0.0009
Maximum	0.0431	0.1232	0.0766	0.0694	0.0550
Variance	0.0001	0.0008	0.0002	0.0001	0.0001
StdDev	0.0106	0.0279	0.0132	0.0083	0.0110
Skewness	-0.7524	-0.2737	-0.1655	-0.3623	-0.2266
Kurtosis	6.0017	2.3426	4.6745	15.7501	1.7036

Source: author's computations.

Furthermore, Figure 4.2 shows the Autocorrelation Function (ACF) calculated with 20 lags for both returns and squared returns of PX50 series. Full-sample (A,B) and in-sample (C,D) data. In line with this, Table 4.3 reports some preliminary test statistics and corresponding p-values for the raw logarithmic returns in terms of both full-sample and in-sample span. Normality is rejected by the Jarque-Bera test for both in-sample and full-sample for all the markets with very low p-values indicating that the underlying markets cannot be explained well with the assumption of normal distribution. ADF and PP test results reject the null hypothesis that there is a unit root in the series. However, this was expected since we are using the logarithmic returns which show the stationary characteristics.

Table 4.3: Preliminary Test Results: refers to raw logarithmic returns. in-sample covers the first 1000 data; Jarque-Bera is the normality test. ADF and PP stands for Augmented Dickey Fuller and Phillips Perron tests for the stationarity. LB_{20} is the Ljung-Box test statistic of autocorrelation computed for 20 lags and applied to raw logarithmic returns. LB^2_{20} is the same test applied to squared raw logarithmic returns. LM_{20} is the Lagrange Multiplier test statistic for the ARCH effect computed for 20 lags. Respective p-values from the tests are given in brackets below each test statistic.

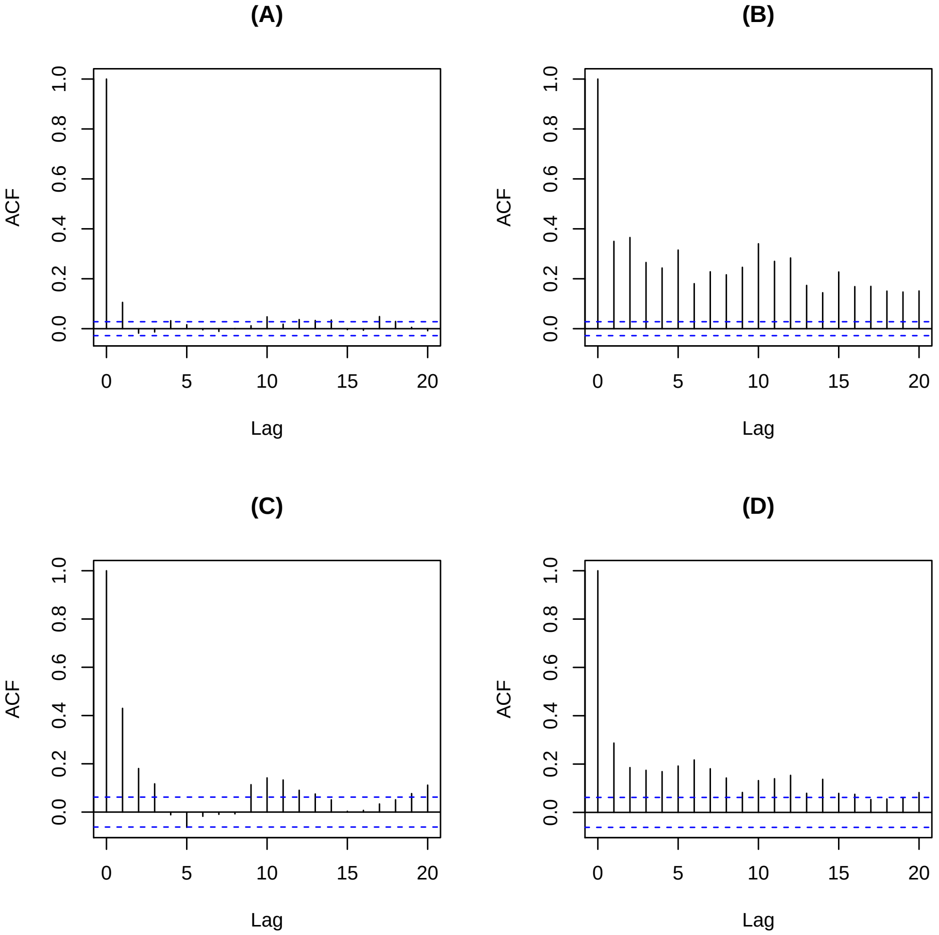
			Full-Sample		
	PX50	BIST100	ATHEX	PSI20	IBEX35
Jarque-Bera	25775.3969	5329.3352	2805.2788	14545.7815	4847.0795
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
ADF	-14.3121	-15.7058	-15.3353	-14.8721	-15.6977
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
PP	-4378.1091	-4949.8960	-4395.0549	-4624.5318	-4514.0419
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
LB_{20}	112.5994	60.1738	91.2993	108.5478	57.5853
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
LB^2_{20}	5791.7243	1813.6310	1809.2944	2172.7490	2873.1232
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
LM_{20}	1426.9643	692.8465	601.6801	660.8598	734.0214
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
			In-Sample		
	PX50	BIST100	ATHEX	PSI20	IBEX35
Jarque-Bera	1595.2242	241.1483	915.0150	10357.9010	129.4928
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
ADF	-8.1324	-9.7739	-9.2102	-9.4199	-9.7379
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
PP	-552.5911	-930.4036	-706.4589	-839.9485	-849.6223
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
LB_{20}	327.0992	28.2408	84.2319	83.8912	33.3212
	(0.0000)	(0.1038)	(0.0000)	(0.0000)	(0.0311)
LB^2_{20}	436.2111	231.8575	444.3989	157.0514	177.8399
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
LM_{20}	236.2086	140.1534	188.8930	140.7415	100.9687
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)

Source: author's computations.

Observed wavy shapes for both plain and squared return series in Figure 4.2, is an indication of autocorrelation. In regard to this, applied Ljung-Box tests with 20 lags to return series implies the rejection of the null hypothesis that series are independently distributed. Logarithmic return series shows strong serial correlation. Additionally, Ljung-Box test with 20 lags applied to squared returns shows the existence of serial autocorrelation in the squared returns. This is implies the temporal dependence structure in the squared values.

Finally, the Lagrange Multiplier test for ARCH effects confirms the test results obtained by applied Ljung-Box tests and rejects the null hypothesis of conditional homoscedasticity in regarding the both full-sample and in-sample observations. Taking into consideration all the previous results obtained from underlying tests, we can conclude that these have shown considerable evidence against normality and iid hypothesis. Moreover, this holds for both full-sample and in-sample observations. Thus, there is a big motivation to use non-normal(fat-tailed) distributions, dynamic models to account for the autocorrelations and to pre-whiten the data in order to adequately use in the second step of DGPD and FHS methods.

Figure 4.2: Correlogram of Returns and Squared Returns for PX50: Full-Sample ACF calculated with 20 lags from both raw logarithmic returns (A) and their squared values (B). In-Sample ACF calculated with 20 lags from both raw logarithmic returns (C) and their squared values (D).



Source: author's computations.

4.3 Dynamic Model Selection

In this section, we will investigate the best fit and specification of the underlined GARCH models for all the markets. In this way we find the most appropriate model specification to use in our backtesting methodology. It would be ideal to find best specification for each window of our rolling estimation using statistical and graphical tools. However, this issue is not feasible since the length of our backtesting period is very large. Therefore we illustrate this section only by evaluating the full-Sample period and use those specifications for our backtesting methodology.

We planned to use 2 different ARMA-GARCH models for each of our markets for the purpose of extracting risk measures. One of them is assumed to have normally distributed innovations while another one is assumed to have t-distributed innovations. One with the normally distributed innovations is of particular importance, because we will save the residuals from this model in order to use in the second step of our FHS and DGPD methods.

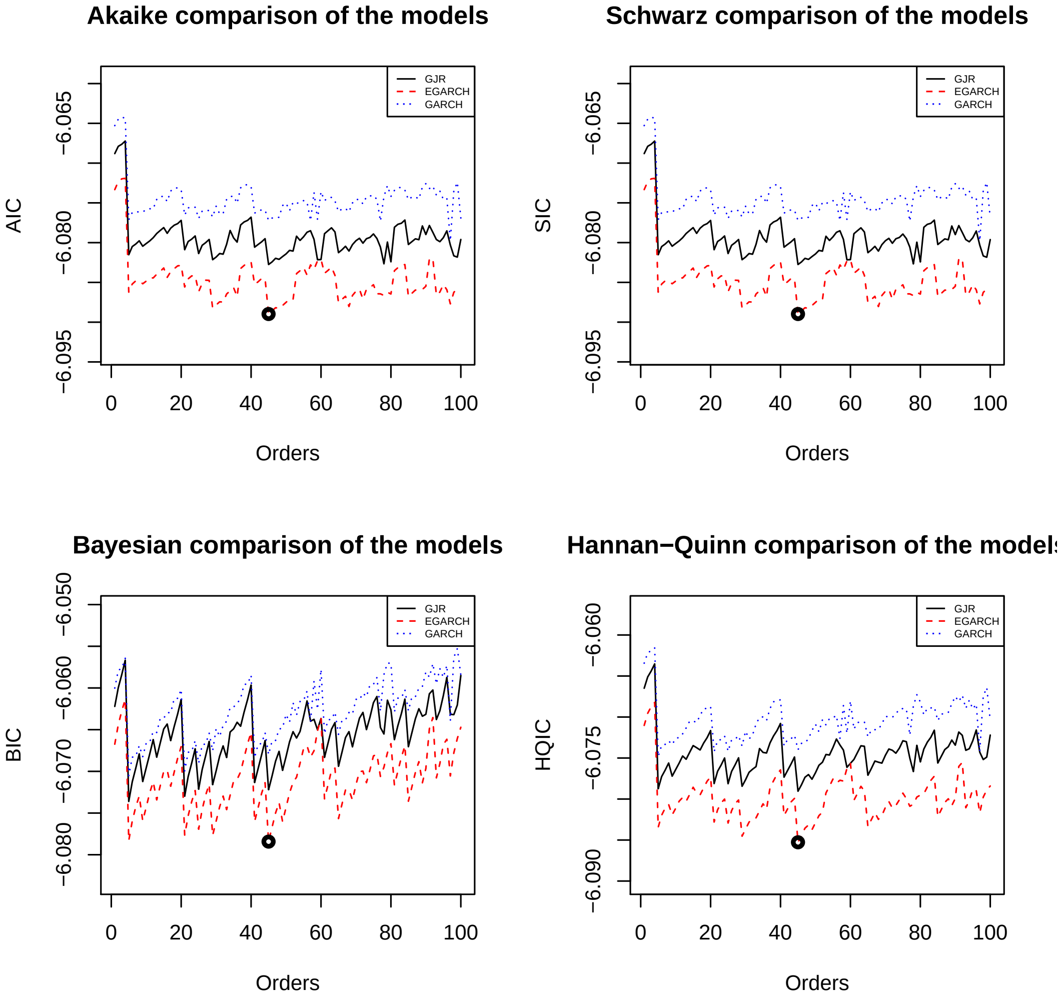
We considered 3 different ARMA-GARCH models as explained in the Section 3.1, namely ARMA-GARCH, ARMA-EGARCH and ARMA-GJR-GARCH model in order to see which model explains better the underlying markets. For this purpose, we evaluated 25 different ARMA lags with 4 different GARCH lags in total 100 different specifications for each considered model. This finally sums up to 300 different models for each of the underlying markets. Regarding ARMA lags we set the maximum lag choice to be 4, which creates vector of 5 $(0, 1, 2, 3, 4)$ for the each lag of AR and MA. Beside these selections we considered the combination of 4 different GARCH lags as follows[2] :

$$(p, q) \rightarrow (1, 1), (1, 2), (2, 1), (2, 2)$$

We base our evaluation according to Akaike (AIC), Bayesian (BIC), Schwarz (SIC) and Hannan-Quinn (HQIC) information criteria and we pick the model that has the lowest value regarding these criterias. Hence, the Figure 4.3 shows the comparison of the competing dynamic models in terms of AIC, BIC, SIC, and HQIC for full-sample observations of PX50 return series. Y-axis shows the information criterias that evaluated while x-axis shows the 100 different specifications. Regular line stands for GJR-GARCH model while dashed and dotted lines stand for EGARCH and GARCH models, respectively. Marked point on the figures stands for the lowest value of the underlying criteria.

[2] Note that initially we assumed normal distribution for the innovations

Figure 4.3: Best GARCH Selection for PX50: Comparison of the competing dynamic models in terms of AIC, BIC, SIC, and HQIC for full-sample observations of PX50 return series. Y-axis shows the information criterias that evaluated while x-axis shows the 100 different specifications($5 \times 5 = 25$ different ARMA orders and 4 different GARCH orders, so that $25 \times 4 = 100$). Regular line stands for GJR-GARCH model while dashed and dotted lines stand for EGARCH and GARCH models, respectively. Marked point on the figures stands for the lowest value of the underlying criteria.



Source: author's computations.

Checking the Figure 4.3, we can see the marked point on the dashed line suggesting the EGARCH model fit as the most appropriate one for PX50. It also lies on the order of 44 which is referring the specification of ARMA(1,2)-EGARCH(1,1) for the PX50 full-sample return series. Therefore we consider 2 dynamic models for PX50 return series in order to use them in our backtesting methodology. First one is ARMA(1,2)-EGARCH(1,1) model with conditionally student-t distributed innovations that parameters are estimated using MLE and we directly include this model in our backtesting methodology and extract the risk measures that we considered. The second one is the ARMA(1,2)-EGARCH(1,1) model with conditionally normal distributed innovations where the parameters are estimated by QMLE. Standardized residuals $\hat{Z}_s$ and one-day ahead predictions of mean $\hat{\mu}_{t+1}$ and standard deviation $\hat{\sigma}_{t+1}$ are extracted and saved from the model estimated by QMLE in order to use them in the second step of our DGPD and FHS methodology.

Similar procedures are applied to other markets as well and information criterias with the corresponding model selection reported in Table 4.4. Bold written values correspond to the picked models for our backtesting methodology. One can see in Table 4.4 that not the all information criterias give the same model specification as the best in other markets. Namely, ARMA(2,2)-GJR(2,1) was found to be the best specified model 3 times out of 4 information criteria evaluation for the BIST100 market and was chosen for the use in extracting risk measures in this market. Similarly, ARMA(0,1)-GJR(1,1) selected 3 times out of 4 information criteria evaluation in ATHEX. AIC and SIC suggests ARMA(4,3)-EGARCH(2,1) and ARMA(2,3)-EGARCH(2,1) for PSI20 and IBEX35 respectively, while BIC and HQIC chose different specifications.

Table 4.4: Best GARCH Specifications: AIC, BIC, SIC and HQIC stands for Akaike, Bayesian, Schwarz and Hannan-Quinn information criterias. Bold written models represent the selected model for the corresponding markets given in the first column.

	AIC	BIC	SIC	HQIC
PX50	**ARMA(1,2)-EGARCH(1,1)**	**ARMA(1,2)-EGARCH(1,1)**	**ARMA(1,2)-EGARCH(1,1)**	**ARMA(1,2)-EGARCH(1,1)**
BIST100	**ARMA(2,2)-GJR(2,1)**	ARMA(0,0)-GJR(1,1)	**ARMA(2,2)-GJR(2,1)**	**ARMA(2,2)-GJR(2,1)**
ATHEX	ARMA(3,4)-GJR(2,2)	**ARMA(0,1)-GJR(1,1)**	**ARMA(0,1)-GJR(1,1)**	**ARMA(0,1)-GJR(1,1)**
PSI20	**ARMA(4,3)-EGARCH(2,1)**	ARMA(2,1)-EGARCH(2,1)	**ARMA(4,3)-EGARCH(2,1)**	ARMA(3,1)-EGARCH(2,1)
IBEX35	**ARMA(2,3)-EGARCH(2,1)**	ARMA(0,0)-EGARCH(2,1)	**ARMA(2,3)-EGARCH(2,1)**	ARMA(0,1)-EGARCH(2,1)

Source: author's computations.

After we fit the chosen dynamic models by MLE under the explicit assumption of conditonal t-distributed innovations and QMLE under the conditional normal distributed innovations for the markets, we obtained standardized residuals $\hat{Z}_s$. We use these residuals to evaluate the adequacy of the selected specifications for the markets. We investigated the obtained standardized residuals from both full-sample and in-sample observations. Parameter estimations and some test statistics applied to standardized residuals from the ARMA(1,2)-EGARCH(1,1) model fit for PX50 is presented in Table 4.5. Significance of all the parameters except μ for full sample fit of PX50 is a first sign of the good fit achieved. Same conclusion can be drawn from in-sample fit as well; however, the second lag of process θ_2 for GARCH-n fit and arch lag α_1 for both GARCH-n and GARCH-t fit remains insignificant for the in-sample data.

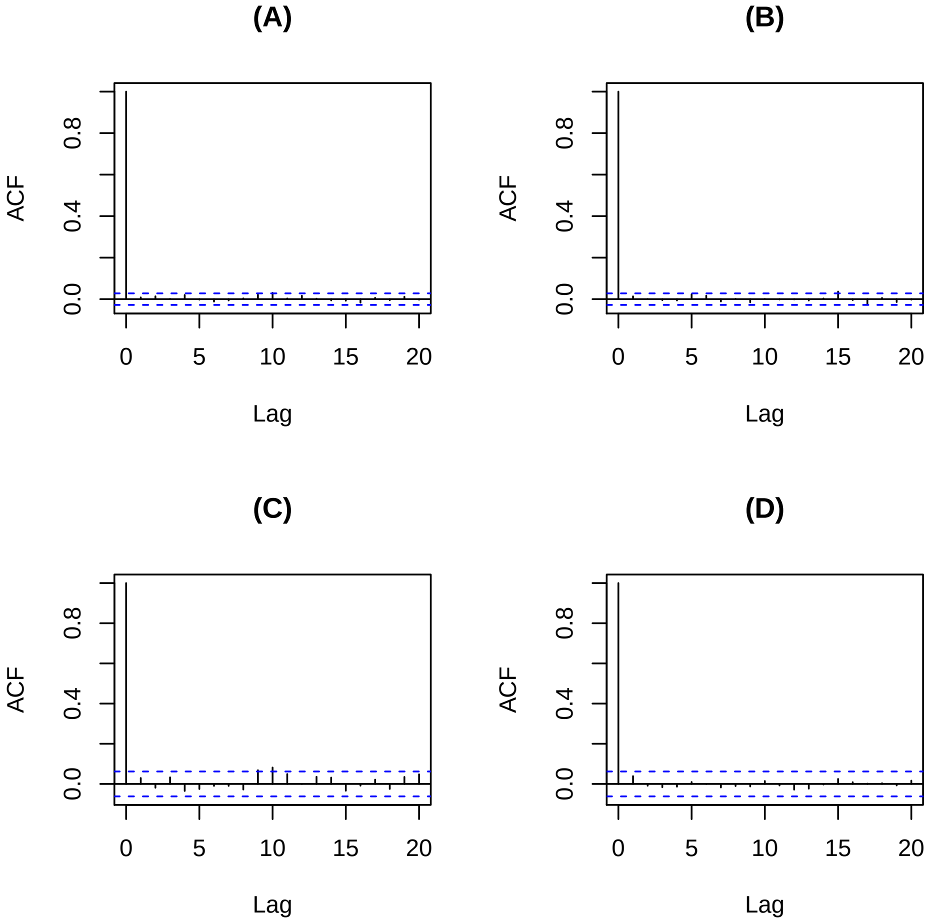
Table 4.5: ARMA(1,2)-EGARCH(1,1) Estimations for PX50: In-Sample covers the first 1000 data. Normal column stands for the normally distributed innovation assumption while t column stands for the student-t distributed innovations. Jarque-Bera is the normality test. LB_{20} is the Ljung-Box test statistic of autocorrelation computed for 20 lags,and applied to standardized residuals. LB_{20}^2 is the same test applied to squared standardized residuals. LM_{20} is the Lagrange Multiplier test statistic for the ARCH effect computed for 20 lags. P-values from the tests are given in brackets below each test statistics.

	Full-Sample		In-Sample	
	Normal	*t*	*Normal*	*t*
μ	0.0003	0.0005	0.0001	0.0001
	(0.2517)	(0.0469)	(0.3568)	(0.2917)
ϕ_1	0.9784	0.9752	0.5606	0.6519
	(0.0000)	(0.0000)	(0.0000)	(0.0000)
θ_1	-0.8541	-0.8596	-0.1773	-0.2632
	(0.0000)	(0.0000)	(0.0417)	(0.0000)
θ_2	-0.1025	-0.0932	0.0289	-0.0424
	(0.0000)	(0.0000)	(0.1209)	(0.0277)
ω	-0.3013	-0.2434	-0.7132	-0.6965
	(0.0000)	(0.0000)	(0.0071)	(0.0018)
α_1	-0.0587	-0.0495	-0.0069	-0.0023
	(0.0000)	(0.0000)	(0.8043)	(0.9460)
β_1	0.9654	0.9730	0.9235	0.9277
	(0.0000)	(0.0000)	(0.0000)	(0.0000)
γ_1	0.2538	0.2539	0.3567	0.4273
	(0.0000)	(0.0000)	(0.0000)	(0.0000)
$dof(\xi)$		7.6733		4.3383
		(0.0000)		(0.0000)
Kurtosis	2.0784	2.2277	5.6405	6.6373
Skewness	-0.2372	-0.2479	-0.7673	-0.8689
Jarque-Bera	935.7	1072.3	1432.2	1972.5
	(0.0000)	(0.0000)	(0.0000)	(0.0000)
LB_{20}	15.265	15.753	27.373	27.141
	(0.7610)	(0.7318)	(0.1251)	(0.1314)
LB_{20}^2	18.451	17.673	5.474	5.059
	(0.5577)	(0.6090)	(0.9994)	(0.9997)
LM_{20}	18.042	17.264	5.266	5.689
	(0.5846)	(0.6358)	(0.9996)	(0.9993)

Source: author's computations.

Moreover, normality is rejected by Jarque-Bera test for all 4 standardized residuals. Positive excess kurtosis and negative sample skewness are also providing evidence against normality. Ljung-Box test results with 20 lags from standardized residuals and their squared values are unable to detect conditional heteroskedasticity and serial correlation. This correction can be clearly seen from the plotted correlograms in Figure 4.4 and Figure A.9 by comparing the same figure obtained from raw returns (see Figure 4.2). Finally, insignificant LM test results for ARCH effects approve the findings of applied Ljung-Box tests that there are no remaining ARCH effects in the standardized residuals. These results are supporting our applied methodology for the dynamic model specification and state that the fitted GARCH models are plausible and they are (roughly) creating an iid residuals. This means that ARMA(1,2)-EGARCH(1,1) model adequately pre-whitens the data for PX50 return series, as this is necessary condition for FHS and DGPD methodology.

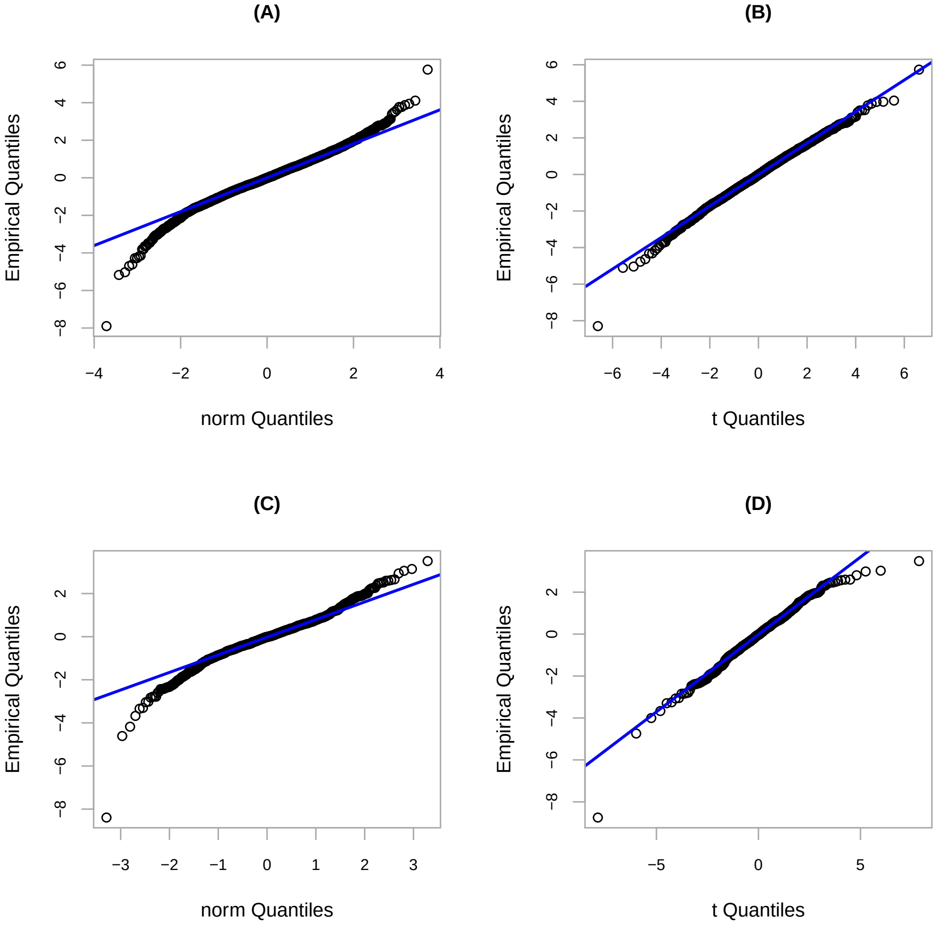
Figure 4.4: Correlogram of Normal-Residuals and Squared Normal-Residuals for PX50: Full-Sample ACF calculated with 20 lags from both normal-residuals (A) and their squared values (B). In-Sample ACF calculated with 20 lags from both normal-residuals (C) and their squared values (D).



Source: author's computations.

As the last part of this section, we also studied QQ-plots of the standardized residuals against the assumed theoretical quantiles (see Figure 4.5). Full-sample and in-sample evidences show almost similar tail behavior. Residuals show strong heavy tail characteristics in both tails. This can be clearly seen from plot **(A)** and **(C)**, where empirical quantiles are higher than the normal quantiles. Therefore, we expect that the estimation of risk from the methods with normality assumption will fail to be accurate. On the other side, clear linear trend in **(B)** and **(D)** states that student-t distribution is more capable to account for heavy tails than the normal distribution, hence give a better risk estimation.

Figure 4.5: QQ-plot of Normal- and t-Residuals for PX50: Quantile-quantile plots of PX50 normal residuals for both full-sample (A) and in-sample (C) span as well as t residuals for both full-sample (B) and in-sample (D) span.



Source: author's computations.

4.4 Tail Analysis for the DGPD model

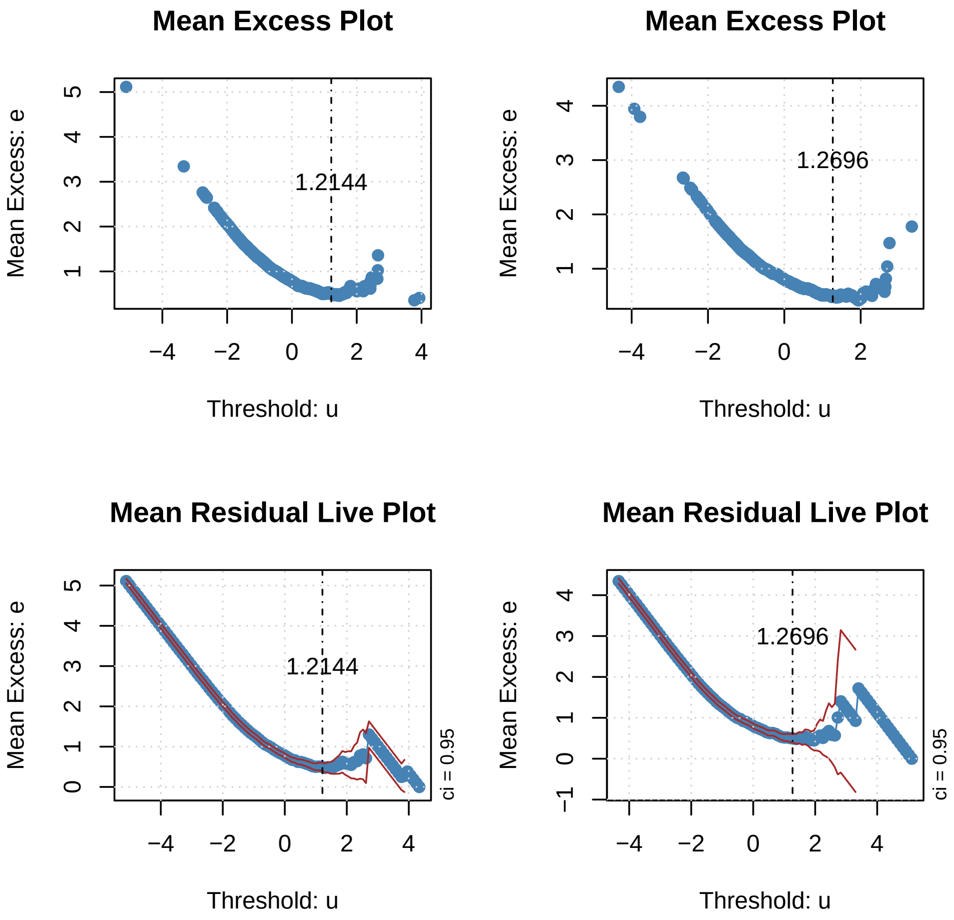
We illustrate this section to show how well our DGPD model is applied. For this reason, we analyzed some of the graphical tools as well as the fit achieved by GPD. For the purpose of simplicity we only report the analysis from the in-sample standardized residuals for the PX50.

As we explained in the methodology part, we set our threshold u selection corresponding to the 100 extreme ($k = 100$) out of 1000 n observations. Therefore we evaluated the GPD fit achieved for the 10%, 100 extreme observations for the PX50 in-sample standardized residuals. Figure 4.6 shows the mean excess and mean residual live plots for both left tail (left panel) and right tail (right panel). Vertical dotted line shows the corresponding threshold that selected for the left tail (1.2144) and for the right tail (1.2696). Ideally these plots are used for the threshold value selection; however we will evaluate only the chosen threshold values. The different threshold values corresponding to 100 exceedances in the left and right tail is a sign of asymmetry in the tails. Linear relationship between mean excess (y-axis) and corresponding threshold values (x-axis) right after the vertical dotted line in all the panels is an indication of the reasonable threshold value selection (1.2144, 1.2696). In this case one must consider the linear trend above the threshold value in order to see the behavior of the empirical distribution. In ideal situations the linear trend between mean excess and corresponding thresholds are interpreted as follows:

- Upward linear relationship with a positive gradient above the threshold indicates a positive shape parameter for the GPD and therefore a heavy tail ($\xi > 0$).
- Downward linear relationship with a negative gradient above the threshold indicates a negative shape parameter for the GPD and therefore a thin tail ($\xi < 0$).
- Horizontal linear relationship above the threshold indicates that GPD shape parameter is approximately equal to zero ($\xi \cong 0$).

In our case it can be clearly seen that the all 4 panels show upward linear trend with a positive gradient above the corresponding threshold values.this suggests that the PX50 in-sample standardized residuals follow a GPD with positive shape parameter ($\xi > 0$) for both tails. Here after, we fit GPD to 100

Figure 4.6: Mean Excess and Mean Residual Live Plots for PX50 Residuals



Source: author's computations.

exceedances of PX50 in-sample standardized residuals for both tails in order to evaluate achieved fit. Hence, MLE is used for this process.

Estimated shape $\hat{\xi}$ and scale $\hat{\beta}$ parameters with their standard errors in brackets for both tails are given in Table 4.6. The table also shows the VaRs and ESs calculated for the $\alpha = 99\%$ confidence level. Decisive parameter is the shape parameter ξ. The higher the it's value, the heavier the tail and hence higher the risk estimates we calculate. Estimated shape parameter is 0.0472 for the left tail and 0.0523 for the right tail. This indicates that the right tail is heavier than the left one for the PX50 standardized residuals. However when we compare tails for the estimated 99% VaR and ES values, they slightly differ from each other. The calculated $\widehat{VaR}^{99\%}$ and $\widehat{ES}^{99\%}$ estimates are 0.0328 and 0.0407 for the left tail and 0.0330 and 0.0408 for the right tail. The reason for this is that the estimated scale parameter for the left tail (0.4816) is smaller than for the right tail (0.4637) as these are used also as a scaling factor in calculating risk measures (Equation 2.9, Equation 2.10).

Table 4.6: GPD Estimates for PX50

	Left Tail	Right Tail
$\hat{\xi}$(shape)	0.0472	0.0523
	(0.0937)	(0.0919)
$\hat{\beta}$(scale)	0.4816	0.4637
	(0.0659)	(0.0629)
u(threshold)	1.2144	1.2696
$\widehat{VaR}^{99\%}$	0.0328	0.0330
$\widehat{ES}^{99\%}$	0.0407	0.0408

Source: author's computations.

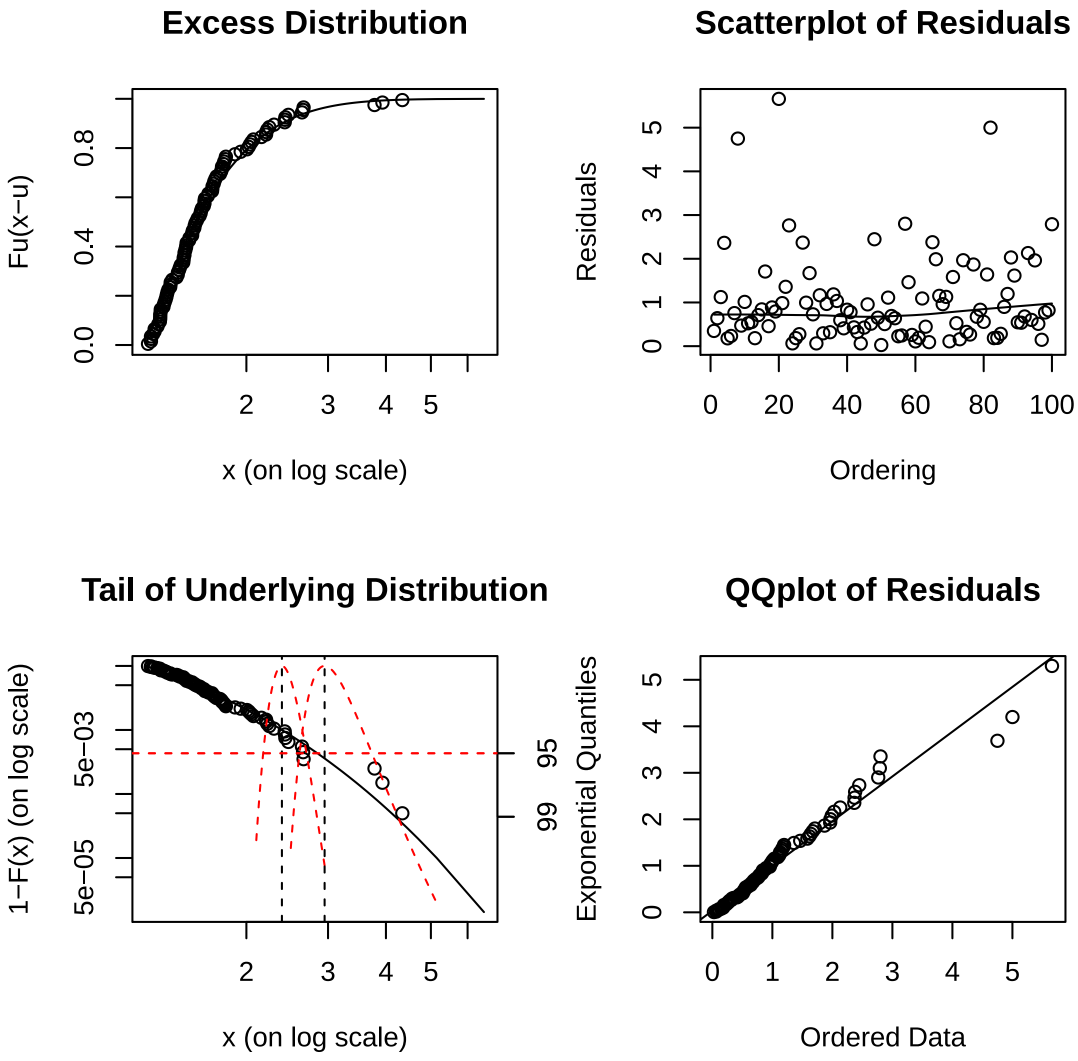
We also studied the behavior of the estimated shape parameter $\hat{\xi}$ and 99% VaR estimations with varying exceedances and their corresponding threshold values. In line with this, the Figure A.10 in Appendix show the respective behavior for both the tails regarding PX50 standardized residuals. We again marked the selected threshold values with dotted line, where the intersection of it with the solid line represents the estimated values we reported in the previous table. We see from all the panels that the shape parameter ξ and VaR estimates are not much sensitive around our dotted line, indicating that the selected thresholds are reasonable.

In addition to the estimated parameters, the GPD fit to the exceedances can be as well investigated by graphical. We demonstrated diagnostic plots for the left tail GPD fit of the PX50 standardized residuals in Figure 4.7 for this purpose[3]. Top left panel shows the excess distribution of the estimated GPD model with a smooth curve. Empirical distribution of 100 exceedances is dotted across this curve. Clearly, the fit of the GPD seems well to the naked eye from this panel. Right panels (top,bottom) show the scatterplot and QQ-plot of the exceedances, respectively. The OLS line in scatterplot is fairly flat and the empirical distribution of the exceedances in QQ-plot lies approximately along a straight line with some minor deviation, suggesting the good fit achieved by the underlying GPD. Tail estimate of the fitted GPD distribution is shown in the left bottom panel. Dots again represent the 100 exceedances and solid curve represent the tail estimation. It is important to see here how the tail estimation from the GPD tail estimation allows extrapolation even if the data is unreliable

[3] Similar figure can be found in Appendix for the right tail GPD fit of the PX50 standardized residuals. Similar conclusion can be drawn from this as well.

and scarce when going further in the tail. We also marked a calculated 99% VaR and ES estimates on this panel. Intersection of the two dashed vertical lines mark the VaR (left) and ES (right) estimates. The dashed curves are showing the confidence intervals for the calculated VaR and ES. It is clear that there is a big uncertainty regarding our coherent measure of risk. However, given a data with such heavy tails, this could be expected. Cautious risk manager shall know about his uncertainty magnitude in terms of extreme phenomena.

Figure 4.7: Diagnostic Plots of the Left tail GPD fit for the PX50 Residuals



Source: author's computations.

Taking into consideration all the findings from this section, we can conclude that the threshold values corresponding to the set 100 exceedances is acceptable for the both left and right tail regarding the PX50 in-sample standardized residuals. In addition, GPD fits the 100 exceedances appropriately for both tails. Therefore we are motivated to apply the very same methodology for other markets in order to go further in this study. However, the same conclusion may not be satisfied for all the markets.

4.5 Backtesting Results

This section presents the empirical results obtained by applying the backtesting methodology to each method selected in order to assess the accuracy performance in each market. As a remainder, we set estimation window n for all methods as 1000 observations, which leaves us really very long span of testing window with approximately around 4000 observations. Here, rolling estimation technique is used, meaning that each time the estimation window is rolled forward. In each step of this rolling window, one day ahead risk measures, VaR and ES are extracted from each single method selected, namely; normal distribution (Normal), Historical simulation method (HS), Static GPD method (GPD), Exponentially weighted moving average (EWMA), 2 different best selected GARCH models that assume normal and t distribution for the innovations (GARCH-n and GARCH-t), respectively, Filtered historical simulation (FHS) and Dynamic GPD method (DGPD). Methods that are in dynamic group are refitted in each step of the rolling window. This rolling estimation technique is applied for all the markets; PX50, BIST100, ATHEX, PSI20, IBEX35. One day ahead risk measures are calculated for 3 different confidence interval α; 95, 99 and 99.5%. Focus is not only on the left tail that corresponds to the short position, but also on the right tail that corresponds to a long position for the underlying markets.

As a result of this largely dimensioned methodology, this section is divided into subsections as Backtesting Value at Risk and Backtesting Expected Shortfall. In addition, Volatile Period Evaluation is presented at the end.

4.5.1 Backtesting Value at Risk

Table A.1, Table A.2 and Table A.3 presents the left tail backtesting results based on one-day ahead VaR estimates calculated with $\alpha = 95, 99, 99.5\%$ confidence level. Similarly, Table A.4, Table A.5 and Table A.6 presents the right tail backtesting results. EV stands for expected number of violations while AV is the actual number of violations. We expect actual number of violations to be closer to those expected number of violations in order to qualitatively comment the accuracy of the predictive performance of the methods. For this purpose, one can check the fraction of the actual violations (fracAV) and/or the violation ratio (VR)[4] given in the 3rd and 4th columns, respectively.

However, ranking methodology for the predictive performance of the methods is based on the quantitative test results. LRuc, LRind and LRcc are the likelihood ratio test statistics from unconditional coverage, independence and conditional coverage tests respectively[5]. The methods are ranked as follows: we first check the significance of the test results[6]. The method with the relatively smallest likelihood ratio statistics is selected as the best method. But, in this case if there is a tie between methods, i.e. the likelihood ratio statistics are the same, we check the distance between EV and AV, and pick the method with smallest distance as the best method. Moreover, if there is also similar distance observed between EV and AV values, we then base our selection on the average VaR values calculated over the backtesting period (given in the last column) and select the one with the smaller average VaR value. There can be also seen cases that none of the methods passes at least one of the underlying tests (for example see IBEX35 left tail results). In this case, same procedure is applied and the method with the relatively smaller test results is selected as the best performer method.

Normal method is the worst method as a result of underestimation of the risk almost in all the markets with all different confidence levels. Surprisingly, accuracy of the Normal method is better than other methods at $\alpha = 95\%$, in PSI20 and IBEX35 for the left tail and in PSI20 for the right tail (Table A.1 and Table A.4). We fail to reject the hypothesis of unconditional coverage in the respective markets. But, when one goes further in the tail, that is 99 and 99.5%, i.e. underestimation of Normal method is clearer with very high values of violation ratios and/or actual fraction of violations. Hypotheses of unconditional coverage, independence and conditional coverage are almost always rejected in the rest of the case. This failure of Normal method is expected since the underlying distributions of the markets are heavy tailed (see Section 4.2); hence the normality assumption is just a rough approximation for the VaR calculation.

[4] Violation ratio (VR) is simply the actual number of violations over expected number of violations, thus closer to 1 the better the methods accuracy.

[5] Critical values at the 5% and 1% significance are 3.8415 and 6.6349 for LRuc and LRind, 5.9915 and 9.2103 for LRcc, respectively.

[6] Notice that we have 30 different cases as a result of 3 different confidence levels $\alpha = 95, 99, 99.5\%$, both right and left tail and 5 different markets, $3 \times 2 \times 5 = 30$.

On the other side, other unconditional methods HS and GPD show improvement over the Normal method since they account for the heaviness of the tail in either natural (HS) or parametric way (GPD). Obviously, when one goes further in the tail, the heaviness of the tail appears to be more important for the risk estimation. HS method is the best performer only once in the left tail of BIST100 at the confidence level $\alpha = 99\%$, for which we fail to reject the hypotheses of unconditional coverage, independence and conditional coverage tests at any significance level (Table A.2). Analogously, GPD method is selected 3 times as the best performer. Those are in left tail of BIST100 and IBEX35 at the confidence level $\alpha = 99.5\%$ (Table A.3) and in the right tail of PX50 at the confidence level $\alpha = 99\%$ (Table A.5). Null hypotheses of the respective tests are again not rejected at any significance level. However, majority of the cases that these unconditional methods have shown bad performance since they always have margin larger than the one from other methods based on violation ratios and actual fraction of violations. This is likely because of the fact that HS and GPD methods do not account for the changing volatility as dynamic methods do. They react slowly to the changing dynamics of the markets (changing volatility and changing market uncertainty), which is confirmed by the significantly high values of the LRind test statistics in many cases. That is, the violations do not fall randomly in time but have tendency to cluster. In addition to this, general bad performance of GPD method can be addressed with the non iid behavior of the exceedances, as EVT assumes exceedances to be iid as a basic building block (see Section 2.3).

When we check the conditional methods which can react to the changing dynamics of the markets, GARCH-n method is the worst method by far. It is selected as a best performer twice at the confidence level $\alpha = 95\%$; in the left tail of PX50 (Table A.1) and in right tail of IBEX35 (Table A.4) with violation ratios equal to 1.0224 and 0.9418, respectively. We fail to reject the hypotheses of all the three tests at any significance level in these two respective cases. One thing to notice for the GARCH-n method is that the hypothesis of independence is rejected only few times (left tail BIST100 at $\alpha = 99\%$ and right tail BIST100 at $\alpha = 95\%$, considering 1% significance for the respective test), which shows the power of conditional methods compare to unconditional ones as the former can avoid violations clustering. However, we reject the hypothesis of unconditional coverage in many of the cases, hence the conditional coverage too[7]. This addresses the issue of normality approximation as a rea-

[7] Notice that the conditional coverage test statistic is a combination of unconditional

son. Systematic underestimation of GARCH-n method or in other words high frequency of violations seems to be a result of assuming conditionally normally distributed innovations. In line with the GARCH-n method, EWMA method shows very poor performance especially in the left tail of all the markets, where hypotheses of all the 3 tests are almost always rejected at the 1% significance. This problem is more pronounced for the unconditional coverage test as a result of systematic underestimation of risk. Conversely, good performance of the EWMA method in the right tail is surprising, especially at the confidence level $\alpha = 95\%$, where we fail to reject all the tests at 1% significance (Table A.4). It is selected as the best method 4 times in the right tail, with the performance being the only method surviving all the tests at $\alpha = 95\%$ in BIST100 and at $\alpha = 99\%$ in PSI20[8] (Table A.5).

Hence, the overall bad performance of the methods with normality approximation (Normal, GARCH-n, and EWMA) shows the importance of accounting fat tail phenomena for the derivation of risk in terms of VaR. Therefore, we continue with the other conditional methods that capture the heaviness of the tail. In this regard, semiparametric FHS method that accounts for the fat tail in a natural way does not seem to be better than EWMA and GARCH-n method. The bad performance of the FHS method is especially in the three markets; namely, BIST100, PSI20 and IBEX35. On the contrary, we select FHS method as a best performer once in both PX50 and ATHEX, in the left tail at $\alpha = 99\%$ and 99.5%, respectively (Table A.2 and Table A.3). We fail to reject the hypotheses of all the tests at any reasonable significance level in the respective cases. This failure of FHS method is surprising since it can capture all the stylized facts (autocorrelation, heteroskedasticity, time varying volatility, fat-tailedness, and asymmetry).

Next method that can capture the stylized facts is GARCH-t method. It is the second best method in the overall evaluation. It has generally good quantitative performance except in IBEX35, which is the most problematic market for the GARCH-t method, where we at least reject the hypothesis of one of the tests. GARCH-t method also shows equal performance in the both tails of all the markets. It was selected 5 times as a best performer out of 30 cases, where 3 of them are in the PSI20 (at the $\alpha = 99\%$ and 99.5% in the left tail, and at the $\alpha = 95\%$ in the right tail) for which we fail to reject

coverage and independence tests, hence the systematic failure of one of the latter will drive conditional coverage test to failure.

[8] Other two are at $\alpha = 95\%$ in ATHEX (Table A.4) and $\alpha = 99.5\%$ in IBEX35 (Table A.6).

the hypotheses of the respective tests at 1% significance. This indicates the capability of GARCH-t method in PSI20 as opposed to other methods for predicting VaR. In addition, GARCH-t method signalizes itself in the left tail of BIST100 (at the $\alpha = 95\%$) as it is the only method passes joint test of correct violations and independence (Table A.1). The good predictive performance of GARCH-t method can be given as an answer to the set objectives in this study. Therefore, one should consider underlying facts about the return series in order to get good predictive accuracy in terms of VaR estimation.

As a last but not the least, DGPD method is the best method in overall evaluation that have been selected 12 times out of 30 cases. It clearly outperforms other methods by showing good performance almost in all the markets. In majority of the cases, it was ranked among the top performing methods. Moreover, in literature, DGPD method was found to show better performance and to dominate other methods especially further in the tail area, since the theory itself concerned with the extremes. In line with this fact, the performance of DGPD method is increasing especially in the right tail of the markets. But, same conclusion does not hold for the left tail, where the accuracy performance of DGPD method is less than the right tail and it is similar regardless of the analyzed confidence level α. Problematic markets are BIST100 and IBEX35 for DGPD method, where we observe significant test results. However, DGPD was best performer 2 times in BIST100 (at $\alpha = 99$ and 99.5% in the right tail) and 3 times in IBEX35 (at $\alpha = 95$ and 99% in the left tail, at 99% in the right tail). Except these two markets, we do not observe any significant test results; hence do not reject the respective hypotheses of unconditional coverage, independence and conditional coverage. We can observe DGPD method with the superior performance at the $\alpha = 95\%$ in the left tail of ATHEX (Table A.1) and at the $\alpha = 99.5\%$ in the right tail of both BIST100 and PSI20 (Table A.6), with the exact match of actual and expected number of violations. However, taking into consideration the different performance of DGPD method in both right and left tail of the markets, we can conclude that DGPD method is not necessarily superior to the rest of the methods. Nevertheless, it is more capable than other methods to give an accurate estimation of risk both qualitatively and quantitatively in general.

Table 4.7 below summarizes above results and presents the best method selected in each case we have considered. Top 3 methods are shaded with gray color starting from the dark to light gray. Analyzing the Table 4.7, it is clear that the DGPD method outperforms other methods with large majority, especially in the right tail of the markets. Moreover, GARCH-t method takes the second place and it is the only competitor of DGPD method. Yet, there is a considerable difference between the performances of these two methods. GARCH-t method shows good performance as opposed to other methods in PSI20; hence it is better for modeling this market for the purpose of risk estimation.

Table 4.7: Best Performer Methods: It presents the best selected methods regarding full-sample. Selection procedure is as follows: We first check the significance of the unconditional coverage, independence and conditional coverage test results. The method with the relatively smallest likelihood ratio statistics is selected as the best method. But, in this case if there is a tie between methods, i.e. the likelihood ratio statistics are the same, we check the distance between expected and actual violations, and pick the method with smallest distance as the best method. Moreover, if there is also similar distance observed between EV and AV values, we then base our selection on the average VaR values calculated over the backtesting period and select the one with the smaller average VaR value.

	Left Tail			Right Tail		
	95%	99%	99.5%	95%	99%	99.5%
***PX*50**	GARCH-n	FHS	DGPD	DGPD	GPD	DGPD
***BIST*100**	GARCH-t	HS	GPD	EWMA	DGPD	DGPD
ATHEX	DGPD	DGPD	FHS	EWMA	DGPD	GARCH-t
***PSI*20**	FHS	GARCH-t	GARCH-t	GARCH-t	EWMA	DGPD
***IBEX*35**	DGPD	DGPD	GPD	GARCH-n	DGPD	EWMA

Source: author's computations.

Overall findings can be summarized as follows:

- Methods which do not account for the heaviness of the tail almost always fail to be appropriate as a result of systematic underestimation of VaR. On the other side, methods that account for the heaviness of the tail either natural or parametric way are shown to have better accuracy. Hence, a fat-tail phenomenon is one of the most important stylized facts that should be considered for the purpose of VaR calculation.

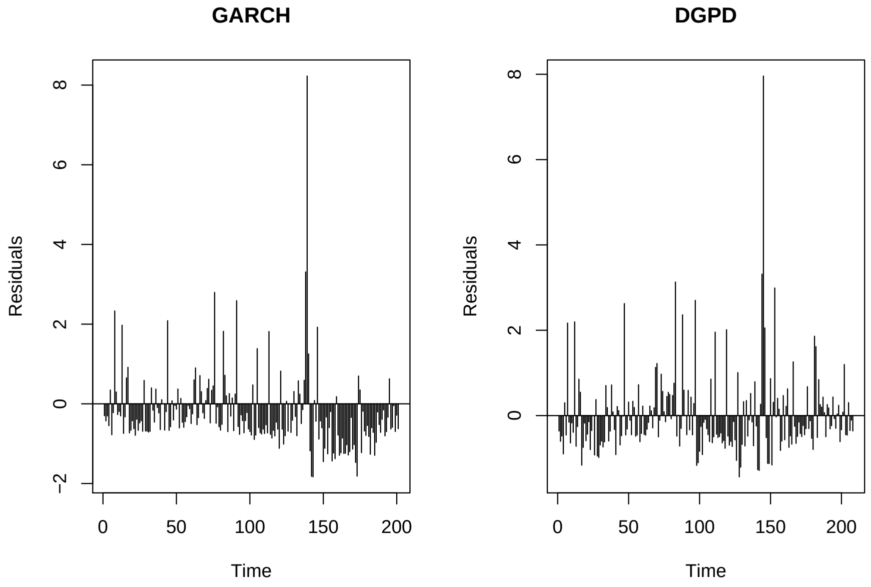
- As opposed to unconditional methods, conditional methods that react quickly to the changes in market dynamics and avoid volatility clustering are better than unconditional ones in terms of their predictive performance for VaR estimation. The performance of conditional methods is increasing, after capturing the fat-tail and asymmetry phenomenon.

- As combination of nonparametric and parametric modelling, semiparametric methods (GPD, FHS, DGPD) are more capable to accurately predict risk in terms of VaR.
- And finally, EVT based methods dominate others as it was found to be more capable than other methods to give an accurate estimation of risk in terms of VaR.

4.5.2 Backtesting Expected Shortfall

This section presents the ES Test results in line with the objectives set in the respective methodology given in Subsection 3.3.2. For reminder, this section aims to test the null hypothesis of correct ES estimates and correctly modeled parameters. Under this hypothesis, exceedance residuals υ_t, shall have an iid properties, i.e. zero mean and unit variance. Accordingly, Figure 4.8 shows the exceedance residuals for PX50 at the 95% confidence level that should have mean zero under this hypothesis[9]. While the left graph refers to the exceedance residuals extracted by GARCH-n method, the left graph shows the respective residuals in the case of DGPD method. Graphical results indicate that the residuals derived under an assumption of conditional normality (GARCH-n) are not likely to have mean zero that can be seen by the heaviness of the negative exceedance residuals in the left graph. On the other side, the balance between the positive and negative exceedance residuals in the right graph (DGPD) visually indicates that they are more reasonably mean zero. Thus, the methods with conditional normality assumption (Normal, EWMA, GARCH-n) seem not to be useful for the purpose of the ES calculation (McNeil & Frey 2000).

Figure 4.8: Exceedance Residuals for PX50–95% Confidence Level



Source: author's computations.

Now we continue with the ES Test results in the Table 4.8 given for both

[9] The number of exceedance residuals derived from GARCH-n method is 201 (left graph), while in the case of DGPD this number is 208 (right graph).

Full-Sample and Sub-Sample. This time we did not include the conventional methods (Normal, HS, GPD, EWMA) in the comparison because of their poor performances that were previously found in Subsection 4.5.1. Note that we also dropped 99.5% confidence level in order to avoid bias that would result from the small number of exceedance residuals included in respective calculation. Thus, the following Table shows the bootstrapped p-values[10] for the test applied to the exceedance residuals that are extracted from the GARCH-n, GARCH-t, FHS and DGPD methods.

Table 4.8: Expected Shortfall Test Results: It presents respective bootstrapped p-values from the one sided t-test for the Expected Shortfall

		Full-Sample				Sub-Sample			
		Left Tail		Right Tail		Left Tail		Right Tail	
		95%	99%	95%	99%	95%	99%	95%	99%
PX50	GARCH-n	0.0002	0.0200	0.0036	0.0380	0.0229	0.0825	0.0249	0.0059
	GARCH-t	0.0000	0.0785	0.0051	0.5424	0.0061	0.0794	0.0547	0.4336
	FHS	0.5571	0.6118	0.7578	0.6367	0.5532	0.2303	0.9596	0.7173
	DGPD	0.4919	0.4993	0.7282	0.9364	0.7496	0.1953	0.9481	0.8466
BIST100	GARCH-n	0.0380	0.5910	0.0033	0.0532	0.0237	0.0198	0.3399	0.3133
	GARCH-t	0.0652	0.5151	0.0010	0.6737	0.0099	0.0492	0.3863	0.3366
	FHS	0.1584	0.4373	0.7642	0.3607	0.1960	0.0585	0.8930	0.5290
	DGPD	0.1284	0.6210	0.8758	0.5054	0.1534	0.4368	0.9193	0.7117
ATHEX	GARCH-n	0.0197	0.0008	0.0017	0.1672	0.2836	0.0000	0.2127	0.5507
	GARCH-t	0.0119	0.1322	0.0034	0.3818	0.3326	0.0000	0.2236	0.4349
	FHS	0.8578	0.4623	0.6849	0.4689	0.0462	0.0009	0.9560	0.8671
	DGPD	0.8202	0.3710	0.6867	0.7332	0.0243	0.0033	0.9498	0.8343
PSI20	GARCH-n	0.0000	0.0155	0.0000	0.2657	0.0001	0.0042	0.0073	0.7873
	GARCH-t	0.0000	0.0388	0.0000	0.2961	0.0003	0.0684	0.0016	0.8198
	FHS	0.8279	0.7873	0.5327	0.5053	0.5523	0.7259	0.5932	0.6507
	DGPD	0.7849	0.8757	0.4654	0.7861	0.4789	0.7722	0.5756	0.8567
IBEX35	GARCH-n	0.0000	0.2214	0.0000	0.1559	0.0015	0.4758	0.1197	0.5254
	GARCH-t	0.0000	0.0441	0.0000	0.0086	0.0001	0.2335	0.1915	0.3182
	FHS	0.6789	0.3579	0.9329	0.7324	0.3457	0.6703	0.9547	0.6704
	DGPD	0.8287	0.5055	0.9245	0.7519	0.2912	0.8127	0.9326	0.7248

Source: author's computations.

Examining Full-Sample, quantitative test results partially confirm the conclusion of the useless of ES calculation under the assumption of conditional normality (GARCH-n). Nearly all p-values from the GARCH-n method reject the null hypothesis of exceedance residuals having mean zero. However, we fail to reject the null hypothesis at the 10% confidence level once in BIST100 left tail ($\alpha = 99\%$), ATHEX right tail ($\alpha = 99\%$) and PSI20 right tail ($\alpha = 99\%$),

[10] Note that we use here bootstrapping method for getting the respective p-values in order to reduce the bias regarding the distributional assumption (McNeil & Frey 2000)

and twice in IBEX35 for both right and left tail ($\alpha = 99\%$). This partial failure of hypothesis rejection can be addressed in the two ways. First, as opposed to the McNeil & Frey (2000) who derived exceedance residuals under the simple GARCH(1,1) method, GARCH-n method that this study considers is different. That is, for instance, ARMA(0,1)-GJR(1,1) for ATHEX and ARMA(2,3)-EGARCH(2,1) for IBEX35 (see Section 4.3 and Table 4.4). Therefore, GARCH-n methods are able to account for the autocorrelation and asymmetry phenomena. Second, the $\alpha = 99\%$ may create bias that would result from the small number of exceedance residuals included in respective calculation. Similarly, following the same examination for GARCH-t results, the difference from the above is that GARCH-t is worse than GARCH-n only in IBEX35 with two more rejection; that is, right and left tail for $\alpha = 99\%$. For the rest of the markets, the results are same or slightly better than in the case of GARCH-n. Hence, even though GARCH-t method differs from GARCH-n due to the assumption of conditionally t-distributed innovations, it is equally useless as the GARCH-n for the purpose of ES calculation. On the other hand, FHS and DGPD method show much more plausible results. The null hypothesis of exceedance residuals having mean zero is never rejected neither at 10% nor at 5% confidence level; which holds for all $\alpha = (95\%, 99\%)$, tails and markets in the Full-Sample. This allows us to conclude that data from all the markets that this study considers strongly supports the zero mean hypothesis.

Examining the Sub-Sample, i.e. the volatile period, our quantitative results for GARCH-n and GARCH-t methods are in line with the respective Full-Sample results but with slightly better performance, probably due to the reduced number of observations. Considering the FHS and DGPD methods, their superior performance found in the Full-Sample results could not be confirmed for the case of Sub-Sample. Namely, ATHEX was found to be the most problematic market in terms of rejecting the null hypothesis of mean zero at the 10% confidence level in the left tail ($\alpha = 95\%$ and 99%) for both FHS and DGPD methods. In addition, we also reject the respective hypothesis in the left tail $\alpha = 99\%$ for FHS method regarding the BIST100. For the remaining markets, this hypothesis was not rejected at all. Furthermore, it seems to be quite strongly supported regarding the PX50, PSI20, IBEX35 series. Thus, we can conclude that the predictive power of FHS and DGPD methods is found to be the best among the considered methods in the case of ES estimation.

4.5.3 Volatile Period Evaluation

In addition to the full sample backtesting evaluation, this part of the study presents the backtesting results only for the high volatile period in order to see methods' performance. For this purpose, 2000 observations from each market around Global Financial Crises (GFC) period are taken as a sub-sample[11] and similar backtesting procedure is applied. Estimation window is again set to 1000 observations, which this time leaves 1000 observations for the testing window. Also the threshold for the extreme value based methods is set for the corresponding 100 extreme observations. For the purpose of simplicity, only the important results from volatile period evaluation is presented. However, we refer reader to following tables for detailed results. Table A.7, Table A.8 and Table A.9 in Appendix presents the left tail backtesting results for the subsample, based on one day ahead VaR estimates calculated with 95, 99, 99.5% confidence, respectively. Similarly, Table A.10, Table A.11 and Table A.12 in Appendix presents the right tail backtesting results for the subsample.

Similar to the results obtained from full-sample, unconditional methods (Normal, HS, GPD) are performing very poor. The hypotheses of unconditional coverage, independence and conditional coverage are rejected nearly in all the markets. One exception is in BIST100, where we fail to reject hypotheses of respective tests for both GPD and HS method at 1% significance, and observe them with the good performance. The poor performance of unconditional methods is inevitable, since we are focused in the volatile period which is characterized with sudden and quick market changes. Conditional GARCH-n method also has a poor performance in general, where we can only observe its good performance at $\alpha = 95\%$. Apparently, it is because of normality approximation.

However, the picture for top performing methods surprisingly differs from the full-sample. EWMA method shows very poor performance considering left tail of the markets, but conversely it has very good performance on the right tail in spite of normality approximation. It dominates both FHS and DGPD method in the right tail of the markets. It is selected 7 times as a best performer method (6 of them in the right tail). Quantitatively, DGPD and FHS methods show very good performance, i.e. better than EWMA, as they survive all the tests in majority of the cases at 1% significance. However, this general good performance does not help both models in specific cases and hence they were

[11] from the beginning of 2004 until the end of the 2011.

selected 4 and 5 times as a best performer, respectively (most of them in the left tail).

GARCH-t method is the best method under the evaluation of volatile period. The hypotheses of the underlying tests are never rejected in the right tail of the markets at 1% significance. It clearly dominates other methods especially in the right tail of the markets. It is selected 8 times as a best performer with 5 of them being in the right tail.

Accordingly, Table 4.9 below presents the best selected methods considering volatile period. Top 3 methods are similarly shaded with gray color starting from the dark to light gray.

Table 4.9: Sub-Sample Best Performer Methods: It presents the best selected methods regarding sub-sample. Selection procedure is as follows: We first check the significance of the unconditional coverage, independence and conditional coverage test results. The method with the relatively smallest likelihood ratio statistics is selected as the best method. But, in this case if there is a tie between methods, i.e. the likelihood ratio statistics are the same, we check the distance between expected and actual violations, and pick the method with smallest distance as the best method. Moreover, if there is also similar distance observed between EV and AV values, we then base our selection on the average VaR values calculated over the backtesting period and select the one with the smaller average VaR value.

	Left Tail			Right Tail		
	95%	99%	99.5%	95%	99%	99.5%
***PX*50**	GARCH-n	FHS	DGPD	GARCH-n	GARCH-t	GARCH-t
***BIST*100**	GARCH-n	HS	GARCH-t	HS	GPD	GARCH-t
ATHEX	DGPD	FHS	FHS	EWMA	EWMA	GARCH-t
***PSI*20**	EWMA	GARCH-t	GARCH-t	FHS	EWMA	EWMA
***IBEX*35**	FHS	DGPD	DGPD	GARCH-t	EWMA	EWMA

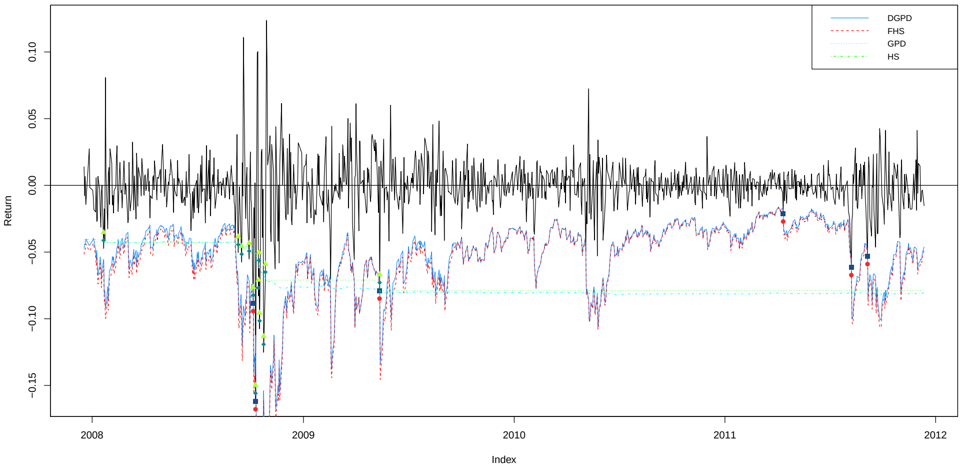
Source: author's computations.

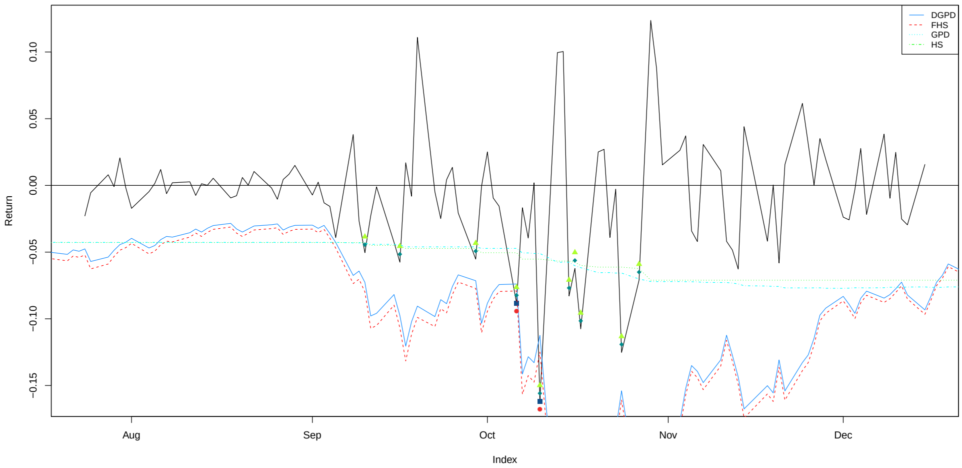
Analyzing the Table 4.9, GARCH-t and EWMA method clearly dominates over semiparametric FHS and DGPD methods, especially in the right tail of markets. Even the Parsimonious EWMA method can show better performance in comparison to more complex DGPD method, which is expected to beat above mentioned method, however the results obtained were indicating poorer performance.

We end this section with Figure 4.9 that shows the out-of sample backtesting performance of DGPD, FHS, GPD and HS methods for PX50 series at

$\alpha = 99.5\%$. Plotted evolution of returns starts from the beginning of 2008 until the beginning of 2012, the period of financial distress that is characterized with high volatility. Superimposed solid, dashed, dotted and dotdashed line represent the evolution of VaR estimates from DGPD, FHS, GPD and HS method, respectively. We marked the dates with different symbols when violation occurs, that is square for DGPD, circle for FHS, diamond for GPD and triangle for HS. We also plotted same figure zoomed on the period of 24/07/2008 - 15/12/2008 (September 2008 financial crash). Figure 4.9 clearly shows how the unconditional methods, HS and GPD cannot react quickly to the changes in the market volatility and tends to create clusters of violations during the stress periods. Instead, conditional methods, DGPD and FHS can react quickly to the changes in volatility and avoid clustering of violations over this period. This can be clearly seen from the zoomed figure, where the DGPD and FHS methods are able to adjust themselves based on the jumps around September 2008 financial crash.

Figure 4.9: Sub-Sample Violation Plot for PX50 at $\alpha = 99.5\%$: Sub-sample VaR backtesting performance of DGPD (solid line), FHS (dashed line), GPD (dotted line) and HS (dotdashed line) superimposed on returns in PX50 for $\alpha = 99.5\%$. Period from 18/12/2007 to 12/12/2011. Square, circle, diamond and triangle symbols points the violations from DGPD, FHS, GPD and HS, respectively. Bottom figure is the same figure but zoomed for the volatile period covering from 24/01/2008 to 15/12/2008.





Source: author's computations.

Chapter 5

Conclusion

The purpose of this study has been to compare the predictive performance of VaR and ES estimates from various univariate methods, with the primary focus on EVT methods. Such comparison was done for five different European stock indexes; namely, PX50 (Prague, Czech Republic), BIST100 (Istanbul, Turkey), ATHEX (Athens, Greece), PSI20 (Lisbon, Portugal) and IBEX35 (Madrid, Spain). For this purpose, we implemented a dynamic setting - so called backtesting procedure and we calculated one-day ahead out-of-sample VaR and ES estimates from each single method, all for 3 different confidence intervals 95, 99, 99.5%. By doing this, we were able to conduct statistical tests for both VaR and ES estimates and to compare their accuracy with the actual risk in the next day. We mainly wanted to answer the following research questions: first, stylized facts such as non-normality, dependence, fat-tailness, asymmetry, etc. are present in the underlying market indices; second, models trying to capture these stylized facts are better than those that do not count for them in terms of their predictive performance for VaR and ES estimation; third, EVT-based method is superior to the non-EVT-based ones in terms of their predictive performance for VaR and ES estimation; and lastly, EVT-based method is significantly better than other methods, especially in the volatile periods; namely, for predicting the extreme risks such as global financial crisis.

This study contributes to the empirical literature on financial risk estimation by providing an extensive and detailed description, subsequent application and detailed comparison of the most popular and most commonly used respective methods. Furthermore, all relevant methods are applied to 5 different stock markets that have not been combined for these purposes before. In addition, as opposed to the largest number of respective studies that usually cover period of several years, our study uses a very long data span (approximately 20 years) that covers periods of several crises that affected chosen markets. Such a large time period coverage allows us to compare the predictive performance of chosen methods in a long span of out-of-sample observations in which the estimates and the actual values are compared.

Results from this study indicate four main findings. First, methods which do not account for the heaviness of the tail almost always fail to be appropriate as a result of systematic VaR underestimation. Furthermore, methods that account for the heaviness of the tail, either in natural or parametric way, are shown to have better accuracy. Hence, a fat-tail phenomenon is one of the most important stylized facts that should be considered for the purpose of VaR calculation. Second, as opposed to unconditional methods, conditional methods that react quickly to the changes in market dynamics and avoid volatility clustering phenomenon are better than the unconditional ones in terms of their predictive performance for VaR estimation. Moreover, the performance of conditional methods is increasing, after capturing the fat-tail and asymmetry phenomenon. Third, as combination of non-parametric and parametric modelling, semi-parametric methods (GPD, FHS, DGPD) are more capable to accurately predict risk in terms of VaR. Finally, EVT based method (DGPD) dominates the others as it was found to be more capable to give an accurate estimation of risk in terms of VaR.

However, when we focus only on the period of global financial crisis, results differ from the above conclusions. To be more precise, the GARCH-t method and parsimonious EWMA method were found to be better than EVT based methods in terms of VaR, especially regarding the high confidence levels (i.e. 99 and 99.5%). This result is very surprising since it is quite opposite from the evidence in empirical literature in which EVT based methods were found to show better predictive performance, especially in the stress periods such as the global financial crisis is.

Regarding the results of ES backtesting, the GARCH-n method was found to be partially useless for the ES calculation. In addition, GARCH-t method that differs from GARCH-n with the assumption of t-distributed innovations was found to be equally useless as GARCH-n method. Finally, semi-parametric FHS and DGPD methods were found to be more capable than other methods for the calculation of ES.

In a nutshell, this study has successfully assessed the potential of EVT in the calculation of market risk. In other words, this study supports the findings from majority of respective empirical studies implying that the EVT, as a sound theory, is one of the most appropriate risk measurement tools. In addition, this study contributes to the overall empirical literature on risk estimation by finding that GARCH family of methods, after accounting for asymmetry and fat tail phenomena by incorporating it into its main equation, can be equally useful and sometimes even better for. However, we have to be cautious not to generalize this findings to all countries and financial indices.

Bibliography

Andersen, H. S. & D. S. Pedersen (2010): *Extreme Value Theory with Applications in Quantitative Risk Management*. Master's thesis, Aarhus School of Business, Aarhus University, Master of Sience in Finance.

Andjelic, G., I. Milosev, & V. DJakovic (2010): "Extreme value theory in emerging markets." *Ekonomski anali* **55(185)**: pp. 63–105.

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

Assaf, A. (2009): "Extreme observations and risk assessment in the equity markets of {MENA} region: Tail measures and value-at-risk." *International Review of Financial Analysis* **18(3)**: pp. 109 – 116.

Avdulaj, K. (2012): "The Extreme Value Theory and Copulas as a Tool to Measure Market Risk." *Bulletin of the Czech Econometric Society* **19(29)**.

Bali, T. G. (2007): "A generalized extreme value approach to financial risk measurement." *Journal of Money, Credit and Banking* **39(7)**: pp. 1613–1649.

Balkema, A. A. & L. De Haan (1974): "Residual life time at great age." *The Annals of Probability* pp. 792–804.

on Banking Supervision, B. C. & B. for International Settlements (2004): *International convergence of capital measurement and capital standards: a revised framework*.

Barone-Adesi, G., F. Bourgoin, & K. Giannopoulos (1998): "Market risk: Don't look back." *RISK-LONDON-RISK MAGAZINE LIMITED-* **11**: pp. 100–103.

BARONE-ADESI, G. & K. GIANNOPOULOS (1996): "A simplified approach to the conditional estimation of value at risk." *Futures and Options World* pp. 68–72.

BARONE-ADESI, G., K. GIANNOPOULOS, & L. VOSPER (2002): "Backtesting Derivative Portfolios with Filtered Historical Simulation (FHS)." *European Financial Management* **8(1)**: pp. 31–58.

BLACK, F. (1976): "Studies of stock price volatility changes." *Proceedings of the Business and Economics Section of the American Statistical Association* pp. 177–181.

BOLLERSLEV, T. (1986): "Generalized autoregressive conditional heteroskedasticity." *Journal of Econometrics* **31(3)**: pp. 307–327.

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

BROOKS, C., A. CLARE, J. DALLE MOLLE, & G. PERSAND (2005): "A comparison of extreme value theory approaches for determining value at risk." *Journal of Empirical Finance* **12(2)**: pp. 339–352.

CAMPBELL, J. Y., A. W. LO, & A. C. MACKINLAY (1997): *The econometrics of financial markets.* Princeton, NJ: Princeton University Press.

CAMPBELL, S. D. (2005): "A review of backtesting and backtesting procedures." *Finance and Economics Discussion Series 2005-21*, Board of Governors of the Federal Reserve System (U.S.).

CHAN, K. F. & P. GRAY (2006): "Using extreme value theory to measure value-at-risk for daily electricity spot prices." *International Journal of Forecasting* **22(2)**: pp. 283 – 300.

CHRISTOFFERSEN, P. (1998): "Evaluating interval forecasts." *International Economic Review* **39**: pp. 841 – 862.

CHRISTOFFERSEN, P., J. HAHN, & A. INOUE (2001): "Testing and comparing value-at-risk measures." *Journal of Empirical Finance* **8(3)**: pp. 325 – 342.

CIFTER, A. (2011): "Value-at-risk estimation with wavelet-based extreme value theory: Evidence from emerging markets." *Physica A: Statistical Mechanics and its Applications* **390(12)**: pp. 2356 – 2367.

DANIELSSON, J., B. N. JORGENSEN, G. SAMORODNITSKY, M. SARMA, & C. G. DE VRIES (2013): "Fat tails, VaR and subadditivity." *Journal of Econometrics* **172(2)**: pp. 283–291.

DANIELSSON, J. & C. G. D. VRIES (1998): "Value-at-Risk and Extreme Returns." *Tinbergen Institute Discussion Papers 98-017/2*, Tinbergen Institute.

DANIELSSON, J. & C. G. D. VRIES (2000): "Value-at-Risk and Extreme Returns." *Annales d'Economie et de Statistique* **(60)**: pp. 239–270.

DEKKERS, A. L. & L. DE HAAN (1989): "On the estimation of the extreme-value index and large quantile estimation." *The Annals of Statistics* pp. 1795–1832.

DIEBOLD, F. X., T. SCHUERMANN, & J. D. STROUGHAIR (1998): "Pitfalls and Opportunities in the Use of Extreme Value Theory in Risk Management." *New York University, Leonard N. Stern School Finance Department Working Paper Seires 98-081*, New York University, Leonard N. Stern School of Business.

DJAKOVIC, V., G. ANDJELIC, & J. BOROCKI (2011): "Performance of extreme value theory in emerging markets: an empirical treatment." *African Journal of Business Management* **5(2)**: pp. 340–369.

EMBRECHTS, P., C. KLÜPPELBERG, & T. MIKOSCH (1997): *Modelling extremal events: for insurance and finance.* Springer.

EMBRECHTS, P., S. I. RESNICK, & G. SAMORODNITSKY (1999): "Extreme value theory as a risk management tool." *North American Actuarial Journal* **3(2)**: pp. 30–41.

ENGLE, R. (2001): "GARCH 101: The Use of ARCH/GARCH Models in Applied Econometrics." *Journal of Economic Perspectives* **15(4)**: pp. 157–168.

ENGLE, R. F. (1982): "Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation." *Econometrica: Journal of the Econometric Society* **50(4)**: pp. 987–1007.

ENGLE, R. F. & S. MANGANELLI (2001): "Value at risk models in finance." *Working Paper Series 0075*, European Central Bank.

ERGEN, I. (2010): "Var prediction for emerging stock markets: Garch filtered skewed t distribution and garch filtered evt method." *Technical report*, Supervision Regulation and Credit, Policy Analysis Unit Federal Reserve Bank of Richmond Baltimore, MD.

FAMA, E. F. (1965): "The behavior of stock-market prices." *Journal of business* **38**: pp. 34–105.

FINGER, C. (2005): "Back to backtesting." *RiskMetrics Monthly Research* .

FISHER, R. A. & L. H. C. TIPPETT (1928): "Limiting forms of the frequency distribution of the largest or smallest member of a sample." In "Mathematical Proceedings of the Cambridge Philosophical Society," volume 24, pp. 180–190. Cambridge Univ Press.

FURIO, D. & F. J. CLIMENT (2013): "Extreme value theory versus traditional GARCH approaches applied to financial data: a comparative evaluation." *Quantitative Finance* **13(1)**: pp. 45–63.

GENCAY, R. & F. SELCUK (2004): "Extreme value theory and value-at-risk: Relative performance in emerging markets." *International Journal of Forecasting* **20(2)**: pp. 287 – 303. Forecasting Economic and Financial Time Series Using Nonlinear Methods.

GLOSTEN, L. R., R. JAGANNATHAN, & D. E. RUNKLE (1993): " On the Relation between the Expected Value and the Volatility of the Nominal Excess Return on Stocks." *Journal of Finance* **48(5)**: pp. 1779–1801.

GNEDENKO, B. (1943): "Sur la distribution limite du terme maximum d'une serie aleatoire." *Annals of mathematics* **44**: pp. 423–453.

GOLDBERG, L. R., G. MILLER, & J. WEINSTEIN (2008): "Beyond value at risk: Forecasting portfolio loss at multiple horizons." *Journal of Investment Management* **6(2)**: p. 73.

HILL, B. M. (1975): "A simple general approach to inference about the tail of a distribution." *The annals of statistics* **3(5)**: pp. 1163–1174.

HUISMAN, R., K. G. KOEDIJK, C. J. M. KOOL, & F. PALM (2001): "Tail-index estimates in small samples." *Journal of Business & Economic Statistics* **19(2)**: pp. 208–216.

HULL, J. & A. WHITE (1998): "Incorporating volatility updating into the historical simulation method for value-at-risk." *Journal of Risk* **1(1)**: pp. 5–19.

JENKINSON, A. F. (1955): "The frequency distribution of the annual maximum (or minimum) values of meteorological elements." *Quarterly Journal of the Royal Meteorological Society* **81(348)**: pp. 158–171.

JORION, P. (2007): *Value at risk: the new benchmark for managing financial risk*, volume 2. McGraw-Hill New York.

JP MORGAN RISKMETRICS, R. (1996): "Riskmetrics® tm-technical document." *Morgan Guaranty Trust Company, Reuters Ltd, 4th edition, New York* .

KEARNS, P. & A. PAGAN (1997): "Estimating the density tail index for financial time series." *Review of Economics and Statistics* **79(2)**: pp. 171–175.

KËLLEZI, E. & M. GILLI (2000): "Extreme Value Theory for Tail-Related Risk Measures." *FAME Research Paper Series rp18*, International Center for Financial Asset Management and Engineering.

KËLLEZI, E. & M. GILLI (2006): "An Application of Extreme Value Theory for Measuring Financial Risk." *Computational Economics* **27(2)**: pp. 207–228.

KOKOSZKA, P. *et al.* (2003): "Garch processes: structure and estimation." *Bernoulli* **9(2)**: pp. 201–227.

KUESTER, K., S. MITTNIK, & M. S. PAOLELLA (2006): "Value-at-Risk Prediction: A Comparison of Alternative Strategies." *Journal of Financial Econometrics* **4(1)**: pp. 53–89.

KUPIEC, P. H. (1995): "Techniques for verifying the accuracy of risk measurement models." *The Journal of Derivatives* **3(2)**: pp. 73–184.

LEE, T.-H., Y. BAO, & B. SALTOGLU (2006): "Evaluating predictive performance of value-at-risk models in emerging markets: a reality check." *Journal of Forecasting* **25(2)**: pp. 101–128.

LONGIN, F. M. (1999): "From Value at Risk to Stress Testing: The Extreme Value Approach." *Technical Report 2161*, C.E.P.R. Discussion Papers.

LONGIN, F. M. (2000): "From value at risk to stress testing: The extreme value approach." *Journal of Banking & Finance* **24(7)**: pp. 1097–1130.

LUX, T. (2001): "The limiting extremal behaviour of speculative returns: an analysis of intra-daily data from the frankfurt stock exchange." *Applied Financial Economics* **11(3)**: pp. 299–315.

MAHONEY, J. M. (1995): "Empirical-based versus model-based approaches to Value-at-Risk: an examination of foreign exchange and global equity portfolios." *Proceedings* pp. 199–220.

MANDELBROT, B. (1963): "The variation of certain speculative prices." *Journal of Business* **(4)**: pp. 394–419.

MAPA, D. S. & O. Q. SUAISO (2009): "Measuring market risk using extreme value theory." *MPRA Paper 21246*, University Library of Munich, Germany.

MCNEIL, A. J. & R. FREY (2000): "Estimation of tail-related risk measures for heteroscedastic financial time series: an extreme value approach." *Journal of Empirical Finance* **7(3–4)**: pp. 271 – 300. Special issue on Risk Management.

MCNEIL, A. J., R. FREY, & P. EMBRECHTS (2005): *Quantitative Risk Management: Concepts, Techniques and Tools.* Princeton University Press, Princeton, NJ.

MUTU, S., P. BALOGH, & D. MOLDOVAN (2011): "The efficiency of value at risk models on central and eastern european stock markets." *Development, energy, environment, economics* **10**: pp. 382–388.

NEFTCI, S. (2000): "Value at risk calculations, extreme events, and tail estimation." *The Journal of Derivatives* **7(3)**: pp. 23–37.

NELSON, D. B. (1991): "Conditional Heteroskedasticity in Asset Returns: A New Approach." *Econometrica* **59(2)**: pp. 347–70.

NIEPPOLA, O. (2009): *Backtesting value-at-risk models.* Master's thesis, Helsinki School of Economics, Department of Economics.

NYSTROM, K. & J. SKOGLUND (2002): "Univariate extreme value theory, garch and measures of risk." *Preprint, Swedbank* .

OZUN, A., A. CIFTER, & S. YILMAZER (2007): "Filtered Extreme Value Theory for Value-At-Risk Estimation." *MPRA Paper 3302*, University Library of Munich, Germany.

PICKANDS III, J. (1975): "Statistical inference using extreme order statistics." *the Annals of Statistics* pp. 119–131.

POWNALL, R. A. & K. G. KOEDIJK (1999): "Capturing downside risk in financial markets: the case of the asian crisis." *Journal of International Money and Finance* **18(6)**: pp. 853–870.

ROCCO, M. (2011): "Extreme value theory for finance: a survey." *Questioni di Economia e Finanza (Occasional Papers) 99*, Bank of Italy, Economic Research and International Relations Area.

SAMANTA, R. & B. LEBARON (2005): "Extreme Value Theory and Fat Tails in Equity Markets." *Computing in economics and finance 2005*, Society for Computational Economics.

SINGH, A. K., D. E. ALLEN, & P. J. ROBERT (2012): "Risk and dependence analysis of australian stock market-the case of extreme value theory." *25th Australasian Finance and Banking Conference 2012* .

SINGH, A. K., D. E. ALLEN, & P. J. ROBERT (2013): "Extreme market risk and extreme value theory." *Mathematics and Computers in Simulation* **94(0)**: pp. 310 – 328.

TRZPIOT, G. & J. MAJEWSKA (2010): "Estimation of Value at Risk: Extreme value and robust approaches." *Operations Research and Decisions* **1**: pp. 131–143.

ZAKOIAN, J.-M. (1994): "Threshold heteroskedastic models." *Journal of Economic Dynamics and Control* **18(5)**: pp. 931–955.

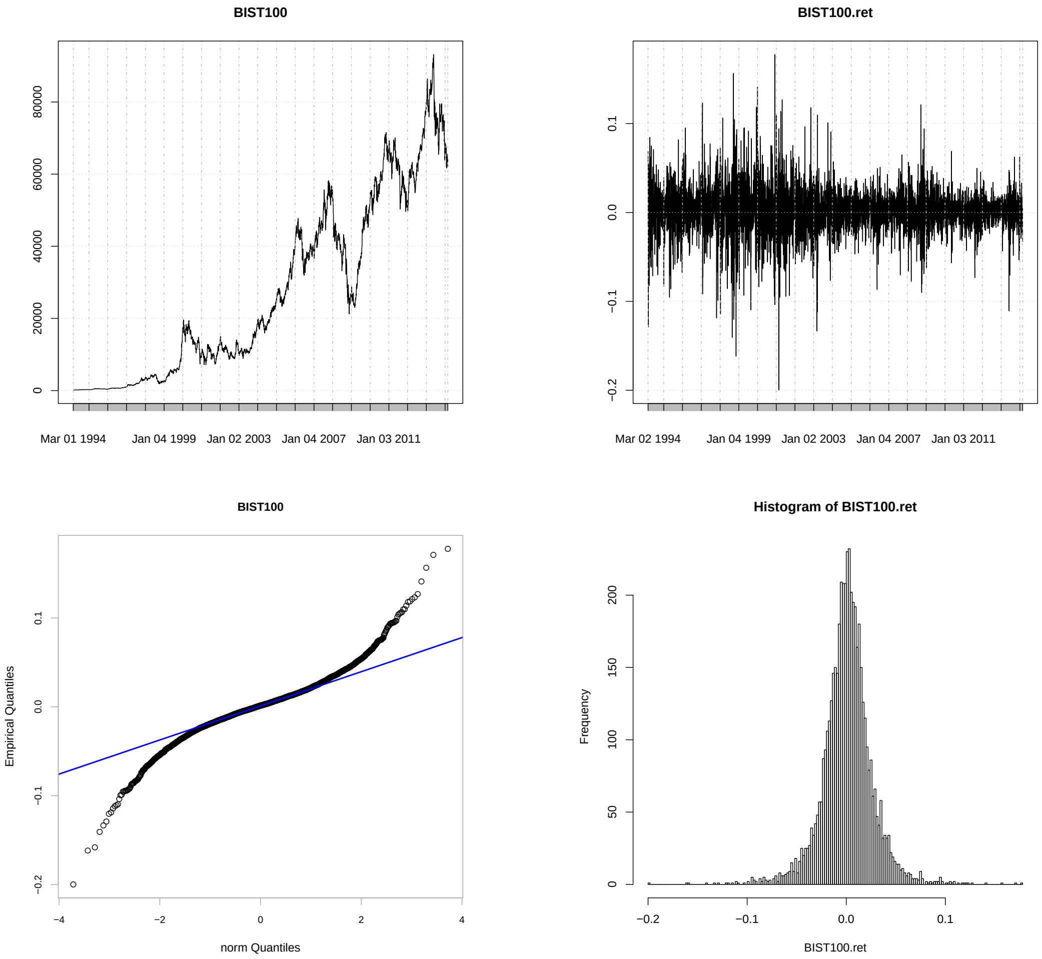
ZIKOVIC, S. & B. AKTAN (2009): "Global financial crisis and VaR performance in emerging markets: A case of EU candidate states - Turkey and Croatia." *Zbornik radova Ekonomskog fakulteta u Rijeci/Proceedings of Rijeka Faculty of Economics* **27(1)**: pp. 149–170.

Appendix A

Tables and Figures

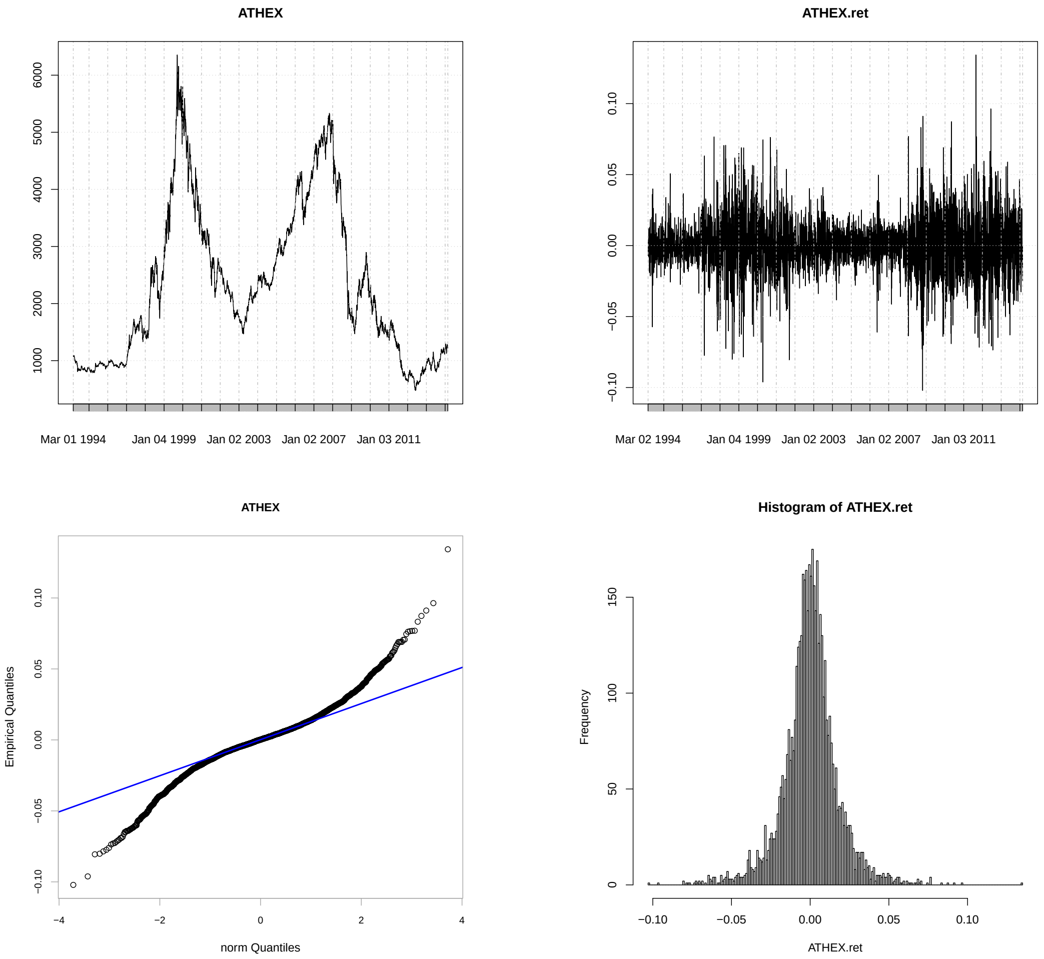
Preliminary Plots

Figure A.1: Preliminary Plots for BIST100



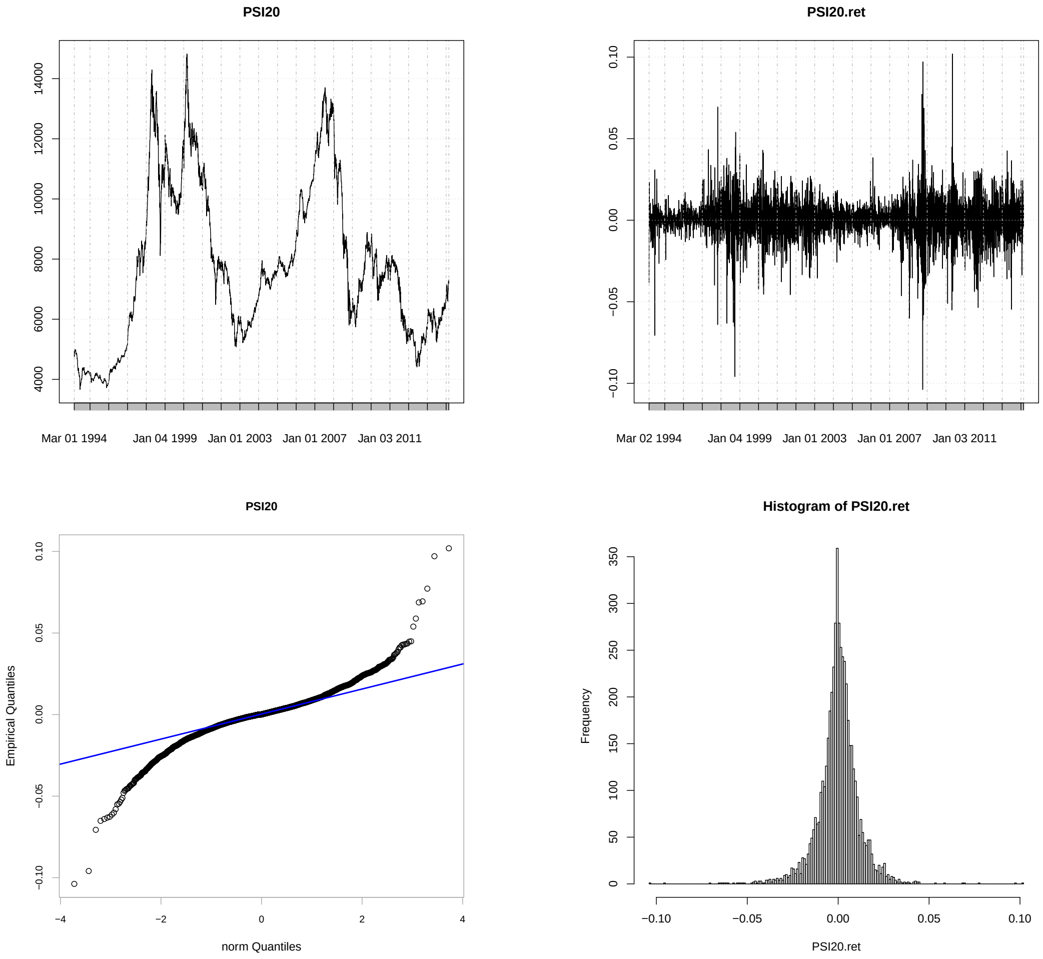
Source: author's computations.

Figure A.2: Preliminary Plots for ATHEX



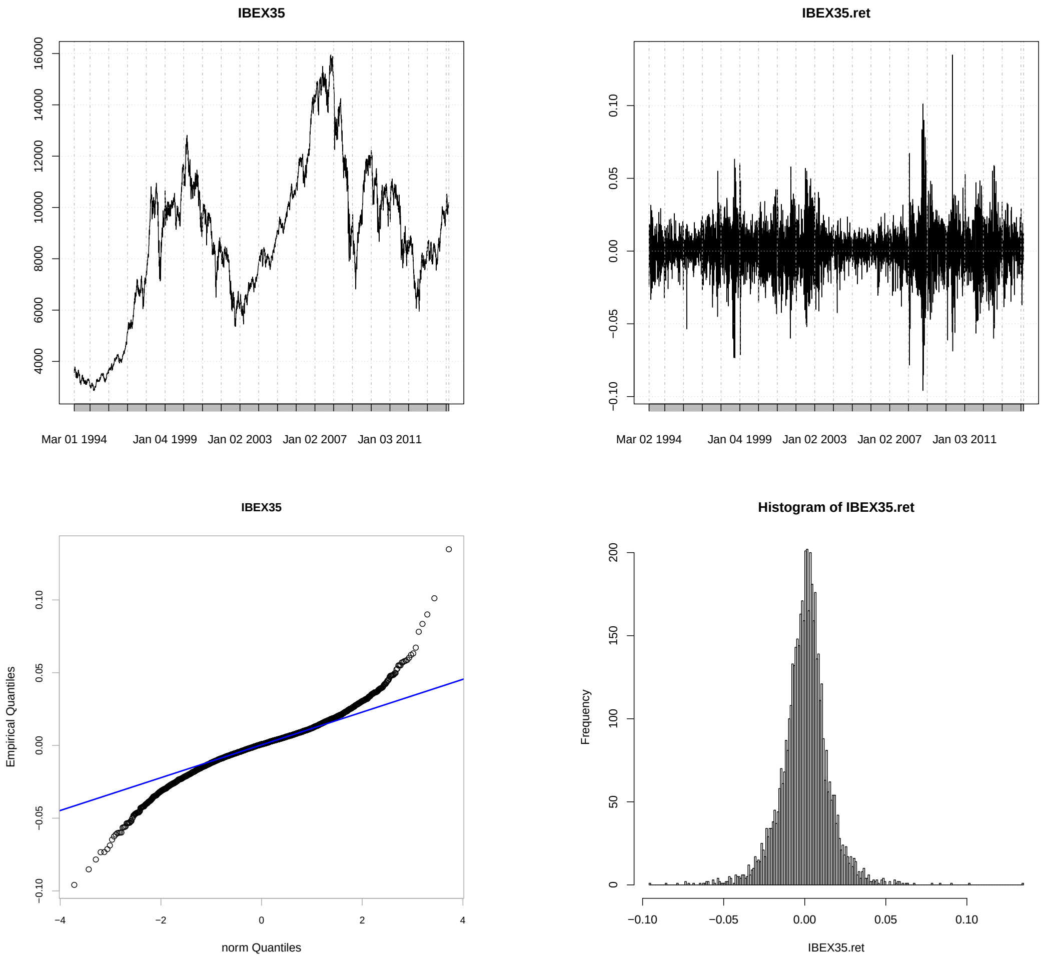
Source: author's computations.

Figure A.3: Preliminary Plots for PSI20



Source: author's computations.

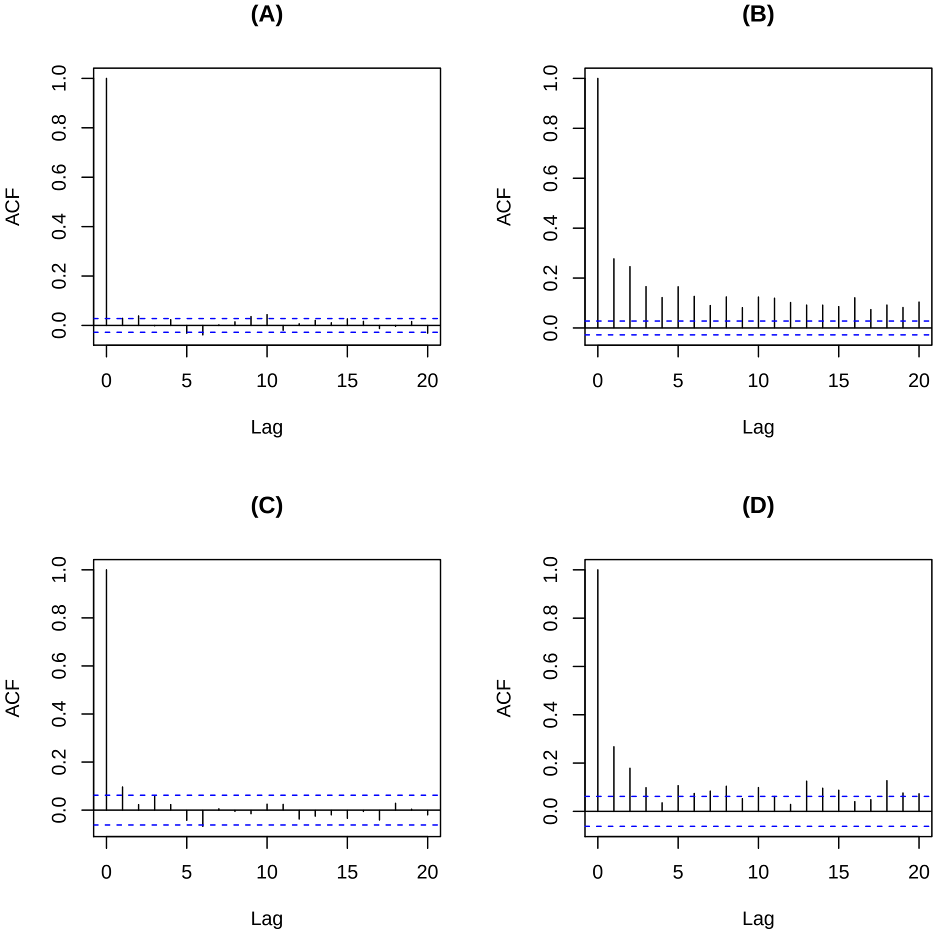
Figure A.4: Preliminary Plots for IBEX35



Source: author's computations.

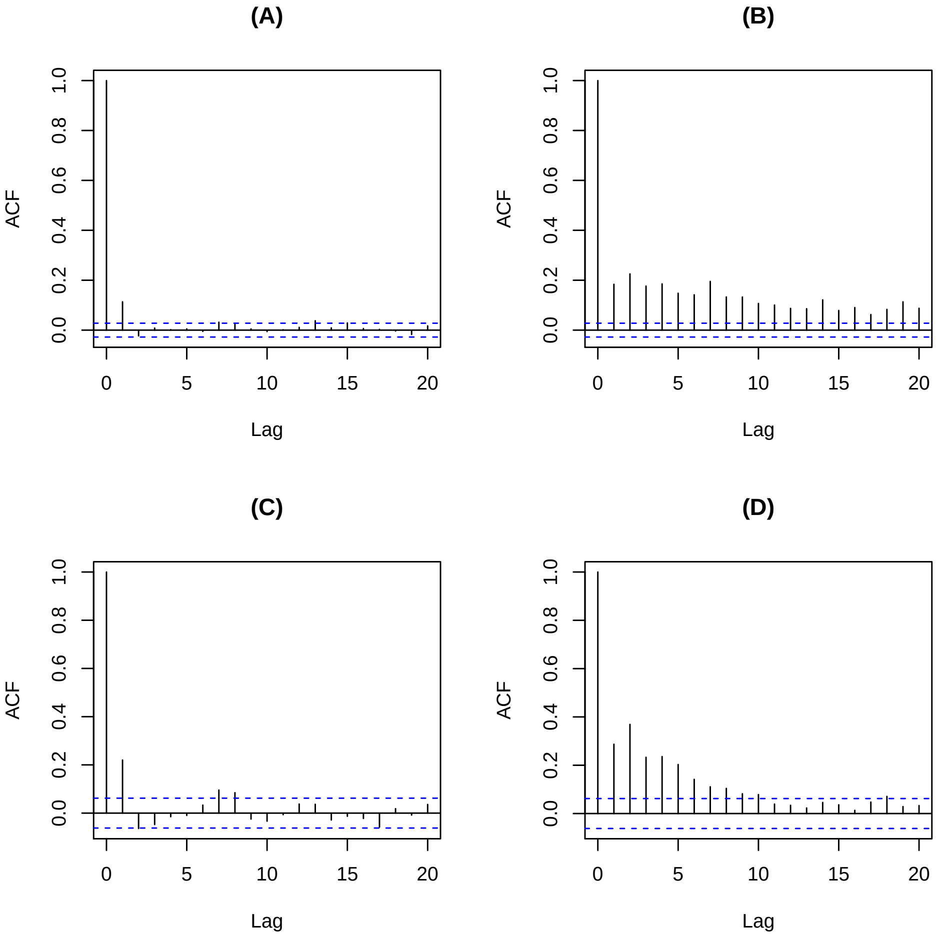
Correlogram Plots

Figure A.5: Correlogram of Returns and Squared Returns for BIST100: Full-Sample ACF calculated with 20 lags from both raw logarithmic returns (A) and their squared values (B). In-Sample ACF calculated with 20 lags from both raw logarithmic returns (C) and their squared values (D).



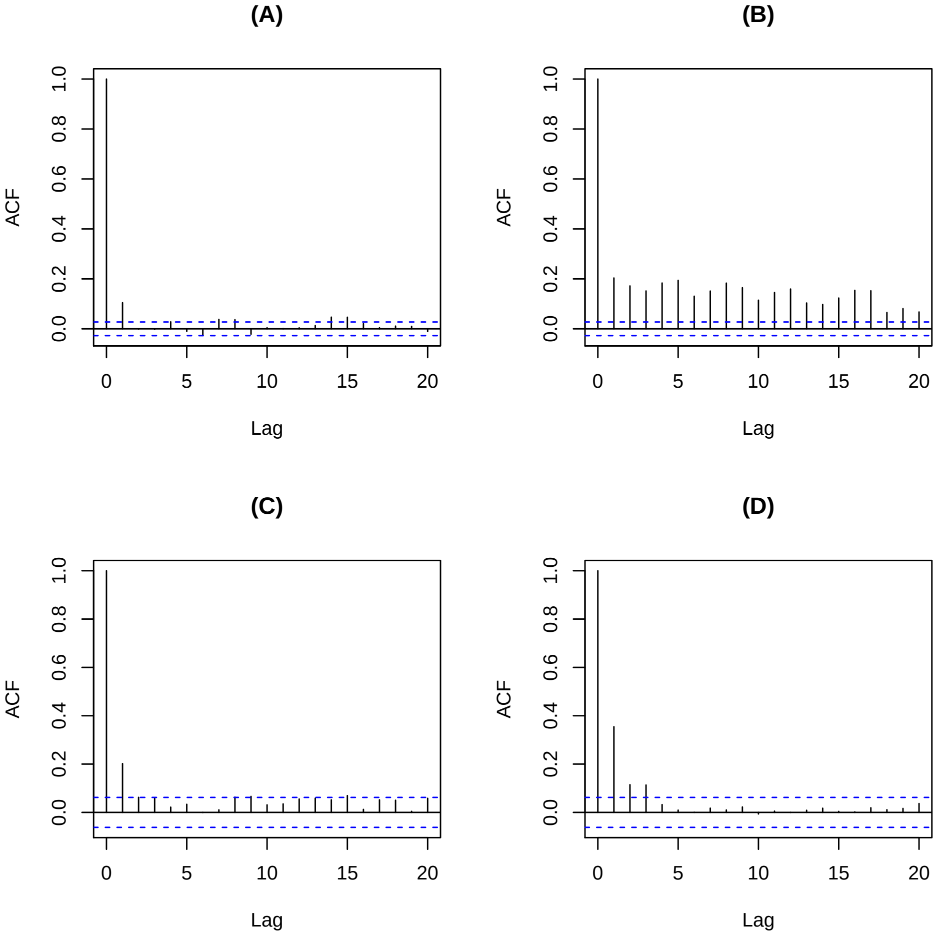
Source: author's computations.

Figure A.6: Correlogram of Returns and Squared Returns for ATHEX: Full-Sample ACF calculated with 20 lags from both raw logarithmic returns (A) and their squared values (B). In-Sample ACF calculated with 20 lags from both raw logarithmic returns (C) and their squared values (D).



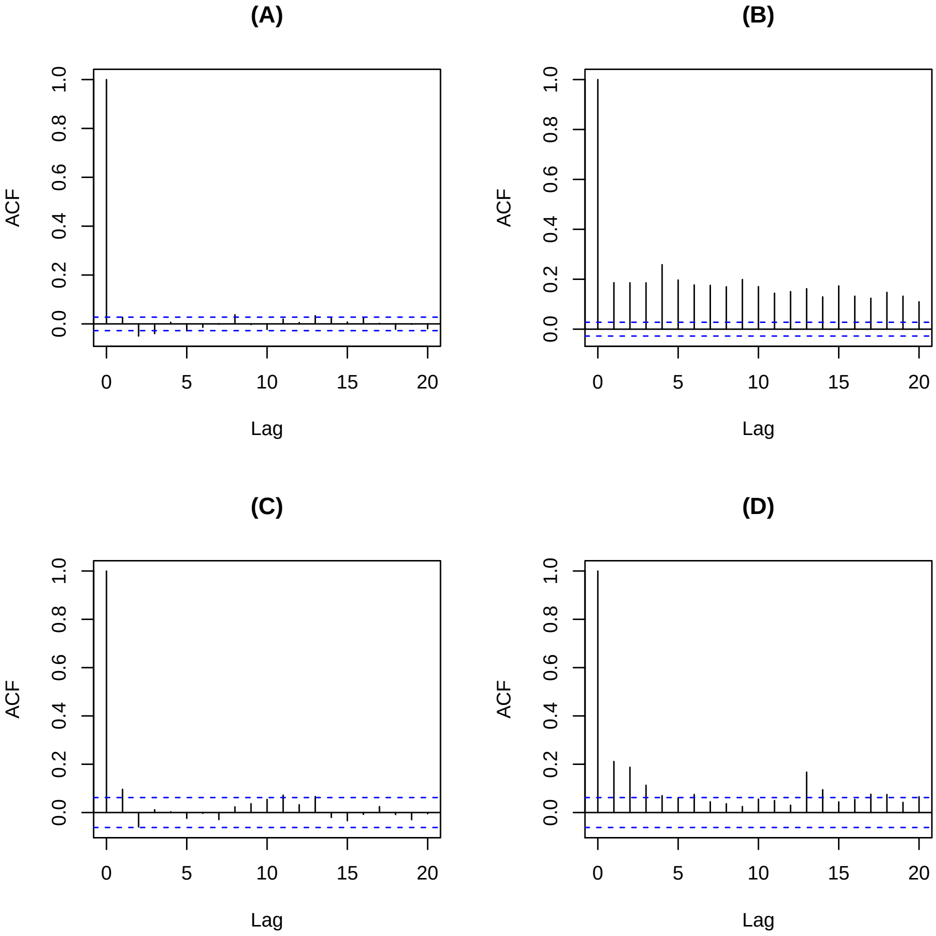
Source: author's computations.

Figure A.7: Correlogram of Returns and Squared Returns for PSI20: Full-Sample ACF calculated with 20 lags from both raw logarithmic returns (A) and their squared values (B). In-Sample ACF calculated with 20 lags from both raw logarithmic returns (C) and their squared values (D).



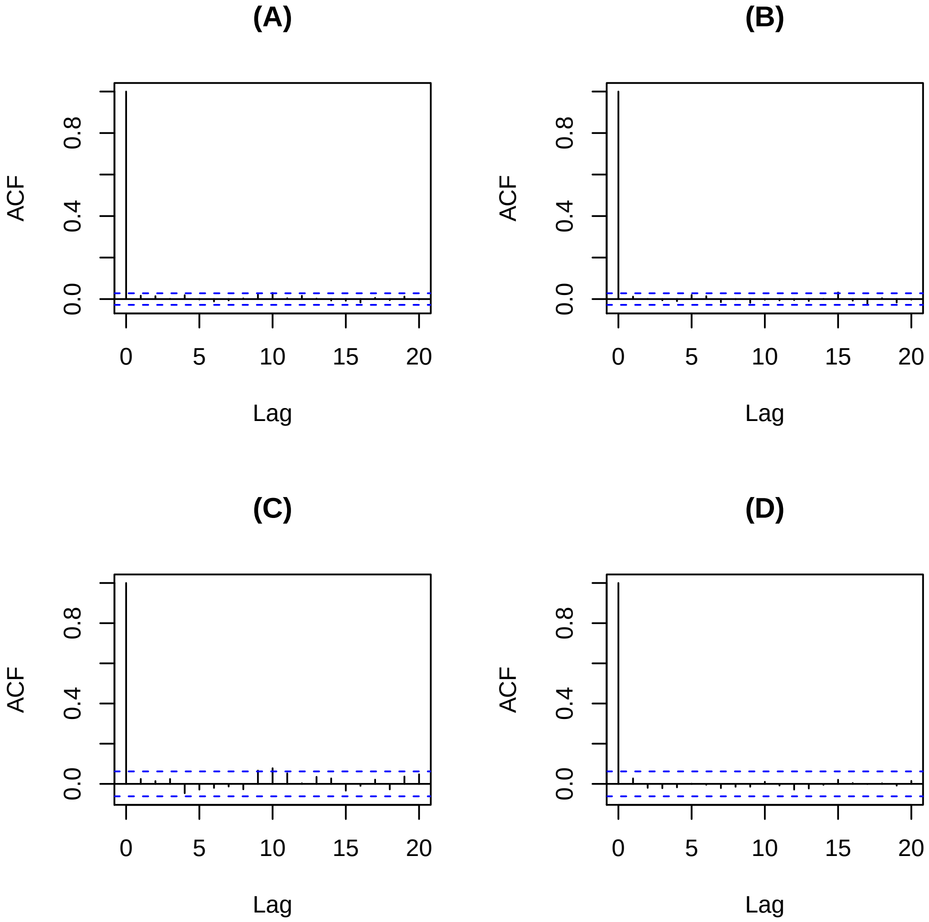
Source: author's computations.

Figure A.8: Correlogram of Returns and Squared Returns for IBEX35: Full-Sample ACF calculated with 20 lags from both raw logarithmic returns (A) and their squared values (B). In-Sample ACF calculated with 20 lags from both raw logarithmic returns (C) and their squared values (D).



Source: author's computations.

Figure A.9: Correlogram of t-Residuals and Squared t-Residuals for PX50: Full-Sample ACF calculated with 20 lags from both t-residuals (A) and their squared values (B). In-Sample ACF calculated with 20 lags from both t-residuals (C) and their squared values (D).



Source: author's computations.

Full-sample Backtesting Tables

Table A.1: Left Tail Backtesting Results at $\alpha = 95\%$: EV and AV stand for the expected and actual number of violations, respectively. fracAV is the fraction of the actual number of violations over the all observations. VR stands for the violation ratios calculated. LRuc, LRind and LRcc are the likelihood ratio test statistics from the unconditional coverage, independence and conditional coverage tests, respectively. * stands for 1% significance while ** stands for the 5% significance for the underlying tests. Critical values at the 5% and 1% significance are 3.8415 and 6.6349 for LRuc and LRind, 5.9915 and 9.2103 for LRcc, respectively. Finally average is the mean of the VaRs calculated from the methods over the backtesting period.

	Models	EV	AV	fracAV	VR	LRuc	LRind	LRcc	average
PX50	Normal	196	213	0.0542	1.0834	1.4036	49.8669*	51.2706*	0.0235
	HS	196	222	0.0565	1.1292	3.3218	47.2579*	50.5798*	0.0222
	GPD	196	224	0.0570	1.1394	3.8542**	46.0584*	49.9126*	0.0223
	EWMA	196	221	0.0562	1.1276	3.0700	6.7020*	9.7720*	0.0212
	GARCH-n	196	201	0.0511	1.0224	0.1029	1.2956	1.3985	0.0212
	GARCH-t	196	214	0.0544	1.0885	1.5776	1.5619	3.1395	0.0207
	FHS	196	208	0.0529	1.0580	0.6834	1.3939	2.0774	0.0208
	DGPD	196	208	0.0529	1.0580	0.6834	1.3939	2.0774	0.0210
BIST100	Normal	199	169	0.0424	0.8475	5.1333**	15.1374*	20.2707*	0.0406
	HS	199	179	0.0449	0.8977	2.2718	23.5310*	25.8028*	0.0391
	GPD	199	181	0.0454	0.9077	1.8419	20.2801*	22.1220*	0.0392
	EWMA	199	224	0.0562	1.1256	3.0773	10.8509*	13.9282*	0.0360
	GARCH-n	199	187	0.0469	0.9378	0.8282	5.3637**	6.1919**	0.0373
	GARCH-t	199	195	0.0489	0.9779	0.1029	4.0705**	4.1734	0.0361
	FHS	199	214	0.0537	1.0732	1.1002	7.1566*	8.2567**	0.0369
	DGPD	199	214	0.0537	1.0732	1.1002	7.1566*	8.2567**	0.0370
ATHEX	Normal	199	232	0.0582	1.1644	5.3943**	42.2841*	47.6784*	0.0286
	HS	199	244	0.0612	1.2246	9.9037*	51.6691*	61.5728*	0.0277
	GPD	199	231	0.0580	1.1593	5.0771**	45.5849*	50.6620*	0.0282
	EWMA	199	240	0.0602	1.2060	8.2579*	8.4927*	16.7506*	0.0281
	GARCH-n	199	196	0.0492	0.9837	0.0561	0.0293	0.0854	0.0291
	GARCH-t	199	207	0.0519	1.0389	0.3135	0.1161	0.4296	0.0285
	FHS	199	198	0.0497	0.9937	0.0083	0.4517	0.4599	0.0286
	DGPD	199	199	0.0499	0.9987	0.0003	0.0890	0.0893	0.0287
PSI20	Normal	206	227	0.0551	1.1014	2.1629	50.1760*	52.3390*	0.0188
	HS	206	263	0.0638	1.2761	15.2657*	53.4271*	68.6928*	0.0178
	GPD	206	259	0.0628	1.2567	13.2636*	55.9551*	69.2186*	0.0180
	EWMA	206	260	0.0631	1.2621	13.7518*	18.2650*	32.0168*	0.0175
	GARCH-n	206	232	0.0563	1.1257	3.2981	1.1436	4.4416	0.0180
	GARCH-t	206	234	0.0568	1.1354	3.8164	0.9961	4.8125	0.0176
	FHS	206	230	0.0558	1.1160	2.8164	1.9936	4.8099	0.0180
	DGPD	206	231	0.0560	1.1208	3.0526	1.8939	4.9465	0.0180
IBEX35	Normal	201	229	0.0568	1.1351	3.7194	18.5179*	22.2374*	0.0244
	HS	201	251	0.0622	1.2441	11.7842*	20.4669*	32.2511*	0.0237
	GPD	201	247	0.0612	1.2243	10.0128*	21.9386*	31.9513*	0.0239
	EWMA	201	252	0.0625	1.2537	12.2511*	0.0583	12.3094*	0.0228
	GARCH-n	201	247	0.0612	1.2243	10.0128*	0.0087	10.0041*	0.0224
	GARCH-t	201	261	0.0647	1.2937	14.7232*	0.0698	14.793*	0.0222
	FHS	201	246	0.0610	1.2193	9.5774*	0.0082	9.5856*	0.0229
	DGPD	201	237	0.0587	1.1747	6.1538**	0.2649	6.4187**	0.0229

Source: author's computations.

Table A.2: Left Tail Backtesting Results at $\alpha = 99\%$: EV and AV are the expected and actual number of violations respectively. fracAV is the fraction of the actual number of violations over the all observations. VR stands for the violation ratios calculated. LRuc, LRind and LRcc are the likelihood ratio test statistics from the unconditional coverage, independence and conditional coverage tests, respectively. * stands for 1% significance while ** stands for the 5% significance for the underlying tests. Critical values at the 5% and 1% significance are 3.8415 and 6.6349 for LRuc and LRind, 5.9915 and 9.2103 for LRcc, respectively. Finally average is the mean of the VaRs calculated from the methods over the backtesting period.

	Models	EV	AV	fracAV	VR	LRuc	LRind	LRcc	average
PX50	Normal	39	85	0.0216	2.1617	40.2342*	35.9538*	76.1880*	0.0333
	HS	39	51	0.0130	1.2970	3.2045	16.8849*	20.0894*	0.0407
	GPD	39	49	0.0125	1.2462	2.2326	17.8099*	20.0425*	0.0412
	EWMA	39	72	0.0183	1.8462	22.0255*	9.4203*	31.4457*	0.0302
	GARCH-n	39	68	0.0173	1.7294	17.3491*	2.3935	19.7426*	0.0302
	GARCH-t	39	53	0.0135	1.3479	4.3353**	1.4484	5.7837	0.0327
	FHS	39	44	0.0112	1.1190	0.5418	0.9959	1.5377	0.0343
	DGPD	39	45	0.0114	1.1445	0.7919	1.0420	1.8339	0.0344
BIST100	Normal	39	74	0.0186	1.8556	23.5479*	12.3924*	35.9403*	0.0579
	HS	39	40	0.0100	1.0030	0.0004	3.3547	3.3551	0.0678
	GPD	39	40	0.0100	1.0030	0.0004	3.3547	3.3551	0.0686
	EWMA	39	72	0.0181	1.8462	21.0960*	13.0595*	34.1555*	0.0515
	GARCH-n	39	67	0.0168	1.6800	15.4683*	11.1878*	26.6561*	0.0533
	GARCH-t	39	50	0.0125	1.2538	2.4008	2.0266	4.4274	0.0578
	FHS	39	41	0.0103	1.0281	0.0315	16.5843*	16.6158*	0.0629
	DGPD	39	40	0.0100	1.0030	0.0004	11.8384*	11.8388*	0.0630
ATHEX	Normal	39	106	0.0266	2.6600	76.2186*	17.9566*	94.1752*	0.0405
	HS	39	61	0.0153	1.5307	9.7553*	13.0211*	22.7764*	0.0495
	GPD	39	65	0.0163	1.6311	13.4651*	11.6601*	25.1252*	0.0483
	EWMA	39	77	0.0193	1.9744	27.4881*	0.1647	27.6528*	0.0400
	GARCH-n	39	67	0.0168	1.6813	15.5097*	0.5709	16.0806*	0.0411
	GARCH-t	39	53	0.0133	1.3300	3.9718**	1.4288	5.4007	0.0442
	FHS	39	52	0.0130	1.3049	3.4141	0.1364	3.5505	0.0442
	DGPD	39	52	0.0130	1.3049	3.4141	0.1364	3.5505	0.0442
PSI20	Normal	41	112	0.0272	2.7171	83.5797*	30.1367*	113.7164*	0.0266
	HS	41	71	0.0172	1.7225	17.8712*	6.9234*	24.7947*	0.0332
	GPD	41	72	0.0175	1.7467	18.9877*	9.8811*	28.8687*	0.0333
	EWMA	41	87	0.0211	2.1220	38.9308*	3.9140**	42.8448*	0.0250
	GARCH-n	41	81	0.0197	1.9651	30.2642*	0.0998	30.3640*	0.0254
	GARCH-t	41	46	0.0112	1.1160	0.5396	0.3703	0.9100	0.0279
	FHS	41	59	0.0143	1.4313	6.8340*	1.7136	8.5476**	0.0271
	DGPD	41	55	0.0133	1.3343	4.2116**	1.4876	5.6993	0.0277
IBEX35	Normal	40	98	0.0243	2.4287	59.4617*	6.2870**	65.7488*	0.0346
	HS	40	58	0.0144	1.4374	6.8689*	14.2588*	21.1277*	0.0409
	GPD	40	59	0.0146	1.4622	7.6208*	13.8811*	21.5020*	0.0408
	EWMA	40	83	0.0206	2.0750	34.8844*	0.8347	35.7190*	0.0324
	GARCH-n	40	86	0.0213	2.1314	39.3857*	0.4705	39.8562*	0.0317
	GARCH-t	40	65	0.0161	1.6109	12.8359*	0.0020	12.8379*	0.0336
	FHS	40	61	0.0151	1.5118	9.2274*	1.8727	11.1001*	0.0338
	DGPD	40	57	0.0141	1.4126	6.1519**	1.6335	7.7855**	0.0352

Source: author's computations.

Table A.3: Left Tail Backtesting Results at $\alpha = 99.5\%$: EV and AV are the expected and actual number of violations respectively. fracAV is the fraction of the actual number of violations over the all observations. VR stands for the violation ratios calculated. LRuc, LRind and LRcc are the likelihood ratio test statistics from the unconditional coverage, independence and conditional coverage tests, respectively. * stands for 1% significance while ** stands for the 5% significance for the underlying tests. Critical values at the 5% and 1% significance are 3.8415 and 6.6349 for LRuc and LRind, 5.9915 and 9.2103 for LRcc, respectively. Finally average is the mean of the VaRs calculated from the methods over the backtesting period

	Models	EV	AV	fracAV	VR	LRuc	LRind	LRcc	average
PX50	Normal	19	64	0.0163	3.2553	62.9024*	24.8458*	87.7482*	0.0369
	HS	19	27	0.0069	1.3733	2.4653	18.0032*	20.4685*	0.0501
	GPD	19	25	0.0064	1.2716	1.3418	12.5810*	13.9227*	0.0510
	EWMA	19	58	0.0148	3.0526	49.1924*	1.1519	50.3442*	0.0335
	GARCH-n	19	49	0.0125	2.4924	31.0375*	1.2367	32.2742*	0.0335
	GARCH-t	19	33	0.0084	1.6785	7.5484*	0.5586	8.1070**	0.0380
	FHS	19	24	0.0061	1.2208	0.8993	0.2948	1.1940	0.0404
	DGPD	19	22	0.0056	1.1190	0.2695	0.2476	0.5171	0.0407
BIST100	Normal	19	53	0.0133	2.6580	37.7780*	4.2996**	42.0776*	0.0642
	HS	19	22	0.0055	1.1033	0.2069	0.2441	0.4510	0.0791
	GPD	19	20	0.0050	1.0030	0.0002	2.8801	2.8803	0.0821
	EWMA	19	55	0.0138	2.8947	41.7973*	11.0210*	52.8182*	0.0571
	GARCH-n	19	44	0.0110	2.2066	21.6748*	0.4280	22.1028*	0.0591
	GARCH-t	19	24	0.0060	1.2036	0.7798	0.2906	1.0704	0.0677
	FHS	19	24	0.0060	1.2036	0.7798	13.1596*	13.9394*	0.0725
	DGPD	19	23	0.0058	1.1535	0.4496	13.6813*	14.1310*	0.0754
ATHEX	Normal	19	87	0.0218	4.3664	123.4554*	22.3106*	145.7661*	0.0448
	HS	19	42	0.0105	2.1079	18.6115*	6.6748*	25.2863*	0.0564
	GPD	19	35	0.0088	1.7566	9.3435*	19.7435*	29.0870*	0.0568
	EWMA	19	51	0.0128	2.6842	33.9589*	0.1631	34.1221*	0.0443
	GARCH-n	19	42	0.0105	2.1079	18.6115*	0.5298	19.1412*	0.0455
	GARCH-t	19	25	0.0063	1.2547	1.2015	0.3157	1.5172	0.0512
	FHS	19	22	0.0055	1.1041	0.2100	0.2443	0.4543	0.0505
	DGPD	19	24	0.0060	1.2045	0.7860	0.2908	1.0768	0.0507
PSI20	Normal	20	92	0.0223	4.4639	133.7362*	20.7420*	154.4781*	0.0295
	HS	20	37	0.0090	1.7952	10.5861*	0.8926	11.4787*	0.0400
	GPD	20	32	0.0078	1.5526	5.4091*	1.3171	6.7262**	0.0409
	EWMA	20	62	0.0150	3.1000	54.2075*	0.9528	55.1603*	0.0277
	GARCH-n	20	54	0.0131	2.6201	37.5190*	1.4337	38.9527*	0.0281
	GARCH-t	20	19	0.0046	0.9219	0.1298	0.1760	0.3058	0.0326
	FHS	20	32	0.0078	1.5526	5.4091*	0.5007	5.9098	0.0303
	DGPD	20	29	0.0070	1.4071	3.0453	0.4109	3.4563	0.0315
IBEX35	Normal	20	73	0.0181	3.6183	82.8063*	12.8603*	95.6667*	0.0383
	HS	20	33	0.0082	1.6357	6.8672*	1.1913	8.0585**	0.0493
	GPD	20	31	0.0077	1.5366	5.0109**	1.3831	6.3940**	0.0486
	EWMA	20	56	0.0139	2.8000	43.0123*	3.8326	46.8449*	0.0360
	GARCH-n	20	56	0.0139	2.7757	43.0123*	1.5763	44.5886*	0.0351
	GARCH-t	20	38	0.0094	1.8835	12.5480*	0.7226	13.2706*	0.0383
	FHS	20	34	0.0084	1.6853	7.8880*	0.5779	8.4658**	0.0384
	DGPD	20	35	0.0087	1.7348	8.9681*	0.6125	9.5806*	0.0400

Source: author's computations.

Table A.4: Right Tail Backtesting Results at $\alpha = 95\%$: EV and AV are the expected and actual number of violations respectively. fracAV is the fraction of the actual number of violations over the all observations. VR stands for the violation ratios calculated. LRuc, LRind and LRcc are the likelihood ratio test statistics from the unconditional coverage, independence and conditional coverage tests, respectively. * stands for 1% significance while ** stands for the 5% significance for the underlying tests. Critical values at the 5% and 1% significance are 3.8415 and 6.6349 for LRuc and LRind, 5.9915 and 9.2103 for LRcc, respectively. Finally average is the mean of the VaRs calculated from the methods over the backtesting period.

	Models	EV	AV	fracAV	VR	LRuc	LRind	LRcc	average
PX50	Normal	196	179	0.0455	0.9105	1.7076	13.6930*	15.4006*	0.0238
	HS	196	223	0.0567	1.1343	3.5833	28.4020*	31.9853*	0.0213
	GPD	196	226	0.0575	1.1495	4.4244**	27.0840*	31.5084*	0.0214
	EWMA	196	166	0.0422	0.8469	5.2811**	0.6871	5.9682	0.0223
	GARCH-n	196	170	0.0432	0.8647	3.9622**	0.0437	4.0060	0.0220
	GARCH-t	196	172	0.0437	0.8749	3.3769	0.0153	3.3921	0.0218
	FHS	196	207	0.0526	1.0529	0.5697	0.0811	0.6508	0.0208
	DGPD	196	207	0.0526	1.0529	0.5697	0.0295	0.5992	0.0209
BIST100	Normal	199	141	0.0354	0.7071	19.9677*	11.9928*	31.9605*	0.0429
	HS	199	173	0.0434	0.8676	3.8442**	13.7571*	17.6012*	0.0398
	GPD	199	171	0.0429	0.8576	4.4643**	14.3610*	18.8253*	0.0400
	EWMA	199	188	0.0471	0.9447	0.6988	2.7562	3.4550	0.0387
	GARCH-n	199	164	0.0411	0.8225	7.0233*	8.2482*	15.2715*	0.0397
	GARCH-t	199	178	0.0446	0.8927	2.5042	6.9592*	9.4634*	0.0386
	FHS	199	185	0.0464	0.9278	1.1206	8.7121*	9.8327*	0.0382
	DGPD	199	181	0.0454	0.9077	1.8419	6.3694**	8.2113**	0.0385
ATHEX	Normal	199	211	0.0529	1.0590	0.7162	142.7715*	143.4877*	0.0286
	HS	199	245	0.0615	1.2296	10.3371*	175.1761*	185.5132*	0.0272
	GPD	199	246	0.0617	1.2346	10.7791*	166.0176*	176.7967*	0.0275
	EWMA	199	198	0.0497	0.9950	0.0083	0.0628	0.0711	0.0289
	GARCH-n	199	190	0.0477	0.9536	0.4588	1.2735	1.7323	0.0291
	GARCH-t	199	196	0.0492	0.9837	0.0561	0.8675	0.9236	0.0283
	FHS	199	205	0.0514	1.0289	0.1731	2.5356	2.7087	0.0282
	DGPD	199	200	0.0502	1.0038	0.0030	2.0699	2.0729	0.0282
PSI20	Normal	206	197	0.0478	0.9558	0.4290	11.9230*	12.3520*	0.0189
	HS	206	249	0.0604	1.2082	8.8392*	33.2751*	42.1143*	0.0172
	GPD	206	247	0.0599	1.1984	8.0671*	31.8652*	39.9323*	0.0172
	EWMA	206	193	0.0468	0.9369	0.8947	0.0832	0.9778	0.0184
	GARCH-n	206	198	0.0480	0.9607	0.3393	2.9369	3.2762	0.0180
	GARCH-t	206	205	0.0497	0.9947	0.0062	2.1409	2.1471	0.0178
	FHS	206	208	0.0505	1.0092	0.0184	2.7085	2.7268	0.0175
	DGPD	206	208	0.0505	1.0092	0.0184	2.7085	2.7268	0.0174
IBEX35	Normal	201	207	0.0513	1.0260	0.1426	10.4267*	10.5693*	0.0248
	HS	201	240	0.0595	1.1896	7.2156*	10.1423*	17.3579*	0.0230
	GPD	201	242	0.0600	1.1995	7.9678*	9.6616*	17.6294*	0.0229
	EWMA	201	176	0.0436	0.8756	3.6086	2.2994	5.9081	0.0238
	GARCH-n	201	190	0.0471	0.9418	0.7340	1.7364	2.4704	0.0226
	GARCH-t	201	192	0.0476	0.9517	0.5037	5.9589**	6.4626**	0.0227
	FHS	201	221	0.0548	1.0954	1.8778	2.7228	4.6006	0.0216
	DGPD	201	216	0.0535	1.0706	1.0367	4.4180**	5.4546	0.0218

Source: author's computations.

Table A.5: Right Tail Backtesting Results at $\alpha = 99\%$: EV and AV are the expected and actual number of violations respectively. fracAV is the fraction of the actual number of violations over the all observations. VR stands for the violation ratios calculated. LRuc, LRind and LRcc are the likelihood ratio test statistics from the unconditional coverage, independence and conditional coverage tests, respectively. * stands for 1% significance while ** stands for the 5% significance for the underlying tests. Critical values at the 5% and 1% significance are 3.8415 and 6.6349 for LRuc and LRind, 5.9915 and 9.2103 for LRcc, respectively. Finally average is the mean of the VaRs calculated from the methods over the backtesing period.

	Models	EV	AV	fracAV	VR	LRuc	LRind	LRcc	average
PX50	Normal	39	56	0.0142	1.4242	6.3168**	1.3231	7.6400**	0.0336
	HS	39	52	0.0132	1.3225	3.7504	1.7109	5.4613	0.0351
	GPD	39	39	0.0099	0.9919	0.0026	3.4776	3.4802	0.0366
	EWMA	39	55	0.0140	1.4103	5.6192**	0.0648	5.6840	0.0313
	GARCH-n	39	46	0.0117	1.1699	1.0870	1.0891	2.1761	0.0310
	GARCH-t	39	30	0.0076	0.7630	2.4301	0.4613	2.8914	0.0338
	FHS	39	48	0.0122	1.2208	1.8083	0.2478	2.0561	0.0311
	DGPD	39	35	0.0089	0.8901	0.4978	0.6287	1.1265	0.0314
BIST100	Normal	39	53	0.0133	1.3290	3.9518**	4.2996**	8.2515**	0.0602
	HS	39	47	0.0118	1.1785	1.2145	2.3314	3.5458	0.0666
	GPD	39	40	0.0100	1.0030	0.0004	3.3547	3.3551	0.0671
	EWMA	39	63	0.0158	1.6154	11.5104*	5.4016**	16.9120*	0.0542
	GARCH-n	39	52	0.0130	1.3039	3.3956	0.1369	3.5325	0.0557
	GARCH-t	39	35	0.0090	0.9027	0.3942	0.6559	1.0501	0.0603
	FHS	39	42	0.0105	1.0532	0.1119	0.8941	1.0060	0.0583
	DGPD	39	40	0.0100	1.0030	0.0004	0.8106	0.8109	0.0590
ATHEX	Normal	39	82	0.0206	2.0577	34.4938*	7.0431*	41.5370*	0.0405
	HS	39	58	0.0146	1.4555	7.3208*	3.4545	10.7753*	0.0456
	GPD	39	58	0.0146	1.4555	7.3208*	3.4545	10.7753*	0.0460
	EWMA	39	63	0.0158	1.6154	11.5457*	2.0241	13.5697*	0.0407
	GARCH-n	39	52	0.0130	1.3049	3.4141	1.3751	4.7891	0.0411
	GARCH-t	39	36	0.0090	0.9034	0.3883	0.6564	1.0447	0.0441
	FHS	39	51	0.0128	1.2798	2.8953	0.1631	3.0584	0.0426
	DGPD	39	43	0.0108	1.0790	0.2452	0.4769	0.7221	0.0440
PSI20	Normal	41	65	0.0158	1.5769	11.7891*	16.0300*	27.8191*	0.0267
	HS	41	48	0.0116	1.1645	1.0699	18.8149*	19.8848*	0.0280
	GPD	41	42	0.0102	1.0189	0.0148	11.3364*	11.3512*	0.0294
	EWMA	41	53	0.0129	1.2927	3.1191	0.1335	3.2526	0.0258
	GARCH-n	41	56	0.0136	1.3586	4.8135**	0.0701	4.8837	0.0254
	GARCH-t	41	36	0.0087	0.8734	0.6975	4.1861**	4.8836	0.0281
	FHS	41	47	0.0114	1.1402	0.7832	9.6571*	10.4404*	0.0265
	DGPD	41	41	0.0099	0.9947	0.0012	7.1133*	7.1145**	0.0272
IBEX35	Normal	40	72	0.0178	1.7844	20.3382*	1.7373	22.0755*	0.0350
	HS	40	65	0.0161	1.6109	12.8359*	0.7096	13.5455*	0.0375
	GPD	40	53	0.0131	1.3135	3.6464	1.6768	5.3231	0.0389
	EWMA	40	33	0.0082	0.8250	1.4420	1.1913	2.6333	0.0335
	GARCH-n	40	49	0.0121	1.2144	1.7532	5.1300**	6.8831**	0.0320
	GARCH-t	40	35	0.0087	0.8674	0.7501	4.3110**	5.0611	0.0341
	FHS	40	57	0.0141	1.4126	6.1519**	3.6688	9.8207*	0.0310
	DGPD	40	51	0.0126	1.2639	2.6203	0.1721	2.7924	0.0322

Source: author's computations.

Table A.6: Right Tail Backtesting Results at $\alpha = 99.5\%$: EV and AV are the expected and actual number of violations respectively. fracAV is the fraction of the actual number of violations over the all observations. VR stands for the violation ratios calculated. LRuc, LRind and LRcc are the likelihood ratio test statistics from the unconditional coverage, independence and conditional coverage tests, respectively. * stands for 1% significance while ** stands for the 5% significance for the underlying tests. Critical values at the 5% and 1% significance are 3.8415 and 6.6349 for LRuc and LRind, 5.9915 and 9.2103 for LRcc, respectively. Finally average is the mean of the VaRs calculated from the methods over the backtesting period.

	Models	EV	AV	fracAV	VR	LRuc	LRind	LRcc	average
PX50	Normal	19	35	0.0089	1.7803	9.7536*	4.2219**	13.9755*	0.0372
	HS	19	20	0.0051	1.0173	0.0059	8.4736*	8.4795**	0.0465
	GPD	19	19	0.0048	0.9664	0.0225	8.8865*	8.9091**	0.0446
	EWMA	19	32	0.0081	1.6842	6.5365**	1.2456	7.7822**	0.0346
	GARCH-n	19	26	0.0066	1.3225	1.8648	0.3461	2.2110	0.0342
	GARCH-t	19	17	0.0043	0.8647	0.3791	0.1476	0.5268	0.0392
	FHS	19	27	0.0069	1.3733	2.4653	0.3734	2.8387	0.0347
	DGPD	19	20	0.0051	1.0173	0.0059	0.2045	0.2104	0.0357
BIST100	Normal	19	37	0.0093	1.8556	11.6995*	3.8830**	15.5825*	0.0665
	HS	19	27	0.0068	1.3541	2.2605	6.1778**	8.4383**	0.0792
	GPD	19	23	0.0058	1.1535	0.4496	7.4179*	7.8675**	0.0807
	EWMA	19	38	0.0095	2.0000	12.9716*	3.7002	16.6717*	0.0599
	GARCH-n	19	33	0.0083	1.6550	7.1725*	0.5507	7.7232**	0.0615
	GARCH-t	19	15	0.0038	0.7523	1.3458	0.1133	1.4591	0.0702
	FHS	19	23	0.0058	1.1535	0.4496	0.2668	0.7165	0.0679
	DGPD	19	19	0.0048	0.9529	0.0452	0.1819	0.2272	0.0673
ATHEX	Normal	19	60	0.0151	3.0113	52.5407*	3.1524	55.6931*	0.0448
	HS	19	29	0.0073	1.4555	3.6394	1.5761	5.2155	0.0558
	GPD	19	34	0.0085	1.7064	8.2382*	1.0847	9.3229*	0.0544
	EWMA	19	43	0.0108	2.2632	20.1379*	0.9381	21.0760*	0.0450
	GARCH-n	19	29	0.0073	1.4555	3.6394	0.4252	4.0646	0.0455
	GARCH-t	19	18	0.0045	0.9034	0.1932	0.1633	0.3566	0.0511
	FHS	19	28	0.0070	1.4053	2.9193	0.3963	3.3156	0.0474
	DGPD	19	22	0.0055	1.1041	0.2100	0.2443	0.4543	0.0511
PSI20	Normal	20	51	0.0124	2.4745	31.8628*	17.4028*	49.2656*	0.0296
	HS	20	32	0.0078	1.5526	5.4091**	5.0299**	10.4390*	0.0337
	GPD	20	29	0.0070	1.4071	3.0453	5.7590**	8.8043**	0.0355
	EWMA	20	30	0.0073	1.5000	3.7668	0.4399	4.2067	0.0285
	GARCH-n	20	31	0.0075	1.5041	4.5554**	1.4162	5.9716	0.0282
	GARCH-t	20	16	0.0039	0.7763	1.1232	0.1247	1.2479	0.0328
	FHS	20	26	0.0063	1.2615	1.3077	1.9972	3.3049	0.0304
	DGPD	20	20	0.0049	0.9704	0.0183	2.9403	2.9586	0.0313
IBEX35	Normal	20	53	0.0131	2.6270	36.9990*	1.6768	38.6758*	0.0387
	HS	20	37	0.0092	1.8340	11.2997*	3.9226**	15.2222*	0.0484
	GPD	20	35	0.0087	1.7348	8.9681*	4.3110**	13.2791*	0.0469
	EWMA	20	19	0.0047	0.9500	0.0701	3.0933	3.1635	0.0370
	GARCH-n	20	30	0.0074	1.4870	4.1793**	5.4287**	9.6079*	0.0354
	GARCH-t	20	22	0.0055	1.0905	0.1612	7.8127*	7.9739**	0.0388
	FHS	20	34	0.0084	1.6853	7.8880*	1.1028	8.9907**	0.0342
	DGPD	20	28	0.0069	1.3879	2.7198	0.3913	3.1112	0.0364

Source: author's computations.

Sub-Sample Backtesting Tables

Table A.7: Sub-Sample Left Tail Backtesting Results at $\alpha = 95\%$: EV and AV are the expected and actual number of violations respectively. fracAV is the fraction of the actual number of violations over the all observations. VR stands for the violation ratios calculated. LRuc, LRind and LRcc are the likelihood ratio test statistics from the unconditional coverage, independence and conditional coverage tests, respectively. * stands for 1% significance while ** stands for the 5% significance for the underlying tests. Critical values at the 5% and 1% significance are 3.8415 and 6.6349 for LRuc and LRind, 5.9915 and 9.2103 for LRcc, respectively. Finally average is the mean of the VaRs calculated from the methods over the backtesing period.

	Models	EV	AV	fracAV	VR	LRuc	LRind	LRcc	average
PX50	Normal	50	72	0.072	1.44	9.0221*	23.5724*	32.5944*	0.0300
	HS	50	80	0.080	1.60	16.1581*	22.7815*	38.9396*	0.0261
	GPD	50	79	0.079	1.58	15.1675*	20.7007*	35.8682*	0.0266
	EWMA	50	65	0.065	1.30	4.3455**	8.8222*	13.1676*	0.0285
	GARCH-n	50	59	0.059	1.18	1.6162	3.1168	4.7330	0.0270
	GARCH-t	50	64	0.064	1.28	3.8054**	5.0437**	8.8492**	0.0261
	FHS	50	62	0.062	1.24	2.8260	3.9146**	6.7406**	0.0259
	DGPD	50	60	0.060	1.20	1.9842	4.5660**	6.5502**	0.0264
BIST100	Normal	50	53	0.053	1.06	0.1860	5.0232**	5.2092	0.0322
	HS	50	54	0.054	1.08	0.3287	4.6680**	4.9967	0.0322
	GPD	50	53	0.053	1.06	0.1860	5.0232**	5.2092	0.0325
	EWMA	50	65	0.065	1.30	4.3455**	1.7059	6.0514**	0.0293
	GARCH-n	50	58	0.058	1.16	1.2843	3.3983	4.6826	0.0296
	GARCH-t	50	62	0.062	1.24	2.8260	2.3534	5.1794	0.0292
	FHS	50	59	0.059	1.18	1.6162	0.6347	2.2509	0.0306
	DGPD	50	61	0.061	1.22	2.3877	0.4231	2.8107	0.0307
ATHEX	Normal	50	102	0.102	2.04	44.3415*	6.8223*	51.1638*	0.0291
	HS	50	103	0.103	2.06	45.8908*	6.4048**	52.2956*	0.0288
	GPD	50	98	0.098	1.96	38.3643*	6.9777*	45.3420*	0.0293
	EWMA	50	58	0.058	1.16	1.2843	0.1100	1.3943	0.0355
	GARCH-n	50	71	0.071	1.42	8.2609*	2.6695	10.9305*	0.0343
	GARCH-t	50	72	0.072	1.44	9.0221*	1.2279	10.2499*	0.0337
	FHS	50	46	0.046	0.92	0.3457	0.7901	1.1358	0.0398
	DGPD	50	47	0.047	0.94	0.1932	0.8988	1.0919	0.0398
PSI20	Normal	50	92	0.092	1.84	30.0817*	13.9319*	44.0136*	0.0201
	HS	50	105	0.105	2.10	49.0543*	12.1966*	61.2509*	0.0190
	GPD	50	106	0.106	2.12	50.6681*	13.5414*	64.2095*	0.0192
	EWMA	50	75	0.075	1.50	11.4835*	0.9683	12.4518*	0.0219
	GARCH-n	50	77	0.077	1.54	13.2692*	0.1602	13.4295*	0.0214
	GARCH-t	50	83	0.083	1.66	19.2915*	0.0464	19.2451*	0.0207
	FHS	50	80	0.080	1.60	16.1581*	0.0126	16.1707*	0.0212
	DGPD	50	81	0.081	1.62	17.1758*	0.0168	17.1590*	0.0212
IBEX35	Normal	50	90	0.090	1.80	27.5100*	2.8667	30.3767*	0.0256
	HS	50	100	0.100	2.00	41.3084*	4.7499**	46.0584*	0.0242
	GPD	50	99	0.099	1.98	39.8252*	5.1056**	44.9308*	0.0247
	EWMA	50	78	0.078	1.56	14.2045*	0.1010	14.3055*	0.0288
	GARCH-n	50	75	0.075	1.50	11.4835*	0.0621	11.5456*	0.0271
	GARCH-t	50	83	0.083	1.66	19.2915*	0.1110	19.4025*	0.0267
	FHS	50	72	0.072	1.44	9.0221*	0.3249	9.3470*	0.0277
	DGPD	50	74	0.074	1.48	10.6337*	0.0271	10.6608*	0.0276

Source: author's computations.

Table A.8: Sub-Sample Left Tail Backtesting Results at $\alpha = 99\%$: EV and AV are the expected and actual number of violations respectively. fracAV is the fraction of the actual number of violations over the all observations. VR stands for the violation ratios calculated. LRuc, LRind and LRcc are the likelihood ratio test statistics from the unconditional coverage, independence and conditional coverage tests, respectively. * stands for 1% significance while ** stands for the 5% significance for the underlying tests. Critical values at the 5% and 1% significance are 3.8415 and 6.6349 for LRuc and LRind, 5.9915 and 9.2103 for LRcc, respectively. Finally average is the mean of the VaRs calculated from the methods over the backtesing period.

	Models	EV	AV	fracAV	VR	LRuc	LRind	LRcc	average
PX50	Normal	10	36	0.036	3.6	40.9161*	13.9748*	54.8909*	0.0424
	HS	10	20	0.020	2.0	7.8272*	12.4937*	20.3209*	0.0546
	GPD	10	19	0.019	1.9	6.4725**	13.3323*	19.8048*	0.0550
	EWMA	10	20	0.020	2.0	7.8272*	12.4937*	20.3209*	0.0405
	GARCH-n	10	23	0.023	2.3	12.4853*	1.0830	13.5683*	0.0382
	GARCH-t	10	20	0.020	2.0	7.8272*	0.8164	8.6436*	0.0411
	FHS	10	15	0.015	1.5	2.1892	0.4569	2.6461	0.0453
	DGPD	10	16	0.016	1.6	3.0766	0.5203	3.5969	0.0442
BIST100	Normal	10	22	0.022	2.2	10.8382*	2.8484	13.6866*	0.0457
	HS	10	14	0.014	1.4	1.4374	1.7425	3.1799	0.0545
	GPD	10	14	0.014	1.4	1.4374	1.7425	3.1799	0.0547
	EWMA	10	21	0.021	2.1	9.2840*	6.9991*	16.2832*	0.0419
	GARCH-n	10	18	0.018	1.8	5.2251**	8.7839*	14.0091*	0.0422
	GARCH-t	10	13	0.013	1.3	0.8306	6.7294*	7.5600**	0.0451
	FHS	10	16	0.016	1.6	3.0766	10.2038*	13.2804*	0.0461
	DGPD	10	11	0.011	1.1	0.0978	8.0930*	8.1908**	0.0483
ATHEX	Normal	10	49	0.049	4.9	79.3020*	9.2921*	88.5941*	0.0409
	HS	10	28	0.028	2.8	21.9880*	7.3294*	29.3174*	0.0537
	GPD	10	32	0.032	3.2	30.9342*	5.4888**	36.4230*	0.0510
	EWMA	10	18	0.018	1.8	5.2251**	0.9518	6.1770**	0.0503
	GARCH-n	10	25	0.025	2.5	16.0430*	1.2822	17.3252*	0.0486
	GARCH-t	10	21	0.021	2.1	9.2840*	0.9010	10.1850*	0.0505
	FHS	10	9	0.009	0.9	0.1045	0.1635	0.2680	0.0585
	DGPD	10	8	0.008	0.8	0.4337	0.1290	0.5628	0.0600
PSI20	Normal	10	47	0.047	4.7	72.8713*	13.6105*	86.4817*	0.0284
	HS	10	32	0.032	3.2	30.9342*	0.0004	30.9338*	0.0361
	GPD	10	33	0.033	3.3	33.3374*	0.6557	33.9932*	0.0367
	EWMA	10	27	0.027	2.7	19.9292*	1.6109	21.5401*	0.0313
	GARCH-n	10	29	0.029	2.9	24.1202*	1.2361	25.3563*	0.0302
	GARCH-t	10	16	0.016	1.6	3.0766	0.5203	3.5969	0.0332
	FHS	10	19	0.019	1.9	6.4725**	3.8485**	10.3210*	0.0318
	DGPD	10	17	0.017	1.7	4.0910**	1.1192	5.2101	0.0331
IBEX35	Normal	10	50	0.050	5.0	82.5822*	2.1577	84.7399*	0.0361
	HS	10	27	0.027	2.7	19.9292*	7.8546*	27.7838*	0.0451
	GPD	10	27	0.027	2.7	19.9292*	7.8546*	27.7838*	0.0456
	EWMA	10	19	0.019	1.9	6.4725**	0.7360	7.2085**	0.0409
	GARCH-n	10	25	0.025	2.5	16.0430*	1.2822	17.3252*	0.0383
	GARCH-t	10	15	0.015	1.5	2.1892	0.4569	2.6461	0.0422
	FHS	10	14	0.014	1.4	1.4374	0.3976	1.8350	0.0439
	DGPD	10	12	0.012	1.2	0.3798	0.2915	0.6713	0.0445

Source: author's computations.

Table A.9: Sub-Sample Left Tail Backtesting Results at $\alpha = 99.5\%$: EV and AV are the expected and actual number of violations respectively. fracAV is the fraction of the actual number of violations over the all observations. VR stands for the violation ratios calculated. LRuc, LRind and LRcc are the likelihood ratio test statistics from the unconditional coverage, independence and conditional coverage tests, respectively. * stands for 1% significance while ** stands for the 5% significance for the underlying tests. Critical values at the 5% and 1% significance are 3.8415 and 6.6349 for LRuc and LRind, 5.9915 and 9.2103 for LRcc, respectively. Finally average is the mean of the VaRs calculated from the methods over the backtesing period.

	Models	EV	AV	fracAV	VR	LRuc	LRind	LRcc	average
PX50	Normal	5	27	0.027	5.4	47.5556*	12.0396*	59.5952*	0.0470
	HS	5	12	0.012	2.4	7.0606*	13.8406*	20.9012*	0.0704
	GPD	5	12	0.012	2.4	7.0606*	13.8406*	20.9012*	0.0722
	EWMA	5	16	0.016	3.2	15.3429*	1.3053	16.6482*	0.0449
	GARCH-n	5	18	0.018	3.6	20.2842*	0.6599	20.9441*	0.0424
	GARCH-t	5	10	0.010	2.0	3.8881**	0.2020	4.0901	0.0478
	FHS	5	6	0.006	1.2	0.1889	0.0724	0.2613	0.0556
	DGPD	5	6	0.006	1.2	0.1889	0.0724	0.2613	0.0529
BIST100	Normal	5	19	0.019	3.8	22.9280*	3.8485**	26.7764*	0.0506
	HS	5	8	0.008	1.6	1.5291	0.1290	1.6581	0.0639
	GPD	5	8	0.008	1.6	1.5291	3.8416**	5.3707	0.0640
	EWMA	5	15	0.015	3.0	13.0592*	10.9999*	24.0591*	0.0465
	GARCH-n	5	10	0.010	2.0	3.8881**	2.9651	6.8532**	0.0469
	GARCH-t	5	5	0.005	1.0	0.0000	0.0503	0.0503	0.0518
	FHS	5	4	0.004	0.8	0.2159	0.0321	0.2480	0.0570
	DGPD	5	5	0.005	1.0	0.0000	0.0503	0.0503	0.0559
ATHEX	Normal	5	43	0.043	8.6	110.5216*	12.8498*	123.3714*	0.0453
	HS	5	19	0.019	3.8	22.9280*	3.8485**	26.7764*	0.0588
	GPD	5	17	0.017	3.4	17.7537*	15.1894*	32.9431*	0.0610
	EWMA	5	9	0.009	1.8	2.5963	0.1635	2.7597	0.0558
	GARCH-n	5	16	0.016	3.2	15.3429*	0.5203	15.8632*	0.0538
	GARCH-t	5	10	0.010	2.0	3.8881**	0.2020	4.0901	0.0574
	FHS	5	3	0.003	0.6	0.9391	0.0181	0.9571	0.0663
	DGPD	5	2	0.002	0.4	2.3439	0.0080	2.3519	0.0679
PSI20	Normal	5	43	0.043	8.6	110.5216*	4.0867**	114.6083*	0.0314
	HS	5	16	0.016	3.2	15.3429*	0.5203	15.8632*	0.0445
	GPD	5	15	0.015	3.0	13.0592*	0.4569	13.5161*	0.0459
	EWMA	5	17	0.017	3.4	17.7537*	1.1192	18.8728*	0.0347
	GARCH-n	5	16	0.016	3.2	15.3429*	0.5203	15.8632*	0.0334
	GARCH-t	5	6	0.006	1.2	0.1889	0.0724	0.2613	0.0391
	FHS	5	11	0.011	2.2	5.3823**	0.2447	5.6270	0.0358
	DGPD	5	7	0.007	1.4	0.7146	0.0987	0.8133	0.0380
IBEX35	Normal	5	40	0.040	8.0	97.6012*	5.1521	102.7533*	0.0400
	HS	5	17	0.017	3.4	17.7537*	0.5880	18.3417*	0.0527
	GPD	5	15	0.015	3.0	13.0592*	0.4569	13.5161*	0.0555
	EWMA	5	9	0.009	1.8	2.5963	0.1635	2.7597	0.0454
	GARCH-n	5	15	0.015	3.0	13.0592*	0.4569	13.5161*	0.0424
	GARCH-t	5	8	0.008	1.6	1.5291	0.1290	1.6581	0.0491
	FHS	5	9	0.009	1.8	2.5963	0.1635	2.7597	0.0491
	DGPD	5	6	0.006	1.2	0.1889	0.0724	0.2613	0.0515

Source: author's computations.

Table A.10: Sub-Sample Right Tail Backtesting Results at $\alpha = 95\%$: EV and AV are the expected and actual number of violations respectively. fracAV is the fraction of the actual number of violations over the all observations. VR stands for the violation ratios calculated. LRuc, LRind and LRcc are the likelihood ratio test statistics from the unconditional coverage, independence and conditional coverage tests, respectively. * stands for 1% significance while ** stands for the 5% significance for the underlying tests. Critical values at the 5% and 1% significance are 3.8415 and 6.6349 for LRuc and LRind, 5.9915 and 9.2103 for LRcc, respectively. Finally average is the mean of the VaRs calculated from the methods over the backtesing period.

	Models	EV	AV	fracAV	VR	LRuc	LRind	LRcc	average
PX50	Normal	50	54	0.054	1.08	0.3287	6.8871*	7.2157**	0.0298
	HS	50	80	0.080	1.60	16.1581*	12.2416*	28.3997*	0.0236
	GPD	50	81	0.081	1.62	17.1758*	11.6609*	28.8367*	0.0237
	EWMA	50	38	0.038	0.76	3.2937	0.2028	3.4965	0.0295
	GARCH-n	50	50	0.050	1.00	0.0000	2.1577	2.1577	0.0275
	GARCH-t	50	46	0.046	0.92	0.3457	1.4647	1.8104	0.0273
	FHS	50	57	0.057	1.14	0.9889	0.8902	1.8791	0.0254
	DGPD	50	57	0.057	1.14	0.9889	0.8902	1.8791	0.0254
BIST100	Normal	50	43	0.043	0.86	1.0807	2.0715	3.1521	0.0329
	HS	50	52	0.052	1.04	0.0832	3.3580	3.4412	0.0311
	GPD	50	53	0.053	1.06	0.1860	3.0767	3.2627	0.0307
	EWMA	50	41	0.041	0.82	1.8120	0.0576	1.8697	0.0317
	GARCH-n	50	45	0.045	0.90	0.5438	3.4603	4.0041	0.0316
	GARCH-t	50	46	0.046	0.92	0.3457	3.1706	3.5163	0.0311
	FHS	50	45	0.045	0.90	0.5438	0.4408	0.9846	0.0307
	DGPD	50	43	0.043	0.86	1.0807	0.6539	1.7346	0.0310
ATHEX	Normal	50	87	0.087	1.74	23.8361*	2.5058	26.3419*	0.0279
	HS	50	114	0.114	2.28	64.3238*	1.1024	65.4262*	0.0247
	GPD	50	114	0.114	2.28	64.3238*	0.5349	64.8587*	0.0247
	EWMA	50	49	0.049	0.98	0.0212	1.1354	1.1566	0.0364
	GARCH-n	50	65	0.065	1.30	4.3455**	0.0009	4.3445	0.0343
	GARCH-t	50	66	0.066	1.32	4.9184**	0.0197	4.9381	0.0338
	FHS	50	58	0.058	1.16	1.2843	0.7154	1.9996	0.0349
	DGPD	50	58	0.058	1.16	1.2843	0.7154	1.9996	0.0351
PSI20	Normal	50	65	0.065	1.30	4.3455**	4.6924**	9.0378**	0.0199
	HS	50	99	0.099	1.98	39.8252*	8.1594*	47.9845*	0.0159
	GPD	50	97	0.097	1.94	36.9261*	7.4134*	44.3395*	0.0161
	EWMA	50	31	0.031	0.62	8.7393*	0.9194	9.6587*	0.0234
	GARCH-n	50	45	0.045	0.90	0.5438	5.7579**	6.3017**	0.0213
	GARCH-t	50	47	0.047	0.94	0.1932	4.9907**	5.1839	0.0210
	FHS	50	51	0.051	1.02	0.0209	1.9389	1.9599	0.0202
	DGPD	50	51	0.051	1.02	0.0209	3.6538	3.6748	0.0200
IBEX35	Normal	50	74	0.074	1.48	10.6337*	0.3996	11.0333*	0.0255
	HS	50	96	0.096	1.92	35.5107*	1.4989	37.0096*	0.0217
	GPD	50	96	0.096	1.92	35.5107*	1.4989	37.0096*	0.0216
	EWMA	50	41	0.041	0.82	1.8120	0.3466	2.1586	0.0300
	GARCH-n	50	47	0.047	0.94	0.1932	0.0189	0.2120	0.0271
	GARCH-t	50	49	0.049	0.98	0.0212	2.3899	2.4111	0.0275
	FHS	50	66	0.066	1.32	4.9184**	0.5978	5.5161	0.0248
	DGPD	50	67	0.067	1.34	5.5238**	0.4890	6.0127**	0.0246

Source: author's computations.

Table A.11: Sub-Sample Right Tail Backtesting Results at $\alpha = 99\%$: EV and AV are the expected and actual number of violations respectively. fracAV is the fraction of the actual number of violations over the all observations. VR stands for the violation ratios calculated. LRuc, LRind and LRcc are the likelihood ratio test statistics from the unconditional coverage, independence and conditional coverage tests, respectively. * stands for 1% significance while ** stands for the 5% significance for the underlying tests. Critical values at the 5% and 1% significance are 3.8415 and 6.6349 for LRuc and LRind, 5.9915 and 9.2103 for LRcc, respectively. Finally average is the mean of the VaRs calculated from the methods over the backtesing period.

	Models	EV	AV	fracAV	VR	LRuc	LRind	LRcc	average
PX50	Normal	10	21	0.021	2.1	9.2840*	3.1570	12.4410*	0.0422
	HS	10	23	0.023	2.3	12.4853*	2.5621	15.0474*	0.0442
	GPD	10	15	0.015	1.5	2.1892	5.6028**	7.7921**	0.0471
	EWMA	10	15	0.015	1.5	2.1892	0.4569	2.6461	0.0415
	GARCH-n	10	14	0.014	1.4	1.4374	0.3976	1.8350	0.0388
	GARCH-t	10	9	0.009	0.9	0.1045	0.1635	0.2680	0.0423
	FHS	10	15	0.015	1.5	2.1892	1.5123	3.7016	0.0383
	DGPD	10	14	0.014	1.4	1.4374	0.3976	1.8350	0.0382
BIST100	Normal	10	12	0.012	1.2	0.3798	0.2915	0.6713	0.0464
	HS	10	12	0.012	1.2	0.3798	0.2915	0.6713	0.0477
	GPD	10	11	0.011	1.1	0.0978	0.2447	0.3425	0.0501
	EWMA	10	15	0.015	1.5	2.1892	0.4569	2.6461	0.0443
	GARCH-n	10	18	0.018	1.8	5.2251**	0.6599	5.8851**	0.0443
	GARCH-t	10	14	0.014	1.4	1.4374	0.3976	1.8350	0.0470
	FHS	10	12	0.012	1.2	0.3798	0.2915	0.6713	0.0480
	DGPD	10	12	0.012	1.2	0.3798	0.2915	0.6713	0.0478
ATHEX	Normal	10	33	0.033	3.3	33.3374*	0.6557	33.9932*	0.0397
	HS	10	30	0.030	3.0	26.3235*	0.0103	26.3338*	0.0424
	GPD	10	28	0.028	2.8	21.9880*	0.0569	22.0449*	0.0438
	EWMA	10	17	0.017	1.7	4.0910**	0.5880	4.6790	0.0513
	GARCH-n	10	19	0.019	1.9	6.4725**	0.7360	7.2085*	0.0486
	GARCH-t	10	18	0.018	1.8	5.2251**	0.6599	5.8851	0.0506
	FHS	10	18	0.018	1.8	5.2251**	0.6599	5.8851	0.0527
	DGPD	10	18	0.018	1.8	5.2251**	0.6599	5.8851	0.0524
PSI20	Normal	10	27	0.027	2.7	19.9292*	4.3202**	24.2494*	0.0282
	HS	10	22	0.022	2.2	10.8382*	6.4794**	17.3175*	0.0296
	GPD	10	21	0.021	2.1	9.2840*	6.9991*	16.2832*	0.0309
	EWMA	10	9	0.009	0.9	0.1045	0.1635	0.2680	0.0328
	GARCH-n	10	12	0.012	1.2	0.3798	2.2846	2.6643	0.0301
	GARCH-t	10	5	0.005	0.5	3.0937	5.8006**	8.8943**	0.0335
	FHS	10	8	0.008	0.8	0.4337	3.8416**	4.2753	0.0319
	DGPD	10	7	0.007	0.7	1.0156	4.3841**	5.3998	0.0321
IBEX35	Normal	10	33	0.033	3.3	33.3374*	2.2528	35.5902*	0.0361
	HS	10	33	0.033	3.3	33.3374*	2.2528	35.5902*	0.0374
	GPD	10	24	0.024	2.4	14.2214*	1.1804	15.4019*	0.0407
	EWMA	10	9	0.009	0.9	0.1045	0.1635	0.2680	0.0421
	GARCH-n	10	13	0.013	1.3	0.8306	1.9987	2.8293	0.0384
	GARCH-t	10	7	0.007	0.7	1.0156	4.3841**	5.3998	0.0429
	FHS	10	20	0.020	2.0	7.8272*	0.6676	8.4948**	0.0355
	DGPD	10	17	0.017	1.7	4.0910**	1.1192	5.2101	0.0367

Source: author's computations.

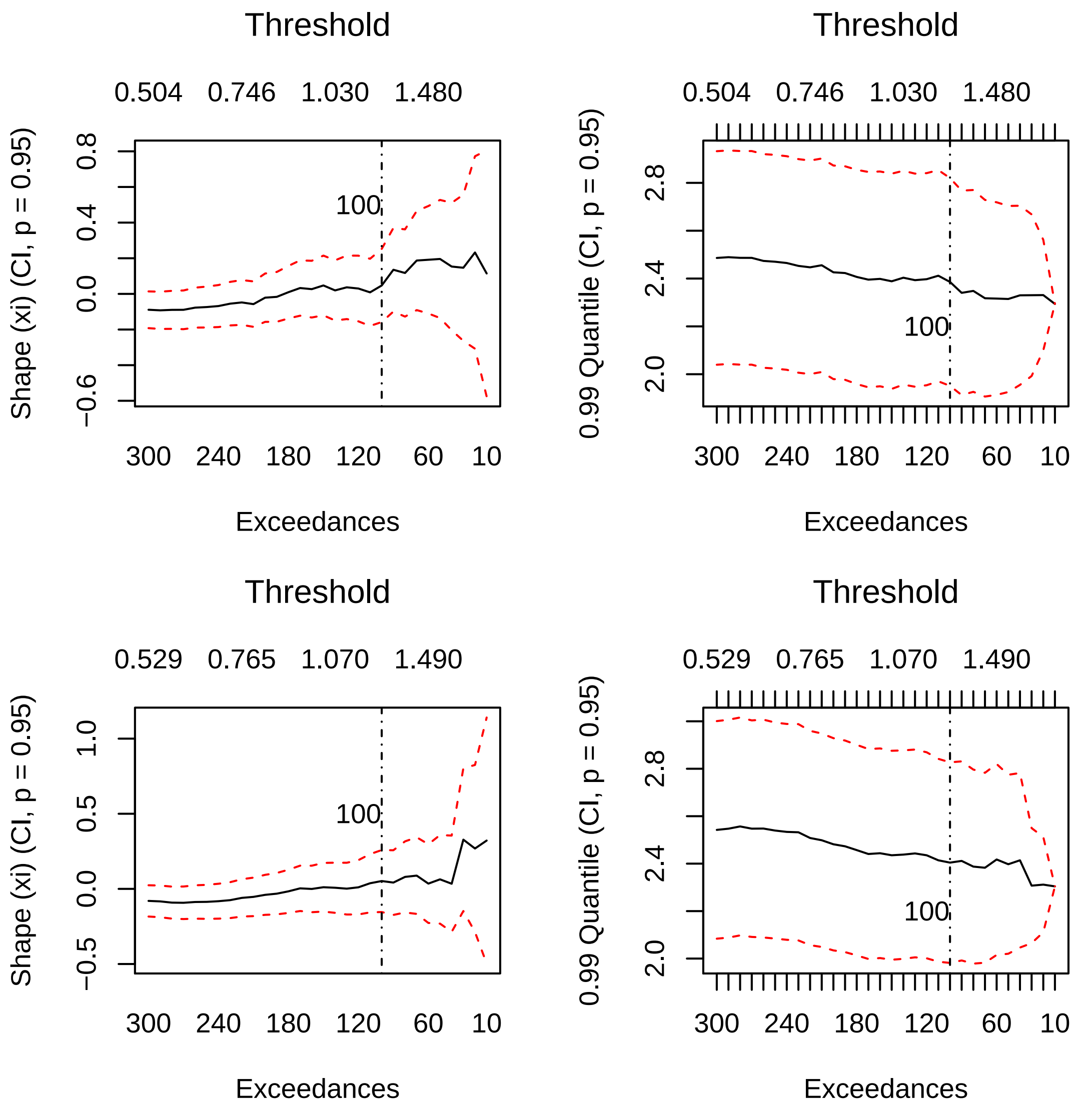
Table A.12: Sub-Sample Right Tail Backtesting Results at $\alpha = 99.5\%$: EV and AV are the expected and actual number of violations respectively. fracAV is the fraction of the actual number of violations over the all observations. VR stands for the violation ratios calculated. LRuc, LRind and LRcc are the likelihood ratio test statistics from the unconditional coverage, independence and conditional coverage tests, respectively. * stands for 1% significance while ** stands for the 5% significance for the underlying tests. Critical values at the 5% and 1% significance are 3.8415 and 6.6349 for LRuc and LRind, 5.9915 and 9.2103 for LRcc, respectively. Finally average is the mean of the VaRs calculated from the methods over the backtesing period.

	Models	EV	AV	fracAV	VR	LRuc	LRind	LRcc	average
PX50	Normal	5	13	0.013	2.6	8.9078*	6.7294*	15.6372*	0.0467
	HS	5	8	0.008	1.6	1.5291	10.8142*	12.3433*	0.0703
	GPD	5	9	0.009	1.8	2.5963	9.7902*	12.3865*	0.0619
	EWMA	5	7	0.007	1.4	0.7146	0.0987	0.8133	0.0459
	GARCH-n	5	8	0.008	1.6	1.5291	0.1290	1.6581	0.0429
	GARCH-t	5	4	0.004	0.8	0.2159	0.0321	0.2480	0.0490
	FHS	5	9	0.009	1.8	2.5963	0.1635	2.7597	0.0436
	DGPD	5	9	0.009	1.8	2.5963	0.1635	2.7597	0.0435
BIST100	Normal	5	8	0.008	1.6	1.5291	0.1290	1.6581	0.0514
	HS	5	8	0.008	1.6	1.5291	0.1290	1.6581	0.0540
	GPD	5	7	0.007	1.4	0.7146	0.0987	0.8133	0.0593
	EWMA	5	9	0.009	1.8	2.5963	0.1635	2.7597	0.0490
	GARCH-n	5	8	0.008	1.6	1.5291	0.1290	1.6581	0.0489
	GARCH-t	5	4	0.004	0.8	0.2159	0.0321	0.2480	0.0537
	FHS	5	7	0.007	1.4	0.7146	0.0987	0.8133	0.0540
	DGPD	5	6	0.006	1.2	0.1889	0.0724	0.2613	0.0547
ATHEX	Normal	5	26	0.026	5.2	44.1766*	0.1421	44.3187*	0.0441
	HS	5	15	0.015	3.0	13.0592*	0.4569	13.5161*	0.0560
	GPD	5	16	0.016	3.2	15.3429*	0.5203	15.8632*	0.0536
	EWMA	5	13	0.013	2.6	8.9078*	0.3425	9.2503*	0.0568
	GARCH-n	5	12	0.012	2.4	7.0606*	0.2915	7.3521**	0.0538
	GARCH-t	5	8	0.008	1.6	1.5291	0.1290	1.6581	0.0575
	FHS	5	13	0.013	2.6	8.9078*	0.3425	9.2503*	0.0568
	DGPD	5	9	0.009	1.8	2.5963	0.1635	2.7597	0.0593
PSI20	Normal	5	21	0.021	4.2	28.5322*	6.9991*	35.5313*	0.0312
	HS	5	17	0.017	3.4	17.7537*	4.6551**	22.4088*	0.0386
	GPD	5	17	0.017	3.4	17.7537*	4.6551**	22.4088*	0.0401
	EWMA	5	5	0.005	1.0	0.0000	0.0503	0.0503	0.0363
	GARCH-n	5	7	0.007	1.4	0.7146	4.3841**	5.0988	0.0333
	GARCH-t	5	4	0.004	0.8	0.2159	0.0321	0.2480	0.0393
	FHS	5	4	0.004	0.8	0.2159	0.0321	0.2480	0.0358
	DGPD	5	4	0.004	0.8	0.2159	0.0321	0.2480	0.0374
IBEX35	Normal	5	26	0.026	5.2	44.1766*	1.3883	45.5649*	0.0399
	HS	5	17	0.017	3.4	17.7537*	0.5880	18.3417*	0.0562
	GPD	5	16	0.016	3.2	15.3429*	0.5203	15.8632*	0.0523
	EWMA	5	6	0.006	1.2	0.1889	0.0724	0.2613	0.0466
	GARCH-n	5	8	0.008	1.6	1.5291	3.8416**	5.3707	0.0425
	GARCH-t	5	5	0.005	1.0	0.0000	5.8006**	5.8006	0.0498
	FHS	5	13	0.013	2.6	8.9078*	1.9987	10.9065*	0.0392
	DGPD	5	11	0.011	2.2	5.3823**	2.6047	7.9870**	0.0418

Source: author's computations.

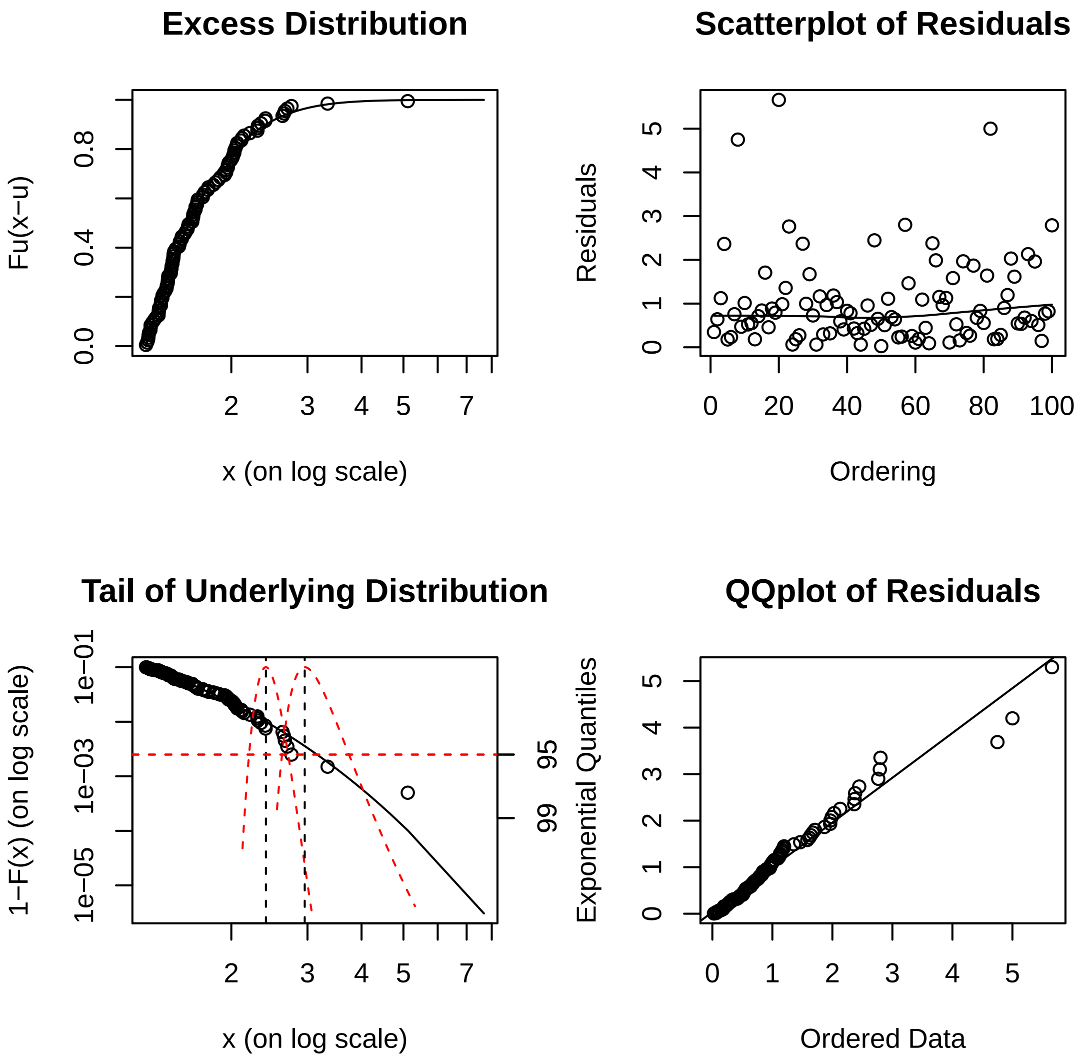
Plots for Tail Analysis

Figure A.10: Tail Sensitivity Plots for PX50 Residuals



Source: author's computations.

Figure A.11: Diagnostic Plots of the Right tail GPD fit for PX50 Residuals



Source: author's computations.