Essays on Measuring Model Risk of Risk Measures

Thesis submitted in partial fulfilment of the requirements

for the degree of Doctor of Philosophy

HENLEY BUSINESS SCHOOL

THE UNIVERSITY OF READING

ICMA Centre

Ning Zhang

March 2020

To my parents and brother, with love

Declaration

I confirm that this is my own work and the use of all material from other sources has been properly and fully acknowledged.

Ning Zhang

Acknowledgements

I am deeply indebted to my advisor Dr Emese Lazar for her patient guidance and continuous support. Emese has led me the way towards academia and taught me to think outside the box, ask questions critically, and write research papers properly. Without her help, the research would not have been possible. Apart from being my supervisor, Emese has been a very inspiring coauthor with perceptive insights into this research. I would also like to thank the other coauthor, Professor Radu Tunaru, for his insightful comments and suggestions.

I sincerely thank the ICMA Centre for its generous financial support. I am very grateful to my family, friends, and colleagues for being very kind and supportive.

Finally, I would like to thank the examiners, Professor Michael Clements and Professor Andreas Kaeck, for their precious time to review this thesis, as well as for their valuable comments.

Abstract

The thesis contributes to the quantitative measurement of model risk of popular models for market risk measures (focusing on Value-at-Risk and Expected Shortfall, denoted by VaR and ES) and volatility forecasting in several ways, and it consists of three main chapters.

The first main contribution is the introduction of measurement of the model risk of ES as the optimal correction needed to pass several ES backtests. We investigate the properties of our proposed measures of model risk from a regulatory perspective. The empirical results show that for the DJIA index, the smallest corrections are required for the ES estimates built using GARCH models. Furthermore, the 2.5% ES requires smaller corrections for model risk than the 1% VaR, which advocates the replacement of VaR with ES as recommended by the Basel Committee. Also, if the model risk of VaR is taken into account, then the corrections made to the ES estimates reduce by 50% on average.

The second main contribution is the development of a new scoring function-based model risk estimation methodology for measuring the joint model risk of the pair of risk measures, VaR and ES, at a given significance level. A simulation

study is carried out to illustrate and analyze the proposed model risk measure across various market risk models. The newly proposed technique accounts for a large proportion of true model risk for a wide set of models popular in the risk management literature. An empirical analysis illustrates its application for different asset classes. The RiskMetrics model and Historical Simulation have the highest level of joint model risk and the highest ES model risk for various assets among all models considered.

The third main contribution is the introduction of a new model risk estimation methodology for volatility models based on the QLIKE loss function. The reliability of the proposed measure has been verified via simulations and compared with the theoretical model risk measure. The efficiency of volatility models can be improved after adjusting variance estimates for model risk. In an empirical study based on several assets, among the models considered, the RiskMetrics method, RW1000 and the ARCH-type models are the most affected by model risk. We find that after crises, model risk increases for poorly fitting volatility models.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation for the Thesis

Measuring, forecasting and controlling financial risk have been tremendously important amongst academics, policymakers, regulators and finance practitioners, as suggested by Christoffersen (2012), Andersen et al. (2013), McNeil et al. (2015) and others. The common categories of financial risk, which we have been dealing with for many years, are market risk, credit risk and operational risk. Due to the long-lasting adverse consequences of the 2008 global financial crisis, the Federal Reserve (2011) raises awareness of model risk and provides supervisory guidance on managing model risk (also see the guidelines of the European Banking Authority, 2014). From their point of view, the term model refers to a quantitative approach or system that digests inputs and produce quantitative estimates using statistical, economic, financial techniques and assumptions, and the use of models

invariably comes with cost. Model risk can negatively affect the decisions of regulators and risk managers in evaluating the risks and defining capital adequacy requirements, as well as lead to financial losses for institutional investors who heavily depend on models in making investment decisions.

Regarding the increasingly extensive use of models and the growing complexity of models, model risk is prevailing and inevitable, for example, in pricing models and risk measurement models, so measuring and managing this type of risk are becoming necessary and nontrivial. In order to manage model risk properly like other types of risk, the regulators suggest that banks should identify the sources of model risk and assess the magnitude of model risk. Based on the significant work of Kerkhof et al. (2010) which first distinguishes the sources of model risk in the context of econometric modeling, this thesis is focused on three main components of the total model risk: 1) parameter estimation risk arises when model parameters are inaccurately estimated, which has been frequently discussed in the current literature; see for example, Christoffersen and Gonçalves (2005), Hartz et al. (2006), Escanciano and Olmo (2010a) and Pitera and Schmidt (2018); 2) model misspecification risk[1] arises when the model is misspecified, documented in Cont (2006) who studies the impact of this component on the pricing models different from our focus on the market risk models; 3) identification risk arises when some information is not detected and considered for forecasting.

The focus of this thesis is on measuring the model risk of risk measures in the context of market risk. Market risk, as a significant risk type, refers to the

risk of a financial portfolio due to changes in the market prices of the underlying assets such as stock, foreign exchange, bond and so forth. The statistical risk measures, Value-at-Risk (VaR) and Expected Shortfall (ES), are widely accepted for market risk measurement and management. The VaR measure quantifies the minimum loss of holding a financial portfolio which should be only exceeded with a small critical probability (typically 1% or 2.5%) over some time period (on a daily basis, for example). As required by the Basel Committee on Banking Supervision (2011), market risk should be measured by ES which is defined as the average loss beyond the VaR threshold.

In the risk management literature (e.g. Christoffersen, 2012), VaR and ES are often defined as positive risk measures (we follow this sign convention in Chapter 2), which can be interpreted as positive losses of the financial portfolios. However, VaR and ES are defined as negative measures in the scoring function literature (e.g. Fissler and Ziegel, 2016), and are interpreted as negative log returns. To keep consistency with the scoring function literature, we use negative risk measures in Chapter 3. For the statistical computation of VaR and ES measures, a variety of risk models produce model-dependent risk estimates, meaning that the VaR and ES measures are exposed to model risk that occurs when a potentially not-well suited risk model leads to imperfect risk estimates (the definition of model risk is taken from Barrieu and Scandolo, 2015).

Also, the forecasting of the volatility of financial times series plays a crucial role in estimating risk measures and other applications in the financial world,

for example, in pricing sophisticated derivatives. The current extensive volatility modeling literature covers the family of autoregressive conditional heteroscedasticity (ARCH) models, stochastic volatility models as well as realized volatility models, in a univariate or multivariate setting (see a comprehensive overview of volatility models in Bauwens et al., 2012). Naturally, the use of these financial econometric techniques that compute volatility estimates invariably presents model risk.

To manage model risk more effectively, it is of much interest to quantify the model risk of risk models as well as of volatility models. To the best of our knowledge, the literature on measuring the model risk of risk models or volatility models is limited. Ideally, if for a given model the true values of target variables (risk or volatility estimates) were observable, then one could measure model risk based on the distance between the true values and the estimated ones. However, the difficulty in measuring the model risk of risk or volatility estimates is that the target variables are latent and not observed ex-post, so the measurement of model risk becomes challenging. The ultimate goal of this thesis is to quantify model risk, account for this type of risk as part of capital requirements as requested by the regulatory authorities, and further facilitate the advance of model risk management.

1.2 Overview of the Thesis

This thesis offers several model risk estimation methodologies which are aimed at numerically estimating the model risk of market risk models or univariate volatility models.

Firstly, this thesis quantifies ES model risk as a correction required for ES estimates of a given model in order to pass several ES backtests jointly, which links model error and statistical testing. In terms of ES backtesting, a time series of desirable ES forecasts should have an appropriate frequency of exceptions which refer to the realized observations (e.g., returns) beyond the corresponding VaR (e.g. VaR in returns), the absence of volatility clustering in the tail and a suitable magnitude of the exceptions. Regarding these desirable criteria, we mainly implement the unconditional/conditional coverage test for ES of Du and Escanciano (2016), and the Z_2 test of Acerbi and Szekely (2014) (additionally, the exceedance residual test of McNeil and Frey, 2000 is used as an alternative to the Z_2 test). Such a backtesting-based correction methodology for ES can be a practical method to improve ES estimates, and whilst not perfect, this provides a possibility of measuring ES model risk.

Moreover, Artzner et al. (1999) argue that effectively regulated measures of risk (market and nonmarket risks) should satisfy the coherence properties, namely, monotonicity, translation invariance, subadditivity and positive homogeneity. We examine whether the aforementioned properties hold for our proposed ES model

risk measure from a regulatory perspective. For our chosen measure of ES model risk which considers the unconditional and conditional coverage tests for ES jointly, all the desirable properties hold except for the subadditivity.

Additionally, we analyze the impact of VaR model risk on ES model risk, primarily for two reasons: 1) for a given model, VaR model risk can affect the calculations of ES estimates, as the inaccuracy of VaR estimates is carried over to the estimated ES which is often a by-product of the VaR estimation procedure (see, e.g. Patton et al., 2019); 2) in terms of this proposed backtesting-based correction technique, wrong VaR estimates may distort the backtesting results, thus leading to inappropriate corrections of ES estimates.

The empirical analysis shows that the 2.5% ES is less affected by model risk than the 1% VaR across different models, thus advocating the replacement of the 1% VaR with the 2.5% ES. Also, if VaR model risk is removed first, then the corrections made to the ES estimates reduce by 50% on average.

Secondly, this thesis develops a scoring function-based model risk estimation methodology that fills in a gap between the scoring function literature and the model risk literature. As the optimal risk estimates can be uniquely obtained via minimizing the expected score of a given scoring function within the FZ class (Fissler et al., 2016) that is strictly consistent for the pair of risk functionals (VaR, ES), we estimate the joint (VaR, ES) model risk of a certain risk model as the average distance between the estimated (VaR, ES) and the improved pair of (VaR, ES) estimates based on a given FZ scoring function over a model risk

evaluation window, as well as estimate ES model risk solely as a by-product.

To allow comparisons with true model risk, we illustrate the newly proposed model risk estimation methodology with a simulation study in which three specific FZ scoring functions are used. We find a high similarity between the true and estimated values of joint (VaR, ES) model risk as well as for ES model risk using a wide set of models popular in the risk management literature, as evidenced by correlations varying from 0.8 to 0.987 and an explanatory power of our proposed model risk measures above 50%. Our proposed scoring function-based model risk measures satisfy all coherence properties of a measure of risk except for the subadditivity in the simulated scenarios numerically.

We also conduct an empirical study to highlight the application of the scoring function-based model risk estimation method for different asset classes, showing that the RiskMetrics method and Historical Simulation have a very high level of joint model risk and ES model risk, among all the models considered, particularly during crisis periods. In addition, the models suffering from model risk, which fail the backtests, can survive the backtesting procedure after adjusting the risk estimates for model risk.

Risk models and volatility models share the latent feature of the target prediction(s), so the scoring function-based model risk estimation methodology for market risk models is extended to the analysis of model risk measurement of volatility models.

Thus, this thesis introduces a model risk estimation methodology based on

the MSE or QLIKE loss function to quantify the model risk of volatility models. This analysis not only reinforces the model risk estimation methodology based on scoring functions but contributes to measuring model risk in the volatility forecasting literature. The MSE and QLIKE loss (scoring) functions which are strictly consistent for volatility estimates are considered due to their widespread use in assessing the accuracy of volatility models (Patton, 2011). We estimate the model risk of volatility models based on the distance from the raw volatility estimates to the improved ones obtained by minimizing the expected score of MSE or QLIKE loss function, considering two different optimization strategies: 1) the first one is via making additive adjustments on the volatility estimates and 2) the second is via making multiplicative adjustments on the volatility estimates.

In a simulation analysis, we consider different optimization strategies to improve on variance estimates, compare different lengths of optimization windows and model risk evaluation windows, and then recommend the QLIKE-based model risk estimation methodology with additive adjustments made to the volatility estimates, as we find that the proposed method leads to high correlations, averaging from 0.88 to 0.98, between the estimated and true model risk measures. Particularly the technique based on an optimization window $\tau_2 = 500$ and a model risk evaluation window $n_1 = 250$ is highly consistent with the true model risk measure, and can explain 65% of the true model risk on average across the models. We examine the coherence properties of a reasonable measure of model risk for our proposed technique, and find that the required properties are satisfied.

In an empirical study, we explore the effect of different volatility proxies (the squared return and the 5-min realized variance, respectively) on the proposed QLIKE-based model risk measures, concluding that the model risk measure using the squared return as volatility proxy generally produces a higher level of model risk for the badly fitting models (the RiskMetrics method, RW1000 and the ARCH models), compared with the model risk measure that uses the realized variance. The level of estimated model risk based on the QLIKE loss function is not sensitive to the use of the volatility proxy across various models in general. After adjusting variance estimates for model risk, the degree of predictability of volatility models has been improved as evidenced by an increase in the values of adjusted R^2 of the MZ regressions. In addition, applying our proposed methodology to several asset classes, we find that the RiskMetrics method, the historical volatility measure RW1000 and the ARCH-type models are most affected by model risk, and that the volatility models applied to various assets carry a higher level of model risk during stressed market states than in normal market states, as expected. We also show that model misspecification risk generally plays a more dominant role than parameter estimation risk.

1.3 Original Contributions

The model-dependent estimates of market risk models or univariate volatility models are undoubtedly affected by the model risk of these models per se. With

the growing awareness of model risk management, measuring the magnitude of model risk becomes essential but challenging. This thesis, consisting of three main chapters, contributes to quantifying the model risk of common models in the context of standard market risk measures (VaR and ES) and volatility forecasting.

(1) Firstly, the original contributions of estimating ES model risk include:

• we derive the theoretical formulae for the biases of ES due to estimation and misspecification risk, as well as for the corrections of ES;

• we introduce a backtesting-based correction methodology for ES, and we provide corrections for ES model risk;

• we consider the desirable coherence properties of a measure of risk for our proposed method, via simulations;

• we consider the impact of VaR model risk on the model risk of ES;

• we illustrate the backtesting-based correction methodology using Monte Carlo simulations and an empirical analysis on different asset classes.

(2) Secondly, the original contributions of estimating the joint (VaR, ES) model risk of risk models include:

• we link model risk to the FZ class, showing the sensitivity of model ranking to the FZ class in the presence of model risk;

• we propose a general FZ scoring function-based model risk estimation methodology to estimate the joint (VaR, ES) model risk and the ES model risk;

• we verify the measures of joint model risk and ES model risk via simulations;

• we examine the coherence properties of a reasonable measure of risk for the

aforementioned measures via simulations;

• we apply this methodology to several asset classes and across various models;

• we show that adjusting for model risk has a positive effect on backtesting;

• we compare the two major model risk components, estimation risk and misspecification risk, of market risk models.

(3) Thirdly, the original contributions of estimating the model risk of univariate volatility models include:

• we develop a model risk estimation methodology for volatility estimates based on scoring functions;

• we recommend a model risk estimation method based on the QLIKE loss function using an additive structure, through a simulation analysis;

• we investigate the desirable coherence properties of a measure of risk for our proposed technique via simulations;

• we apply the QLIKE-based model risk estimation method to different asset classes and across various models;

• we consider the effect of volatility proxy on our proposed method empirically;

• we show that the efficiency of volatility models can be improved after adjusting variance estimates for model risk as evidenced by an increase in the adjusted R^2 of the MZ regressions;

• we decompose model risk into estimation risk and misspecification risk across various models, and we reinforce the reliability of the proposed technique via panel regressions of model risk components as endogenous variable.

1.4 Outline of the Thesis

The rest of this thesis is organized as follows: Chapter 2 focuses on ES model risk and proposes a backtesting-based correction methodology for ES; Chapter 3 introduces a scoring function-based model risk estimation method to quantify the joint (VaR, ES) model risk and the individual ES model risk, of market risk models; Chapter 4 develops a model risk estimation methodology for volatility estimates, considering the choice of scoring function and volatility proxy. Chapter 5 summarizes the main findings and discusses further research that builds on the findings presented in this thesis.

For a better reading experience, we make each chapter self-contained. As such, we (re)introduce variables and abbreviations in each chapter. Whenever possible, we endeavour to follow consistent notations throughout this thesis.

Notes

[1]Noticeably, some studies use the term model risk for model misspecification risk; see e.g. Escanciano and Olmo (2010a). To avoid any confusion throughout this thesis, we distinguish between model risk and model misspecification risk; the former refers to the total model risk, while the latter refers to the component misspecification risk.

Chapter 2

Model Risk of Expected Shortfall

2.1 Introduction

For risk forecasts like Value-at-Risk (VaR) and Expected Shortfall (ES)[1], the forecasting process often involves sophisticated models. The model itself is a source of risk in getting inadequate risk estimates, so assessing the model risk of risk measures becomes vital as could be seen during the global financial crisis when the pitfalls of inadequate modelling were revealed. Also, the Basel Committee (2012) advocates the use of the 2.5% ES as a replacement for the 1% VaR that has been popular for many years but has been highly debatable for its underestimation of risk.

Though risk measures are gaining popularity, a concern about the model risk of risk estimation arises. Based on a strand of literature, the model risk of risk measures can be owed to the misspecification of the underlying model (Cont,

2006), the inaccuracy of parameter estimation (Berkowitz and Obrien, 2002), or the use of inappropriate models (Daníelsson et al., 2016; Alexander and Sarabia, 2012). As such, Kerkhof et al. (2010) decompose model risk into estimation risk, misspecification risk and identification risk[2].

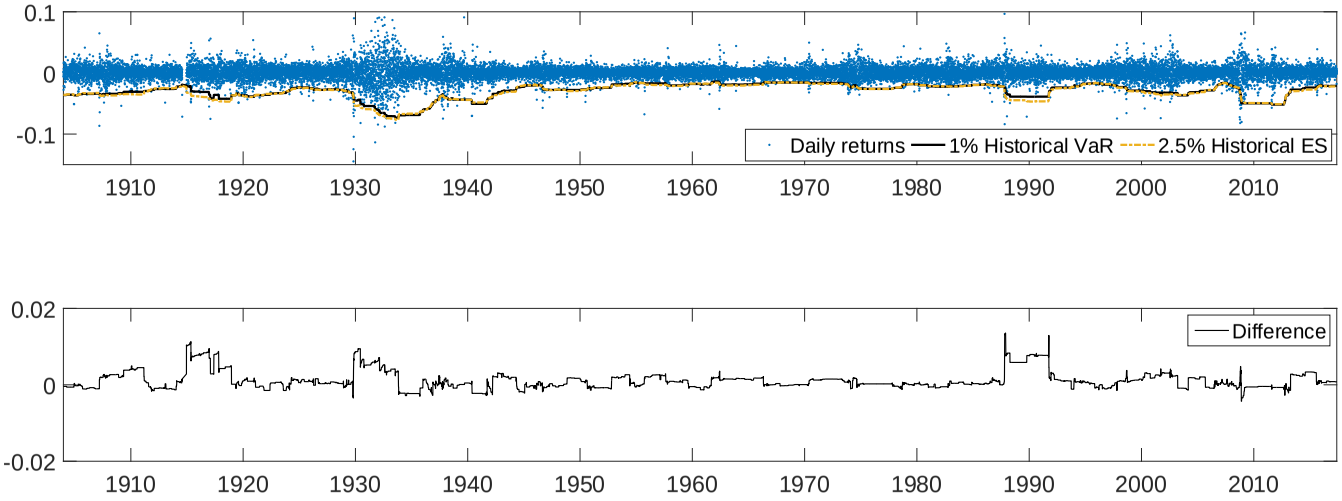
To address these different sources of model risk, several inspiring studies look into the quantification of VaR model risk followed by the adjustments of VaR estimates. One of the earliest works is Hartz et al. (2006), considering estimation error only, where the size of adjustments is based on a data-driven method. Alexander and Sarabia (2012) propose to quantify VaR model risk and correct VaR estimates for estimation and specification errors mainly based on probability shifting. Using Taylor's expansion, Barrieu and Ravanelli (2015) derive the upper bound of the VaR adjustments, only taking specification error into account, whilst Farkas et al. (2016) derive confidence intervals for VaR and Median Shortfall and propose a test for model validation based on extreme losses. Daníelsson et al. (2016) argue that the VaR model risk is significant during the crisis periods but negligible during the calm periods, computing model risk as the ratio of the highest VaR to the lowest VaR across all the models considered. However, this way of estimating VaR model risk is on a relative scale. It has been observed that model risk affects test statistics and so hypothesis testing (West, 1996; Escanciano and Olmo, 2010a)[3]. To take the effect of model risk of risk estimates into account, (1) an approach is to modify the test statistics (West, 1996); (2) an alternative is to modify the risk estimates, which can be carried out in two

different ways: (2.1) based on specific distances as in Kellner et al. (2016) and Huggenberger et al. (2018) or (2.2) based on backtests. Kerkhof et al. (2010) make absolute corrections to VaR forecasts based on regulatory backtesting measures. Similarly, Boucher et al. (2014) suggest a correction for VaR model risk, which ensures various VaR backtests are passed. These studies link model error and statistical testing, and show how backtesting can give corrections for model estimates[4]. Whilst not perfect, such a methodology can be a practical tool to improve risk estimates and provide a proxy for model risk. With the growing literature on ES backtesting (see selected ES backtests in Table 2.B.1, Appendix 2.B), measuring the model risk of ES has become plausible.

Figure 2.1.1 shows the disagreement between the daily historical VaR and ES with significance levels 1% and 2.5%, repectively, based on the DJIA index (Dow Jones Industrial Average index) daily returns from 28/12/1903 to 23/05/2017. During the crisis periods, the difference between the historical ES and VaR becomes wider and more positive, which supports the replacement of the VaR with the ES measure; nevertheless, the clustering of exceptions when ES is violated is still noticeable. In other words, the historical ES does not react to adverse changes immediately when the market returns worsen, and also it does not immediately adjust when the market apparently goes back to normal.

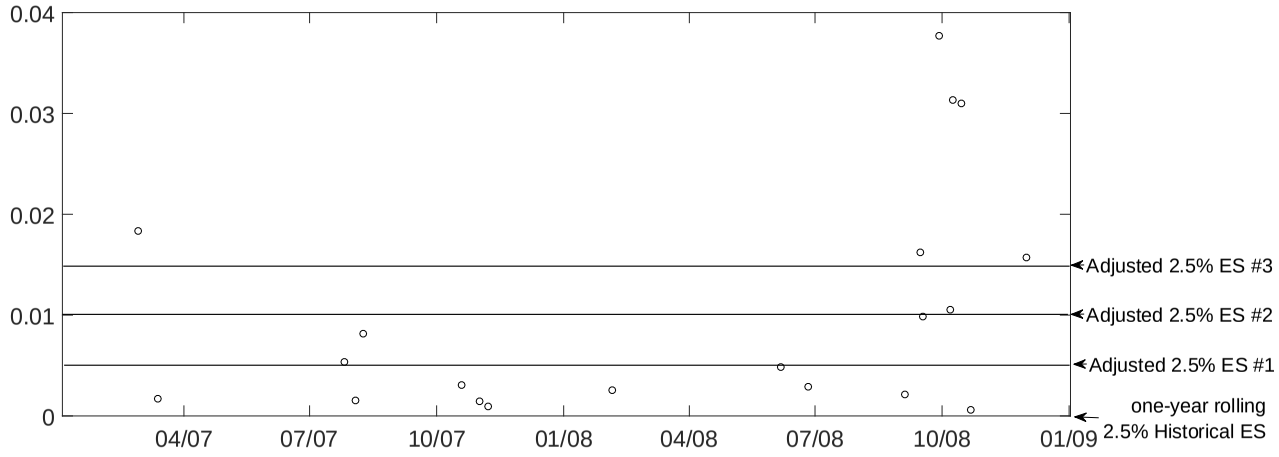
Another example is around the 2008 financial crisis, presented in Figure 2.1.2, which shows the peaked-over-ES ($\alpha = 2.5\%$) and three tiers of corrections (labelled as #1, #2 and #3 on the right-hand side) made to the daily historical ES

Figure 2.1.1: 1% historical VaR and 2.5% historical ES, based on DJIA



The upper panel of this figure shows the DJIA index daily returns, the daily historical VaR estimates ($\alpha = 1\%$) and the daily historical ES estimates ($\alpha = 2.5\%$) from 28/12/1903 to 23/05/2017; the lower panel shows the difference between the 2.5% historical ES and the 1% historical VaR. We use a four-year rolling window to compute the risk estimates.

Figure 2.1.2: Peaked-over-ES and adjustments, based on DJIA



This figure presents the peaked-over-ES and three tiers of adjustments labelled as #1, #2 and #3 for daily historical ES estimates, based on the DJIA index from 01/01/2007 to 01/01/2009. One-year moving window is used to forecast daily historical ES ($\alpha = 2.5\%$).

estimates ($\alpha = 2.5\%$), based on a one-year rolling window. Adjustment #1 with a magnitude of 0.005 (about 18% in relative terms) added to the daily ES estimates can avoid most of the exceptions that occur during this crisis. The higher the adjustment level (#2 and #3), the more the protection from extreme losses, but even an adjustment of 0.015 (adjustment #3) still has several exceptions. However, too much protection is not favorable to risk managers, implying that effective adjustments (not too large or too small) for ES estimates are needed to cover for model risk. In this chapter, we mainly focus on several ES backtests with respect to the following properties[5] of a desirable ES forecast: one referring to the expected number of exceptions, one regarding the absence of violation clustering, and one about the appropriate size of exceptions.

To the best of our knowledge, we are the first to quantify ES model risk as a correction needed to pass various ES backtests (Du and Escanciano, 2016; Acerbi and Szekely, 2014; McNeil and Frey, 2000), and examine whether our chosen measures of model risk satisfy certain desirable properties which would facilitate the regulations concerning these measures. Also, we compare the correction for the model risk of VaR ($\alpha = 1\%$) with that for ES model risk ($\alpha = 2.5\%$) based on different models and different assets, concluding that the 2.5% ES is less affected by model risk than the 1% VaR. Regarding the substantial impact of VaR on ES in terms of the ES calculations and the ES backtesting, if VaR model risk is accommodated for, then the correction made to ES forecasts reduces by 50% on average.

The structure of Chapter 2 is as follows: Section 2.2 analyzes the sources of ES model risk focusing on estimation and specification errors, and performs Monte Carlo simulations to quantify them; Section 2.3 proposes a backtesting-based correction methodology for ES model risk, considers the properties of our chosen measures of model risk and also investigates the impact of VaR model risk on the model risk of ES; Section 2.4 presents the empirical study and Section 2.5 concludes.

2.2 Model risk of Expected Shortfall

2.2.1 Sources of model risk

We first establish a general scheme (see Figure 2.2.1) in which the sources of model risk of risk estimates are shown. Consider a portfolio affected by risk factors, and the goal is to compute risk estimates such as VaR and ES. The first step is the identification of risk factors, and this process is affected by identification risk, which arises when some risk factors are not identified, with a very high risk of producing inaccurate risk estimates. The next step is the specification of risk factor models which, again, will have a large effect on the estimation of risk. This is followed by the estimation of the risk factor model (this, in our view, has a medium effect on the risk estimate). In step 3, the relationship between the portfolio P&L and the risk factors is considered and the formulation of this model will have a high effect on the estimation of the risk. The estimation of this will

have a medium effect on the risk estimation. Step 4 links the risk estimation with the dependency of the P&L series on the risk factors.

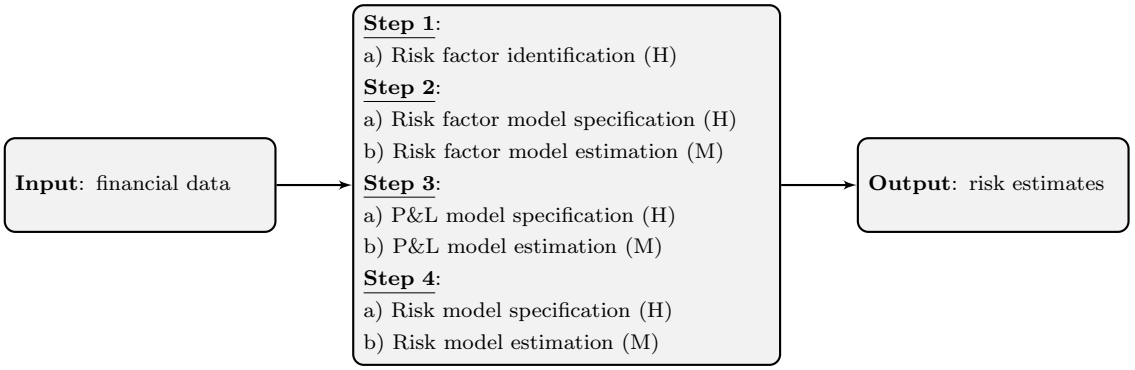
For example, when computing the VaR of a portfolio of derivatives, step 1 would identify the sources of risk, step 2 would specify and estimate the models describing these risk factors (underlying asset returns most importantly), step 3 would model the P&L of the portfolio as a function of the risk factors, and in step 4 the risk model would transform P&L values into risk estimates.

The diagram shows that the main causes of model risk of risk estimates are (1) identification error, (2) model estimation error (for the risk factor model, the P&L model or the risk model), which arises from the estimation of the parameters of the model and (3) model specification error (for the risk factor model, the P&L model or the risk model), which arises when the true model is not known. Other sources of model risk that may give wrong risk estimates are, for example, granularity error, measurement error and liquidity risk (Boucher et al., 2014).

2.2.2 Bias and correction of Expected Shortfall

Most academic research on the adequacy of risk models mainly focuses on two of the sources of model risk: estimation error and specification error. Referring to Boucher et al. (2014), the theoretical results about the two sources of VaR model risk are presented in Appendix 2.A. In a similar vein, we investigate the impact of the earlier mentioned two errors on the ES estimates, deriving the theoretical

Figure 2.2.1: Risk estimation process



This diagram shows the sources of model risk of risk estimates. H and M represent high and medium impacts on risk estimates, respectively.

formulae for estimation and specification errors, as well as correction of ES. VaR[6], for a given distribution function F and a given significance level α, is defined as:

$$VaR_t(\alpha) = -inf\{q : F_t(q) \geq \alpha\}, \tag{2.2.1}$$

where q denotes the quantile of the cumulative distribution F. ES, as an absolute downside risk measure, measures the average losses exceeding VaR, taking extreme losses into account; it is given by:

$$ES_t(\alpha) = \frac{1}{\alpha}\int_0^\alpha VaR_t(u)du \tag{2.2.2}$$

Estimation bias of Expected Shortfall

Assuming that the data generating process (DGP), a model with a cumulative distribution F for the returns, is known and the true parameter values (θ_0) of this 'true' model are also known, the theoretical VaR, denoted by ThVaR(θ_0, α) and the theoretical ES, denoted by ThES(θ_0, α), both at a significance level α, can be computed as:

$$ThVaR(\theta_0, \alpha) = -q_\alpha^F = -F_\alpha^{-1} \tag{2.2.3}$$

$$ThES(\alpha) = \frac{1}{\alpha} \int_0^\alpha ThVaR(\theta_0, u)du \tag{2.2.4}$$

Now, we assume that the DGP is known, but the parameter values are not known. The estimated VaR in this case is denoted by $VaR(\hat{\theta}_0, \alpha)$, where $\hat{\theta}_0$ is an estimate of θ_0. The relationship between the theoretical VaR and the estimated VaR is:

$$ThVaR(\theta_0, \alpha) = VaR(\hat{\theta}_0, \alpha) + bias(\theta_0, \hat{\theta}_0, \alpha) \tag{2.2.5}$$

We also have that:

$$ThVaR(\theta_0, \alpha) - \mathbb{E}(VaR(\hat{\theta}_0, \alpha)) = \mathbb{E}(bias(\theta_0, \hat{\theta}_0, \alpha)) \tag{2.2.6}$$

where $\mathbb{E}[bias(\theta_0, \hat{\theta}_0, \alpha)]$ denotes the mean bias of the estimated VaR from the

theoretical VaR as a result of model estimation error. Based on this, we can write the estimation bias of $\mathrm{ES}(\hat{\theta}_0, \alpha)$, and we have that

$$ThES(\theta_0, \alpha) - \mathbb{E}[ES(\hat{\theta}_0, \alpha)] = \frac{1}{\alpha}\int_0^\alpha \mathbb{E}[bias(\theta_0, \hat{\theta}_0, u)]du, \tag{2.2.7}$$

Ideally, correcting for the estimation bias, the ES estimate, denoted by $ES(\hat{\theta}_0, \alpha)$, can be improved as below:

$$ES^E(\hat{\theta}_0, \alpha) = ES(\hat{\theta}_0, \alpha) + \frac{1}{\alpha}\int_0^\alpha \mathbb{E}[bias(\theta_0, \hat{\theta}_0, u)]du \tag{2.2.8}$$

Specification and estimation biases of Expected Shortfall

However, in most cases the 'true' DGP is not known, and the returns are assumed to follow a different model, given a cumulative distribution ($\hat{F}$) for the returns with estimated parameter values $\hat{\theta}_1$, where θ_0 and $\hat{\theta}_1$ can have different dimensions depending on the models used and their values are expected to be different. This gives the following value for the estimated VaR:

$$VaR(\hat{\theta}_1, \alpha) = -q_\alpha^{\hat{F}} = -\hat{F}_\alpha^{-1} \tag{2.2.9}$$

The relationship between the true VaR and the estimated VaR is given as:

$$ThVaR(\theta_0, \alpha) = VaR(\hat{\theta}_1, \alpha) + bias(\theta_0, \theta_1, \hat{\theta}_1, \alpha) \tag{2.2.10}$$

where θ_1 and $\hat{\theta}_1$ have the same dimension under the specified model, but θ_1 denotes the true parameter values different from the estimated parameter values of $\hat{\theta}_1$. Similarly:

$$ThVaR(\theta_0, \alpha) - \mathbb{E}(VaR(\hat{\theta}_1, \alpha)) = \mathbb{E}(bias(\theta_0, \theta_1, \hat{\theta}_1, \alpha)) \tag{2.2.11}$$

where $\mathbb{E}[bias(\theta_0, \theta_1, \hat{\theta}_1, \alpha)]$ denotes the mean bias of the estimated VaR from the theoretical VaR as a result of model specification and estimation errors. According to equation (2.2.2), the mean estimation and specification biases of ES can be formulated as below:

$$ThES(\theta_0, \alpha) - \mathbb{E}[ES(\hat{\theta}_1, \alpha)] = \frac{1}{\alpha}\int_0^{\alpha} \mathbb{E}[bias(\theta_0, \theta_1, \hat{\theta}_1, v)]dv \tag{2.2.12}$$

Correcting for these biases, the estimated ES, denoted by $ES(\hat{\theta}_1, \alpha)$, can be improved as:

$$ES^{SE}(\hat{\theta}_1, \alpha) = ES(\hat{\theta}_1, \alpha) + \frac{1}{\alpha}\int_0^{\alpha} \mathbb{E}[bias(\theta_0, \theta_1, \hat{\theta}_1, v)]dv \tag{2.2.13}$$

In practice, the choice of the risk model for computing VaR and ES forecasts is usually subjective, along with specification errors (and other sources of model risk). In Appendix 2.C, we give a review of risk forecasting models used in this chapter.

2.2.3 Monte Carlo simulations

In this section, assume a simplified risk estimation process (Figure 2.2.1) so that only one risk factor exists. Thus, the identification risk and the P&L model specification and estimation risks are not modelled, and we are left with the specification and estimation risks for the risk factor model and, consequently, for the risk model, namely steps 2 and 4. Following the theoretical formulae for estimation and specification errors of the ES estimates, Monte Carlo simulations are implemented to investigate the impacts of these two errors on the estimated ES.

We simulate the daily return series assuming a model, thus knowing the theoretical ES. Then, the parameters are estimated using the same model as specified to generate the daily returns, thus giving the value of the estimation bias of ES, as in equation (2.2.7). We also forecast ES based on other models to examine the values of joint estimation and specification biases of ES, as in equation (2.2.12).

In our setup[7], a GARCH(1,1) model with normal disturbances (GARCH(1,1)-N) is assumed to be the 'true' data generating process, given by:

$$r_t = \mu + \varepsilon_t \tag{2.2.14}$$

$$\varepsilon_t = \sigma_t \cdot z_t, \; z_t \sim \mathcal{N}(0,1) \tag{2.2.15}$$

$$\sigma^2{}_t = \omega + \alpha\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2 \tag{2.2.16}$$

Using market data, we first estimate the parameters[8] of this model. Next, we simulate 1,000 paths of 1,000 daily returns, compute one-step ahead ES forecasts under several different models and compare these forecasts with the theoretical ES. The purpose of Monte Carlo simulations is to compute the perfect corrections for the model risk of ES forecasts. The second and third columns in Table 2.2.1 present the annualized ES forecasts and theoretical ES at 5%, 2.5% and 1%.

We compare the theoretical ES given by the data generating process with the estimated ES based on the same specification in Panel A of Table 2.2.1, showing that the mean estimation bias is close to 0 for the 5%, 2.5% and 1% ES estimates. Also, the estimation bias can be reduced by increasing the size of the estimation period as suggested by Du and Escanciano (2016). The standard error of the bias decreases when α increases, as expected. In Panel B, the mean specification and estimation biases are computed from the theoretical ES and the historical ES. The negative values of the bias show that the estimated ES is more conservative than the theoretical ES, whilst the positive values of the bias refer to an estimated ES lower than the theoretical ES. Panel C examines the specification and estimation biases of the Gaussian Normal ES estimates. In this case, the Gaussian Normal ES estimates are more conservative than the theoretical ES. The specification and estimation biases of the ES estimates computed from EWMA are positive as shown in Panel D, which requires a positive adjustment to be added to the EWMA ES estimates.

Table 2.2.1: Simulated bias associated with the ES estimates

Significance level	Mean estimated ES(%)	Theoretical ES(%)	Mean bias(%)	Std. err of bias(%)
Panel A. GARCH(1,1)-N DGP with estimated GARCH(1,1)-N ES: estimation bias				
α=5%	23.82	23.83	0.01	1.73
α=2.5%	28.50	28.51	0.01	1.94
α=1%	34.07	34.08	0.01	2.20
Panel B. GARCH(1,1)-N DGP with historical ES: specification and estimation biases				
α=5%	28.92	23.83	-5.09	15.79
α=2.5%	36.38	28.51	-7.87	18.97
α=1%	45.77	34.08	-11.69	23.16
Panel C. GARCH(1,1)-N DGP with Gaussian Normal ES: specification and estimation biases				
α=5%	26.27	23.83	-2.44	14.86
α=2.5%	31.27	28.51	-2.76	16.84
α=1%	37.23	34.08	-3.15	19.20
Panel D. GARCH(1,1)-N DGP with EWMA ES: specification and estimation biases				
α=5%	21.68	23.83	2.15	2.54
α=2.5%	26.31	28.51	2.20	2.87
α=1%	31.82	34.08	2.26	3.28

This table presents biases between the simulated theoretical ES and the estimated ES, based on the DJIA index from 01/01/1900 to 23/05/2017, downloaded from DataStream. First, we simulate 1,000 paths of 1,000 daily returns according to the DGP of GARCH(1,1)-N. Then we forecast ES based on the GARCH(1,1)-N, historical, Gaussian Normal and EWMA ($\lambda = 0.94$) specifications, for α = 5%, 2.5% and 1%.

The specification and estimation biases in Panel B, C and D are much higher than the estimation bias in Panel A in absolute value, indicating that the specification error has a bigger importance than the estimation error. Overall, our results indicate that an adjustment is needed to correct for the model risk of ES estimates.

2.3 Measuring ES model risk

2.3.1 Backtesting-based correction methodology for ES

If a data generating process is known, then it is straightforward to compute the model risk of ES, as shown in Table 2.2.1. In a realistic setup, the 'true' model is unknown, so it is impossible to measure model risk directly. By correcting the estimated ES and forcing it to pass backtests, model risk is not broken into its components, but the correction would be for all the types of model risk considered jointly. In this way, the backtesting-based correction methodology for ES, proposed in this chapter, provides corrections for all the sources of ES model risk.

Comparing the ex-ante forecasted ES with the ex-post realizations of returns, the accuracy of ES estimates is examined via backtesting. For a given backtest, we can compute the correction needed for the ES forecasts made by a risk model, M_j, so that the adjusted ES passes this backtest. The value of ES corrected via

backtesting, $ES^B_{i,j}$, is written as:

$$ES^B_{i,j}(\hat{\theta}_1, \alpha) = ES_j(\hat{\theta}_1, \alpha) + C^*_{i,j} \tag{2.3.1}$$

The minimum correction is given by:

$$C^*_{i,j} = min\{C_{i,j} | ES_{j,t}(\hat{\theta}_1, \alpha) + C_{i,j} \text{ passes the } i\text{th backtest}, t = 1, ..., T, C_{i,j} \geq 0\}$$

where $\{ES_{j,t}(\hat{\theta}, \alpha), t = 1, ..., T\}$ denotes the forecasted ES made using model M_j during the period from 1 to T. A correction, $C_{i,j} = C_{i,j}(\theta_0, \theta_1, \hat{\theta}_1, \alpha)$, is needed to be made so that the ith backtest of the ES estimates is passed successfully; of these, $C^*_{i,j}$ is the minimum correction required to pass the ith ES backtest. In this chapter, $i \in \{1, 2, 3, 4\}$; $C_{1,j}$, $C_{2,j}$, and $C_{3,j}$ refer to the correction required to pass the unconditional coverage test for ES and the conditional coverage test for ES introduced by Du and Escanciano (2016), and the Z_2 test proposed by Acerbi and Szekely (2014), respectively. Additionally, the exceedance residual test by McNeil and Frey (2000), associated with $C_{4,j}$, is an alternative to the Z_2 test. By learning from past mistakes, we can find the appropriate correction made to the ES forecasts, through which the model risk of ES forecasts can be quantified.

In this chapter, we define model risk as $MR^I : \mathbb{R}^n \times V_M \rightarrow \mathbb{R}^+$, where $MR^I((X_{0,t}), M_j)$ refers to the maximum of the optimal corrections $C^*_{i,j}$ made to ES forecasts of a series of empirical observations $X_{0,t}$ during the period $t = 1, ..., T$,

which ensures that certain backtests I are passed. V_M represents a set of models with $M_j \in V_M$. This definition can be transformed into the following definition of model risk $MR : \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^+$:

$$MR^I\left((X_{0,t}), (v_{j,t}), (e_{j,t})\right) = \max_I(C^*_{i,j}). \tag{2.3.2}$$

In this notation, X, v, and e denote the empirical observations and, respectively, the one-step ahead VaR and ES forecasts made for time t. The subscripts j and i refer to the model j used to build risk forecasts and the ith backtest, accordingly. The superscript I refers to a set of ES backtests used to make corrections for ES model risk. For example, if $I = \{1,2,3\}$, we find the maximum correction needed to pass the unconditional coverage test (UC_{ES} test), the conditional coverage test (CC_{ES} test) and the Z_2 test jointly. Likewise, we also consider $I = \{1,2\}$ or $\{1,2,3,4\}$. Clearly, this representation of model risk shows that it is affected by the data and the risk model used to make VaR and ES forecasts. In the following, for simplification we use the notation $X = (X_{0,t}), v_j = (v_{j,t})$, $e_j = (e_{j,t})$, and $MR^I = MR$ given I.

2.3.2 Backtesting framework for ES

Backtesting, as a way of model validation, checks whether ES forecasts satisfy certain desirable criteria. Here we consider that a good ES forecast should have an appropriate frequency of exceptions, absence of volatility clustering in

the tail and a suitable magnitude of the violations. Regarding these attractive features, we mainly implement the unconditional/conditional coverage test for ES (UC_{ES}/CC_{ES} test), and the Z_2 test (Du and Escanciano, 2016; Acerbi and Szekely, 2014).

Exception frequency test

Based on the seminal work of (Kupiec, 1995), in which the unconditional coverage test (UC_{VaR} test) for VaR considers the number of exceptions, Du and Escanciano (2016) investigate the cumulation of violations and develop an unconditional coverage test statistic for ES. The estimated cumulative violations $\hat{H}_t(\alpha)$ are defined as:

$$\hat{H}_t(\alpha) = \frac{1}{\alpha}(\alpha - \hat{u}_t)\mathbf{1}(\hat{u}_t \leqslant \alpha) \tag{2.3.3}$$

where $\hat{u}_t$ is the estimated probability level corresponding to the daily returns (r_t) in the estimated distribution $(\hat{F}_t)$ with the estimated parameters $(\hat{\theta}_1)$, and Ω_{t-1} denotes all the information available until $t-1$.

$$\hat{u}_t = \hat{F}(r_t, \Omega_{t-1}, \hat{\theta}_1) \tag{2.3.4}$$

The null hypothesis of the unconditional coverage test for ES, H_1, is given by:

$$H_1 : \mathbb{E}\left[H_t(\alpha, \theta_0) - \frac{\alpha}{2}\right] = 0 \tag{2.3.5}$$

Hence, the simple t-test statistic[9] and its distribution is:

$$U_{ES} = \frac{\sqrt{n}\left(1/n\sum_{t=1}^{n}\hat{H}_t(\alpha) - \alpha/2\right)}{\sqrt{\alpha(1/3-\alpha/4)}} \sim N(0,1) \tag{2.3.6}$$

Exception frequency and independence test

The conditional coverage test (CC_{VaR} test) for VaR is a very popular formal backtesting measure (Christoffersen, 1998). Inspired by this, Du and Escanciano (2016) propose a conditional coverage test for ES and give its test statistic. The null hypothesis of the conditional coverage test for ES, H_2, is given by:

$$H_2 : \mathbb{E}\left[H_t(\alpha,\theta_0) - \frac{\alpha}{2}|\Omega_{t-1}\right] = 0 \tag{2.3.7}$$

Du and Escanciano propose a general test statistic to test the mth-order dependence of the violations, following a Chi-squared distribution with m degrees of freedom. In the present context, the first order dependence of the violations is considered, so the test statistic follows $\chi^2(1)$. During the evaluation period from $t = 1$ to $t = n$, the basic test statistic[0], $C_{ES}(1)$, is written as:

$$C_{ES}(1) = \frac{n^3}{(n-1)^2} \cdot \frac{\left(\sum_{t=2}^{n}(\hat{H}_t(\alpha)-\alpha/2)(\hat{H}_{t-1}(\alpha)-\alpha/2)\right)^2}{\left(\sum_{t=1}^{n}(\hat{H}_t(\alpha)-\alpha/2)(\hat{H}_t(\alpha)-\alpha/2)\right)^2} \sim \chi^2(1) \tag{2.3.8}$$

Escanciano and Olmo (2010b) point out that the VaR (and correspondingly, ES) backtesting procedure may not be convincing enough due to estimation risk

and propose a robust backtest. In spite of that, Du and Escanciano (2016) agree with Escanciano and Olmo (2010b) that estimation risk can be ignored and the basic test statistic is robust enough against the alternative hypothesis if the estimation period is much larger than the evaluation period. In this context, the estimation period (1,000) we use is much larger than the evaluation period (250), so the robust test statistic is not considered.

Exception frequency and magnitude test

Acerbi and Szekely (2014) directly backtest ES by using the test statistic (Z_2 test):

$$\mathrm{Z}_2 = \sum_{t=1}^{T} \frac{r_t I_t}{T\alpha ES_{\alpha,t}} + 1 \tag{2.3.9}$$

I_t, an indicator function, is equal to 1 when the forecasted VaR is violated, otherwise, 0. The Z_2 test is non-parametric and only needs the magnitude of the VaR violations ($r_t I_t$) and the predicted ES ($ES_{\alpha,t}$), thus easily implemented and considered a joint backtest of VaR and ES forecasts. The Z_2 score at a certain significance level can be determined numerically based on the simulated distribution of Z_2. If the test statistic is smaller than the Z_2 score[10], the model is rejected. The authors also demonstrate that there is no need to do Monte Carlo simulations to store the predictive distributions due to the stability of the p-values of the Z_2 test statistic across different distribution types. Clift et al. (2016) also support this test statistic (Z_2) by comparing some existing backtesting approaches for ES.

In the Z_2 test, ES is jointly backtested in terms of the frequency and the magnitude of VaR exceptions. Alternatively, we also use a tail losses based backtest for ES, proposed by McNeil and Frey (2000), only taking into account the size of exceptions. The exceedance residual (er_t), conditional on the VaR being violated (I_t), is given below:

$$er_t = (r_t + ES_{\alpha,t}) \cdot I_t \tag{2.3.10}$$

here r_t denotes the return at time t, and $ES_{\alpha,t}$ represents the forecasted ES for time t. The null hypothesis of the backtest is that the exceedance residuals are on average equal to zero against the alternative that their mean is greater than zero. The p-value used for this one-sided bootstrapped test is 0.05.

2.3.3 Properties of measures of model risk

We introduce some basic notations and assumptions: we assume a *r.v.* A defined on a probability space $(\Omega, \mathcal{F}, P)$, and F_A the associated distribution function. If $F_A \equiv F_B$, the cumulative distributions associated with A and B are considered the same and we write $A \sim B$. In the same fashion, we will write $A \sim F$, if $F_A \equiv F$. A measure of risk is a map $\rho : V_\rho \to \mathbb{R}$, defined on some space of *r.v.* V_ρ.

Artzner et al. (1999) propose four desirable properties of measures of risk (market and nonmarket risks), and argue that effectively regulated measures of risk should satisfy the four properties stated below:

1) Monotonicity: $A, B \in V_\rho, A \leq B \Rightarrow \rho(A) \geq \rho(B)$.

2) Translation invariance: $A \in V_\rho, a \in \mathbb{R} \Rightarrow \rho(A + a) = \rho(A) - a$.

3) Subadditivity: $A, B, A + B \in V_\rho \Rightarrow \rho(A + B) \leq \rho(A) + \rho(B)$.

4) Positive homogeneity: $A \in V_\rho, h > 0, h \cdot A \in V_\rho \Rightarrow \rho(h \cdot A) = h \cdot \rho(A)$.

ES is considered coherent as a result of satisfying the above four properties, whilst VaR is not due to the lack of subadditivity (Acerbi and Tasche, 2002). As model risk is becoming essential from a regulatory point of view, we are examining whether the above properties hold for our proposed measure of model risk of ES.

Regarding this measure of model risk, the four desirable properties of risk measures mentioned above are considered below:

1. *Monotonicity*:

   1a) For a given model M_j, and two data series X, Y with $X \leq Y$, it is desirable to have that $MR(X, v_j, e_j) \geq MR(Y, v_j, e_j)$.

   1b) For a data series X, models $M_1, M_2 \in V_M, v_1 < v_2, e_1 < e_2$, it is desirable to have that $MR(X, v_1, e_1) \geq MR(X, v_2, e_2)$.

   The property *1a)* states that the risk estimates (v_j, e_j) of model M_j that is applied to the data series Y are not able to accommodate for bigger losses associated with the data series X and thus should have a higher model risk, which is in line with the argument of Daníelsson and Zhou (2017). The property *1b)* is a natural requirement that, for a given return series, models that

forecast low values of VaR and ES risk estimates should carry a higher model risk (and require higher corrections).

2. *Translation invariance*:

   2a) For a given model M_j, a series of data X, and a constant $a \leq v_j$, it is desirable to have that $MR(X + a, v_j - a, e_j - a) = MR(X, v_j, e_j)$.

   2b) For a given model M_j, a series of data X, and a constant $a \in \mathbb{R}^+$, it is desirable to have that $MR(X + a, v_j, e_j) \geq MR(X, v_j, e_j) - a$.

   2c) For a given model M_j, a series of data X, and a constant $a \in \mathbb{R}^+$, it is desirable to have that $MR(X, v_j + a, e_j + a) \geq MR(X, v_j, e_j) - a$.

   Generally, when shifting the observations with a constant and lowering the values of VaR and ES forecasts by the same amount, the model risk is expected to stay constant in the case of *2a)*. In *2b)* and *2c)*, if the real data or the risk forecasts are shifted with a positive constant (a), the model risk would be larger than (or equal with) the difference between the previous model risk and the size of the shift.

3. *Subadditivity*

   3a) For a given model M_j, $(v_{1j}, e_{1j}), (v_{2j}, e_{2j})$ and $(v_{1+2,j}, e_{1+2,j})$ are estimates based on X_1, X_2 and $X_1 + X_2$, it is desirable to have that:

   $$MR(X_1 + X_2, v_{1+2,j}, e_{1+2,j}) \leq MR(X_1, v_{1j}, e_{1j}) + MR(X_2, v_{2j}, e_{2j}).$$

   The property *3a)* is desirable, since we expect that the model risk is smaller in a diversified portfolio than the sum of the model risks of the individual assets. However, the desirability of subadditivity for measures of risk is an ongoing

discussion. Cont et al. (2010) point out that subadditivity and statistical robustness are exclusive for measure of risks, and that robustness should be a concern to the regulators. Also, Krätschmer et al. (2012, 2014, 2015) argue that robustness may not be necessary in a risk management context. Subadditivity, expressed in this format, is not too important because we rarely use the same model for two different data sets.

4. *Positive homogeneity*

   4a) For a given model M_j, and a data series X, $h > 0, h \cdot X \in V_M$, we have that $MR(h \cdot X, h \cdot v_j, h \cdot e_j) = h \cdot MR(X, v_j, e_j)$.

   The property *4a)* states that the change in the size of the investment is consistent with the change in the size of model risk.

Property: *Assuming model risk is computed as in equation (2.3.2), the following properties will hold:*

(1) For $I = \{1,2\}$, properties 1a), 1b), 2a), 2b), 2c) and 4a).

(2) For $I = \{1,2,3\}$, properties 1a), 1b), 2a) and 4a).

We mainly consider two measures of ES model risk: *(1)* When we compute the model risk of ES in terms of the UC_{ES} and CC_{ES} tests (I=$\{1,2\}$), allowing for the frequency and clustering of exceptions, all properties considered above hold, except for subadditivity; *(2)* when we compute the model risk of ES in terms of the UC_{ES}, CC_{ES} and Z_2 tests (I =$\{1,2,3\}$), allowing for the frequency, clustering and size of exceptions, *2b)* and *2c)* of translation invariance and subadditivity are

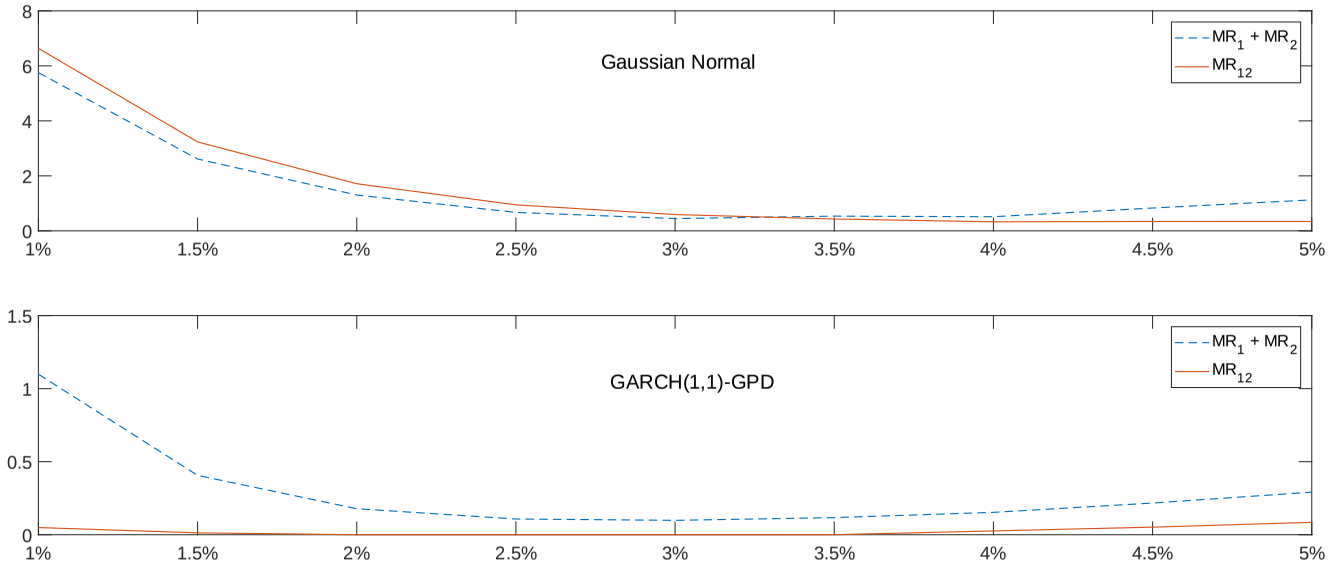
not satisfied, whilst the rest still hold. Due to the nature of the Z_2 test, translation invariance is not guaranteed. This is not necessarily a problem, because shifting data or risk estimates with a constant is not encountered routinely.

Next, let's look at subadditivity in more detail and we are going to give an example why it is not always satisfied for $MR^{I=\{1,2,3\}}$. Inheriting an example from Daníelsson et al. (2013), we consider two independent assets, X_1 and X_2, but with the same distribution, specified as:

$$X = \epsilon + \eta, \quad \epsilon \sim \text{IID}\mathcal{N}(0,1), \quad \eta = \begin{cases} 0 & \text{with a probability } 0.991 \\ -10 & \text{with a probability } 0.009 \end{cases} \tag{2.3.11}$$

Based on this, we generate two series of data with 5,000 observations for X_1 and X_2. Considering the Gaussian Normal or GARCH(1,1)-GPD model used to make one-step ahead VaR and ES forecasts at different significance levels with a rolling window of length 1,000, we measure the model risk of ES forecasts based on the two models by the backtesting-based methodology. Then we compare the model risk of an equally weighted portfolio of $(X_1 + X_2)$, MR^I_{12}, with the sum of model risks of X_1 and X_2, $MR^I_1 + MR^I_2$, shown in Figure 2.3.1. The upper figure shows that the model risk of ES of an equally weighted portfolio based on the Gaussian Normal model is higher than the sum of model risks of ES of the two individual assets at some significance levels such as 2.5%. One possible explanation for this is that the Gaussian Normal model is not appropriate to make ES forecasts at

Figure 2.3.1: Subadditivity of ES model risk measure



This figure presents average values of ES model risk of an equally weighted portfolio, $(X_1 + X_2)$, and the sum of ES model risks of X_1 and X_2, based on the Gaussian Normal ES and the GARCH(1,1)-GPD ES for a series of significance levels.

these alpha levels. In the lower figure where the model used offers a better fit, the model risk of the portfolio is much lower than the sum of model risks based on the GARCH(1,1)-GPD model. Therefore, subadditivity is not guaranteed for our measure of model risk. However, in our applications, similar to the second part of Figure 2.3.1, subadditivity is satisfied when the model fits the data well.

2.3.4 The impact of VaR model risk on the model risk of ES

The backtesting-based correction methodology for ES shows that the correction made to the ES forecasts can be regarded as a barometer of ES model risk. VaR has been an indispensable part of ES calculations and the ES bakctests used in this chapter. For instance, the Z_2 test (Acerbi and Szekely, 2014) is commonly considered as a joint backtest of VaR and ES. For this reason, it is of much interest to explore to what extent the model risk of VaR is transferred to the model risk of ES. On the one hand, ES calculations may be affected by the model risk of VaR, since the inaccuracy of VaR estimates is carried over to the ES estimates as seen in equation (2.2.2). On the other hand, the wrong VaR estimates may have an impact on backtesting, thus leading to inappropriate corrections of ES estimates. As such, the measurement of the ES correction required to pass a backtest is likely to be affected by VaR model risk. To address this, as an additional exercise, we compute the optimal correction of VaR for model risk (estimated at the same

significance level as the corresponding ES) as in Boucher et al. (2014)[11]. Then we use the corrected VaR for ES calculation, estimating ES corrected for VaR model risk. Consequently, based on the backtesting-based correction framework, the optimal correction made to the ES, corrected for VaR model risk, is gauged as a measurement of ES model risk alone.

2.3.5 Monte Carlo simulations of ES model risk

According to the backtesting-based correction methodology for ES, we quantify ES model risk by passing the aforementioned ES backtests based on Monte Carlo simulations, where we simulate 5,000 series of 1,000 returns using a GARCH(1,1)-t model with model parameters taken from Kratz et al. (2018), specified below:

$$r_t = \sigma_t Z_t, \qquad \sigma_t^2 = 2.18 \times 10^{-6} + 0.109 r_{t-1}^2 + 0.890 \sigma_{t-1}^2, \tag{2.3.12}$$

where Z_t follows a standardised Student's t distribution with 5.06 degrees of freedom.

We implement several well known models (see details in Appendix 2.C) for comparison, such as the Gaussian Normal distribution, the Student's t distribution, GARCH(1,1) with normal or standardised Student's t innovations, GARCH(1,1)-GPD, EWMA, Cornish-Fisher expansion as well as the historical method.

It is known that ES considers average extreme losses which VaR disregards. Consequently, it is of interest to investigate the adequacy of ES estimates in

measuring the size of extreme losses and also quantify ES model risk by passing the Z_2 test inasmuch as the Z_2 test considers the frequency and magnitude of exceptions. Table 2.3.1 shows the mean values of the optimal absolute and relative corrections (in the 3rd and 5th columns) made to the daily ES ($\alpha = 2.5\%$), estimated by different methods, in order to pass the Z_2 test without considering the impact of VaR model risk on the ES calculations and ES backtesting, as well as the mean values of the absolute and relative optimal correction (in the 4th and 6th columns) made to the daily ES after correcting VaR model risk. In this simulation study, the data generating process is specified by GARCH(1,1)-t as in equation (2.3.12). Thus, according to the last two rows in Table 2.3.1, ES estimates are only subject to estimation risk measured by the mean of the absolute optimal correction, 0.0001, which is much smaller than the mean values of the optimal corrections associated with the other models, which are different from the DGP. This shows that misspecification risk plays a crucial role in giving accurate ES estimates, and also applies when we correct for VaR model risk. The mean values of the optimal corrections made to the ES estimates generally decrease after excluding the impact of VaR model risk on ES model risk.

2.4 Empirical Analysis

Based on the same set of models used in the previous section, we evaluate the backtesting-based correction methodology for ES using the DJIA index from

Table 2.3.1: Optimal correction for ES based on the Z_2 test, before and after correcting VaR

Model	Mean ES	Abs. C_3 $(*10^{-2})$	Abs. C_3^* $(*10^{-2})$	Rel. C_3	Rel. C_3^*
Historical	0.062	0.45	0.41	7.1%	6.6%
EWMA	0.046	0.73	0.70	15.7%	14.9%
Gaussian Normal	0.047	0.91	0.87	19.5%	18.4%
Student's t	0.060	0.40	0.36	6.6%	6.0%
GARCH(1,1)-N	0.039	0.08	0.08	2.2%	1.9%
Cornish-Fisher	0.097	0.03	0.03	0.3%	0.3%
GARCH(1,1)-GPD	0.046	0.03	0.02	0.7%	0.6%
GARCH(1,1)-t	0.045	0.01	0.01	0.3%	0.3%
DGP	0.046	0.00	0.00	0.1%	0.1%

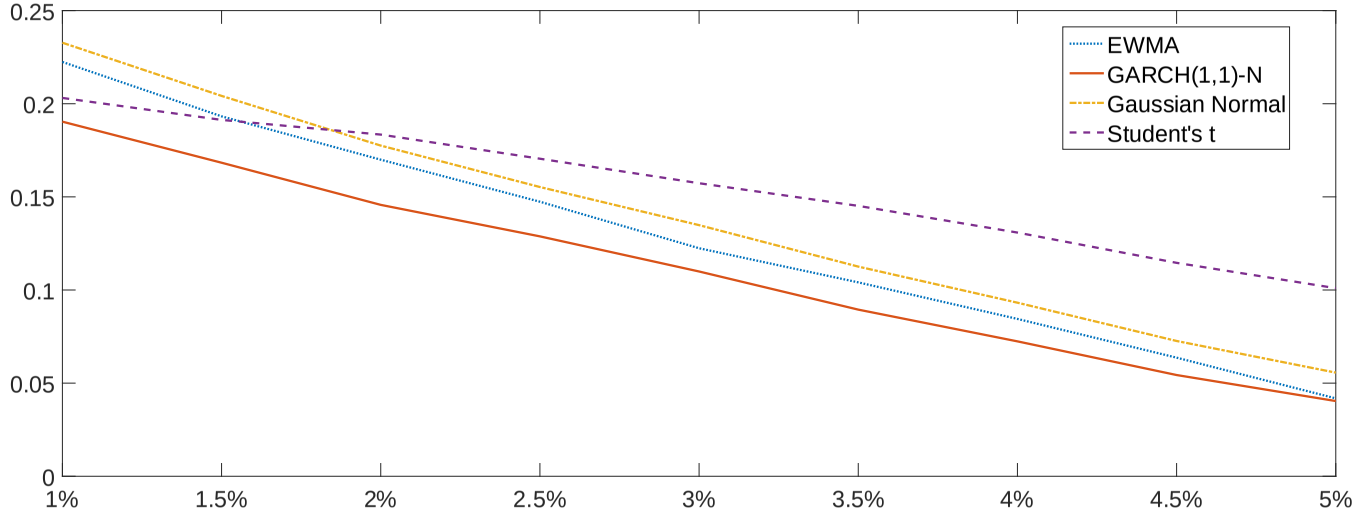
This table presents the mean values of the absolute and relative optimal correction, obtained by passing Z_2 test, made to daily ES ($\alpha = 2.5\%$), estimated by different models. Based on the DGP (GARCH(1,1) with standardised student's t *disturbances), we first simulated 5,000 series of 1,000 daily returns. Then ES estimates are obtained by using different methods with a rolling window of length 1,000. By passing the Z_2 test with a backtesting window of length 250, the optimal correction made to the daily ES are calculated. C_3 represents the optimal corrections made to ES forecasts required to pass the Z_2 test; C_3^* stands for the optimal corrections made to the corrected ES allowing for VaR model risk, required to pass the Z_2 test.*

01/01/1900 to 05/03/2017 (29,486 daily returns in total). Based on equation (2.3.1), we quantify the model risk of ES as the maximum of minimum corrections required to pass the ES backtests[12] and make comparisons among different models, where backtesting is performed over a year. Moreover, we examine this measure of model risk based on different asset classes by using the GARCH(1,1)-GPD model due to its best performance shown in the case of the DJIA index.

Figure 2.4.1 shows the relative corrections made to the daily ES, estimated at different significance levels, of four models: EWMA, GARCH(1,1)-N, Gaussian Normal, and Student's t, when considering the frequency of the exceptions (passing the UC_{ES} test). ES forecasts are computed with a four-year moving window and backtested using the entire sample. The level of relative corrections is decreasing when alpha is increasing, implying that the ES at a smaller significance level may need a larger correction to allow for model risk. Not surprisingly, the dynamic approaches, GARCH(1,1)-N and EWMA, require smaller corrections than the two static models in general, though the Student's t distribution performs better at capturing the fat tails than the EWMA model, for example, at 1% and 1.5% significance levels.

Figure 2.4.2 presents the optimal corrections made to the daily ES forecasts based on various forecasting models with regard to passing the unconditional coverage test for ES (UC_{ES} test), the conditional test for ES (CC_{ES} test) and the magnitude test (Z_2 test), respectively, where ES is estimated at a 2.5% significance level using a four-year moving window[13] and the evaluation period for backtesting

Figure 2.4.1: Relative correction for ES based on the UC test



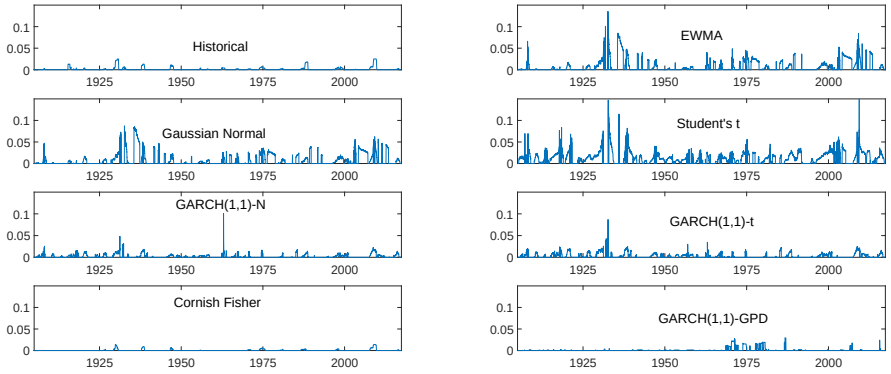
This figure shows the relative correction (computed as the ratio of the absolute correction to the average daily ES) based on the UC test made to the daily ES associated with EWMA, GARCH(1,1)-N, Gaussian Normal, and Student's t *along with a range of* α *levels.*

procedures is one year. This figure shows that a series of dynamic adjustments are needed for the daily ES ($\alpha = 2.5\%$) across all different models, especially during the crisis periods. This is in line with our expectation of model inadequacy in the crisis periods. The smaller the correction, the more accurate the ES estimates, therefore the less the model risk of the ES forecasting model. Among the models considered, the historical, EWMA, Gaussian Normal and Student's t models require larger corrections than the others when considering the three backtests jointly, indicating that they have higher model risk than the others. Particularly, the GARCH(1,1)-GPD performs the best. Also, the Cornish-Fisher expansion, GARCH(1,1)-GPD, and GARCH(1,1)-t models require the smallest adjustments in order to pass the UC_{ES}, CC_{ES}, and Z_2 tests, accordingly. Noticeably, the ES forecasts made by the non-GARCH models need larger corrections in order to pass the Z_2 test that refers to the size of the exceptions, compared with these corrections required by the UC_{ES} and CC_{ES} test particularly during the 2008 financial crisis. Thus, the GARCH(1,1) models are more able to capture the extreme losses, as expected.

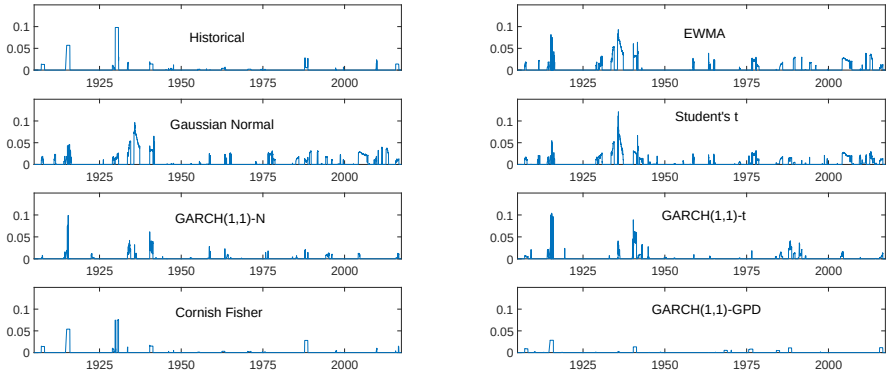
We present the time taken to arrive at the peak of the optimal corrections in Figure 2.4.3, for the UC_{ES}, CC_{ES} and Z_2 tests, which shows that more than a decade is needed to get the highest correction required to cover for model risk (also see Appendix 2.D, Table 2.D.1 for the dates when the highest corrections are required). When considering the UC_{ES} and CC_{ES} tests, the highest values of the optimal corrections made to the daily ES of various models are achieved before the

Figure 2.4.2: Dynamic optimal corrections for the daily ES

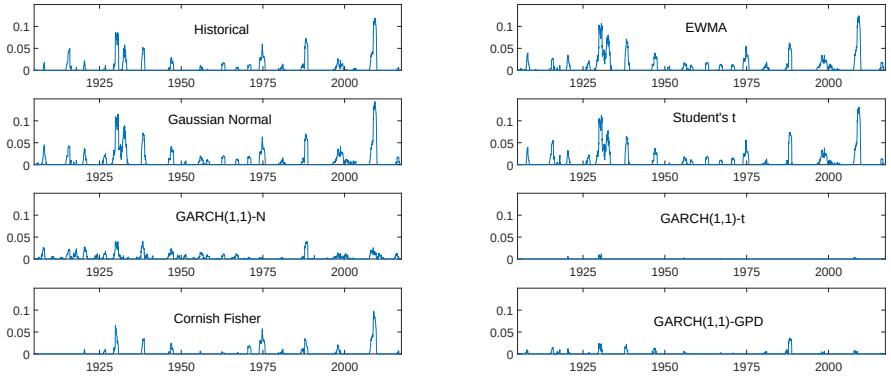
(a) Based on the UC_{ES} test



(b) Based on the CC_{ES} test



(c) Based on the Z_2 test



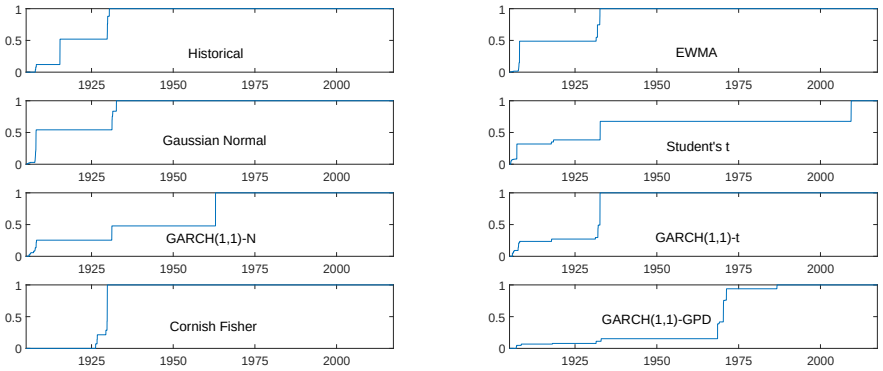
This figure shows dynamic optimal corrections made to the daily ES estimates ($\alpha = 2.5\%$) associated with various models for the DJIA index from 01/01/1900 to 23/05/2017, required to pass the UC_{ES} , CC_{ES} and Z_2 tests, respectively. The parameters are re-estimated using a four-year moving window (1,000 daily returns) and the evaluation window for backtesting is one year.

21st century (except that the highest value of the optimal corrections made to the Student's t ES is found around 2008, required to pass the UC_{ES} test), indicating that based on past mistakes we could have avoided the ES failures using these two tests, for instance, in the 2008 credit crisis. Nevertheless, when considering the three tests jointly, all the models, except for the GARCH models, find the peak values of the optimal corrections around 2008. Therefore, the GARCH models are more favorable than the others in avoiding model risk. This way, we could have been well prepared against the 2008 financial crisis if the GARCH(1,1) models were used to make ES forecasts. This is also supported by the results shown in Appendix 2.D, Figure 2.D.2, which presents extreme optimal corrections of ES forecasts based on different models, required to pass various backtests.

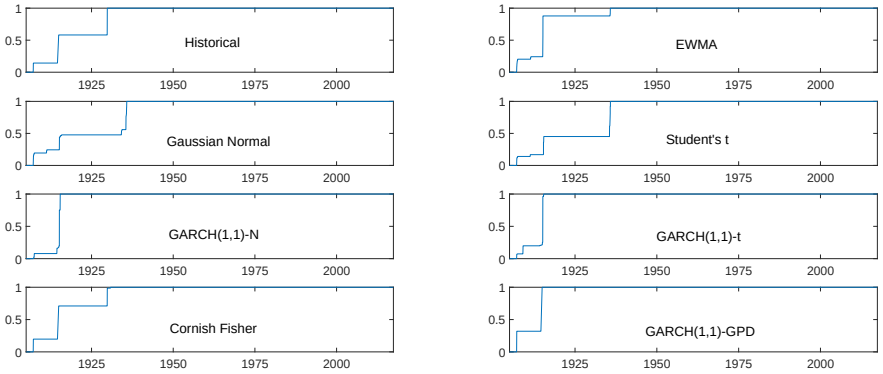
In Table 2.4.1, we measure the model risk of ES forecasts made by various risk models for the DJIA index, and compare the model risk of the 2.5% ES with that of the 1% VaR. Besides, we look into how ES model risk is affected by the model risk of VaR as discussed in section 2.3.4. Panel A and Panel B give the maximum and mean values of the absolute and relative optimal corrections to the daily ES ($\alpha = 2.5\%$) across various risk models with respect to the aforementioned three backtests and an alternative to the Z_2 test. The largest absolute corrections are needed for the Gaussian Normal and Student's t models, whilst the GARCH models perform well in capturing extreme losses. With the requirement of passing the three backtests jointly, the GARCH(1,1)-GPD performs best and requires a correction of 0.0011 made to the daily ES against model risk. We present

Figure 2.4.3: Ratio of dynamic optimal correction to the maximum optimal correction over the entire period

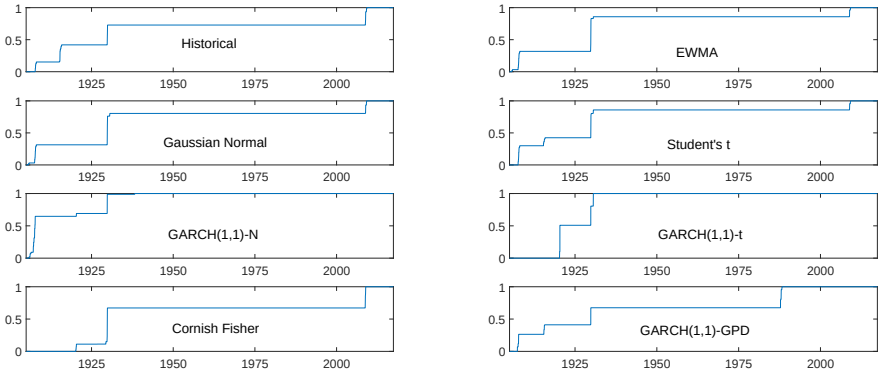
(a) Based on the UC_{ES} test



(b) Based on the CC_{ES} test



(c) Based on the Z_2 test



This figure shows the ratio of dynamic optimal correction over the maximum of the optimal corrections over the entire period, in which the optimal correction is made to the daily ES estimates ($\alpha = 2.5\%$) associated with various models by passing the UC_{ES}, CC_{ES}, Z_2 tests.

Table 2.4.1: Maximum and mean of optimal corrections for ES and VaR

Model	Mean ES (VaR)	Max C_1	Max C_2	Max C_3	Max C_4	Mean C_1	Mean C_2	Mean C_3	Mean C_4
Panel A: Maximum and mean of the absolute optimal corrections ($*10^{-2}$) to the daily ES (α= 2.5%)									
Historical	0.031	2.50	9.80	11.86	8.43	0.13	0.20	0.53	0.11
EWMA	0.024	13.55	9.30	12.41	5.55	0.69	0.37	0.74	0.56
Gaussian Normal	0.025	8.73	9.64	14.33	9.66	0.72	0.42	0.84	0.63
Student's t	0.030	21.84	12.12	13.15	9.14	1.13	0.38	0.73	0.19
GARCH(1,1)-N	0.023	10.11	9.90	4.08	4.79	0.20	**0.08**	0.33	0.30
GARCH(1,1)-t	0.031	8.69	10.41	**1.18**	**3.93**	0.29	0.15	**0.01**	0.10
Cornish-Fisher	0.050	**1.40**	7.60	9.75	22.94	**0.05**	0.14	0.29	0.09
GARCH(1,1)-GPD	0.028	2.95	**2.85**	3.60	4.09	0.11	**0.08**	0.09	**0.04**
Panel B: Maximum and mean of the relative optimal corrections to the daily ES (α= 2.5%)									
Historical	0.031	98.5%	319.0%	436.8%	274.4%	4.5%	6.1%	18.2%	3.9%
EWMA	0.024	318.8%	399.3%	537.5%	295.8%	26.0%	11.6%	30.7%	24.4%
Gaussian Normal	0.025	269.0%	214.3%	672.0%	420.9%	27.4%	13.4%	35.8%	27.5%
Student's t	0.030	479.8%	241.1%	480.8%	337.1%	39.6%	9.8%	25.5%	7.1%
GARCH(1,1)-N	0.023	560.4%	397.2%	133.7%	296.1%	8.4%	3.4%	13.4%	13.4%
GARCH(1,1)-t	0.031	155.0%	317.4%	**23.4%**	**162.0%**	8.7%	4.1%	**0.2%**	3.1%
Cornish-Fisher	0.050	**52.2%**	240.1%	339.0%	182.1%	**1.8%**	**2.2%**	9.8%	**1.5%**
GARCH(1,1)-GPD	0.028	157.7%	**134.4%**	121.4%	192.8%	5.8%	3.0%	2.5%	1.5%
Panel C: Maximum and mean of the relative optimal corrections to the daily VaR (α= 1%)									
Historical	0.030	78.2%	280.9%	213.0%	213.0%	2.9%	7.7%	22.6%	22.6%
EWMA	0.024	101.8%	297.8%	313.7%	313.7%	6.3%	10.8%	42.1%	42.1%
Gaussian Normal	0.024	139.4%	423.5%	305.5%	305.5%	7.3%	14.3%	41.7%	41.7%
Student's t	0.028	89.1%	366.2%	235.3%	235.3%	4.2%	10.0%	28.1%	28.1%
GARCH(1,1)-N	0.022	50.5%	298.1%	434.9%	434.9%	2.3%	6.5%	63.7%	63.7%
GARCH(1,1)-t	0.030	**7.1%**	**173.9%**	236.5%	236.5%	**0.0%**	**1.5%**	32.0%	32.0%
Cornish-Fisher	0.050	36.6%	180.1%	**105.4%**	**105.4%**	0.8%	2.4%	**12.6%**	**12.6%**
GARCH(1,1)-GPD	0.027	22.6%	204.9%	337.3%	337.3%	0.2%	2.5%	43.2%	43.2%
Panel D: Maximum and mean of the relative corrections to the daily ES, corrected for VaR model risk									
Historical	0.032	46.4%	248.6%	190.0%	213.8%	2.4%	5.6%	8.3%	4.0%
EWMA	0.026	68.5%	308.6%	229.1%	295.7%	4.5%	4.3%	15.3%	19.7%
Gaussian Normal	0.026	186.2%	203.1%	249.6%	293.4%	8.0%	4.6%	15.7%	20.9%
Student's t	0.032	165.2%	**132.2%**	208.2%	235.1%	8.1%	3.3%	10.7%	5.7%
GARCH(1,1)-N	0.023	189.4%	421.1%	119.8%	295.8%	6.0%	2.9%	9.4%	12.4%
GARCH(1,1)-t	0.031	171.3%	317.4%	**23.1%**	162.0%	**0.3%**	**2.3%**	**0.2%**	3.1%
Cornish-Fisher	0.052	**23.6%**	176.0%	121.2%	**105.9%**	1.1%	3.1%	4.2%	**1.3%**
GARCH(1,1)-GPD	0.028	147.7%	134.4%	99.8%	192.8%	4.2%	3.0%	2.0%	1.9%

This table presents the maximum and mean of the absolute and relative optimal corrections made to the daily 2.5% ES, the relative optimal corrections made to the daily 1% VaR, as well as the relative optimal corrections made to the corrected ES after VaR model risk is accounted for, using different backtests across various models, based on the DJIA index from 01/01/1900 to 23/05/2017, downloaded from DataStream. Based on various forecasting models, ES and VaR are forecasted with a four-year moving window (1,000 daily returns), and the mean ES and VaR are calculated over the entire sample. In Panel A, B, and D, C_1, C_2, C_3 and C_4 denote the optimal corrections made to the ES estimates, accordingly, required to pass the unconditional coverage test (UC_{ES} test), the conditional coverage test (CC_{ES} test), and the magnitude tests (Z_2 test and the exceedance residual test). In Panel C, C_1, C_2, and C_3 (C_4 is the same as C_3, to be consistent with other panels) represent the optimal corrections made to VaR forecasts, required to pass Kupiec's unconditional coverage test, Christoffersen's conditional coverage test and Berkowitz's magnitude test, respectively. The relative correction is the ratio of the optimal correction over the average daily ES (or VaR); backtesting is done over 250 days.

the relative corrections in Panel B, expressed as the optimal corrections over the average daily ES. When looking at the three backtests jointly, the EWMA, Gaussian Normal and Student's t models face the highest ES model risk with the mean values of the relative corrections at 30.7%, 35.8%, and 39.6%, respectively, thereby needing the largest buffers; whilst the GARCH(1,1)-GPD model has the best performance with a mean value of the relative optimal correction of 5.8%.

Applying the backtesting-based correction methodology to the 1% VaR as in Boucher et al. (2014)[14], we compute the relative corrections made to one-step ahead VaR forecasts by passing three VaR backtests[15], reported in Panel C of Table 2.4.1. The results show that the Cornish-Fisher expansion and GARCH(1,1)-t models outperform the other models, requiring the smallest corrections for VaR model risk. Comparing Panel B and Panel C, it can be seen that the peak values of the relative correction required to pass the UC_{VaR} and CC_{VaR} tests for VaR estimates are generally (with a few exceptions) smaller than the corresponding values for ES estimates, whilst the ES estimates require much smaller corrections than the VaR estimates when considering the Z_2 test or its alternative. That is, the ES measure is more able to measure the size of the extreme losses than the VaR measure, just as Colletaz et al. (2013) and Daníelsson and Zhou (2017) argue. When the three backtests are considered jointly, the 2.5% ES is less affected by model risk than the 1% VaR.

It is interesting to compare our results with those of Daníelsson and Zhou (2017). In their Table 1, they show that VaR estimation has a higher bias than ES

estimation, but a smaller standard error. However, this is based on a simulation study that focuses on estimation risk. The results presented in the empirical part of their paper somewhat contradict their theoretical expectation of VaR being superior to ES, and it can be argued that this is caused by the presence of specification error. So when only estimation error is considered, VaR is superior to ES, but when both estimation error and specification error are considered jointly, our results show that ES outperforms VaR, being less affected by model risk.

Supplementary to the backtesting-based correction methodology for ES, we examine the impact of VaR model risk on the model risk of ES in Panel D, Table 2.4.1. For all the models, the relative optimal corrections (shown in Panel D) required to pass the three ES backtests jointly, made to the daily ES after accommodating for VaR model risk, are smaller than the relative corrections (shown in Panel B) made to the daily ES when VaR is not corrected for model risk. Thus, ES is less affected by model risk, when VaR model risk is removed first. Roughly speaking, the corrections for model risk to the ES estimates reduce by about 50% if the VaR estimates are corrected for model risk. Also, we find further evidence in Table 2.D.3, Appendix 2.D to support the previous result that GARCH models are less affected by model risk, thus are preferred to make risk forecasts, when compared with the other models considered.

Additionally, we apply this proposed methodology to different asset classes (equity, bond and commodity from 31/10/1986 to 07/07/2017), as well as the FX (USD/GBP) and Microsoft (MSFT) shares (adjusted or non-adjusted for

dividends) from 01/01/1987 to 04/10/2017. Panel A and B of Table 2.4.2 report the absolute and relative corrections required for the GARCH(1,1)-GPD ES ($\alpha = 2.5\%$) of various asset classes[16]. The higher the corrections, the more unreliable the ES forecasts of the specified model for the data. We find that commodity ES carries the highest model risk with the highest mean value of the relative optimal correction at 5.2% required to pass the three tests jointly, provided that a GARCH(1,1)-GPD model is used. This is consistent with the statistical properties of the dataset considered, namely that commodity returns are fat-tailed and negatively skewed. Interestingly, in Table 2.D.2 of Appendix 2.D we find that commodity ES does not provide enough buffer against unfavorable extreme events in the global financial crisis, since the largest adjustments are needed in 2008 and 2009, suggesting that commodity ES suffers the highest model risk over the crisis period. However, equity and bond ES could have avoided the failures around 2008. Panel C shows the maximum and mean of the relative optimal corrections made to the 1% VaR, obtained by passing the three VaR backtests. Clearly, for the three different asset classes, the 1% VaR forecasts require much higher corrections than the 2.5% ES forecasts made by the GARCH(1,1)-GPD model, thereby carrying a higher model risk by considering the three backtests jointly as can be seen in the last column.

To get a further insight into the model risk of ES estimates of specific assets, we conduct a case study on the USD/GBP foreign currency and the MSFT stock (adjusted or non-adjusted for dividends) listed in the Nasdaq Stock Market. We

Table 2.4.2: Optimal corrections made to the GARCH(1,1)-GPD 2.5% ES and 1% VaR for different assets.

	Statistics of asset returns			Backtesting-based corrections						
Asset class	Std. dev	Skewness	Kurtosis	Mean ES	Max C_1	Max C_2	Max C_3	Mean C_1	Mean C_2	Mean C_3
Panel A: Maximum and mean of the absolute corrections ($*10^{-2}$) to the daily GARCH (1,1)-GPD ES ($\alpha = 2.5\%$)										
equity	0.012	-0.362	11.923	0.029	2.83	0.33	0.93	0.06	0.00	0.02
bond	0.003	0.017	7.400	0.007	0.33	0.04	0.34	0.01	0.00	0.01
commodity	0.004	-0.439	9.018	0.011	0.65	0.07	2.11	0.03	0.00	0.08
Panel B: Maximum and mean of the relative corrections to the daily GARCH (1,1)-GPD ES ($\alpha = 2.5\%$)										
equity	0.012	-0.362	11.923	0.029	97.0%	10.0%	37.5%	2.2%	0.1%	0.6%
bond	0.003	0.017	7.400	0.007	63.9%	5.5%	56.6%	1.0%	0.2%	1.8%
commodity	0.004	-0.439	9.018	0.011	95.2%	9.7%	123.8%	4.1%	0.4%	5.2%
Panel C: Maximum and mean of the relative corrections to the daily GARCH (1,1)-GPD VaR ($\alpha = 1\%$)										
equity	0.012	-0.362	11.923	0.029	3.6%	3.6%	177.9%	0.0%	0.0%	42.9%
bond	0.003	0.017	7.400	0.007	7.2%	15.6%	120.8%	0.1%	0.6%	31.7%
commodity	0.004	-0.439	9.018	0.010	15.1%	15.1%	235.3%	0.3%	0.6%	29.5%

This table presents the maximum and mean of the absolute and relative corrections made to the daily GARCH(1,1)-GPD ES (α = 2.5%), and the relative corrections made to the daily GARCH(1,1)-GPD VaR (α = 1%) for different asset classes based on different backtests. The empirical data is downloaded from DataStream, from 31/10/1986 to 07/07/2017. For the equity, we use a composite index with 95% "MSCI Europe Index" and 5% "MSCI World Index"; for the bond, we use the "Bank of America Merrill Lynch US Treasury & Agency Index"; for the commodity, we use the "CRB Spot Index". The average daily 2.5% ES (and 1% VaR) of various asset classes is computed based on the GARCH(1,1)-GPD model in a four-year rolling forecasting scheme. C_1, C_2 and C_3 represent the optimal corrections required to pass the UC_{ES}, CC_{ES} and Z_2 tests accordingly; backtesting is done over 250 days. The relative correction is the ratio of the optimal correction over the average daily ES (or VaR).

consider that ES is estimated at a significance level of 2.5%, and we have a position of 1 million dollars in each asset. Table 2.4.3 shows the dollar exposures to the model risk of the GARCH(1,1)-GPD ES when investing in the USD/GBP exchange rate or by purchasing the Microsoft stock, respectively. The average 2.5% ES of the FX and MSFT (adjusted) investments are \$14,291 and \$48,879, accordingly. The mean model risks, considering the three ES backtests jointly, are \$1,371 and \$1,350 for FX and MSFT (adjusted). It is inappropriate to consider a certain ES backtest, since the mean of the dollar exposures for FX with respect to different backtests varies from \$107 to \$1,371. Also, the non-adjusted MSFT equity has a much higher model risk than its counterparts, because the share prices shocked by dividend distributions are more volatile and therefore the risk model used is more vulnerable in this case. These examples show why it is necessary for banks to introduce enough protection against model risk when calculating the risk-based capital requirement introduced in Basel Committee on Banking Supervision (2011).

Our empirical analysis shows that, when forecasting ES, the GARCH(1,1) models are preferred, whilst the static models (e.g. the Gaussian Normal and Student's t models) and EWMA should be avoided. This is in contrast to the recommendations of Boucher et al. (2014) made for the model risk of VaR, namely that the EWMA VaR is preferred. Also, the 2.5% ES is the preferred measure of risk since it is less affected by model risk than the 1% VaR across different models or based on different assets, especially after VaR model risk is removed first. Using

Table 2.4.3: Dollar exposures to model risk of the GARCH(1,1)-GPD ES.

Asset	Mean ES	Max C_1	Max C_2	Max C_3	Mean C_1	Mean C_2	Mean C_3
FX USD/GBP	14,291	11,100	3,300	8,700	1,371	107	152
MSFT (adjusted)	48,879	106,400	19,800	62,200	212	646	1,350
MSFT (non-adjusted)	65,200	2,500	3,500	34,700	6	129	3,168

The table presents dollar exposures to the model risk of GARCH(1,1)-GPD ES (α = 2.5%) of the USD/GBP exchange rate and Microsoft equity, based on various ES backtests. The USD/GBP spot rate and MSFT share prices from 01/01/1987 to 04/10/2017 are downloaded from DataStream and Bloomberg, respectively. All the outcomes are in dollar units, computed by using a four-year moving window and a one-year backtesting period, based on the GARCH(1,1)-GPD model. C_1, C_2 and C_3 represent the dollar values of the optimal corrections required to pass the UC_{ES}, CC_{ES} and Z_2 tests accordingly, when considering a position of 1 million dollars in the asset specified in the first column.

the GARCH(1,1)-GPD model to make ES forecasts of various asset classes, we find that commodity ES carries the highest model risk especially around 2008, compared to equity and bond ES.

2.5 Conclusions

In this chapter, we propose a practical method to quantify ES model risk based on ES backtests. Model risk is considered as an optimal correction required to pass several ES backtests jointly. These ES backtests are tailored to the following characteristics of ES forecasts: 1) the frequency of exceptions; 2) the absence of autocorrelations in exceptions; 3) the magnitude of exceptions. We theoretically examine the desirable properties of model risk from a regulatory perspective. Considering the UC_{ES} and CC_{ES} tests for our chosen measure of model risk,

all the desirable properties hold, whilst subadditivity is not guaranteed and our results show that it is generally satisfied by well-fitting models.

We compare the 2.5% ES with the 1% VaR in terms of model risk across different models and based on different assets. We find that the 2.5% ES is less affected by model risk than the 1% VaR, needing a smaller correction to pass the three ES backtests jointly. Besides, commodity ES carries the highest model risk especially around 2008, compared to equity and bond ES, provided that the GARCH(1,1)-GPD model is used. Moreover, we consider the impact of VaR model risk on ES model risk in terms of the ES calculations and the ES backtests. If VaR model risk is first removed, then ES model risk reduces further by approximately 50%.

Our results are strengthened when the standard deviations of the corrections for model risk are considered: the GARCH(1,1) models not only require the smallest corrections for model risk, but the level of the corrections are the most stable, when compared to the other models considered in our study.

Appendices

2.A Theoretical analysis of estimation and specification errors of VaR

Estimation bias and correction of VaR

Based on equation (2.2.5) and (2.2.6), correcting for the estimation error, the VaR estimate can be written as:

$$VaR^{E}(\hat{\theta}_0, \alpha) = VaR(\hat{\theta}_0, \alpha) + \mathbb{E}(bias(\theta_0, \hat{\theta}_0, \alpha)) \tag{2.A.1}$$

This tells us that the mean bias of the forecasted VaR from the theoretical VaR is caused by estimation error.

Specification and estimation biases and correction of VaR

Based on equation (2.2.10) and (2.2.11), correcting for these biases (specification and estimation biases), the VaR estimate can be written as:

$$VaR^{SE}(\hat{\theta}_1, \alpha) = VaR(\hat{\theta}_1, \alpha) + \mathbb{E}(bias(\theta_0, \theta_1, \hat{\theta}_1, \alpha)) \tag{2.A.2}$$

The mean of the estimation and specification biases for VaR can be considered as a measurement of economic value of the model risk of VaR.

2.B Backtesting measures of VaR and ES

Table 2.B.1: Selected backtesting methodologies for VaR and ES

VaR backtests	ES backtests
Exception Frequency Tests:	***Exception Frequency Tests:***
(1)UC_{VaR} test- Kupiec (1995)	(1)UC_{ES} test- Du and Escanciano (2016)
(2)data-driven- Escanciano and Pei (2012)	
	(2)risk map- Colletaz et al. (2013)
	(3)traffic light- Moldenhauer and Pitera (2019)
Exception Independence Tests:	***Exception Independence Tests:***
(1)independence test-Christoffersen (1998)	
(2)density test- Berkowitz (2001)	
Exception Frequency and Independence Tests:	***Exception Frequency and Independence Tests:***
(1)CC_{VaR} test- Christoffersen (1998)	(1)CC_{ES} test- Du and Escanciano (2016); Costanzino and Curran (2015, 2018)
(2)dynamic quantile-Engle and Manganelli (2004);Patton et al. (2019)	(2)dynamic quantile- Patton et al. (2019)
(3)multilevel test- Campbell (2006)	
(4)multilevel test-Leccadito et al. (2014)	
(5)multinomial test-Kratz et al. (2018)	(3)multinomial test-Kratz et al. (2018); Emmer et al. (2015); Clift et al. (2016)
(6)two-stage test- Angelidis and Degiannakis (2006)	
Exception Duration Tests:	***Exception Duration Tests:***
(1)duration test- Christoffersen and Pelletier (2004)	
(2)duration-based test- Berkowitz et al. (2011)	
(3)GMM duration-based test- Candelon et al. (2010)	
Exception Magnitude Tests:	***Exception Magnitude Tests:***
(1)tail losses- Wong (2010)	(1)tail losses- Wong (2008); Christoffersen (2009); McNeil and Frey (2000)
(2)magnitude test-Berkowitz (2001)	
Exception Frequency and Magnitude Tests:	***Exception Frequency and Magnitude Tests:***
(1)risk map- Colletaz et al. (2013)	
(2)quantile regression- Gaglianone et al. (2011)	
	(1)Z_2 test-Acerbi and Szekely (2014)

2.C Risk forecasting models

In the following, we focus on several commonly discussed models for computing one-step ahead VaR and ES forecasts (Christoffersen, 2012) using a rolling window of length τ at a significance level α.

Historical Simulation

Among all the models considered in this chapter, Historical Simulation[17] is the simplest and easiest to implement, in which the forecasting of risk estimates is model free, based on past return data. VaR is computed as the empirical α-quantile ($\hat{Q}(\cdot)$) of the observed returns $X_t, X_{t+1}, ..., X_{t+\tau-1}$, and its formulation is given below

$$\widehat{VaR}_{t+\tau}^{\alpha} = -\hat{Q}_{\alpha}(X_t, X_{t+1}, ..., X_{t+\tau-1}). \tag{2.C.1}$$

ES is the expected value of the returns in the tail, and it is computed as

$$\widehat{ES}_{t+\tau}^{\alpha} = -\frac{\sum_{i=t}^{i=t+\tau-1} X_i I_{\{X_i<-\widehat{VaR}_{t+\tau}^{\alpha}\}}}{\sum_{i=t}^{i=t+\tau-1} I_{\{X_i<-\widehat{VaR}_{t+\tau}^{\alpha}\}}}, \tag{2.C.2}$$

where $I(\cdot)$ is equal to 1 when the empirical return is smaller than the negative value of VaR, otherwise 0.

Gaussian Normal distribution

Simply assuming that the observed returns follow a normal distribution, the one-step ahead return is $\hat{r}_{t+\tau} = \hat{\mu}_{t+\tau} + \hat{\sigma}_{t+\tau}\Phi_{\alpha}^{-1}$, where $\hat{\mu}_{t+\tau}$ and $\hat{\sigma}_{t+\tau}^2$ are mean and variance of the previous τ observations $X_t, X_{t+1}, ..., X_{t+\tau-1}$, and Φ denotes the cumulative distribution function of the standard normal distribution. In this case, we compute $VaR_{t+\tau}^{\alpha}$ as

$$\widehat{VaR}_{t+\tau}^{\alpha} = -\hat{\mu}_{t+\tau} - \hat{\sigma}_{t+\tau}\Phi_{\alpha}^{-1}. \tag{2.C.3}$$

ES can be derived as

$$\widehat{ES}_{t+\tau}^{\alpha} = -\hat{\mu}_{t+\tau} + \hat{\sigma}_{t+\tau}\frac{\phi\left(\Phi_{\alpha}^{-1}\right)}{\alpha}, \tag{2.C.4}$$

where ϕ denotes the density function of the standard normal distribution.

Student's *t* distribution

Here, we consider a symmetric Student's t, capturing the fatter tails and the more peak in the distribution of the standardised returns as compared with the normal case. Let X denote a Student's t variable with the pdf defined as below:

$$f_{t(d)}(x;d) = \frac{\Gamma((d+1)/2)}{\Gamma(d/2)\sqrt{d\pi}}(1 + x^2/d)^{-(1+d)/2}, \quad for\ d > 2, \tag{2.C.5}$$

where $\Gamma(\cdot)$ is the gamma function and d is the degree of freedom larger than 2. The one-step ahead return is $\hat{r}_{t+\tau} = \hat{\mu}_{t+\tau} + \hat{\sigma}_{t+\tau} t_{\alpha}^{-1}(\hat{d})$, where $t_{\alpha}^{-1}(\hat{d})$ refers to the empirical α-quantile of the standardised returns following a Student's t distribution with estimated parameter $\hat{d}$. VaR can therefore be computed as

$$\widehat{VaR}_{t+\tau}^{\alpha} = -\hat{\mu}_{t+\tau} - \hat{\sigma}_{t+\tau} t_{\alpha}^{-1}(\hat{d}). \tag{2.C.6}$$

ES is given by

$$\widehat{ES}_{t+\tau}^{\alpha} = -\hat{\mu}_{t+\tau} + \hat{\sigma}_{t+\tau} \frac{f_{t(\hat{d})}\left(t_{\alpha}^{-1}(\hat{d})\right)}{\alpha}, \tag{2.C.7}$$

where $\hat{\mu}_{t+\tau}$ and $\hat{\sigma}_{t+\tau}^2$ are mean and variance of the previous τ observations.

GARCH models

The Gaussian Normal and Student's t distributions are fully parametric approaches and belong to the location-scale family with the general expression for the returns $\hat{r}_{t+\tau} = \hat{\mu}_{t+\tau} + \hat{\sigma}_{t+\tau} z_{t+\tau}$, where the mean $\mu_{t+\tau}$ and standard deviation $\sigma_{t+\tau}$ are the location and scale parameters, respectively. $z_{t+\tau}$ is the empirical quantile of the assumed distribution of the standardised returns such as the standard normal distribution in the normal case. The GARCH models play a crucial role in the location-scale family with time-varying conditional variances and a modeled distribution for the standardised residuals, thus being considered dynamic approaches, as opposed to the static models (the Guassian Normal and Student's t distributions). Considering GARCH(1,1) models with the normal or

Student's t disturbances (GARCH(1,1)-N or GARCH(1,1)-t), the time-varying conditional variance is written as

$$\hat{\sigma}^2_{t+\tau} = \omega + \alpha X^2_{t+\tau-1} + \beta \hat{\sigma}^2_{t+\tau-1} \tag{2.C.8}$$

Within the estimation window $t, t+1, ..., t+\tau$, the model parameters $(\mu, \omega, \alpha, \beta; d)$ are estimated via maximum likelihood estimation with the constraints: $\omega, \alpha, \beta > 0$, $\alpha+\beta < 1$, and $d > 2$. For GARCH(1,1)-N, the formulae for computing VaR and ES are the same as equation (2.C.3) and (2.C.4). We can refer to equation (2.C.6) and (2.C.7) to make VaR and ES forecasts using the GARCH(1,1)-t model.

Exponentially Weighted Moving Average

The exponentially weighted moving average method (EWMA) is a special case of the GARCH(1,1) model with normal disturbances, as the conditional variance is expressed as

$$\hat{\sigma}^2_{t+\tau} = (1-\lambda) X^2_{t+\tau-1} + \lambda \hat{\sigma}^2_{t+\tau-1}, \quad \lambda = 0.94. \tag{2.C.9}$$

VaR and ES are computed as in equations (2.C.3) and (2.C.4).

GARCH with Extreme Value Theory

The advantage of extreme value theory is to model the tail distribution, thereby it focuses on the extreme values in the tail. In this chapter, we use the GARCH(1,1) model with standardised t disturbances, combined with the EVT methodol-

ogy (GARCH(1,1)-GPD). First, we obtain the standardised empirical losses via GARCH(1,1), assuming they are distributed as a standardised t distribution.

$$X_{t+\tau} = \hat{\sigma}_{t+\tau} St^{-1}(d), \quad \hat{\sigma}^2_{t+\tau} = \omega + \alpha X^2_{t+\tau-1} + \beta \hat{\sigma}^2_{t+\tau-1}, \tag{2.C.10}$$

where $St^{-1}(d)$ denotes the inverse of the cumulative distribution function of a standardised t distribution with its pdf expressed as

$$f_{\tilde{t}(d)}(\tilde{x}; d) = C(d)(1 + \tilde{x}^2/(d-2))^{-(1+d)/2}, \quad for\ d > 2, \tag{2.C.11}$$

where

$$C(d) = \frac{\Gamma((d+1)/2)}{\Gamma(d/2)\sqrt{\pi(d-2)}}. \tag{2.C.12}$$

$\tilde{x}$ is a standardised random variable distributed as a standardised t distribution with mean 0, variance 1 and degree of freedom larger than 2. Then we fit Generalized Pareto Distribution (GPD) to excesses y over the given threshold u, where

$$GPD(y; \xi, \beta) = \begin{cases} 1 - (1 + \xi y/\beta)^{-1/\xi}, & \text{if } \xi > 0 \\ 1 - \exp(-y/\beta), & \text{if } \xi = 0 \end{cases} \tag{2.C.13}$$

with $\beta > 0$ and $y \geq u$. The tail index parameter ξ controls the shape of the tail. When ξ is positive, the tail distribution is fat-tailed. Consequently, in this

approach VaR could be computed as:

$$\widehat{VaR}_{t+\tau}^{\alpha} = \hat{\sigma}_{t+\tau} VaR_z(\alpha), \tag{2.C.14}$$

where

$$VaR_z(\alpha) = \left(u + \frac{\hat{\beta}}{\hat{\xi}} \left(\left(\frac{\alpha}{k/n} \right)^{-\hat{\xi}} - 1 \right) \right) \tag{2.C.15}$$

with k the number of peaks over the threshold and n the total number of standardised empirical observations. ES is given by

$$\widehat{ES}_{t+\tau}^{\alpha} = \hat{\sigma}_{t+\tau} ES_z(\alpha), \tag{2.C.16}$$

where

$$ES_z(\alpha) = VaR_z(\alpha) \left(\frac{1}{1-\hat{\xi}} + \frac{(\hat{\beta} - \hat{\xi} u)}{(1-\hat{\xi}) VaR_z(\alpha)} \right). \tag{2.C.17}$$

Cornish-Fisher expansion

The Cornish-Fisher expansion (Christoffersen, 2012) allows for skewness and kurtosis to make VaR and ES forecasts by using the sample moments without any assumption on the returns.

$$\widehat{VaR}_{t+\tau}^{\alpha} = -\hat{\sigma}_{t+\tau} CF_{\alpha}^{-1} \tag{2.C.18}$$

where $\hat{\sigma}^2_{t+\tau}$ is the variance of the previous τ observations, and CF_α^{-1} is expressed below:

$$CF_\alpha^{-1} = \Phi_\alpha^{-1} + \frac{\hat{\zeta}_1}{6}\left[(\Phi_\alpha^{-1})^2 - 1\right] + \frac{\hat{\zeta}_2}{24}\left[(\Phi_\alpha^{-1})^3 - 3\Phi_\alpha^{-1}\right] - \frac{\hat{\zeta}_1^2}{36}\left[2(\Phi_\alpha^{-1})^3 - 5\Phi_\alpha^{-1}\right] \tag{2.C.19}$$

ES is formulated as

$$\widehat{ES}^{\alpha}_{t+\tau} = -\hat{\sigma}_{t+\tau} ES_{CF(\alpha)} \tag{2.C.20}$$

where

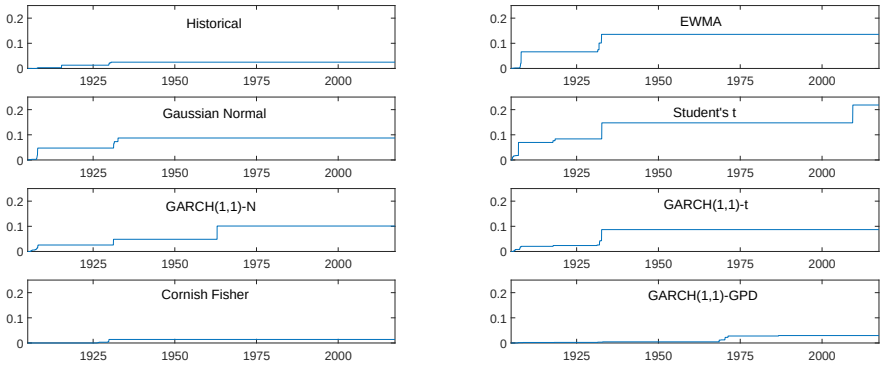
$$ES_{CF(\alpha)} = \frac{-\phi(CF_\alpha^{-1})}{\alpha}\left[1 + \frac{\hat{\zeta}_1}{6}(CF_\alpha^{-1})^3 + \frac{\hat{\zeta}_2}{24}\left[(CF_\alpha^{-1})^4 - 2(CF_\alpha^{-1})^2 - 1\right]\right] \tag{2.C.21}$$

$\hat{\zeta}_1$ and $\hat{\zeta}_2$ represent the skewness and excess kurtosis of the standardised returns, calculated based on the past τ observations.

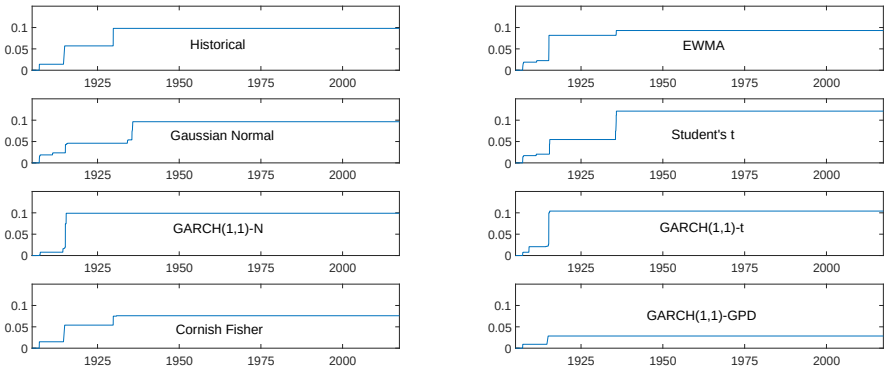
2.D Empirical results

Figure 2.D.1: Historical maximum of optimal adjustments for ES estimates

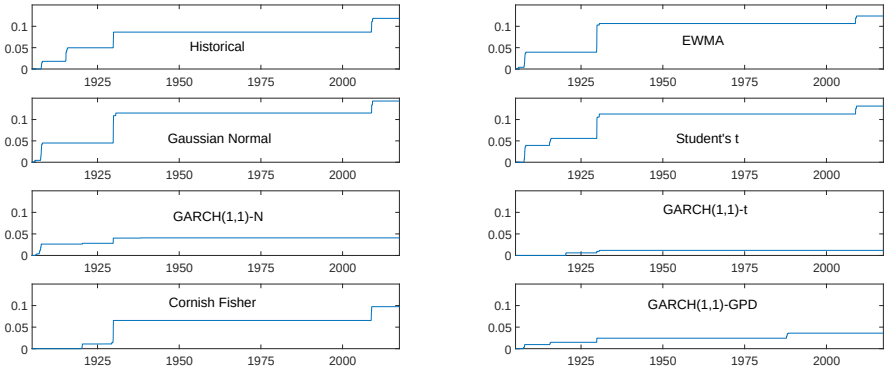
(a) Based on the UC_{ES} test



(b) Based on the CC_{ES} test



(c) Based on the Z_2 test



This figure presents the historical maximum of required optimal adjustments made to the daily ES estimates ($\alpha = 2.5\%$) of various models for the DJIA index from 01/01/1900 to 23/05/2017, obtained by passing the UC_{ES}, CC_{ES} and Z_2 tests, respectively.

Table 2.D.1: Dates associated with the highest values of the absolute minimum corrections made to the daily ES of various models

		UC_{ES} test		CC_{ES} test		Z_2 test	
Model		Date	C_1	Date	C_2	Date	C_3
Historical	1	16/06/1930	0.0250	29/10/1929	0.0980	20/04/2009	0.1186
	2	11/09/2009	0.0240	14/12/1914	0.0570	30/03/2009	0.1176
	3	20/11/2008	0.0230	30/10/1930	0.0300	05/03/2009	0.1172
	4	12/12/1929	0.0220	13/12/1915	0.0280	19/05/2009	0.1167
EWMA	1	15/08/1932	0.1355	15/10/1935	0.0930	20/04/2009	0.1241
	2	08/08/1932	0.1196	18/10/1935	0.0898	05/03/2009	0.1238
	3	09/11/1931	0.1010	17/10/1935	0.0897	30/03/2009	0.1229
	4	22/06/1931	0.0744	16/10/1935	0.0893	05/05/2009	0.1225
Gaussian Normal	1	17/08/1932	0.0873	15/10/1935	0.0964	20/04/2009	0.1433
	2	13/09/1935	0.0861	18/10/1935	0.0927	05/03/2009	0.1431
	3	12/09/1935	0.0859	17/10/1935	0.0925	30/03/2009	0.1421
	4	16/09/1935	0.0850	16/10/1935	0.0921	05/05/2009	0.1418
Student's t	1	29/05/2009	0.2184	25/10/1935	0.1212	05/03/2009	0.1315
	2	15/09/1932	0.1475	04/10/1935	0.1118	20/04/2009	0.1308
	3	11/10/1932	0.1324	28/10/1935	0.1041	30/03/2009	0.1300
	4	08/09/1932	0.1206	29/10/1935	0.1005	02/03/2009	0.1299
GARCH(1,1)-N	1	14/12/1962	0.1011	02/06/1915	0.0990	29/03/1938	0.0408
	2	19/12/1962	0.0990	10/06/1915	0.0775	29/10/1929	0.0403
	3	27/03/1931	0.0484	01/03/1915	0.0744	14/04/1988	0.0397
	4	26/03/1931	0.0471	02/03/1915	0.0721	08/08/1930	0.0396
GARCH(1,1)-t	1	24/08/1932	0.0869	08/06/1915	0.1041	08/08/1930	0.0118
	2	25/08/1932	0.0854	25/05/1915	0.1022	28/10/1928	0.0095
	3	26/08/1932	0.0812	03/03/1915	0.1002	12/12/1929	0.0086
	4	02/02/1932	0.0427	09/06/1915	0.0999	21/07/1930	0.0084
Cornish-Fisher	1	06/11/1929	0.0140	28/10/1930	0.0760	01/12/2008	0.0975
	2	29/10/1929	0.0130	29/10/1929	0.0750	08/12/2008	0.0951
	3	10/02/1930	0.0120	14/12/1914	0.0540	29/12/2008	0.0933
	4	28/10/1929	0.0110	19/10/1987	0.0280	20/11/2008	0.0915
GARCH(1,1)-GPD	1	24/09/1986	0.0295	14/12/1914	0.0285	14/04/1988	0.0360
	2	26/09/1986	0.0294	07/05/1915	0.0284	25/03/1988	0.0358
	3	23/09/1986	0.0293	15/12/1914	0.0283	08/01/1988	0.0344
	4	21/11/1986	0.0292	14/05/1940	0.0132	10/03/1988	0.0343

This table presents the dates associated with the highest values of the absolute minimum corrections made to the daily 2.5% ES of various models based on different ES backtests. The calculations are based on the DJIA index daily returns from the 1st January 1900 to the 23rd May 2017, downloaded from DataStream. We make the 2.5% one-step ahead ES forecasts based on various models with a four-year moving window and backtest ES estimates in the evaluation period of 250 days. C_1, C_2 and C_3 denote the optimal corrections required to pass the unconditional coverage test (UC_{ES} test), the conditional coverage test (CC_{ES} test) and the magnitude test (Z_2 test), respectively.

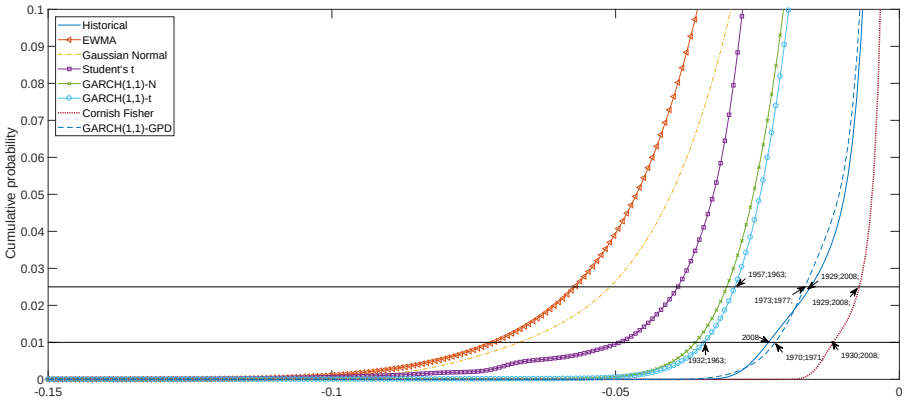
Table 2.D.2: Dates associated with the highest values of the absolute minimum corrections made to the GARCH(1,1)-GPD ES for different assets

		UC_{ES} test		CC_{ES} test		Z_2 test	
Asset		Dates	C_1	Dates	C_2	Dates	C_3
equity	1	30/10/2001	0.0283	27/08/2002	0.0033	21/01/2008	0.0093
	2	26/10/2001	0.0282	05/09/2002	0.0028	12/02/2008	0.0063
	3	22/10/2001	0.0281	19/09/2002	0.0027	10/10/2008	0.0057
bond	1	05/07/2013	0.0033	14/05/1999	0.0004	05/08/1994	0.0034
	2	01/08/2013	0.0027	21/04/1995	0.0001	16/09/1994	0.0033
	3	09/08/2013	0.0026	15/08/1991	0.0000	06/05/1994	0.0032
commodity	1	30/04/1993	0.0065	20/12/1994	0.0007	17/02/2009	0.0211
	2	28/04/1993	0.0064	19/12/1994	0.0005	20/02/2009	0.0198
	3	26/04/1993	0.0063	07/03/2008	0.0004	19/11/2008	0.0190

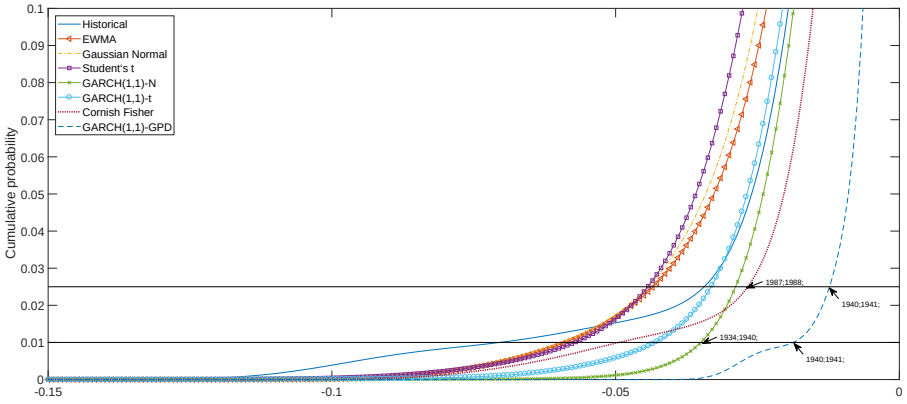
This table presents the dates regarding the highest values of the absolute minimum corrections made to the GARCH(1,1)-GPD ES ($\alpha = 2.5\%$) for different assets by passing different ES backtests. The empirical data is downloaded from DataStream. For the equity, we use a composite index with 95% "MSCI Europe Index" and 5% "MSCI World Index"; for the bond, we use the "Bank of America Merrill Lynch US Treasury & Agency Index"; for the commodity, we use the "CRB Spot Index", from 31/10/1986 to 07/07/2017. We compute the GARCH(1,1)-GPD ES of different assets at a 2.5% coverage level by using a four-year moving window and backtest ES estimates in the evaluation period of 250 days. The variables C_1, C_2 and C_3 denote the optimal corrections required to pass the unconditional coverage test (UC_{ES} test), the conditional coverage test (CC_{ES} test) and the magnitude test (Z_2 test), respectively.

Figure 2.D.2: Left tail of the cumulative distribution of the negative of required optimal adjustments made to the daily ES estimates

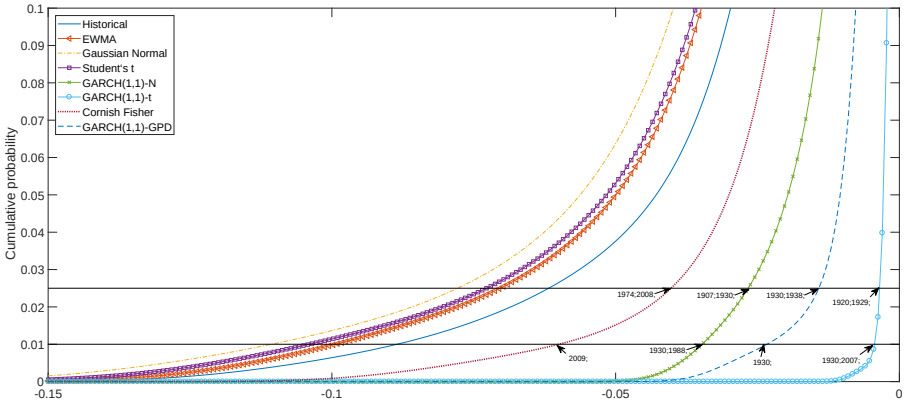
(a) Based on the UC_{ES} test



(b) Based on the CC_{ES} test



(c) Based on the Z_2 test



This figure shows the left tail of the cumulative distribution (using Gaussian Kernel smoothing) of the negative of required optimal adjustments made to the daily ES estimates ($\alpha = 2.5\%$) for the DJIA index from 01/01/1900 to 23/05/2017, in order to pass the UC_{ES} (panel a), CC_{ES} (panel b), and Z_2 (panel c) tests, respectively.

Table 2.D.3: Means and standard deviations of optimal corrections for the 2.5% ES and 1% VaR

Model	Mean C_1	Mean C_2	Mean C_3	Std. dev of C_1	Std. dev of C_2	Std. dev of C_3
Panel A: Means ($*10^{-2}$) and standard deviations of the absolute optimal corrections made to the daily ES ($\alpha = 2.5\%$).						
Historical	0.13	0.20	0.53	0.0039	0.0108	0.0157
EWMA(λ=0.94)	0.69	0.37	0.74	0.0133	0.0108	0.0179
Gaussian Normal	0.72	0.42	0.84	0.0135	0.0111	0.0200
Student's t	1.13	0.38	0.73	0.0125	0.0098	0.0186
GARCH(1,1)-N	0.20	0.08	0.33	0.0039	0.0038	0.0067
GARCH(1,1)-t	0.29	0.15	0.01	0.0051	0.0063	0.0006
Cornish-Fisher	0.05	0.14	0.29	0.0019	0.0076	0.0104
GARCH(1,1)-GPD	0.11	0.08	0.09	0.0039	0.0035	0.0038
Panel B: Means and standard deviations of the relative optimal corrections made to the daily ES ($\alpha = 2.5\%$).						
Historical	4.5%	6.1%	18.2%	0.1215	0.3050	0.5010
EWMA(λ=0.94)	26.0%	11.6%	30.7%	0.4263	0.3034	0.6769
Gaussian Normal	27.4%	13.4%	35.8%	0.4339	0.3095	0.7991
Student's t	39.6%	9.8%	25.5%	0.3823	0.2167	0.5933
GARCH(1,1)-N	8.4%	3.4%	13.4%	0.1530	0.1471	0.2415
GARCH(1,1)-t	8.7%	4.1%	0.2%	0.1430	0.1556	0.0138
Cornish-Fisher	1.8%	2.2%	9.8%	0.0586	0.1085	0.3373
GARCH(1,1)-GPD	5.8%	3.0%	2.5%	0.2087	0.1169	0.0952
Panel C: Means and standard deviations of the relative optimal corrections made to the daily VaR ($\alpha = 1\%$), by passing VaR backtests.						
Historical	2.9%	7.7%	22.6%	0.0978	0.3168	0.3425
EWMA	6.3%	10.8%	42.1%	0.1565	0.3065	0.5226
Gaussian Normal	7.3%	14.3%	41.7%	0.1830	0.4392	0.5100
Student's t	4.2%	10.0%	28.1%	0.1275	0.3822	0.3974
GARCH(1,1)-N	2.3%	6.5%	63.7%	0.0601	0.2271	0.7828
GARCH(1,1)-t	0.0%	1.5%	32.0%	0.0019	0.1134	0.4904
Cornish-Fisher	0.8%	2.4%	12.6%	0.0366	0.0989	0.2040
GARCH(1,1)-GPD	0.2%	2.5%	43.2%	0.0155	0.1461	0.6180
Panel D: Means and standard deviations of the relative optimal corrections made to the daily ES ($\alpha = 2.5\%$), after VaR model risk is first removed.						
Historical	2.4%	5.6%	8.3%	0.0648	0.2495	0.2437
EWMA	4.5%	4.3%	15.3%	0.1029	0.2460	0.3306
Gaussian Normal	8.0%	4.6%	15.7%	0.1835	0.1801	0.3545
Student's t	8.1%	3.3%	10.7%	0.1879	0.1183	0.2834
GARCH(1,1)-N	6.0%	2.9%	9.4%	0.1142	0.1479	0.1834
GARCH(1,1)-t	0.3%	2.3%	0.2%	0.0323	0.1349	0.0133
Cornish-Fisher	1.1%	3.1%	4.2%	0.0317	0.0965	0.1462
GARCH(1,1)-GPD	4.2%	3.0%	2.0%	0.1736	0.1167	0.0750

This table presents means and standard deviations of the absolute and relative corrections made to the daily 2.5% ES, the relative corrections made to the daily 1% VaR, and the relative corrections required for the 2.5% ES after VaR model risk is excluded first, based on the UC_{ES}, CC_{ES} and Z_2 backtests. The calculations are based on the DJIA index from 01/01/1900 to 23/05/2017, downloaded from DataStream.

2.E Simulated Bias

Similar to Table 2.2.1, we conduct a simulation study to show the impacts of estimation and specification biases on the ES forecasts in Table 2.E.1. Assuming a different data generating process, Markov Switching with 2 regimes combined with GARCH(1,1) with normal innovations (denoted by MS(2)-GARCH(1,1)-N) introduced by Klaassen (2002), we simulate 1000 paths of 1000 daily returns, thus computing the theoretical ES forecasts. The specification of the data generating process for the daily returns is given as below:

$$r_t = \sqrt{h_{s_t}} Z_t, \quad Z_t \sim \text{IID}\mathcal{N}(0,1), \quad s_t = \{1,2\}, \tag{2.E.1}$$

s_t denotes the possible states of the market at time t, 1 and 2, in which the conditional variance dynamics follow a GARCH(1,1) process and are specified as:

$$h_{s_t} = \omega_{s_t} + \alpha_{s_t} r_{t-1}^2 + \beta_{s_t} \sum_{i=1}^{2} p_{ij} h_{i,t-1}, \tag{2.E.2}$$

where p_{ij} represents the probability of state j at time t conditional that the market is in state i at time *t-1*, and $h_{i,t-1}$ is the conditional variance in state i at time $t-1$. The constraints on the parameters are $\omega_{s_t}, \alpha_{s_t}$ and $\beta_{s_t} > 0$ in order to ensure the positivity of the variance dynamics. The results are based on the DJIA index from 03/01/2000 to 30/12/2011, the estimated parameters are $\omega_1 = 1.1198e^{-04}$, $\alpha_1 = 0.0025$, and $\beta_1 = 0.9152$; $\omega_2 = 8.2761e^{-07}$, $\alpha_2 = 0.0677$, $\beta_2 = 0.9152$ with

the probabilities $p_{11} = 0.7726$ and $p_{22} = 0.9938$. We run simulations using these parameters and make one-step ahead ES forecasts as equation (2.C.4) for the simulated data series using the MS(2)-GARCH(1,1)-N model, historical method, Gaussian Normal distribution as well as the EWMA model, thereby giving the corresponding estimation and specification biases in Table 2.E.1.

Table 2.E.1: Simulated bias associated with the ES estimates

Significance level	Mean estimated ES(%)	Theoretical ES(%)	Mean bias(%)	Std. err of bias(%)
Panel A. MS(2)-GARCH(1,1)-N DGP with estimated MS(2)-GARCH(1,1)-N ES: estimation bias				
α=5%	37.22	37.08	-0.15	6.90
α=2.5%	42.19	42.02	-0.17	7.82
α=1%	48.10	47.91	-0.19	8.92
Panel B. MS(2)-GARCH(1,1)-N DGP with historical ES: specification and estimation biases				
α=5%	48.21	37.08	-11.13	21.15
α=2.5%	58.76	42.02	-16.74	24.63
α=1%	72.15	47.91	-24.24	29.05
Panel C. MS(2)-GARCH(1,1)-N DGP with Gaussian Normal ES: specification and estimation biases				
α=5%	42.62	37.08	-5.55	20.42
α=2.5%	48.31	42.02	-6.29	23.14
α=1%	55.07	47.91	-7.17	26.37
Panel D. MS(2)-GARCH(1,1)-N DGP with EWMA ES: specification and estimation biases				
α=5%	42.46	37.08	-5.39	19.23
α=2.5%	48.12	42.02	-6.10	21.78
α=1%	54.86	47.91	-6.96	24.83

This table presents the bias between the simulated theoretical ES and the estimated ES, based on the DJIA index from 01/01/1900 to 23/05/2017, downloaded from DataStream. First, we simulate 1,000 paths of 1,000 daily returns according to the DGP of MS(2)-GARCH(1,1)-N. Then we forecast ES based on the MS(2)-GARCH(1,1)-N, historical, Gaussian Normal and EWMA ($\lambda = 0.94$) specifications, for α = 5%, 2.5% and 1%.

Notes

[1]Alternatives are Median Shortfall (So and Wong, 2012), and expectiles (Bellini and Bignozzi, 2015).

[2]Estimation risk refers to the uncertainty of parameter estimates. Misspecification risk is the risk associated with inappropriate assumptions of the risk model, whilst identification risk refers to the risk that future sources of risk are not currently known and included in the model.

[3]When it comes to backtesting risk estimates, Escanciano and Olmo (2010a), in their Theorem 1 of the first paper, show how estimation risk and specification risk (which they call model risk) affect the test statistic (S_p) of the unconditional coverage backtest for VaR:

$$S_p = \frac{1}{\sqrt{P}} \sum_{t=R+1}^{n} \left[I_{t,\alpha}(\theta_0) - F_{W_{t-1}}(m_\alpha(W_{t-1},\theta_0)) \right] + \underbrace{\mathbb{E}\left[g'_\alpha(W_{t-1},\theta_0) f_{W_{t-1}}(m_\alpha(W_{t-1},\theta_0)) \right] \frac{1}{\sqrt{P}} \sum_{t=R+1}^{n} H(t-1)}_{\text{Estimation risk}} + \underbrace{\frac{1}{\sqrt{P}} \sum_{t=R+1}^{n} \left[F_{W_{t-1}}(m_\alpha(W_{t-1},\theta_0)) - \alpha \right]}_{\text{Model risk}} + o_P(1)\,.$$

[4]Rather than calibrating model risk based on statistical significance testing, assessing model risk concerning the space of possible models is of prominent importance in the Bayesian model averaging literature. Brock et al. (2003, 2007) study the role of model risk in policy evaluation and propose the model averaging method. However, this technique is difficult to use in the applications we have in mind, since it requires the specification of prior probabilities over the model space. Additionally, risk assessment of a particular model typically calls for quantification of risk by means of a single number representing the required capital reserve.

[5]Similar characteristics of a desirable VaR estimate are considered by Boucher et al. (2014).

[6]The values of VaR and ES are considered positive in this chapter.

[7]We also consider a different model, MS(2)-GARCH(1,1)-N, as the data generating process, and give simulated biases in Table 2.E.1, Appendix 2.E.

[8]The parameters of GARCH(1,1)-N estimated from the DJIA index (1st Jan 1900 to 23rd May 2017) are : $\mu = 4.4521e^{-04}$; $\omega = 1.3269e^{-06}$; $\alpha = 0.0891$; and $\beta = 0.9017$.

[9] we use the p-value = 0.05 in this chapter. For different $p-$values, the results are essentially similar to those presented in this chapter.

[10]The critical value related to the 5% significance level for the Z_2 test is -0.7, which is stable for different distribution types (Acerbi and Szekely, 2014).

[11]To find the optimal correction of VaR accommodating for model risk, two VaR backtests are considered. The VaR backtests are Kupiec' s unconditional coverage test (Kupiec, 1995), and Christoffersen's conditional coverage test (Christoffersen, 1998). We do not include Berkowitz's magnitude test (Berkowitz, 2001), because in principle it is very similar to the magnitude test for ES (it checks the size of exceptions).

[12]The UC_{ES} and CC_{ES} tests for all the distribution-based ES are examined in the setting proposed by Du and Escanciano (2016), whilst the Cornish-Fisher expansion and the historical method are entertained in the same setting but in a more general way. ES for the asymmetric and fat-tailed distributions (Broda and Paolella, 2011) can also be examined using these backtests.

[13]The results computed using a five-year moving window and a three-year moving window are very similar to those required here (available from the authors on request).

[14]Boucher et al. (2014) only present the results for the 5% VaR.

[15]The three VaR backtests are Kupiec's unconditional coverage test (Kupiec, 1995), Christoffersen's conditional coverage test (Christoffersen, 1998) and Berkowitz's magnitude test (Berkowitz, 2001).

[16]See the data source in the note to Table 2.4.2.

[17]Other varieties of Historical Simulation, such as Filtered Historical Simulation, are found in (Christoffersen, 2012).

Chapter 3

Scoring Function-Based Model Risk of Risk Models

3.1 Introduction

Managing financial risk is paramount to corporate companies. The measurement of different types of risk is required to satisfy investors and regulators. The most used statistical risk measures, Value-at-Risk (VaR) and Expected Shortfall (ES)[1], are of particular interest in assessing market risk which refers to the risk arising from a change in the value of a financial position due to the unexpected price movements of primary risk factors such as stock prices, commodity prices or interest rates. As required by the Basel Committee on Banking Supervision (2019), market risk should be measured by ES which is defined as the average loss beyond the VaR threshold.

Nevertheless, the statistical computation of these risk measures not only depends on model choice, meaning that the VaR and ES measures are subject to model risk, but the true risk is not observable ex-post, so it is challenging to perform backtesting. Hence, the decisions taken by managers may be impaired by model risk and accounting for this additional risk is requested by the Federal Reserve and the European Banking Authority. Financial companies tend to be conservative in managing model risk by adding an extra capital buffer, irrespective of the value of model risk. To this end, quantifying the model risk of the risk estimation methods and incorporating it into the regulatory capital have become nontrivial and significant in the advance of model risk management.

In this chapter we propose an improved methodology to measure and compare the two main model risk components, parameter estimation risk and model specification risk, of market risk models by analyzing the pair (VaR, ES) based on the FZ scoring functions introduced by Fissler and Ziegel (2016). We first show that in the presence of model risk, the ordering of competing (VaR, ES) models is sensitive to the choice of FZ scoring function. Secondly, we put forward a general FZ scoring function-based model risk computation methodology to estimate the joint (VaR, ES) model risk and the ES model risk of a certain model, at a given significance level. Thirdly, in a simulation study, we verify the above proposed measurement of the joint (VaR, ES) model risk and, separately, of the ES model risk alone, by using several specific FZ scoring functions which are positively homogeneous of degree 0, 0.5 and -1. Lastly, we highlight that our

proposed scoring function-based measures of joint model risk and ES model risk satisfy all the desirable coherence properties except for the subadditivity property which is not always satisfied.

Jorion (1996) signals early on the existence of risk in estimating VaR. With the increasingly intensive use of complex risk models, the concern among academics[2] about model risk has grown after the global financial crisis in 2008, and it has prompted a line of research in the accuracy of risk models. The performance of VaR models has been investigated and then further improved in several strands of recent studies: 1) the quantification of model risk of a given market risk model developed around a reference model (see e.g., Kerkhof et al., 2010; Lönnbark, 2013; Glasserman and Xu, 2014; Barrieu and Scandolo, 2015; Daníelsson et al., 2016); 2) the computation of model risk based on numerical algorithms like the bootstrapping technique, leading to more computational burden (see e.g., Christoffersen and Gonçalves, 2005); 3) the calculation of model uncertainty associated with the backtesting procedures for which Escanciano and Olmo (2010a,b) proposed robust test statistics allowing for parameter estimation risk (also see the estimation bias correction of Pitera and Schmidt, 2018) and model misspecification risk. Furthermore, Boucher et al. (2014) suggest a correction to VaR estimates required to pass several backtests tailored to some criteria; considering the accuracy of ES models, Lazar and Zhang (2019) develop a similar methodology to adjust ES estimates that would pass certain ES backtests. Although these studies take the model uncertainty of VaR or ES models into account and compute a

backtesting-based correction for risk forecasts, their approaches do not quantify model risk numerically as such.

In addition, the subject of modeling and backtesting ES has generated a lot of interest recently, Acerbi and Szekely (2014), Colletaz et al. (2013), Du and Escanciano (2016), Emmer et al. (2015), Fissler et al. (2016), Kratz et al. (2018), and Kellner and Rösch (2016) being major contributions to this topic. Since the estimation of ES is often a by-product of the VaR estimation procedure, referred to the more recent literature (e.g., Patton et al., 2019), the model risk of the ES is closely linked to that of VaR at a given significance level. Hence, we are motivated to measure directly the magnitude of model risk of joint (VaR, ES) forecasts at a certain significance level.

Market risk models may carry three sources of model risk (Kerkhof et al., 2010 and Boucher et al., 2014): 1) misspecification error, arising when the model is misspecified; 2) estimation risk, occurring due to the inaccurate parameter estimation for the model; 3) non-nested information sets of two different models leading to identification problems, when not all the information is detected and considered for forecasting. The current scoring function literature documents that scoring functions work well to estimate model parameters for financial risk models (Patton et al., 2019) and to rank the predictive performance of competing models (e.g., Ehm et al., 2016 and Nolde and Ziegel, 2017a). Patton (2019) links forecast evaluation to specific sources of model risk, arguing that since VaR models may be impacted by misspecification error, estimation error and nonnested information

sets (that is, identification risk), the ranking of VaR models may be sensitive to the choice of the generalized piecewise linear (GPL) scoring function which is strictly consistent for VaR. Motivated by Patton (2019), we bridge the gap between the scoring function literature and model risk literature, proposing a methodology to estimate the model risk of the pair (VaR, ES) forecasts, based on the FZ scoring functions discussed in Fissler and Ziegel (2016).

The coherence properties that a risk measure should satisfy as introduced by Artzner et al. (1999) are important from a regulatory perspective. Here we analyze the coherence properties of our scoring function-based model risk estimation methodology via simulations. In particular, the subadditivity property, which has been a major concern of the VaR measure and the main theoretical advantage of the ES measure (Garcia et al., 2007 and Daníelsson et al., 2013), is revisited for the model risk measure proposed in this chapter.

The structure of Chapter 3 continues as follows. Section 3.2 is focused on the sensitivity of ranking (VaR, ES) models to the choice of the FZ scoring function in relation to the major sources of model risk. Section 3.3 proposes an FZ scoring function-based model risk measure of (VaR, ES) risk measures, illustrating its effectiveness via simulations and Section 3.4 examines its properties in a realistic simulation study. Section 3.5 applies our proposed model risk measure to a set of real-world financial data and Section 3.6 concludes.

3.2 Model risk in relation to scoring functions

3.2.1 Scoring functions

We start with some background information and introduce notations that we follow from Nolde and Ziegel (2017a). A risk measure ρ is defined on some space of random variables, for example, a random variable R taking values in an observation domain $B \subseteq \mathbb{R}$. F_R denotes the cumulative distribution function of the return R assumed to have a finite mean. A series of risk measure estimates $\Theta_1(R), ..., \Theta_T(R)$ take values in an action domain A $\subseteq \mathbb{R}^k$, where $\Theta_i(R) = (\rho_1(R), ..., \rho_k(R))$ is a k-dimensional vector of risk measures, for $i = 1, \ldots, T$. The emphasis in our study being on VaR and ES measures (that is, $k = 2$), let v_α denote the VaR measure and e_α for the ES measure at a given significance level $\alpha \in (0, 1)$, such as $\alpha = 2.5\%$, recommended by the Basel Committee on Banking Supervision (2019). VaR and ES at an α critical level are computed as:

$$v_\alpha(F) = \inf\{r \in \mathbb{R} : F_R(r) \geq \alpha\}, \quad e_\alpha(F) = \frac{1}{\alpha}\int_0^\alpha v_u(F)du. \tag{3.2.1}$$

Hence, v_α and e_α have negative values, following the sign convention of Ziegel et al. (2020). Without loss of generality, we shall omit henceforth the subscript α from v_α and e_α.

Definition 1. A scoring function[3] is a map $S : A \times B \to \mathbb{R}$. For a given family of probability measures $\mathcal{P}$, the scoring function S is considered *consistent*

for the vector of risk measure(s), $\Theta(R)$, with respect to the class $\mathcal{P}$, if for all $Y = (Y_1, ..., Y_k)$, any R and all $P \in \mathcal{P}$:

$$\mathbb{E}_P\left[\mathrm{S}\left(\Theta(\mathrm{R}), \mathrm{R}\right)\right] \leq \mathbb{E}_P\left[\mathrm{S}\left(Y, \mathrm{R}\right)\right]$$

When there is no equality in the above condition for all $Y \neq \Theta(R)$, S is called *strictly consistent* for the vector of risk measures $\Theta(R)$ which are called *elicitable*.

Gneiting (2011) proves that VaR is elicitable, since it can be uniquely obtained by minimizing the expected score given by the GPL scoring function which is strictly consistent for VaR, but at the same time ES is not elicitable (see Ziegel, 2016). However, Fissler and Ziegel (2016) argue that VaR and ES are jointly elicitable under the assumption that the conditional distributions of returns are continuous, and formally provide a class of scoring functions strictly consistent for this pair of risk functionals $\Theta = [v, e]$. For a critical level α, considering two increasing continuously differentiable functions G_1 and $G_2 = \mathcal{G}_2'$ such that $\mathbb{E}[G_1(z)]$ exists and $\lim_{z\to-\infty} G_2(z) = 0$, and a realization of return denoted by r, the class of strictly consistent scoring functions[4] for the pair of risk measures (v, e) is given below (hereafter, FZ scoring functions):

$$\begin{aligned} S_{FZ}(r, v, e; \alpha, G_1, G_2) \;=\; & \left(\mathbb{1}_{\{r\leq v\}} - \alpha\right)\left(G_1(v) - G_1(r)\right) \\ & + G_2(e)\left(\frac{1}{\alpha}\mathbb{1}_{\{r\leq v\}}(v - r) - (v - e)\right) - \left(\mathcal{G}_2(e) - \mathcal{G}_2(r)\right) \end{aligned} \tag{3.2.2}$$

Definition 2. A scoring function S within the FZ family is called *positively homogeneous* of some order $b \in \mathbb{R}$ if for all $v = (v_1, ..., v_n)$, $e = (e_1, ..., e_n)$ and all r

$$S(\lambda r, \lambda v, \lambda e; \alpha, G_1, G_2) = \lambda^b S(r, v, e; \alpha, G_1, G_2), \text{ for all } \lambda > 0.$$

Upon rescaling the data, positive homogeneity will ensure that the same parameter estimates are derived, or the same orderings of models are obtained, using the same form of scoring function (see details in Efron, 1991 and Patton, 2011). This is a desirable feature for forecast ranking (Patton, 2011). Nolde and Ziegel (2017a) streamline the full class of FZ family in (3.A.2) such that the resulting scoring differences are positively homogeneous of degree b. This is equivalent to:

$$\begin{aligned}
&\text{if } b < 0: && G_1(z) = -c_0, && \mathcal{G}_2(z) = c_1(-z)^b + c_0,\\
&\text{if } b = 0: && G_1(z) = d_0 \mathbb{1}_{\{z \le 0\}} + d_0' \mathbb{1}_{\{z > 0\}}, && \mathcal{G}_2(z) = -c_1 \log(-z) + c_0,\\
&\text{and if } b \in (0,1): && G_1(z) = (d_1' \mathbb{1}_{\{z > 0\}} - d_1 \mathbb{1}_{\{z \le 0\}})|z|^b + c_0, && \mathcal{G}_2(z) = -c_1(-z)^b + c_0;
\end{aligned} \tag{3.2.3}$$

for constants[5] $c_0, d_0, d_0' \in \mathbb{R}$ with $d_0 \le d_0'$, $d_1, d_1' \ge 0$ and $c_1 > 0$. For $b \ge 1$, there is no positively homogeneous scoring function. The computational assumption that ES is strictly negative is being used. Throughout this chapter, we use the notation "$FZ0$" coined by Patton et al. (2019) for the 0-homogeneous case: if

and only if $G_1(z) = 0$ and $G_2(z) = -1/z$ in (3.A.2), it can be written as:

$$S_{FZ0}(r, v, e; \alpha) = -\frac{1}{\alpha e}\mathbb{1}_{\{r \leq v\}}(v - r) + \frac{v}{e} + \log(-e) - 1.$$

3.2.2 Model risk in relation to the FZ class

In his seminal paper, Patton (2019) investigates the sensitivity of ranking risk models to different scoring functions, making three assumptions with respect to identification risk, estimation risk as well as misspecification risk, accordingly: 1) the information sets of the forecasters are nested, so $\mathcal{F}_t^B \subseteq \mathcal{F}_t^A$ or $\mathcal{F}_t^A \subseteq \mathcal{F}_t^B$ for all t, and they do not lead to identical optimal forecasts for all t; 2) if the forecasts are based on models, then the models are free from estimation error; and 3) if the forecasts are based on models, then the models are correctly specified for the statistical functional(s) of interest. One of his major findings is that, if any of the assumptions above is not satisfied by the VaR models (in other words, the VaR risk measure models come with identification risk, parameter estimation risk or misspecification risk), then the ordering of VaR models may be sensitive to the choice of scoring function within the GPL class. Inspired by this, we draw a connection between the FZ scoring functions for the pair (VaR, ES) at an α critical level and different sources of model risk in the following proposition, similar to Proposition 4 in Patton (2019):

Proposition 1 *(a) Under the aforementioned three assumptions, the ranking of*

two risk models (A, B) by comparing the expected score of $FZ0$ scoring function, $S_{FZ0,A}$ and $S_{FZ0,B}$, is sufficient for their ranking by any (strictly consistent) FZ scoring function S defined in (3.A.2)*. That is, for all $S \in S_{FZ}$ and the pair (VaR, ES) measures estimated at an α significance level,*

$$\mathbb{E}\left[S_{FZ0,B}\right] \lesseqgtr \mathbb{E}\left[S_{FZ0,A}\right] \Longrightarrow \mathbb{E}\left[S(r_t, \hat{v}^B_{\alpha,t}, \hat{e}^B_{\alpha,t})\right] \lesseqgtr \mathbb{E}\left[S(r_t, \hat{v}^A_{\alpha,t}, \hat{e}^A_{\alpha,t})\right]. \quad (3.2.4)$$

(b) If any of Assumptions 1,2 and 3 fails to hold, then the ranking of these two risk models may be sensitive to the choice of the FZ scoring function.

Proposition 1*(a)* shows that conditioning on the absence of model risk will warrant that the ranking of risk models by the $FZ0$ scoring function is consistent with the ordering given by any other FZ scoring function; Proposition 1*(b)* states that if model risk is present, the ranking of risk estimation procedures may be affected by the choice of (strictly consistent) scoring function. Proofs adapted from Patton (2019) are given in the supplemental appendix.

3.2.3 Sensitivity of model ranking to the FZ class in the presence of model risk

Regarding Proposition 1*(b)*, we provide three examples showing that the ranking of risk models is sensitive to the choice of FZ scoring function when model risk is present. Specifically, we compute the average loss (score) difference between two competing models. This technique is widely accepted, see Gneiting (2011), Nolde

and Ziegel (2017a) and Patton (2019). For instance, if model A has a smaller expected loss than model B, implying a negative average loss difference between model A and model B, then model A dominates model B. Nolde and Ziegel (2017a) find that model comparison based on the expected score of a given FZ scoring function should be made on a sample large enough in order to reduce the effect of the data on the stability of ranking competing models. As in Nolde and Ziegel (2017b), we use a window length of 2,000 to compute the expected score (of FZ class) in the following simulation and empirical study so that the quality of risk measures could be evaluated without outliers' effect. Our calculations confirm that 2,000 data points are sufficient to achieve stability (for the sample average score to converge to the true mean score).

(i) First, consider the case characterized by the presence of identification risk when non-nested information sets are applied to two competing risk models, based on the positively homogeneous FZ class specified in (3.2.3). We first simulate 10,000 daily stock returns according to the AR(1)-GARCH(1,1) model specified below:

$$r_t = \mu_t + \sigma_t \varepsilon_t, \quad \varepsilon_t \sim iid\ N(0,1), \tag{3.2.5}$$

$$\text{where} \quad \mu_t = 0.03 + 0.05 r_{t-1}, \quad \sigma_t^2 = 0.05 + 0.88\sigma_{t-1}^2 + 0.05\sigma_{t-1}^2\varepsilon_{t-1}^2.$$

Next, we compute daily VaR and ES estimates at an α significance level, free of estimation and misspecification risk, based on the non-nested information sets

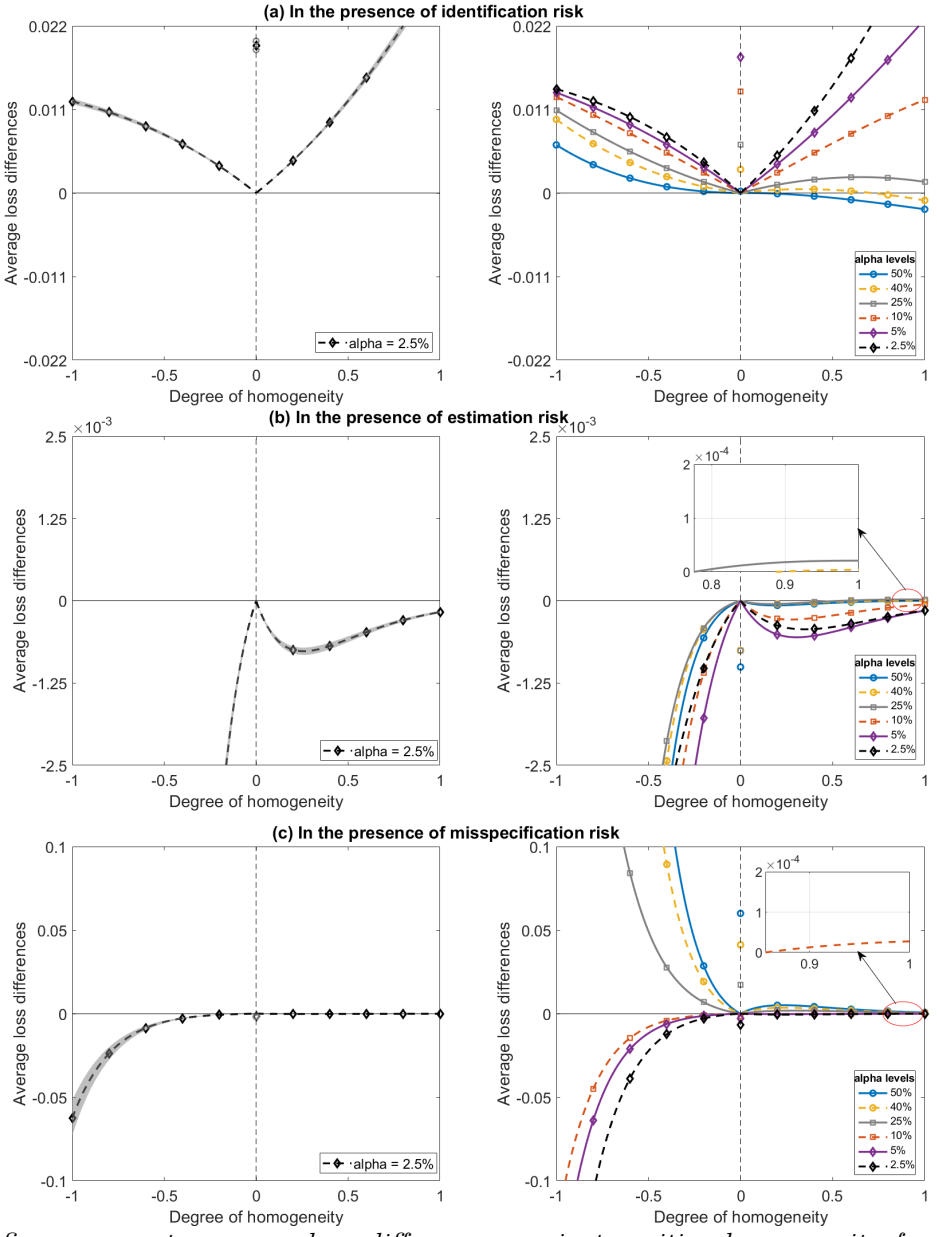
(thus leading to identification risk):

$$\hat{v}_t^A = \mu_t + \bar{\sigma}\Phi^{-1}(\alpha), \qquad \hat{e}_t^A = \mu_t + \frac{\bar{\sigma}}{\alpha}\int_0^\alpha \Phi^{-1}(x)dx; \tag{3.2.6}$$
$$\hat{v}_t^B = \bar{\mu} + \sigma_t\Phi^{-1}(\alpha), \qquad \hat{e}_t^B = \bar{\mu} + \frac{\sigma_t}{\alpha}\int_0^\alpha \Phi^{-1}(x)dx;$$

where the unconditional mean and volatility are $\bar{\mu} = 0.0316$ and $\bar{\sigma} = 0.8452$, respectively, given the known parameters in (3.2.5). The first pair of risk functional only utilizes the information on the conditional mean (μ_t) associated with the AR(1) process, referred to as the mean forecast A; while the second pair only employs the variance (σ_t^2) associated with the GARCH(1,1) process, referred to as the volatility forecast B.

The right figure of panel (a) in Figure 3.2.1 shows average FZ loss differences between the mean forecast and volatility forecast along with the degree of homogeneity, [-1,1), when one-step ahead VaR and ES measures are computed using a rolling window of length 1,000 at multiple significance levels based on the simulated data. The positive loss differences suggest that the volatility forecast fits the simulated data better than the mean forecast, indicating that at more extreme α levels, the volatility forecast performs better in capturing the variations of the data than the mean forecast. The average loss differences can be negative for less extreme α levels (for example, $\alpha = 50\%$), showing that the mean forecast is favorable in making risk forecasts. This is in line with Figure 5 in Patton (2019). The switching sign of the average loss differences for different FZ scoring functions is

Figure 3.2.1: Sensitivity of model ranking to the choice of scoring function



This figure presents average loss differences against positive homogeneity from -1 to 1. Panel (a): loss differences between the mean forecast and the volatility forecast in (3.2.6)*; panel (b): loss differences between the estimations based on estimation windows of 1,000 and 500; panel (c): loss differences between NEVT and NFHS. In all panels: the left figure presents average loss differences with 95% confidence intervals in grey for daily risk measures at* $\alpha = 2.5\%$ *when 1,000 paths of 2,000 returns are simulated, and the right figure presents average loss differences for daily risk measures at multiple levels in a simulated path of 10,000 (panel a) or 2,000 (panel b and c) returns.*

clearly identified where VaR and ES are estimated at $\alpha = 40\%$ and 50%. This is consistent with Proposition 1*(b)*, arguing that the ordering of competing risk models is affected by the choice of the FZ scoring function. Nevertheless, at the regulatory coverage level of $\alpha = 2.5\%$ for VaR and ES, the ranking of these two models based on various FZ scoring functions, though subject to identification risk, is still consistent as shown in the right figure of panel (a), which may support the use of $FZ0$ scoring function for forecast evaluation. In order to exclude the effect of noise of the simulated data on switching signs, the left figure of panel (a) presents the average loss differences within the 95% confidence intervals in grey, based on simulating 1,000 paths of 2,000 daily returns.

(ii) Secondly, consider the sensitivity of ranking two risk models affected by parameter estimation risk, in our case when the length of estimation windows varies, to the choice of scoring function. We choose a simple GARCH(1,1) process with Student's t innovations to simulate financial returns, using the model parameters in Kratz et al. (2018) who fitted a t-GARCH(1,1) model on the daily log-returns of S&P500 index between 2000-2012. The model is written as:

$$\begin{aligned} r_t &= \sigma_t Z_t, \quad Z_t \sim iid\ St(5.06), \\ \text{where} \quad \sigma_t^2 &= 2.18 \times 10^{-6} + 0.109 r_{t-1}^2 + 0.890 \sigma_{t-1}^2, \end{aligned} \tag{3.2.7}$$

$\{Z_t\}_{t\in\mathbb{N}}$ is an i.i.d sequence of Student's t distributed random variables with degrees of freedom equal to 5.06. Then we use the same GARCH(1,1) specification

with Student's t disturbances to compute the daily VaR and ES estimates at multiple levels with different parameter estimation windows, L_1=1,000 and L_2=500, where $L = \{L_1, L_2\}$:

$$\hat{v}_t^L = \hat{\sigma} St_\alpha^{-1}(\hat{d}), \qquad \hat{e}_t^L = \frac{\hat{\sigma}}{\alpha} \cdot f_{St(\hat{d})}\left(St_\alpha^{-1}(\hat{d})\right). \tag{3.2.8}$$

The values of average loss differences between the estimations based on the longer estimation window and the shorter one are mostly negative for all the significance levels considered, and they converge to zero as the degree of homogeneity increases, occasionally turning positive, as can be seen from the right figure of panel (b) in Figure 3.2.1, in which a path of 2,000 returns is generated by the DGP in (3.2.7). On the left figure of panel (b) we simulate 1,000 paths of 2,000 returns, compute the daily risk estimates at $\alpha = 2.5\%$, and generate the average loss differences with the 95% confidence intervals in grey. Panel (b) provides evidence that the choice of scoring function may have an effect on the ranking of the (VaR, ES) models subject to estimation risk.

(iii) Thirdly, we show that the ranking of two misspecified risk models (with parameter estimation risk present) may be affected by the choice of FZ scoring function. In this case, we implement the same DGP as in (3.2.7). In order to compute the daily VaR and ES measures at a certain significance level, we fit a GARCH(1,1) model with normal innovations to the simulated return data and then obtain the standardized residuals which tend to be fat-tailed, using a

rolling window scheme with a window length of 1,000. Subsequently, we employ different risk estimation models on the standardized residuals. First, we apply the generalized Pareto distribution (GPD) parameter estimation procedure under the extreme value theory, developed by McNeil and Frey (2000), to model the tail distribution of these residuals, with the threshold chosen as the 12% quantile as in Nolde and Ziegel (2017a). This is denoted by NEVT. Specifically, we fit the distribution of exceedances (y) beyond the threshold (u) with $GPD(y;\xi,\beta)$ shown as below, where ξ and β are the shape and scale parameters with $\beta > 0$, respectively:

$$GPD(y;\xi,\beta) = \begin{cases} 1-(1+\xi y/\beta)^{-1/\xi}, & \text{if } \xi > 0, \\ 1-\exp(-y/\beta), & \text{if } \xi = 0; \end{cases} \qquad for\ all\ y \geq u.$$

Then for a given significance level, the pair of risk estimates (VaR, ES) obtained with NEVT is analytically given by:

$$\hat{v}_t^{EVT} = -\hat{\sigma} v_{evt}(\alpha), \qquad \hat{e}_t^{EVT} = -\hat{\sigma} e_{evt}(\alpha),$$

$$\text{where} \quad v_{evt}(\alpha) = \left(u + \frac{\hat{\beta}}{\hat{\xi}} \left(\left(\frac{\alpha}{k/\tau} \right)^{-\hat{\xi}} - 1 \right) \right),$$

$$\text{and} \quad e_{evt}(\alpha) = v_{evt}(\alpha) \cdot \left(\frac{1}{1-\hat{\xi}} + \frac{(\hat{\beta} - \hat{\xi}u)}{(1-\hat{\xi})v_{evt}(\alpha)} \right).$$

In the above formulae, k is the number of exceedances and τ represents the total number of standardized empirical observations. All model parameters are estimated using a rolling window of length 1,000. Secondly, we use Filtered Historical Simulation (FHS) to estimate the lower tail of the innovations without assuming any conditional distribution for the data, for which we perform bootstrapping 10,000 times (Ruiz and Pascual, 2002), which is referred to as NFHS. Then, we combine the estimated GARCH variances with the upper $(1-\alpha)$-percentile of the standardized residuals ($\hat{Z}_t$) to compute daily VaR and ES measures:

$$\hat{v}_t^{FHS} = -\hat{\sigma} v_{fhs}(\alpha), \qquad \hat{e}_t^{FHS} = -\hat{\sigma} e_{fhs}(\alpha);$$

$$\text{where} \quad v_{fhs} = percentile\left\{ \{-\hat{Z}_{i,t}\}_{i=1}^{N}, 100(1-\alpha) \right\},$$

$$\text{and} \quad e_{fhs} = \frac{1}{N\alpha} \sum_{i=1}^{N} (-\hat{Z}_{i,t}) \mathbb{1}(-\hat{Z}_{i,t} > v_{fhs}).$$

The right figure of panel (c) in Figure 3.2.1 shows the advantage of EVT in modeling the tail distribution especially at small α levels. We can differentiate two misspecified risk models using below zero-homogeneity scoring functions due to

the non-zero values of average loss differences, whilst this is not the case when using above zero-homogeneity scoring functions. The left figure of panel (c) presents the 95% confidence intervals of average loss differences for risk measures at 2.5% level.

Overall, the above three realistic simulation-based scenarios show that in the presence of model risk, the ordering of competing (VaR, ES) risk models is sensitive to the choice of FZ scoring function with degrees of homogeneity between -1 and 1, indicating that model risk indeed matters in making model comparisons based on the expected FZ score. This provides a possible explanation for the results of Fissler et al. (2019) who describe the properties of scoring functions using different definitions of order sensitivity. They conclude that the FZ scoring function is order-sensitive on line segments (see their Definition 3.3), meaning that the scoring function is linearly increasing between the true functional value (true risk) and any risk functional. In other words, denoting the vector of true (VaR, ES) by z, for any given vector η, the scoring function of an estimate $z + s\eta$ where $s \in [0, \infty]$ is linearly increasing in s. Although the true value of (VaR, ES) is never known in practice, the optimal risk estimates can be uniquely obtained via minimizing the expected score of a given scoring function due to the joint elicitability of (VaR, ES) (Fissler and Ziegel, 2016). In addition, the relationship between the expected score and the size of model risk is not clear. In the remaining part of Chapter 3, our purpose is to quantify the model risk of a (VaR, ES) risk model via minimizing the expected score of a certain FZ scoring function.

3.3 Scoring function-based model risk measure

In this section, we assess the model risk of a set of widely known risk models considered also in Nolde and Ziegel (2017a) (a review of these risk models is provided in Appendix 3.B): the nonparametric method is Historical Simulation (HS); the semi-parametric methods include the GARCH(1,1) models with the normal, standardized Student's t, and skewed t innovations combined with the Filtered Historical Simulation technique (NFHS, TFHS, and SKTFHS); the parametric methods include the GARCH(1,1) processes with the normal, standardized Student's t, and skewed t distributed innovations (NFP, TFP, and SKTFP), as well as these models combined with the Extreme Value Theory (EVT) methodology (NEVT, TEVT, and SKTEVT). We also consider the newly proposed semiparametric models[6] based on the $FZ0$ minimization of Patton et al. (2019): the one-factor GAS model (denoted by FZ1F), the GARCH model via FZ minimization (denoted by GFZ) as well as the hybrid GAS/GARCH model (denoted by Hybrid for brevity), and add the EWMA model ($\lambda = 0.94$, also called *RiskMetrics*) to the parametric approaches. We use these risk estimation methods to compute the ex-ante one step ahead VaR and ES estimates at a given significance level using rolling windows of length 1,000.

In the following, our scoring function-based model risk framework is constructed based on: 1) a time series of observed ex-post realizations of returns $r_t, ..., r_{t+T}$ and 2) for a given significance level α, a time series of ex-ante daily

(VaR, ES) measures $(\hat{v}_t^j, \hat{e}_t^j), ..., (\hat{v}_{t+T}^j, \hat{e}_{t+T}^j)$ made by risk model $j \in \{1, ..., m\}$ of $m \geq 1$ competing risk models.

3.3.1 Scoring function-based joint model risk measure

Ideally, if the pair of true risk measures were known, it would be straightforward to compute the distance from the estimated risk measures to the true ones, thus measuring the size of model risk.

Definition 3. For the risk functional(s) Z and all F: $\mathcal{F} \to D \subseteq \mathbb{R}^2$, consider $\hat{z}_t^j = (\hat{v}_t^j, \hat{e}_t^j)$, and $z_t = (v_t, e_t)$, such that $\hat{z}_t^j$ and $z_t \in D$. $\{\hat{z}_i^j\}_{t \leq i \leq t+n}$ is a time series of risk estimates made by model $j \in \{1, ..., m\}$, where we have m competing models over the model risk evaluation period, with $0 \leq n \leq T$, whilst $\{z_i\}_{t \leq i \leq t+n}$ is the time series of true values of risk measures Z given by the true model over the same period. The joint model risk measure $p_{[t,t+n]}^j$ of a risk model j over the evaluation window $[t, t+n]$ is defined as:

$$p_{[t,t+n]}^j = \frac{1}{n+1} \cdot \sum_{i=t}^{t+n} \sqrt{(\hat{v}_i^j - v_i)^2 + (\hat{e}_i^j - e_i)^2}. \tag{3.3.1}$$

However, the true values of risk measures $z_i = (v_i, e_i)$ are unknown in practice. As such, we propose a pragmatic method to estimate the joint model risk of a (VaR, ES) risk model: first, we calculate the optimum multipliers[7] $\{x_{1,i}^j\}$ and $\{x_{2,i}^j\}$ by minimizing the expected score over the multiplier estimation window[8] i from $t+k$ to $t+\tau+k$ with window length of $\tau+1$, where $k = 0 : T-\tau$. For any

risk model j, the time series of $\{\hat{v}_i^j\}$ and $\{\hat{e}_i^j\}$ are estimated at an α significance level, then these time-varying multipliers are the solution to the minimization exercise:

$$(x^j_{1,t+\tau+k}, x^j_{2,t+\tau+k}) = \underset{(X_1,X_2)}{\arg\min} \frac{1}{\tau+1} \cdot \sum_{i=t+k}^{t+\tau+k} S_{FZ}\left(r_i, X_1 \cdot \hat{v}_i^j, X_2 \cdot \hat{e}_i^j; \alpha\right) \quad (3.3.2)$$

In the above, $\{r_i\}$ is the daily return series, X_1 and X_2 are multipliers of $\hat{v}_i^j$ and $\hat{e}_i^j$, respectively, and we use the restrictions that $X_1 \cdot \hat{v}_i^j > X_2 \cdot \hat{e}_i^j$, with $X_1, X_2 > 0$. Next, we approximate the joint (VaR, ES) model risk $\rho^j_{[t+\tau,t+\tau+n]}$ of risk model j as the average distance between $\hat{z}_i^j$ and $z_i^{j,min}$ across the model risk evaluation window $[t+\tau, t+\tau+n]$. Here $z_i^{j,min} = (x^j_{1,i} \cdot \hat{v}_i^j, x^j_{2,i} \cdot \hat{e}_i^j)$ is an improved pair of risk estimates after the estimated multipliers (obtained via FZ minimization) are applied:

$$\rho^j_{[t+\tau,t+\tau+n]} = \frac{1}{n+1} \cdot \sum_{i=t+\tau}^{t+\tau+n} \sqrt{(\hat{v}_i^j - x^j_{1,i} \cdot \hat{v}_i^j)^2 + (\hat{e}_i^j - x^j_{2,i} \cdot \hat{e}_i^j)^2}. \quad (3.3.3)$$

The solution in (3.3.2) shows that if the multiplier is larger (smaller) than 1, the corresponding risk estimate is underestimated (overestimated). To this extent, $\rho^j_{[t+\tau,t+\tau+n]}$ in (3.3.3) provides an approximation of the true joint model risk $p^j_{[t+\tau,t+\tau+n]}$ in (3.3.1). For simplicity, we will henceforth omit the subscripts of $p^j_{[t+\tau,t+\tau+n]}$ and $\rho^j_{[t+\tau,t+\tau+n]}$, and use p^j and ρ^j for the true joint model risk measure and our proposed joint model risk measure estimate of a risk model j.

In order to gauge the degree of similarity between the theoretical and estimated measures of model risk, we first compute Pearson's linear correlation coefficient $\mathcal{C}^{\mathcal{M}} = Correl(p^{\mathcal{M}}, \rho^{\mathcal{M}})$ between the two series to see whether our FZ scoring function-based joint model risk estimate $\rho^{\mathcal{M}}$ approximates the true joint model risk measure $p^{\mathcal{M}}$ across the set of risk models $\mathcal{M}$ which is the set of risk models discussed in Appendix 3.B. As the correlation only considers the strength of the linear relationship between the true and estimated joint model risk across the set of models $\mathcal{M}$, we also use the $\tau_x^{\mathcal{M}} = \tau_x(p^{\mathcal{M}}, \rho^{\mathcal{M}})$ correlation coefficient from Emond and Mason (2002) that extends the nonparametric Kendall's τ_b measure, in order to estimate the possibly nonlinear association between the true and estimated joint model risk measures over a set of models $\mathcal{M}$. The values of τ_x can vary from -1 (perfect inversion) to 1 (perfect agreement), with a value of 0 indicating that the true joint model risk measure and the corresponding scoring function-based joint model risk measure estimate are independent.

In addition, we also compute the proportion ψ^j of true joint model risk (p^j) explained by our joint model risk estimate (ρ^j) over the model risk evaluation period, defined as $\psi^j = \rho^j / p^j$.

3.3.2 Scoring function-based individual model risk measure

In order to consider individual VaR and ES model risk, assuming that the true VaR and ES (v, e) are known, we calculate the average absolute biases $(\mathcal{B}_v^j, \mathcal{B}_e^j)$ of the estimated VaR and ES from the true risk measures over the evaluation period from t to $t+n$ as follows:

$$\mathcal{B}_v^j = \frac{1}{n+1} \cdot \sum_{i=t}^{t+n} \left| \hat{v}_i^j - v_i \right|, \quad \mathcal{B}_e^j = \frac{1}{n+1} \cdot \sum_{i=t}^{t+n} \left| \hat{e}_i^j - e_i \right|. \tag{3.3.4}$$

In practice, we can derive the individual VaR and ES model risk measures (ρ_v^j, ρ_e^j) of model j built upon the optimum multipliers $(x_{1,i}, x_{2,i})$ in (3.3.2) assigned to the estimated VaR and ES $(\hat{v}_i, \hat{e}_i)$ over the model risk evaluation period from $t+\tau$ to $t+\tau+n$:

$$\rho_v^j = \frac{1}{n+1} \cdot \sum_{i=t+\tau}^{t+\tau+n} |\hat{v}_i^j - x_{1,i} \cdot \hat{v}_i^j|, \quad \rho_e^j = \frac{1}{n+1} \cdot \sum_{i=t+\tau}^{t+\tau+n} |\hat{e}_i^j - x_{2,i} \cdot \hat{e}_i^j|. \tag{3.3.5}$$

Similar to $\mathcal{C}^{\mathcal{M}}$ and $\tau_x^{\mathcal{M}}$, we compute $\mathcal{C}_v^{\mathcal{M}}$, $\mathcal{C}_e^{\mathcal{M}}$, $\tau_{x,v}^{\mathcal{M}}$ and $\tau_{x,e}^{\mathcal{M}}$ to measure the level of correlation between our scoring function-based individual model risk measures in (3.3.5) and the true values of individual model risk computed in equation (3.3.4), over a certain evaluation period across a set of models $(\mathcal{M})$ considered:

$$\mathcal{C}_v^{\mathcal{M}} = Correl(\mathcal{B}_v^{\mathcal{M}}, \rho_v^{\mathcal{M}}), \quad \mathcal{C}_e^{\mathcal{M}} = Correl(\mathcal{B}_e^{\mathcal{M}}, \rho_e^{\mathcal{M}}).$$

$$\tau_{x,v}^{\mathcal{M}} = \tau_x(\mathcal{B}_v^{\mathcal{M}}, \rho_v^{\mathcal{M}}), \quad \tau_{x,e}^{\mathcal{M}} = \tau_x(\mathcal{B}_e^{\mathcal{M}}, \rho_e^{\mathcal{M}}).$$

For model j, $\psi_v^j = \rho_v^j/\mathcal{B}_v^j$ ($\psi_e^j = \rho_e^j/\mathcal{B}_e^j$) shows the proportion of true model risk captured by our model risk measure in terms of the VaR (ES) measure over the evaluation period.

3.3.3 Simulation study

To verify how our FZ scoring function-based model risk estimation methodology works in a simulation setting (which allows comparisons with the true model risk), as in Dimitriadis and Bayer (2019), we implement three different positively homogeneous FZ scoring functions of degree $b = 0, 0.5$ and -1, presented in Table 3.3.1. These are natural examples of scoring functions, and we denote them by S_1, S_2 and S_3. These degrees of homogeneity correspond to $\mathcal{G}_2(z) = -\log(-z)$, $\mathcal{G}_2(z) = -\sqrt{-z}$ as well as $\mathcal{G}_2(z) = -1/z$, respectively. In order to put the emphasis on the ES (Ziegel et al., 2017), we fix $G_1(z) = 0$ in (3.2.3).

Table 3.3.1: Three *FZ* scoring functions with different degrees of positive homogeneity

Positive homogeneity (b)	FZ scoring function
0	$S_1 = -\frac{1}{\alpha e}\mathbb{1}\{r \leq v\}(v-r) + \frac{v}{e} + \log(-e) - 1$
0.5	$S_2 = \frac{1}{2\sqrt{-e}}\left(\frac{1}{\alpha}\mathbb{1}\{r \leq v\}(v-r) - (v-e)\right) + \sqrt{-e}$
-1	$S_3 = \frac{1}{e^2}\left(\frac{1}{\alpha}\mathbb{1}\{r \leq v\}(v-r) - (v-e)\right) + \frac{1}{e}$

In our simulation study, the GARCH(1,1) model with Student's t distributed residuals, specified in (3.2.7), is used as the data generating process, denoted by

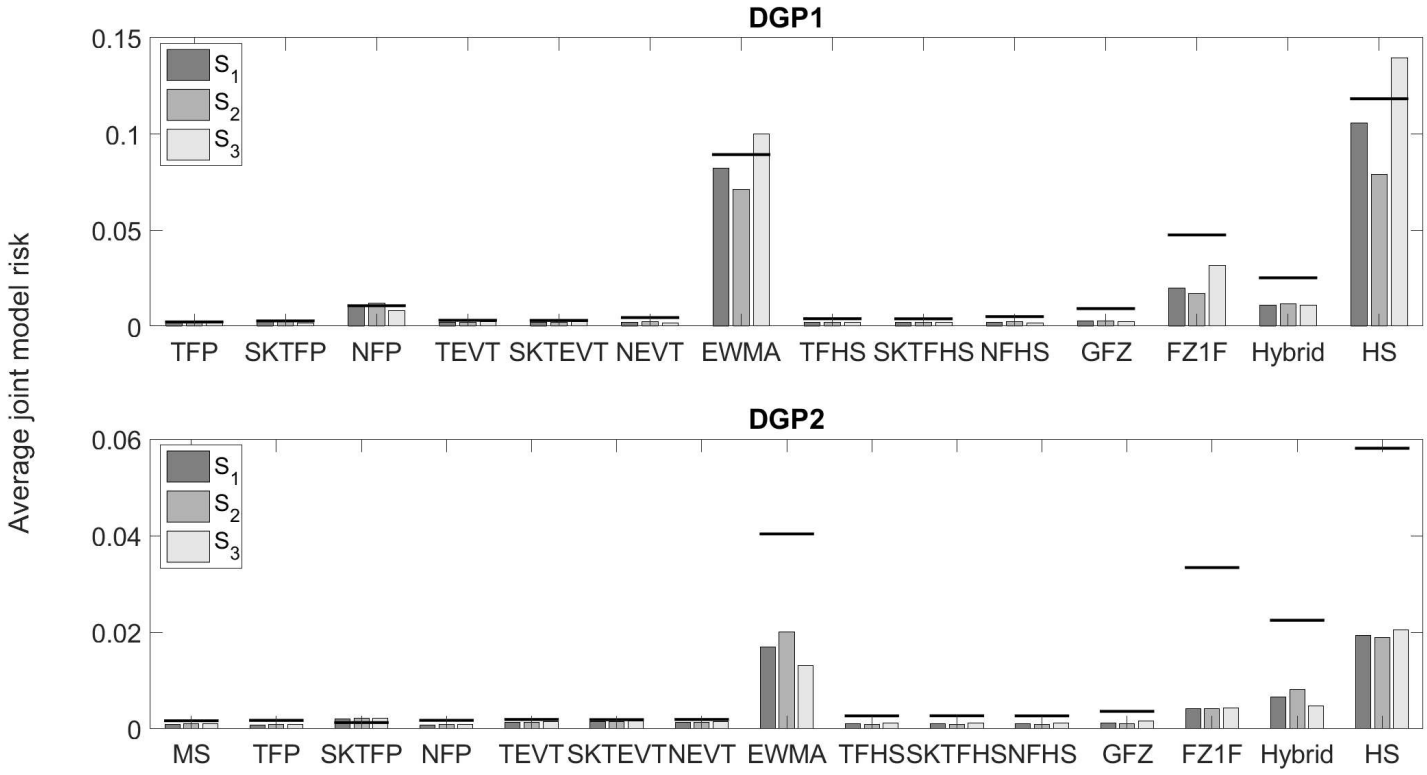
DGP1. We also adopt the Markov Switching GARCH(1,1) model with normal disturbances (Klaassen, 2002) as DGP2 shown below:

$$r_t = \sqrt{\hat{h}_{s_t}} Z_t, \quad Z_t \sim iid\ N(0,1), \quad s_t = \{1,2\}, \tag{3.3.6}$$

$$\text{where} \quad \hat{h}_{s_t} = \hat{\omega}_{s_t} + \hat{\beta}_{s_t} r_{t-1}^2 + \hat{\gamma}_{s_t} \sum_{i=1}^{2} p_{ij} \hat{h}_{i,t-1}, \quad \hat{\omega}_{s_t}, \hat{\beta}_{s_t} \text{ and } \hat{\gamma}_{s_t} > 0;$$

$s_t = 1$ or 2 shows the possible market state at time t; p_{ij} denotes the probability of state j at time t conditional that the market is in state i at time $t-1$; $\hat{h}_{i,t-1}$ denotes the conditional variance in state i at time $t-1$. The model parameters used for simulation in DGP2 are: $\hat{\omega}_1 = 1.8960e^{-04}, \hat{\beta}_1 = 0.15841$ and $\hat{\gamma}_1 = 0.41507; \hat{\omega}_2 = 2.4130e^{-05}, \hat{\beta}_2 = 0.56147$ and $\hat{\gamma}_2 = 0.41507; p_{11} = 0.4323$ and $p_{22} = 0.9992$, estimated from the S&P500 daily returns from 2001/01/01 to 2018/05/20 (4492 observations). Then we simulate 5,000 returns by each data generating process and compute daily VaR and ES measures using rolling windows of length 1,000 across the set of models. With respect to the FZ scoring functions (S_1, S_2 and S_3), Figure 3.3.1 compares the mean values of true joint model risk (shown in horizontal lines) with the average estimated joint model risk (shown in bars) over the time period, for various (VaR, ES) risk models based on the simulated data generated by the aforementioned two data generating processes, DGP1 and DGP2. Regarding the FZ scoring function-based joint model risk measure in (3.3.3), at 2.5% critical level we calculate the estimates for the

Figure 3.3.1: Average estimated S_1, S_2 and S_3-based joint model risk of various (VaR, ES) risk models

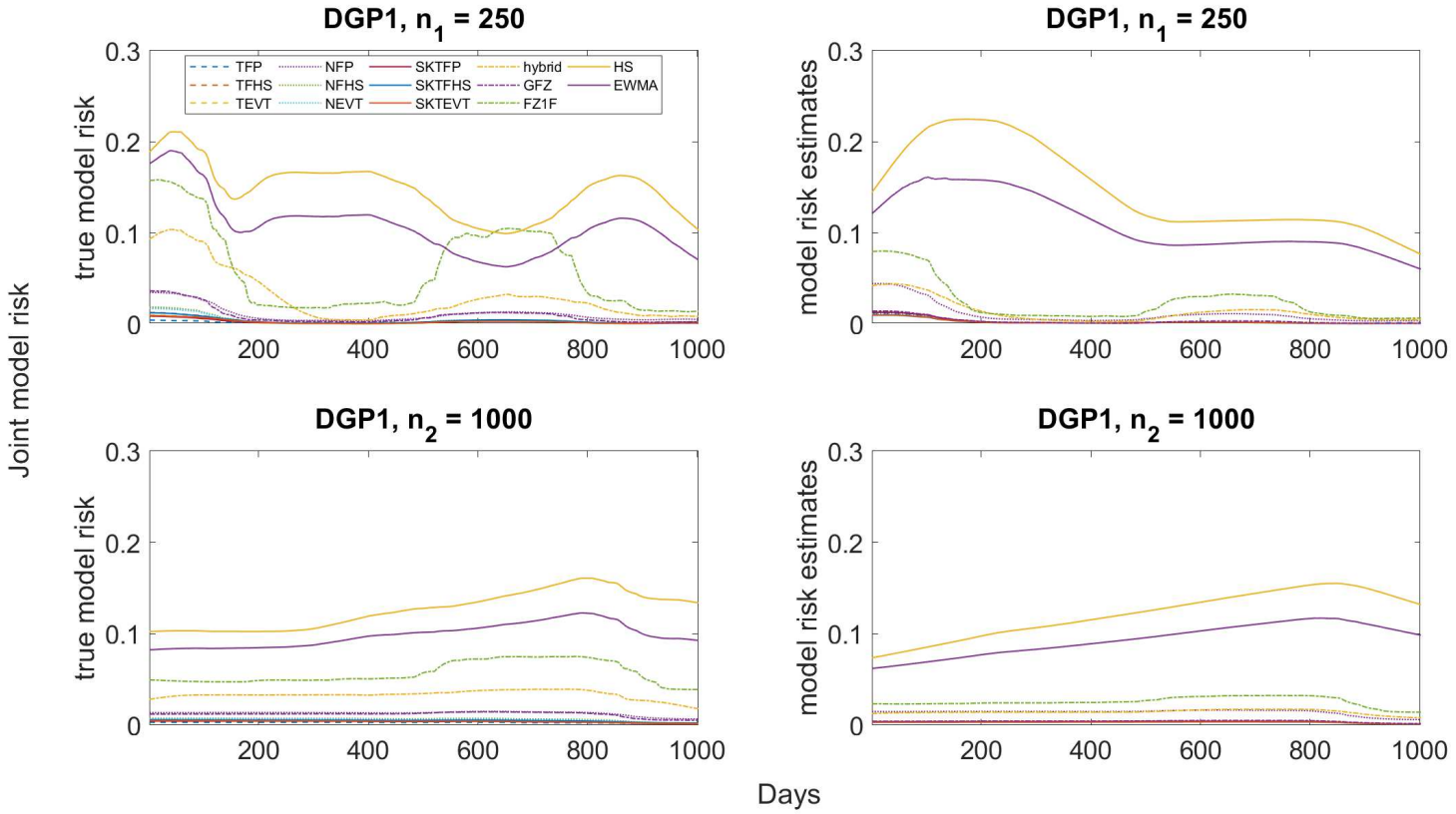


This figure shows average values of the estimated and true joint model risk, illustrated in bars and horizontal lines, respectively. The daily VaR and ES are estimated at 2.5% level under DGP1 and DGP2 (TFP and MS are the estimated DGP1 and DGP2, respectively). The model risk evaluation window length is $n_1 = 250$.

given set of models using a rolling window with the length of $n_1 = 250$. This choice of the window length[9] for calculating model risk follows the supervisory requirement by the Basel Committee on Banking Supervision (2019) that a one-year backtesting period (i.e. 250 trading days) is needed to confirm the quality of a model. We compute the true values of joint model risk using (3.3.1). The magnitude of joint model risk based on our methodology is generally smaller than its true value, with a few exceptions. Historical Simulation (HS) is the worst method in forecasting the daily 2.5% VaR and ES under the two different DGPs since it has the highest level of joint model risk among all the models examined. The parametric approaches seem to perform best (except for the EWMA model, which is the second-worst due to high persistence to the shocks), followed by the semiparametric models. This is in line with the results reported in Table 5 of Patton et al. (2019), showing that the parametric models estimated by Maximum Likelihood estimation outperform the $FZ0$ minimization-based semiparametric models in computing risk estimates.

From a dynamic perspective, over two different model risk evaluation windows $n_1 = 250$ and $n_2 = 1{,}000$, Figure 3.3.2 gives the evolution of true joint model risk (on the left side) and the $FZ0$-based joint model risk estimates (on the right side) of the daily (VaR, ES) at 2.5% coverage level across various risk models, for returns simulated via DGP1. In terms of the joint model risk measures in the right panel, the measure based on the shorter window ($n_1 = 250$) indicates more variation of joint model risk than the measure based on the longer eval-

Figure 3.3.2: Dynamic true and *FZ0*-based estimated model risk of various (VaR, ES) models



This figure shows the evolution of true joint model risk on the left side and the FZ0-based joint model risk estimates on the right side. The daily VaR and ES are estimated at $\alpha = 2.5\%$ and the joint model risk is computed over two model risk evaluation windows ($n_1 = 250$ and $n_2 = 1,000$) based on returns simulated by DGP1.

Figure 3.3.3: Joint model risk computed over several model risk evaluation windows

This figure shows the true joint model risk (on the left side) and FZ0-based joint model risk (on the right side) of the EWMA model ($\lambda = 0.94$) at $\alpha = 2.5\%$ over several model risk evaluation windows $n_1 = 250$, $n_2 = 1,000$ and $n_3 = 100,000$, based on simulated data by DGP1.

uation window (n_2 = 1,000). The estimated model risk illustrated in the right subpanels presents a high resemblance to the dynamics of true model risk in the left subpanels. As expected, the joint model risk computed over a window of $n_1 = 250$ is more sensitive to market events as indicated by the more volatile solid line in Figure 3.3.3 for the EWMA model. Beyond that, using a long model risk evaluation window such as $n_3 = 100,000$ is not sensible due to the possible structural breaks in the data generating process.

To get a closer look, we report three measures of similarity between the true joint model risk and our joint model risk estimates[10] in Panel A of Table 3.3.2, taking into account the risk models and time consistency simultaneously. The first measure is the average correlation ($\bar{\mathcal{C}}^{\mathcal{M}}$) between the true and estimated joint model risk of various (VaR, ES) models over time. These correlations are at least 0.946 (0.800) under DGP1 (DGP2), suggesting the proposed FZ-based joint model risk measures closely related to the true joint model risk. Considering scoring functions with different levels of homogeneity, we find that S_3, with positive homogeneity parameter b equal to -1, offers the highest correlations, whilst S_2 ($b = 0.5$) leads to the lowest correlations in general. Also, the correlations are higher when the model risk evaluation window is longer. The second measure we report is the average explanatory power ($\bar{\psi}^{\mathcal{M}}$) over a set of models $\mathcal{M}$. We find that the joint model risk measure is able to capture on average more than 50% of the true model risk of joint risk estimates under these two data generating processes. As a third measure we look at the degree of similarity $\bar{\tau}_x^{\mathcal{M}}$, and find

Table 3.3.2: Measures of similarity between the true and estimated model risk

	DGP1						DGP2					
Evaluation window	$n_1 = 250$			$n_2 = 1000$			$n_1 = 250$			$n_2 = 1000$		
A: Joint model risk	$\bar{\mathcal{C}}^{\mathcal{M}}$	$\bar{\psi}^{\mathcal{M}}$	$\bar{\tau}_x^{\mathcal{M}}$	$\bar{\mathcal{C}}^{\mathcal{M}}$	$\bar{\psi}^{\mathcal{M}}$	$\bar{\tau}_x^{\mathcal{M}}$	$\bar{\mathcal{C}}^{\mathcal{M}}$	$\bar{\psi}^{\mathcal{M}}$	$\bar{\tau}_x^{\mathcal{M}}$	$\bar{\mathcal{C}}^{\mathcal{M}}$	$\bar{\psi}^{\mathcal{M}}$	$\bar{\tau}_x^{\mathcal{M}}$
S_1	0.958	69.2%	0.662	0.974	74.6%	0.576	0.826	53.8%	0.757	0.900	57.2%	0.708
S_2	0.946	66.7%	0.673	0.972	79.2%	0.536	0.800	58.1%	0.733	0.883	62.0%	0.730
S_3	0.966	83.9%	0.637	0.981	66.8%	0.630	0.861	59.1%	0.713	0.908	61.7%	0.644
B: ES model risk	$\bar{\mathcal{C}}_e^{\mathcal{M}}$	$\bar{\psi}_e^{\mathcal{M}}$	$\bar{\tau}_{x,e}^{\mathcal{M}}$	$\bar{\mathcal{C}}_e^{\mathcal{M}}$	$\bar{\psi}_e^{\mathcal{M}}$	$\bar{\tau}_{x,e}^{\mathcal{M}}$	$\bar{\mathcal{C}}_e^{\mathcal{M}}$	$\bar{\psi}_e^{\mathcal{M}}$	$\bar{\tau}_{x,e}^{\mathcal{M}}$	$\bar{\mathcal{C}}_e^{\mathcal{M}}$	$\bar{\psi}_e^{\mathcal{M}}$	$\bar{\tau}_{x,e}^{\mathcal{M}}$
S_1	0.959	79.3%	0.589	0.980	86.6%	0.307	0.804	48.1%	0.828	0.887	42.7%	0.867
S_2	0.943	74.8%	0.637	0.975	90.6%	0.341	0.776	55.8%	0.816	0.873	51.3%	0.867
S_3	0.968	94.8%	0.600	0.987	68.8%	0.629	0.852	52.0%	0.802	0.905	43.9%	0.865
C: VaR model risk	$\bar{\mathcal{C}}_v^{\mathcal{M}}$	$\bar{\psi}_v^{\mathcal{M}}$	$\bar{\tau}_{x,v}^{\mathcal{M}}$	$\bar{\mathcal{C}}_v^{\mathcal{M}}$	$\bar{\psi}_v^{\mathcal{M}}$	$\bar{\tau}_{x,v}^{\mathcal{M}}$	$\bar{\mathcal{C}}_v^{\mathcal{M}}$	$\bar{\psi}_v^{\mathcal{M}}$	$\bar{\tau}_{x,v}^{\mathcal{M}}$	$\bar{\mathcal{C}}_v^{\mathcal{M}}$	$\bar{\psi}_v^{\mathcal{M}}$	$\bar{\tau}_{x,v}^{\mathcal{M}}$
S_1	0.849	39.5%	0.824	0.897	34.1%	0.876	0.500	65.0%	0.527	0.771	73.5%	0.433
S_2	0.791	42.9%	0.788	0.920	42.6%	0.724	0.564	66.0%	0.493	0.838	76.1%	0.406
S_3	0.912	46.1%	0.852	0.945	44.8%	1.000	0.505	72.3%	0.447	0.660	80.5%	0.412

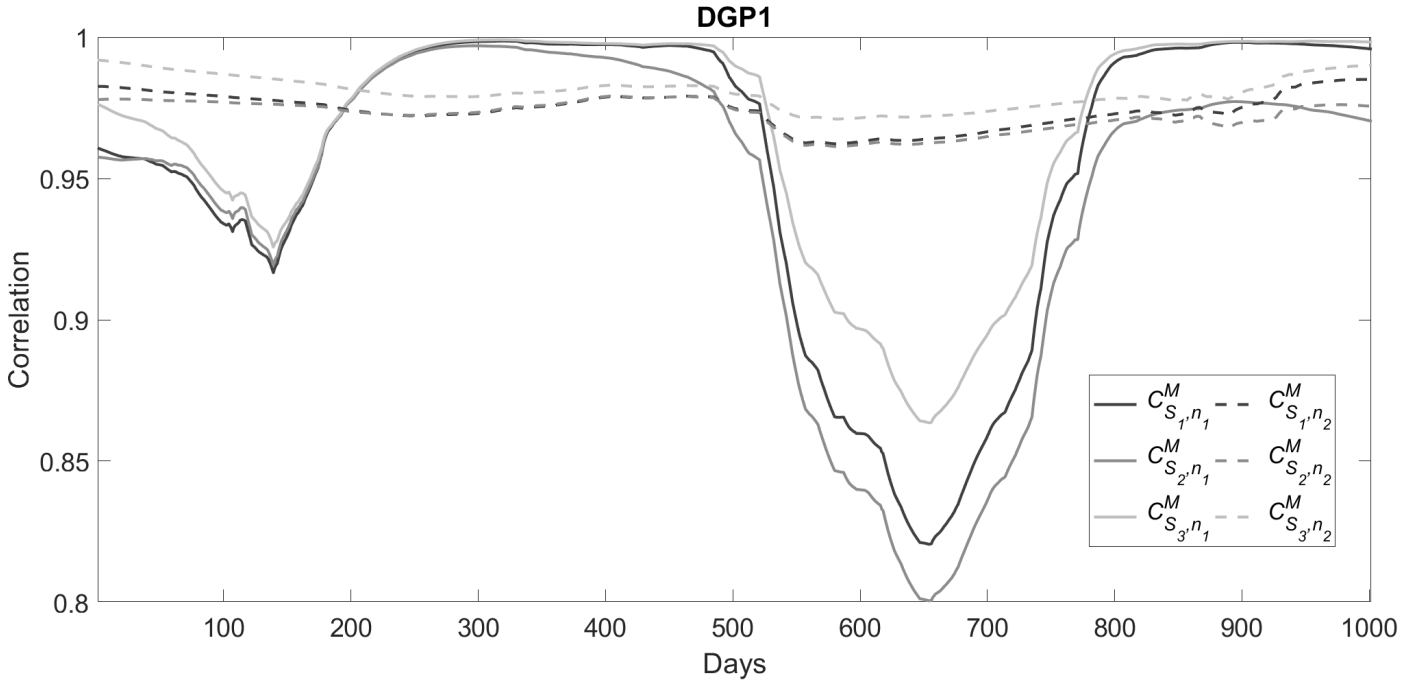
*This table presents the results of three measures of similarity between the true and estimated model risk. We generate a path of 5,000 daily returns using DGP1 and DGP2 (the GARCH(1,1)-*t *and MS(2)-GARCH(1,1)-N models specified in (3.2.7) and (3.3.6)). The estimated model risk measures are based on three FZ scoring functions (S_1, S_2 and S_3) with positive homogeneity $b = 0, 0.5$ and -1. The daily risk measures are estimated at 2.5% level. For two model risk evaluation windows (n_1 and n_2), Panel A shows the average correlation ($\bar{\mathcal{C}}^{\mathcal{M}}$) between the true and estimated values of joint model risk of various models, the average explanatory power ($\bar{\psi}^{\mathcal{M}}$) of our proposed joint model risk measure and the degree of similarity ($\bar{\tau}_x^{\mathcal{M}}$) between true and estimated joint model risk; Panels B and C report the values of measures of similarity for ES model risk and VaR model risk, respectively.*

that, our FZ scoring function-based joint model risk measure exhibits a significant resemblance to the true joint model risk measure, with $\bar{\tau}_x^{\mathcal{M}}$ equal to at least 0.536 (0.644) under DGP1 (DGP2).

Figure 3.3.4 depicts the dynamic correlation between the true and estimated joint model risk of a series of models when the estimated joint model risk is calculated based on S_1, S_2 and S_3 in two different evaluation windows $n_1 = 250$ and $n_2 = 1{,}000$, under DGP1. The joint model risk measure based on the -1 homogeneous FZ scoring function (S_3) in a given evaluation window exhibits the strongest dynamic correlation with the true joint model risk, followed by the S_1-based joint model risk measure. Moreover, Figure 3.3.5 shows the distributions of the $FZ0$-based joint (VaR, ES) model risk over an evaluation window of $n_1 = 250$ for the selected models when simulating 5,000 paths of 1,000 returns by DGP1. One-step ahead VaR and ES are calculated at 2.5%. We find that the EWMA model is the worst-performing in making risk estimates as evidenced by the highest mean and largest dispersion of joint model risk estimates among the selected models.

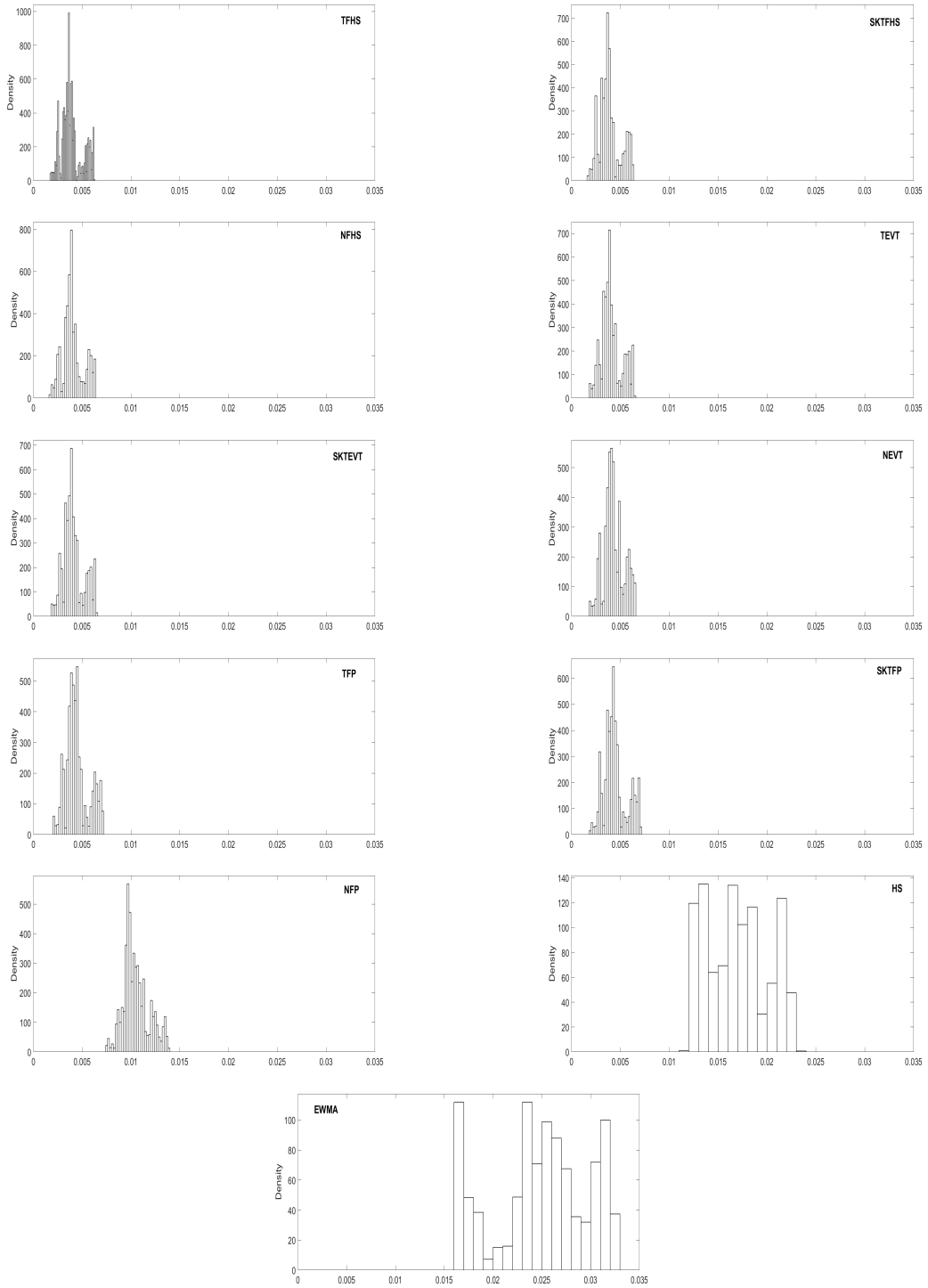
For chief risk officers of companies as well as for regulators it would be very useful to be able to disentangle the effects of different sources of model risk. Hence, we decompose the joint model risk into estimation risk and specification risk in the simulation study, illustrated in panel (a) of Figure 3.3.6. To evaluate the estimation risk of a given model, we first estimate the parameters of a certain model based on the return series of S&P500 index from 03/01/2000 to

Figure 3.3.4: Dynamic correlation ($\mathcal{C}^{\mathcal{M}}$) between true joint model risk and S_1, S_2 and S_3-based joint model risk



This figure displays the dynamic correlation between the true and estimated joint model risk of a series of models, in which the joint model risk is computed over model risk evaluation windows ($n_1 = 250$ and $n_2 = 1,000$), under DGP1.

Figure 3.3.5: Histograms of *FZ0*-based joint model risk estimates of selected risk models



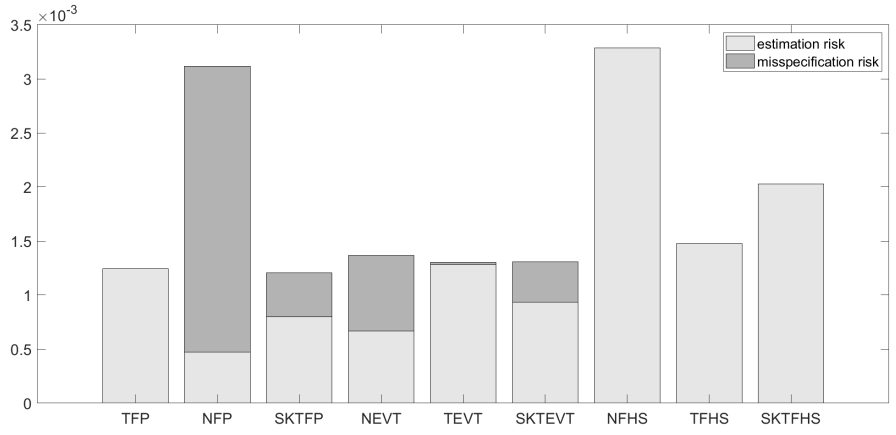
This figure shows the histograms of the FZ0*-based joint model risk over an evaluation window of* $n_1 = 250$ *for the selected models, based on the simulated 5,000 paths of 1,000 returns under DGP1. The plots are ordered from left to right and top to bottom by the ascending values of mean and standard deviation.*

31/12/2007, simulate 5,000 paths of 1,000 returns and then re-estimate the model in a rolling window of 1,000 to make one-step ahead risk estimates at 2.5% level. Subsequently, we compute the $FZ0$-based estimation risk in an evaluation window of $n_1 = 250$ for a given set of risk models. The $FZ0$-based joint model risk of a certain (VaR, ES) model, comprised of estimation risk and misspecification risk, is calculated based on the simulated data generated by the GARCH(1,1) process with standardized Student's t innovations (TFP). Comparing the GARCH(1,1) models with the fully parametric normal and skewed t disturbances (NFP and SKTFP), we notice that SKTFP is less misspecified than NFP but has a larger estimation risk due to the higher number of parameters included in the model. Comparing the FHS, the EVT and the fully parametric method (FP) applied to the distribution of the standardized residuals extracted from the GARCH(1,1) models, FHS has the highest estimation risk, followed by EVT and FP, as displayed in panel (b) of Figure 3.3.6.

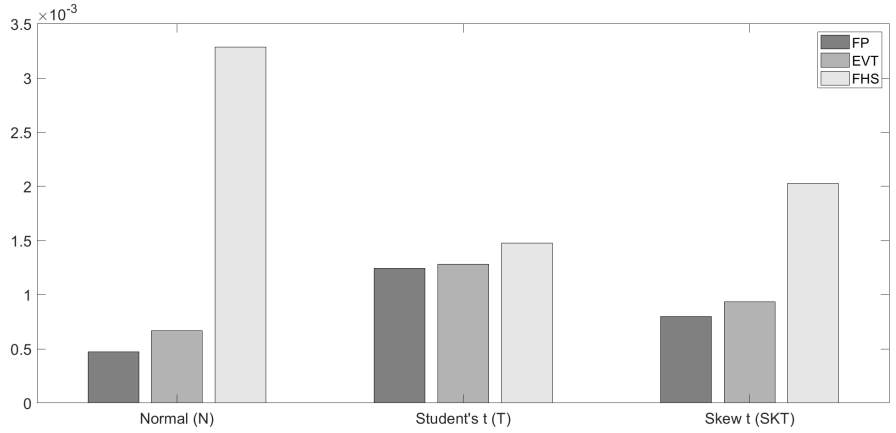
Regarding the FZ-based individual ES model risk measure, the average correlations $\bar{\mathcal{C}}_e^{\mathcal{M}}$ between the true and estimated values of ES model risk of various models are shown in Panel B of Table 3.3.2, which share similar values to the average correlations ($\bar{\mathcal{C}}^{\mathcal{M}}$) between the true and estimated joint model risk in Panel A, Table 3.3.2. Supplemental to $\bar{\mathcal{C}}_e^{\mathcal{M}}$, the values of $\bar{\tau}_{x,e}^{\mathcal{M}}$ are above 0.5 generally, signaling a high degree of similarity between the true and FZ-based ES model risk measures. Generally, for both DGP1 and DGP2, the FZ-based ES model risk measures can explain more than half of the true model risk of ES estimates

Figure 3.3.6: *FZ0*-based joint model risk

(a) Estimation risk and misspecification risk



(b) Estimation risk



This figure shows the components of FZ0-based joint model risk of (VaR, ES) models at $\alpha = 2.5\%$. We simulate 5,000 paths of 1,000 returns using model parameters estimated from the S&P500 index from 03/01/2000 to 31/12/2007. Panel (a): the data generating process is the GARCH(1,1) process with Student's t *innovations (TFP). Panel (b): the FHS, EVT and FP estimation methods are implemented for the normal, Student's* t *and skewed* t *distributions of the standardized residuals.*

as shown by the average explanatory power ($\bar{\psi}_e^{\mathcal{M}}$) in Panel B. Hence, the ES model risk measure via FZ minimization is almost as efficient as the FZ-based joint model risk measure. Nevertheless, the true and estimated VaR model risk are less correlated under DGP1 and DGP2 as seen in Panel C of Table 3.3.2, indicating that the FZ-based VaR model risk measure is less adequate than the FZ-based ES model risk measure.

The simulation study confirms that our FZ scoring function-based joint model risk and ES model risk measures are practical tools to measure the model risk of risk models.

3.4 Properties of model risk measures

To facilitate the regulation of model risk measures, similarly to other measures of risks (e.g., market risk), we investigate whether the proposed FZ scoring function-based model risk measure has the coherence properties of an acceptable positive measure $\rho(\cdot)$ of risk (McNeil et al., 2015), where X and Y are returns of two different financial assets:

i) Subadditivity: $\rho(X+Y) \leq \rho(X) + \rho(Y)$.

ii) Positive homogeneity: for any positive number $h \in \mathbb{R}^+$, $\rho(h \cdot X) = h \cdot \rho(X)$.

iii) Monotonicity: for random variables of payoffs X and Y with $X \leq Y$, $\rho(X) \geq \rho(Y)$.

iv) Translation invariance: for any cash position represented by $a \in \mathbb{R}$, $\rho(X+a) =$

$\rho(X) - a$.

Concerning the above axioms, we consider the likewise properties for our proposed model risk measures which are defined as positive measures in this chapter, in contrast to negative VaR and ES estimates following the sign convention of risk measures used in the scoring function literature. Let $\rho^{MR}(X, \hat{v}_X^j, \hat{e}_X^j)$ be the joint model risk of a risk model j with $\hat{v}_X^j$ and $\hat{e}_X^j$ being the VaR and ES estimates[11] made by model j for the financial asset X, and let $\rho_e^{MR}(X, \hat{v}_X^j, \hat{e}_X^j)$ be the scoring functioned-based ES model risk (similar notation for asset Y or a portfolio $X + Y$). The following properties are considered for the joint model risk and ES model risk measures:

1. **Subadditivity** $\rho^{MR}(X+Y, \hat{v}_{X+Y}^j, \hat{e}_{X+Y}^j) \leq \rho^{MR}(X, \hat{v}_X^j, \hat{e}_X^j) + \rho^{MR}(Y, \hat{v}_Y^j, \hat{e}_Y^j)$.

   This property lines up with the diversification effect that the joint model risk of a certain risk estimation model fitted to a diversified portfolio of different assets is lower than the sum of the joint model risk of the same risk model applied to each asset.

2. **Positive homogeneity**

   For any $h \in \mathbb{R}^+$, $\rho^{MR}(h \cdot X, h \cdot \hat{v}_X^j, h \cdot \hat{e}_X^j) = h \cdot \rho^{MR}(X, \hat{v}_X^j, \hat{e}_X^j)$.

   The joint model risk will be scaled by the same size as long as all the inputs are rescaled by a positive number h.

3. **Monotonicity**

   If $|\hat{v}_X^1 - v_X| \geq |\hat{v}_X^2 - v_X|$ and $|\hat{e}_X^1 - e_X| \geq |\hat{e}_X^2 - e_X|$, $\rho^{MR}(X, \hat{v}_X^1, \hat{e}_X^1) \geq \rho^{MR}(X, \hat{v}_X^2, \hat{e}_X^2)$.

As expected, the pair of risk functional $(\hat{v}_X^1, \hat{e}_X^1)$ for the first model is more distant from the perfect risk estimates (v_X, e_X) than the corresponding risk estimates of the second model $(\hat{v}_X^2, \hat{e}_X^2)$, then it will have higher joint model risk.

4. **Translation invariance**

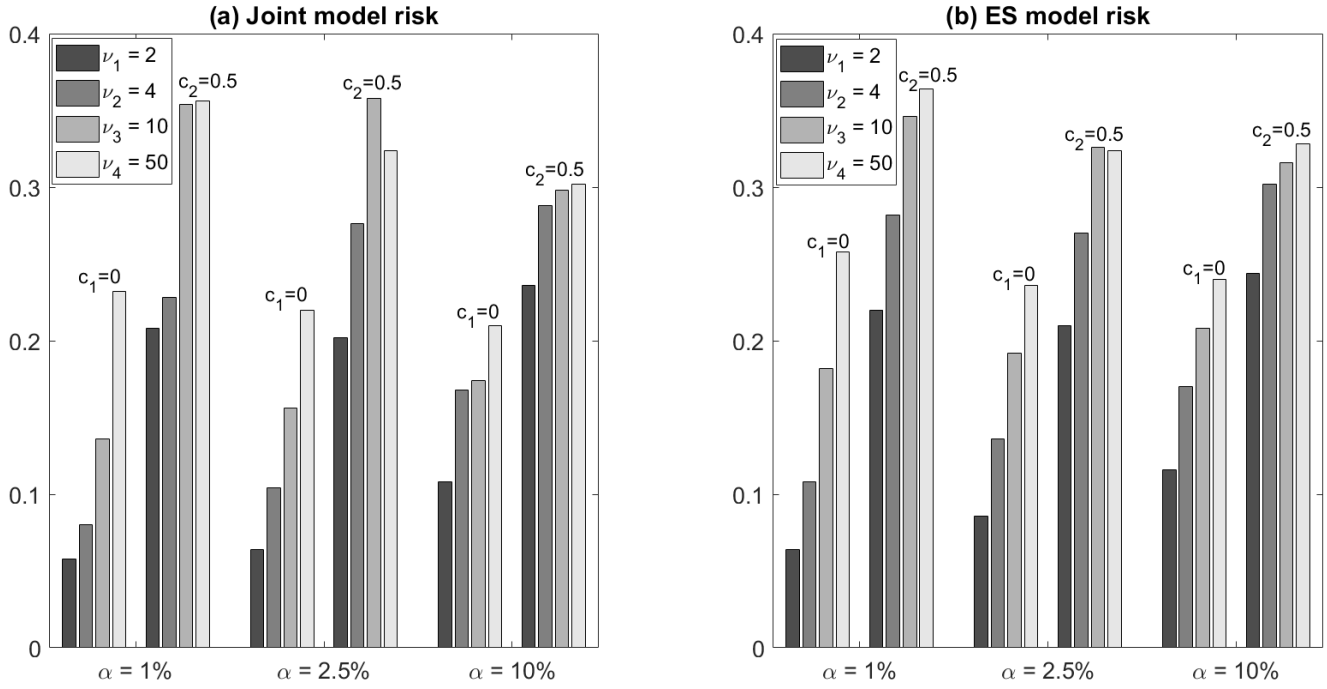
   For a constant $a > \hat{e}_X^j$, $\rho^{MR}(X + a, \hat{v}_X^j - a, \hat{e}_X^j - a) = \rho^{MR}(X, \hat{v}_X^j, \hat{e}_X^j)$.

   This is also called risk free condition, stating that the joint model risk is expected to be unaffected when a constant a is added to X and risk estimates are adjusted with the same amount.

We verify via simulations that our proposed model risk estimation methodology satisfies the properties of positive homogeneity, monotonicity as well as translation invariance. We place a particular focus on the subadditivity property as Daníelsson et al. (2013) do for the VaR measure. Given a certain risk model, we examine whether our model risk measure applied to a portfolio consisting of two assets is lower than the sum of model risk of individual assets (if not, we have a violation of subadditivity) in the following simulated scenarios displayed in Figure 3.4.1, which presents the percentage of subadditivity violations of joint model risk and ES model risk measures.

Consider that assets X and Z are independent but share the same Student's t distribution with degrees of freedom $\nu_1 = 2$, $\nu_2 = 4$, $\nu_3 = 10$, and $\nu_4 = 50$, and asset Y defined as $Y = cX + \sqrt{1 - c^2}Z$, so X and Y are correlated with correlation

Figure 3.4.1: Subadditivity violation rates of *FZ0*-based model risk measures of HS



This figure shows subadditivity violation rates of FZ0-based joint model risk and ES model risk measures of Historical Simulation. We employ the Historical Simulation method (HS) to estimate the 1%, 2.5% and 10% daily VaR and ES and compute the joint model risk and individual ES model risk over 250 days, for equally weighted portfolios of assets X and Y having correlations $c_1 = 0$ and $c_2 = 0.5$, where $Y = cX + \sqrt{1-c^2}Z$ and X and Z are independent but share a Student's t distribution with $\nu = 2, 4, 10$ and 50 degrees of freedom. If our model risk measure applied to this portfolio is larger than the sum of model risk of individual assets, then we have a violation of subadditivity.

coefficient c. Consider two cases: in the first one X and Y are independent ($c_1 = 0$); in the second case X and Y are correlated with $c_2 = 0.5$. We then simulate 500 paths of 3,250 returns by the Student's t distribution with different degrees of freedom for the two risky assets, X and Y. Equally weighted portfolios $(X + Y)$ are constructed based on the simulated data. We calculate the 1%, 2.5% and 10% daily VaR and ES for the two assets and portfolios, using the Historical Simulation method (HS), and then compute the $FZ0$-based model risk, including joint model risk and ES model risk, in a model risk evaluation window of 250 days. As highlighted in Figure 3.4.1, the portfolio with higher correlations leads to a higher subadditivity violation rate. The ES model risk measure has a higher violation rate of subadditivity than the corresponding joint model risk measure. Generally, the closer the return distributions of assets are to the normal distribution (the Student's t distribution converges to the normal distribution as the degree of freedom increases to the infinity), the higher the rate of subadditivity violations. In general, the higher the α level, the higher the violation rate. Thus, the subadditivity property is not guaranteed to be satisfied by the model risk measures.

3.5 Empirical Investigation

Here, we focus on several types of assets using daily market data from 01/01/1980 to 20/02/2019, downloaded from DataStream: 1) BARCLAYS equity price (BAR-

CLAYS); 2) S&P500 Index (S&P500); 3) Gold bullion price (GOLD); 4) Standard & Poor's Goldman Sachs Commodity Index total return (GSCI); and 5) FX EUR/USD rate (EUR/USD). First, we compute the daily log returns and construct the out-of-sample daily VaR and ES measures using rolling windows of 1,000 observations for the set of models $\mathcal{M}$. Next, we find the optimized multipliers for the pair of risk by minimizing the FZ scoring functions in a multiplier estimation window of the length 2,000 ($\tau = 1,999$) as in Nolde and Ziegel (2017b). We consider the S_1 and S_3-based model risk measures in the empirical analysis since the S_2-based model risk measure does not cover well the true model risk, as illustrated in Figure 3.3.4. We apply our scoring function-based model risk measures to market data using a model risk evaluation window with length of 250 ($n = 249$), as the Basel Committee on Banking Supervision (2019) recommends the prior 12 months (around 250 trading days) as the backtesting period for risk measures.

Table 3.5.1 presents the dollar values of annualized average S_1 and S_3-based joint model risk of (VaR, ES) risk models at 2.5% level, assuming that an investor holds a position of 1 million dollars in each asset. The EWMA model performs the worst, leading to misestimation of risk averaging $398,250. The GFZ, FZ1F and Hybrid models proposed by Patton et al. (2019) are less affected by joint model risk than the GARCH(1,1) model with normal innovations (NFP) and the Historical Simulation (HS), in line with the results in their paper. The GARCH(1,1) models combined with the Extreme Value Theory (TEVT and SKTEVT) as well

as the SKTFP model carry the lowest average joint model risk. Given a certain asset, the S_1 and S_3-based joint model risk measures are able to identify the same risk model in general as having the lowest level of joint model risk, but the joint model risk measure based on S_3 is more conservative due to the larger values of joint model risk compared with the measure based on S_1.

Panel (a) of Figure 3.5.1 captures the dynamic joint model risk based on $FZ0$ (S_1) of the daily 2.5% VaR and ES measures across various risk models applied to the daily log return series for BARCLAYS from 01/01/1980 to 20/02/2019. This signals the rising and significant joint model risk of various (VaR, ES) models during the crisis periods, confirming the discussion on model risk in Daníelsson et al. (2016). Our method reveals the dynamics of model risk corresponding to the market events, though there is a a-year delay due to the model risk evaluation window of 250 trading days. Out of all the models studied, the EWMA and HS methods are the least reactive to the market and thus display the highest joint model risk. We also find that the GARCH based models adapt to the price movements more quickly and exhibit lower joint model risk. Misspecification risk generally contributes more than estimation risk to the ($FZ0$-based annualized) joint model risk, and usually peaks during the turmoil periods, when the GARCH(1,1) model with normal innovations (NFP) produces daily risk estimates at 2.5% for BARCLAYS, as shown in panel (b) of Figure 3.5.1.

Additionally, Figure 3.5.2 displays the dynamics of optimized multipliers required for the daily 2.5% VaR and ES of several risk models, obtained via $FZ0$

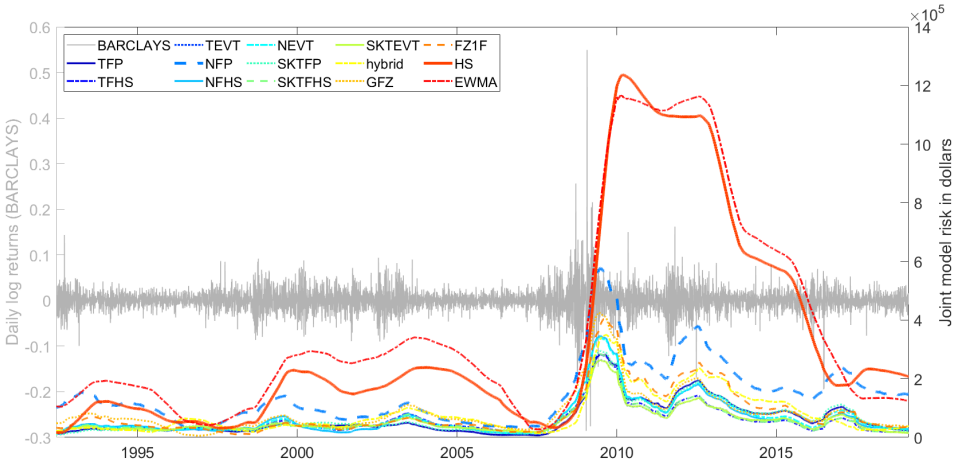
Table 3.5.1: Dollar values of annualized average joint model risk of daily risk measures

	BARCLAYS		S&P500		GOLD		GSCI		EUR/USD	
Models	S_1	S_3	S_1	S_3	S_1	S_3	S_1	S_3	S_1	S_3
TFP	62,099	61,513	44,490	55,004	23,278	27,927	29,972	34,804	14,102	15,543
SKTFP	64,743	61,477	**27,852**	**37,274**	**21,427**	**24,199**	28,172	34,211	11,439	13,657
NFP	146,183	153,737	93,619	102,047	84,570	89,799	66,434	69,237	20,857	24,718
TEVT	56,103	60,275	30,444	46,681	23,807	34,344	25,912	**29,115**	8,960	13,751
SKTEVT	**56,040**	**59,892**	30,403	46,994	23,932	34,339	**25,896**	29,428	9,021	13,648
NEVT	66,767	75,972	34,716	48,865	23,755	32,341	27,824	30,478	8,032	11,282
EWMA	376,179	398,250	192,143	220,933	121,455	126,816	134,078	134,784	35,400	38,013
TFHS	57,629	66,090	29,151	45,511	23,627	35,399	27,692	29,935	7,596	11,159
SKTFHS	57,485	65,429	29,090	45,757	23,912	35,297	27,706	30,351	7,538	10,987
NFHS	64,805	76,618	33,274	47,079	22,938	32,377	27,894	29,374	**7,021**	**9,992**
GFZ	83,641	102,651	33,157	41,246	30,128	40,397	38,906	44,526	8,383	12,131
FZ1F	81,548	109,173	46,991	55,404	42,938	50,348	45,075	49,714	11,197	13,929
Hybrid	79,876	69,212	37,586	42,613	41,991	45,635	40,287	40,142	9,766	10,115
HS	329,307	373,073	130,446	157,866	68,320	72,707	103,058	111,516	18,930	21,343

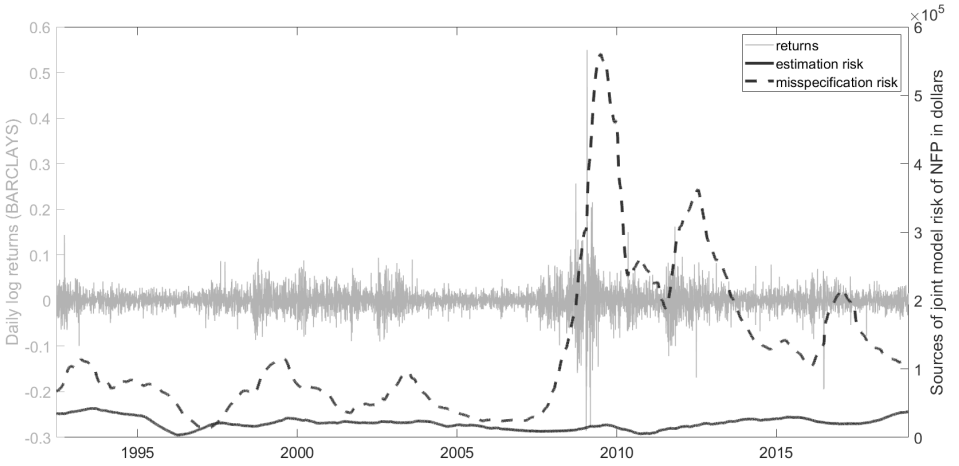
This table presents the dollar values of annualized average S_1 and S_3-based joint model risk of the daily VaR and ES estimated at 2.5% level across various models, based on daily data for all assets from 01/01/1980 to 20/02/2019. We simply use the "square root of time" rule to compute the annualized model risk in returns by multiplying the daily model risk in returns with $\sqrt{250}$, and obtain the dollar values of annualized model risk assuming an investment of 1 million. The values in bold are the lowest dollar values of annualized average joint model risk.

Figure 3.5.1: Dynamic *FZ0*-based annualized joint model risk in dollars for BARCLAYS

(a) Comparison of joint model risk of various models

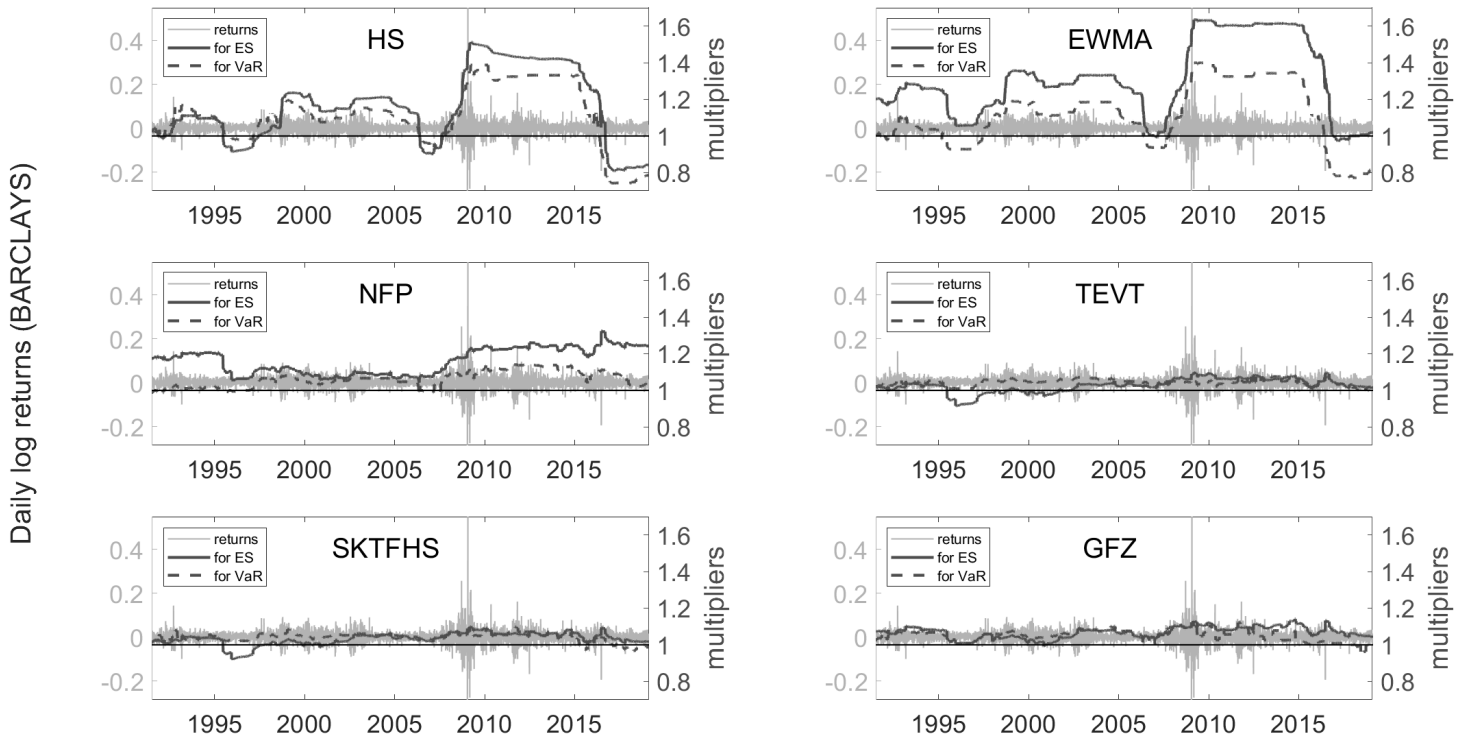


(b) Decomposition of joint model risk of NFP



This figure shows the dynamic FZ0-based joint model risk of daily (VaR, ES) estimates at $\alpha = 2.5\%$ for various models in panel (a) as well as the decomposition of FZ0-based joint model risk of the GARCH(1,1) model with normal innovations (NFP), based on the log returns of BARCLAYS from 01/01/1980 to 20/02/2019. The model risk evaluation period is 250. The average absolute ES across various models is about 9×10^5 dollars.

Figure 3.5.2: Dynamic optimized multipliers for the daily 2.5% VaR and ES, obtained via *FZ0* minimization



This figure displays the dynamics of optimized multipliers required for the daily 2.5% VaR and ES of several risk models, obtained via FZ0 minimization, based on daily returns of BARCLAYS returns from 01/01/1980 to 20/02/2019.

minimization, based on daily returns of BARCLAYS. The models examined in this figure tend to underestimate risk due to the multipliers for the risk estimates being larger than 1 most of the time.

The Basel Committee on Banking Supervision (2019) requires using a 10-day trading period for large cap equities and a 60-day trading period for exchange rates for risk calculation purposes, so we also compute risk estimates over a 10-day and 60-day period, for BARCLAYS and EUR/USD, accordingly. In Table 3.5.2 we report the dollar values of annualized average joint model risk of the 10-day (60-day) risk measures for BARCLAYS (EUR/USD) with the average annualized absolute values of ES for easy comparison, assuming that an investor holds a position of 1 million dollars in each asset. We extrapolate to multi-day risk estimates from daily risk measures using the "square root of time" rule[12], as recommended by the Basel Committee on Banking Supervision (2019) and following the practice of companies. The GARCH model with FZ minimization (GFZ) has the least model risk as compared to the average absolute values of the corresponding ES, whilst the EWMA model is the most affected by model risk when applied to these two assets. Comparing Table 3.5.1 and Table 3.5.2, the joint model risk of the 10-day risk measures for BARCLAYS is about 1-4 times as high as the joint model risk of daily risk measures. For the exchange rate, the joint model risk of the 60-day risk measures is about 2-4 times the joint model risk estimate of daily risk measures, so the dependence of the joint model risk estimate on the time horizon is not linear.

Table 3.5.2: Dollar values of annualized average joint model risk of multi-day risk measures

	BARCLAYS			EUR/USD		
Models	Avg. $\lvert ES\rvert$	S_1	S_3	Avg. $\lvert ES\rvert$	S_1	S_3
TFP	896,532	205,370	224,626	244,629	39,789	46,938
SKTFP	894,279	204,047	224,574	241,536	39,194	46,672
NFP	800,890	**109,065**	**123,362**	217,672	35,371	45,582
TEVT	901,541	210,180	229,640	235,270	31,383	37,881
SKTEVT	901,627	210,368	229,855	235,230	31,312	37,896
NEVT	903,136	190,638	199,306	234,576	32,074	39,447
EWMA	878,577	313,232	292,048	224,115	77,304	90,682
TFHS	899,498	202,466	219,875	236,164	31,090	37,378
SKTFHS	899,558	202,818	220,279	236,124	30,933	37,510
NFHS	902,109	183,341	188,568	235,521	31,826	39,587
GFZ	887,250	125,872	129,999	236,154	**20,894**	**20,485**
FZ1F	941,150	207,973	199,803	232,316	26,357	30,977
Hybrid	922,855	195,070	194,253	242,548	25,881	27,073
HS	1,063,175	314,956	284,969	253,760	48,129	51,934

This table reports the dollar values of annualized average joint model risk of the 10-day (60-day) risk measures at $\alpha = 2.5\%$ for BARCLAYS (EUR/USD) from 01/01/1980 to 20/02/2019, as compared to the average of absolute ES, assuming an investment of 1 million dollars in each asset. The values in bold are the lowest dollar values of average annualized joint model risk.

In Table 3.5.3 we present the backtesting results (1 for pass and 0 for failure) of several ES and VaR backtests for various risk models applied to the daily return series of BARCLAYS from 03/01/2000 to 01/01/2002 before and after the optimum multipliers, obtained based on $FZ0$, are used to improve the precision of the 2.5% daily VaR and ES estimates. With respect to VaR backtests, we use the likelihood ratio unconditional coverage test (UC_{VaR}) developed by Kupiec (1995) and the likelihood ratio conditional coverage test (CC_{VaR}) by Christoffersen (1998), which remain widely used (Nieto and Ruiz, 2016) amongst academics and practitioners. We also include the dynamic quantile regression-based test (DQ) for VaR, proposed by Engle and Manganelli (2004), considered a more effective VaR evaluation method (see Berkowitz et al. 2011). To backtest ES, we apply the exceedance residual test (ER) of McNeil and Frey (2000), the Z_2 test of Acerbi and Szekely (2014) as well as the unconditional/conditional coverage test (UC_{ES} and CC_{ES}) of Du and Escanciano (2016) to assess the frequency, magnitude and independence of excessive losses (see a detailed description of these ES backtests in Section 2.3.2). Our results show that adjusting for model risk does have a positive effect on backtesting, and the models suffering from model risk which fail the backtests can survive the backtesting procedure after adjusting the risk estimates for model risk, as indicated by 0^* in Table 3.5.3.

As expected, when increasing the α levels of the VaR and ES estimates, the joint model risk of the risk models decreases. This is illustrated in Table 3.5.4 which reports the average annualized $FZ0$-based joint model risk of risk measures

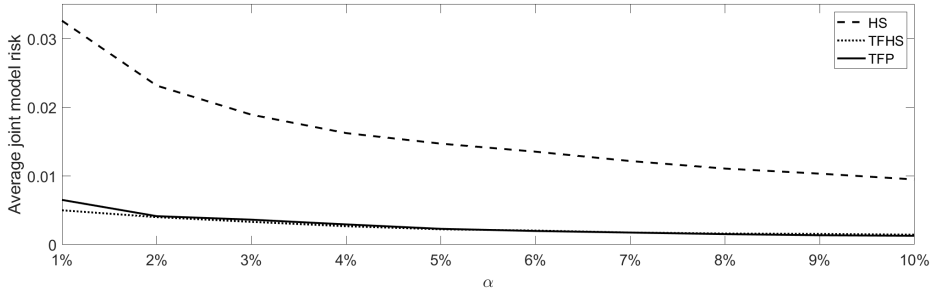
Table 3.5.3: Backtesting results before and after correcting for model risk

	ES backtests				VaR backtests		
Models	ER	Z_2	UC_{ES}	CC_{ES}	UC_{VaR}	CC_{VaR}	DQ
TFP	1	1	1	1	0^*	0^*	1
SKTFP	1	0	0	1	0	0	0
NFP	1	1	0	1	1	0^*	1
TEVT	1	0^*	1	1	0^*	0	0
SKTEVT	1	0^*	0	1	0^*	0^*	0^*
NEVT	1	0^*	0^*	1	0^*	0	0
EWMA	0^*	1	1	1	1	1	1
TFHS	1	0^*	1	1	0^*	0	0
SKTFHS	1	0^*	0	1	0^*	0^*	0^*
NFHS	1	0^*	0^*	1	0^*	0	1
GFZ	0^*	0	0	1	0	0	0
FZ1F	1	1	1	1	1	1	1
Hybrid	1	0^*	0	1	0^*	0	0^*
HS	1	1	1	1	1	0^*	0^*

This table presents the backtesting results of the 2.5% ES and VaR across various risk models for the daily returns of BARCLAYS from 03/01/2000 to 01/01/2002, before and after the optimized multipliers based on the $FZ0$ scoring function are applied to improve the risk estimates: 1 for pass, 0 for failure. 0^* *indicates that the model suffering from model risk which fails the backtest can survive the backtesting procedure after adjusting for model risk; The ES backtests are McNeil and Frey's exceedance residual test (ER), Acerbi and Szekely's Z_2 test (Z_2) as well as Du and Escanciano's unconditional/conditional coverage test (UC_{ES} and CC_{ES}); the VaR backtests are Kupiec's unconditional coverage test (UC_{VaR}), Christoffersen's conditional coverage test (CC_{VaR}) as well as Engle and Manganelli's dynamic quantile test (DQ). The backtest significance level is 5%.*

at 1%, 2.5% and 5% coverage levels across various models applied to the daily return series of BARCLAYS from 01/01/1980 to 20/02/2019. This is also highlighted in Figure 3.5.3 showing the average joint model risk of TFP, TFHS and HS against α levels. TFHS has less joint model risk than TFP at low coverage levels, since TFHS is better at capturing the extreme losses in the tail, while HS is the worst, as expected.

Figure 3.5.3: Average *FZ0*-based joint model risk along with multiple α levels



This figure shows the average joint model risk of models TFP, TFHS and HS against α levels, computed over 250 days, based on BARCLAYS from 01/01/1980 to 20/02/2019.

Beyond the scoring function-based joint model risk measure, we also examine the FZ-based ES model risk measure. Figure 3.5.4 displays the ratio of the $FZ0$-based ES model risk over the average of absolute ES at 2.5% critical level with a model risk evaluation window of 250 across various risk models applied to the daily return series of BARCLAYS from 01/01/1980 to 20/02/2019.

After the Lehman Brothers' collapse in 2008, the model risk of the 2.5% ES computed with the HS model, the EWMA model and the GARCH(1,1) with normally distributed innovations (NFP) inflate to more than 60%, 40% and 20%,

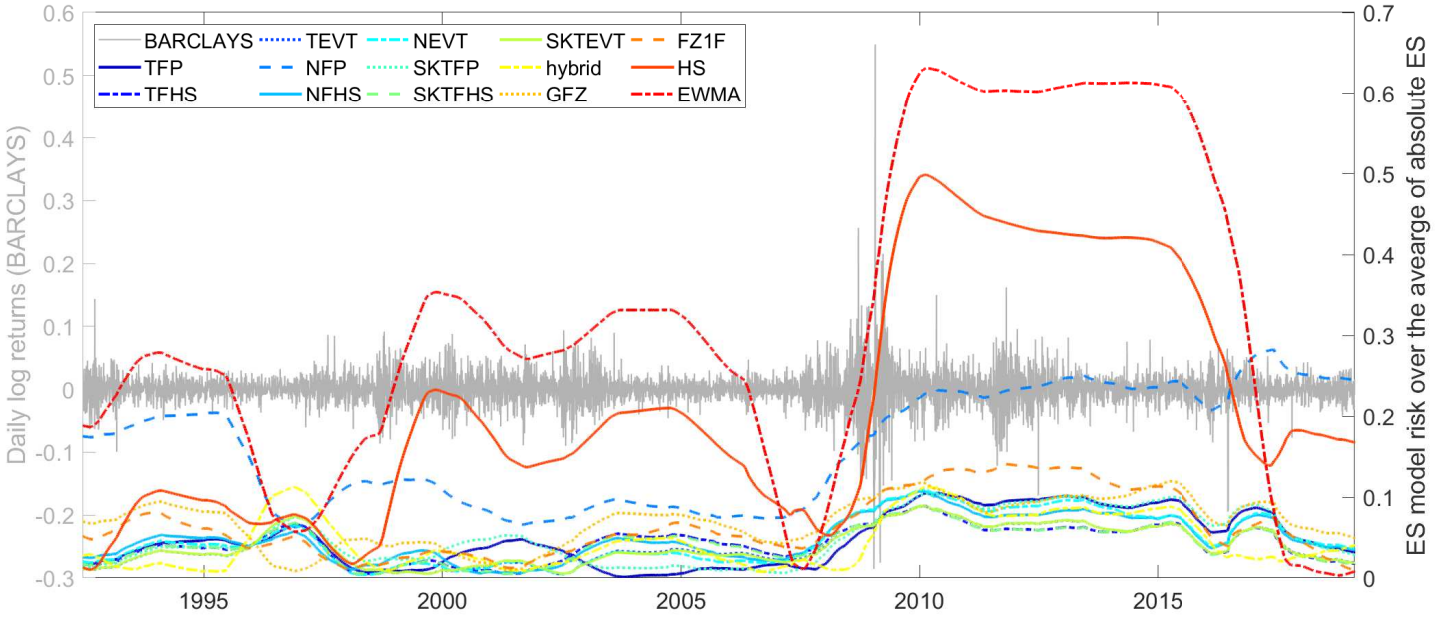
Table 3.5.4: Average annualized joint model risk of daily risk measures at several α levels

	Parametric						
α	TFP	SKTFP	NFP	TEVT	SKTEVT	NEVT	EWMA
1%	0.103	0.103	0.262	0.081	0.081	0.097	0.645
2.5%	0.062	0.065	0.146	0.056	**0.056**	0.067	0.376
5%	0.036	0.038	0.078	0.039	0.039	0.040	0.236

	Semi-parametric					Nonparametric
α	TFHS	SKTFHS	NFHS	GFZ	FZ1F	Hybrid	HS
1%	0.079	**0.078**	0.093	0.161	0.123	0.250	0.516
2.5%	0.058	0.057	0.065	0.084	0.082	0.080	0.329
5%	0.035	**0.035**	0.037	0.049	0.045	0.047	0.233

This table presents the average annualized FZ0-based joint model risk of various (VaR, ES) models at α = 1%, 2.5% and 5%, based on the return series of BARCLAYS from 01/01/1980 to 20/02/2019, over a model risk window of 250. Bold numbers represent the lowest values of average joint model risk for a given α.

Figure 3.5.4: Ratio of *FZ0*-based ES model risk over the average of absolute ES at 2.5% level



This figure shows the ratio of FZ0*-based ES model risk to the average of absolute ES at 2.5% level for a series of models, based on BARCLAYS returns from 01/01/1980 to 20/02/2019. The model risk window is 250.*

respectively, of the average of absolute ES in the evaluation window, whilst the ES model risk is only around 10% of the average absolute ES for the other models, see Figure 3.5.4. Moreover, in Table 3.5.5 we report the average ratio of ES model risk associated with the 0 and -1 homogeneous FZ scoring functions (S_1 and S_3) over the absolute value of average ES at 2.5% level for various assets over the same set of models and evaluation window of $n_1 = 250$. Generally, the FZ-based ES model risk of risk models applied to the S&P500 Index is much higher compared to other asset classes. Except for the worst-performing models (EWMA and HS), the average ratio of ES model risk over the average absolute ES varies between 2% to 10%. Similarly to the joint model risk, the S_3-based ES model risk measure is generally higher than the S_1-based ES model risk measure.

3.6 Conclusions

In this chapter, we disentangle the components of model risk of financial market risk models based on strictly consistent FZ scoring functions applied to the risk functionals (VaR, ES). We show that, when model risk is present, the ordering of (VaR, ES) models is sensitive to the FZ specification function, although the model ranking is not sensitive to the choice of homogeneous FZ scoring function when the pair of (VaR, ES) is estimated at small critical levels (e.g., 2.5%). Instead of focusing on model comparison, we quantify the joint model risk of (VaR, ES) risk models and also the ES model risk solely, based on the FZ scoring

Table 3.5.5: Average ratio of ES model risk over the average of absolute ES

	BARCLAYS		S&P500		GOLD		GSCI		EUR/USD	
Models	S_1	S_3	S_1	S_3	S_1	S_3	S_1	S_3	S_1	S_3
TFP	4.8%	4.7%	8.3%	10.3%	4.4%	5.0%	4.6%	5.4%	5.1%	5.4%
SKTFP	5.1%	5.0%	5.3%	6.8%	4.8%	5.0%	4.2%	5.4%	3.4%	3.3%
NFP	16.1%	17.0%	24.0%	25.8%	23.8%	24.9%	14.1%	14.3%	8.8%	10.3%
TEVT	4.0%	4.4%	5.8%	8.5%	4.6%	7.1%	4.2%	4.3%	2.7%	4.0%
SKTEVT	4.0%	4.4%	5.8%	8.6%	4.7%	7.1%	4.1%	4.3%	2.7%	4.0%
NEVT	4.8%	5.6%	6.5%	8.9%	4.4%	6.4%	4.3%	4.2%	2.6%	3.8%
EWMA	31.8%	33.9%	40.3%	46.0%	29.1%	30.1%	26.5%	26.6%	14.2%	15.3%
TFHS	4.2%	5.0%	5.9%	8.6%	4.6%	6.9%	4.7%	4.7%	2.4%	3.5%
SKTFHS	4.2%	4.9%	5.9%	8.7%	4.6%	6.9%	4.7%	4.8%	2.5%	3.5%
NFHS	4.9%	6.0%	6.5%	8.8%	4.2%	6.1%	4.6%	4.4%	2.3%	3.4%
GFZ	6.9%	8.6%	6.4%	8.1%	5.0%	7.5%	7.1%	7.6%	3.1%	4.7%
FZ1F	6.5%	9.3%	8.7%	10.7%	7.9%	9.3%	7.5%	8.6%	3.7%	5.0%
Hybrid	4.5%	4.3%	6.0%	7.2%	7.3%	7.8%	6.2%	6.0%	2.8%	3.1%
HS	21.0%	23.5%	20.8%	24.6%	12.1%	12.7%	15.3%	17.3%	6.4%	7.2%

This table reports the average ratio of S_1 and S_3-based ES model risk of ES measures at 2.5% level over the absolute value of average ES for various assets and a set of models. The ES model risk is computed in a model risk evaluation window of $n_1 = 250$, based on historical daily data for all assets from 01/01/1980 to 20/02/2019.

functions. The proposed model risk methodology framework is confirmed with a simulation study in which we use three specific FZ scoring functions which are 0, 0.5 and -1 positively homogenous. We find a high similarity between the true and estimated model risk of (VaR, ES) risk measures as well as for the ES model risk, across various risk models, with correlations varying from 0.8 to 0.987, with an explanatory power above 50%.

In our simulation analysis, the newly proposed measures of joint model risk and ES model risk satisfy numerically all coherence properties of a measure of risk, except for the subadditivity property. This essential property, sometimes called the diversification of risk property, is not always satisfied numerically but it holds true in most of our simulated scenarios when risk measures are estimated at small α levels (e.g., 1% and 2.5%). The empirical results point out that the EWMA model and Historical Simulation have a very high level of joint model risk and ES model risk among all models considered, particularly during extreme events. In addition, the backtesting performance of these models is improved upon adjusting for model risk.

The scoring function-based model risk methodology could facilitate other extensions for quantifying the model risk of risk models. For instance, by replacing the FZ class with the GPL class, the individual VaR model risk can be examined in a similar manner and this may invite further research. Other interesting future research could consider the model risk of using the "square root of time" rule and also consider theoretical proofs for the properties of model risk measures.

Appendices

3.A Proofs

The following proofs are adapted from Patton (2019).

Proof. Proof of Proposition 1a). Here we show that under assumptions 1)-3), $\mathbb{E}\left[S^{\alpha}_{FZ0}(r_t, \hat{v}^B_t, \hat{e}^B_t)\right] \geq \mathbb{E}\left[S^{\alpha}_{FZ0}(r_t, \hat{v}^A_t, \hat{e}^A_t)\right]$ implies that $\mathcal{F}^B_t \subseteq \mathcal{F}^A_t$ for all t, which in turn implies $\mathbb{E}\left[S^{\alpha}_{FZ}(r_t, \hat{v}^B_t, \hat{e}^B_t)\right] \geq \mathbb{E}\left[S^{\alpha}_{FZ}(r_t, \hat{v}^A_t, \hat{e}^A_t)\right]$, where S_{FZ} is any loss function from the FZ class.

First, we prove that $\mathbb{E}\left[S^{\alpha}_{FZ0}(r_t, \hat{v}^B_t, \hat{e}^B_t)\right] \geq \mathbb{E}\left[S^{\alpha}_{FZ0}(r_t, \hat{v}^A_t, \hat{e}^A_t)\right] \Longrightarrow \mathcal{F}^B_t \subseteq \mathcal{F}^A_t$ for all t.

Starting from $\mathbb{E}\left[S^{\alpha}_{FZ0}(r_t, \hat{v}^B_t, \hat{e}^B_t)\right] \geq \mathbb{E}\left[S^{\alpha}_{FZ0}(r_t, \hat{v}^A_t, \hat{e}^A_t)\right]$ we assume that $\mathcal{F}^A_t \subseteq \mathcal{F}^B_t$ for a t. This implies that $\mathbb{E}\left[S^{\alpha}_{FZ0}(r_t, \hat{v}^A_t, \hat{e}^A_t)|\mathcal{F}^B\right] \geq \mathbb{E}\left[S^{\alpha}_{FZ0}(r_t, \hat{v}^B_t, \hat{e}^B_t)|\mathcal{F}^B\right]$ *a.s.* for t, since $(\hat{v}^A_t, \hat{e}^A_t) \in \mathcal{F}^A \subseteq \mathcal{F}^B$, and thus $\mathbb{E}\left[S^{\alpha}_{FZ0}(r_t, \hat{v}^A_t, \hat{e}^A_t)\right] \geq \mathbb{E}\left[S^{\alpha}_{FZ0}(r_t, \hat{v}^B_t, \hat{e}^B_t)\right]$ by the Law of Iterated Expectations.

The inequality $\mathbb{E}\left[S^{\alpha}_{FZ0}(r_t, \hat{v}^A_t, \hat{e}^A_t)\right] \geq \mathbb{E}\left[S^{\alpha}_{FZ0}(r_t, \hat{v}^B_t, \hat{e}^B_t)\right]$ can be satisfied if and only if $\mathbb{E}\left[S^{\alpha}_{FZ0}(r_t, \hat{v}^A_t, \hat{e}^A_t)|\mathcal{F}^B\right] = \mathbb{E}\left[S^{\alpha}_{FZ0}(r_t, \hat{v}^B_t, \hat{e}^B_t)|\mathcal{F}^B\right]$ a.s. for t.

The FZ class is restricted to the assumption that any pair of cumulative distri-

bution function F_t^i, $i \in \{A, B\}$ from this class, are strictly increasing with unique α-quantiles (Fissler and Ziegel, 2016). Let $\hat{v}_t^i$ be the unique solution to $\alpha = F_t^i(v_t^i)$, and $\hat{e}_t^i = \frac{1}{\alpha}\int_{-\infty}^{\hat{v}_t^i} x f_r^i(x)dx$, for $i \in \{A, B\}$. The necessity and sufficiency of strict consistency of the FZ class for joint VaR and ES estimation (see details in Fissler and Ziegel, 2016), including $FZ0$, implies that $\hat{v}_t^i$ is the solution to the following minimization problem:

$$\hat{v}_t^i = \arg\min_{\hat{v}} \mathbb{E}\left[S_{FZ0}^{\alpha}(r_t, \hat{v}, \hat{e})|\mathcal{F}_t^i, \hat{e} = \frac{1}{\alpha}\int_{-\infty}^{\hat{v}} x f_r^i(x)dx\right], \qquad for\ i \in \{A, B\} \tag{3.A.1}$$

Corroborating this with $\mathbb{E}\left[S_{FZ0}^{\alpha}(r_t, \hat{v}_t^A, \hat{e}_t^A)|\mathcal{F}^B\right] = \mathbb{E}\left[S_{FZ0}^{\alpha}(r_t, \hat{v}_t^B, \hat{e}_t^B)|\mathcal{F}^B\right]$ a.s. for all t, leads to $(\hat{v}_t^A, \hat{e}_t^A) = (\hat{v}_t^B, \hat{e}_t^B)$. However, the last equality is in contradiction with assumption 1) that the nested information sets do not give identical optimal forecasts.

Next, we are going to prove that $\mathcal{F}_t^B \subseteq \mathcal{F}_t^A \Longrightarrow \mathbb{E}\left[S_{FZ}^{\alpha}(r_t, \hat{v}_t^B, \hat{e}_t^B)\right] \geq \mathbb{E}\left[S_{FZ}^{\alpha}(r_t, \hat{v}_t^A, \hat{e}_t^A)\right]$, where S_{FZ} is any loss function of the FZ class. From the necessary condition of strictly consistent FZ scoring functions (Fissler and Ziegel, 2016), for any $S^{\alpha} \in S_{FZ}^{\alpha}$, the solution to the minimization of expected losses based on any strict consistent function with the FZ class is defined in equation (3.A.1). It is straightforward to show that $\alpha = \mathbb{E}\left[\mathbb{1}\{r_t \leq \hat{v}_t^i\}|\mathcal{F}_t^i\right]$ and that this condition holds for all possible distributions of r_t. Since $\mathcal{F}_t^B \subseteq \mathcal{F}_t^A$ for all t, $\mathbb{E}\left[S^{\alpha}(r_t, \hat{v}_t^B, \hat{e}_t^B)|\mathcal{F}_t^B\right] \geq \mathbb{E}\left[S^{\alpha}(r_t, \hat{v}_t^A, \hat{e}_t^A)|\mathcal{F}_t^A\right]$, for any $S^{\alpha} \in S_{FZ}^{\alpha}$. Applying the Law of Iterated Expectations, we have $\mathbb{E}\left[S^{\alpha}(r_t, \hat{v}_t^B, \hat{e}_t^B)\right] \geq \mathbb{E}\left[S^{\alpha}(r_t, \hat{v}_t^A, \hat{e}_t^A)\right]$, for all $S^{\alpha} \in S_{FZ}^{\alpha}$. □

Proof. Proof of Proposition 1b). In the following, we give the analytical proofs for Proposition *1b)* under three scenarios.

(i) First, when the information sets are non-nested violating assumption 1), we consider a simple example below: $Y = -(X + Z)$, where X is uniformly distributed as $Unif(0, 10)$, Z has a triangular distribution $Tri(0, 12)$, and X and Z are independent.

Given that $\alpha = 50\%$, we assume that the risk estimates based on model A condition on X and those based on model B condition on Z. Since X and Z are independent, then:

$$\hat{v}^a = -(X + Median[Z]), \quad Median[Z] = M[Z] \approx 3.51,$$

$$\hat{e}^a = \mathbb{E}\left[Y \mid Y \leq \hat{v}^a\right] = -\mathbb{E}\left[X + Z \mid Z \geq M[Z]\right] = -\mathbb{E}[X] - \mathbb{E}\left[Z \mid Z \geq M[Z]\right],$$

$$\hat{v}^b = -(Z + Median[X]), \quad Median[X] = M[X] = 5,$$

$$\hat{e}^b = \mathbb{E}\left[Y \mid Y \leq \hat{v}^b\right] = -\mathbb{E}\left[X + Z \mid X \geq M[X]\right] = -\mathbb{E}[Z] - \mathbb{E}\left[X \mid X \geq M[X]\right].$$

For a critical level α, considering two increasing continuously differentiable functions G_1 and G_2 such that $\mathbb{E}[G_1(X)]$ exists, $\lim_{x\to-\infty} G_2(x) = 0$ and $\mathcal{G}_2' = G_2$, and a realization denoted by r, the class of FZ scoring functions is as follows:

$$\begin{aligned} S_{FZ}(r, v, e; \alpha, G_1, G_2) \;=\; & \left(\mathbb{1}_{\{r\leq v\}} - \alpha\right)\left(G_1(v) - G_1(r)\right) \\ & + G_2(e)\left(\frac{1}{\alpha}\mathbb{1}_{\{r\leq v\}}(v - r) - (v - e)\right) - \left(\mathcal{G}_2(e) - \mathcal{G}_2(r)\right) \end{aligned} \tag{3.A.2}$$

Next, within the FZ class we calculate the expected losses, $\bar{S}_0$ and $\bar{S}_1$, depending on $G_1(z)$ and $\mathcal{G}_2(z)$ as in (3.A.2). Let $G_1(z) = 0$, $G_2(z) = -1/z$, and $\mathcal{G}_2(z) = -\log(-z)$, and thus $\bar{S}_0$ is the expected loss. Also, taking $G_1(z) = 0$, $G_2(z) = -1/z^2$, and $\mathcal{G}_2(z) = -1/z$, allows the computation of $\bar{S}_1$. The expected loss $\bar{S}_0$ (the same expression is obtained for $\bar{S}_1^A$) associated with model A is:

$$\begin{aligned}\bar{S}_0^A(y, \hat{v}^a, \hat{e}^a; \alpha, G_1, G_2) = \frac{G_2(e)}{\alpha}\mathbb{E}\left[\mathbb{1}\left\{-(X+Z) \leq -(X+M[Z])\right\} \cdot (Z - M[Z])\right] \\ + eG_2(e) - \mathcal{G}_2(e) - G_2(e)\mathbb{E}[v] \\ = \frac{G_2(e)}{\alpha}\mathbb{E}\left[\mathbb{1}\left\{Z \geq M[Z]\right\} \cdot (Z - M[Z])\right] + eG_2(e) \\ - \mathcal{G}_2(e) - G_2(e)\mathbb{E}[-(X + M[Z])].\end{aligned}$$

For model B, the expected loss $\bar{S}_0^B$ (the same expression is obtained for $\bar{S}_1^B$) is computed below:

$$\begin{aligned}\bar{S}_0^B(y, \hat{v}^b, \hat{e}^b; \alpha, G_1, G_2) = \frac{G_2(e)}{\alpha}\mathbb{E}\left[\mathbb{1}\left\{X \geq M[X]\right\} \cdot (X - M[X])\right] + eG_2(e) \\ - \mathcal{G}_2(e) - G_2(e)\mathbb{E}[-(Z + M[X])].\end{aligned}$$

Finally we obtain that:

$$\bar{S}_0^A = 2.489 \ < \ \bar{S}_0^B = 2.532; \qquad \bar{S}_1^A = -0.170 \ > \ \bar{S}_1^B = -0.191.$$

(ii) Secondly, our setup is that based on the nested information sets, the

risk models (A and B) are subject to estimation error, though they are correctly specified: we have that $Y = -(X+Z)$, where X and Z are independent, uniformly distributed as $Unif(-10, 0)$ and $Unif(0, 12)$, respectively.

Then, the probability density function of Y is easily derived as:

$$f_Y(y) = \begin{cases} \frac{12+y}{120}, & for \ -12 < y < -2, \\ \frac{1}{12}, & for \ -2 < y < 0, \\ \frac{10-y}{120}, & for \ 0 < y < 10. \end{cases}$$

We compute risk estimates at $\alpha = 50\%$. model A gives optimal risk estimates without any conditioning information as below:

$$\hat{v}^a = Median[Y] = M_y = -1, \quad \hat{e}^a = \mathbb{E}\left[Y \mid Y \leq \hat{v}^a\right] = \mathbb{E}\left[Y \mid Y \leq M_y\right].$$

Model B conditions on Z, and makes the risk estimates by estimating $Median[X]$. Forecaster B estimates $Median[X]$ using $n = 1$ observation of X. Since X and Z are independent, the risk estimates are predicted as:

$$\hat{v}^b = -(\tilde{X} + Z),$$
$$\hat{e}^b = \mathbb{E}\left[Y \mid Y \leq \hat{v}^b\right] = -\mathbb{E}\left[X + Z \mid X \geq \tilde{X}\right] = -\mathbb{E}\left[X \mid X \geq \tilde{X}\right] - \mathbb{E}\left[Z\right].$$

To compute $\hat{e}^b$, we will use the result below, since $\tilde{X}$ and X have the same

distribution:

$$\mathbb{E}\left[X \mid X \geq \tilde{X}\right] = \mathbb{E}\left[X \cdot \mathbf{1}\{X \geq \tilde{X}\}\right] = \mathbb{E}\left[\mathbb{E}\left[\mathbf{1}\{X \geq \tilde{X}\} \mid X\right] X\right]$$
$$= \mathbb{E}\left[F_{\tilde{x}}(X)X\right] = \mathbb{E}\left[F_x(X)X\right].$$

Therefore, $\hat{e}^b = -\mathbb{E}\left[F_x(X) \cdot X\right] - \mathbb{E}[Z]$. For $X \sim Unif(L, U)$, we have $\mathbb{E}\left[F_x(X)X\right] = \frac{1}{6} \cdot (L + 2U)$.

The expected loss $\bar{S}_0$ (the same expression is obtained for $\bar{S}_1^A$) associated with model A:

$$\bar{S}_0^A(y, \hat{v}^a, \hat{e}^a; \alpha, G_1, G_2) = \frac{G_2(e)}{\alpha}\mathbb{E}\left[\mathbb{1}\left\{Y \leq M_y\right\} \cdot (M_y - Y)\right]$$
$$+ eG_2(e) - \mathcal{G}_2(e) - G_2(e)M_y.$$

For model B, we calculate the expected loss $\bar{S}_0^B$ (the same expression is obtained for $\bar{S}_1^B$) as follows:

$$\bar{S}_0^B(y, \hat{v}^b, \hat{e}^b; \alpha, G_1, G_2) = \frac{G_2(e)}{\alpha}\mathbb{E}\left[\mathbb{1}\left\{X \geq \tilde{X}\right\} \cdot (X - \tilde{X})\right] + eG_2(e)$$
$$- \mathcal{G}(e) - G_2(e)\mathbb{E}[-(\tilde{X} + Z)]$$

To compute $\bar{S}_0^B$ and $\bar{S}_1^B$, since $\tilde{X}$ and X have the same distribution, we will use

that

$$\mathbb{E}\left[\tilde{X} \mid X \geq \tilde{X}\right] = \mathbb{E}\left[\tilde{X} \cdot \mathbf{1}\{X \geq \tilde{X}\}\right] = \mathbb{E}\left[\mathbb{E}\left[\mathbf{1}\{X \geq \tilde{X}\} \mid \tilde{X}\right] \tilde{X}\right]$$
$$= \mathbb{E}\left[(1 - F_x(\tilde{X})) \cdot \tilde{X}\right] = \mathbb{E}[X] - \mathbb{E}\left[F_x(X)X\right].$$

Thus we get that:

$$\bar{S}_0^A = 1.853 \ > \ \bar{S}_0^B = 1.466; \qquad \bar{S}_1^A = -0.852 \ < \ \bar{S}_1^B = -0.231.$$

(iii) Finally, we consider the case of misspecified models, although these models are without estimation error given the nested information sets. For simplicity, assume that the DGP is $Y = -X, X \sim Unif(0, 10)$. The parameters of the linear models A and B (subject to misspecification error) are different from (0,1), and let $(\beta_0, \beta_1) = (0.33, 0.67)$ and $(\gamma_0, \gamma_1) = (-0.25, 1.25)$ for A and B, respectively. In this example, models A and B are conditioning on the same information set and they are free of estimation error, predicting the risk estimates at $\alpha = 50\%$ as follows:

$$\hat{v}^a = -\beta_0 - \beta_1 X, \quad \hat{e}^a = -\mathbb{E}\left[X \cdot \mathbf{1}\left\{(1 - \beta_1)X \geq \beta_0\right\}\right],$$
$$\hat{v}^b = -\gamma_0 - \gamma_1 X, \quad \hat{e}^b = -\mathbb{E}\left[X \cdot \mathbf{1}\left\{(1 - \gamma_1)X \geq \gamma_0\right\}\right].$$

In the following, we will use that:

$$\mathbb{E}\left[\mathbf{1}\left\{(1-\beta_1)X \geq \beta_0\right\}\right] = \begin{cases} 1 - F_x\left(\frac{\beta_0}{1-\beta_1}\right), & for\ \beta_1 < 1, \\ F_x\left(\frac{\beta_0}{1-\beta_1}\right), & for\ \beta_1 > 1, \\ \mathbf{1}\{\beta_0 \leq 0\}, & for\ \beta_1 = 1. \end{cases}$$

$$\mathbb{E}\left[X \cdot \mathbf{1}\left\{(1-\beta_1)X \geq \beta_0\right\}\right] = \begin{cases} \mathbb{E}\left[X \cdot \mathbf{1}\{X \geq \frac{\beta_0}{1-\beta_1}\}\right], & for\ \beta_1 < 1, \\ \mathbb{E}\left[X \cdot \mathbf{1}\{X \leq \frac{\beta_0}{1-\beta_1}\}\right], & for\ \beta_1 > 1, \\ \mathbb{E}[X] \cdot \mathbf{1}\{\beta_0 \leq 0\}, & for\ \beta_1 = 1. \end{cases}$$

A similar expression is obtained for model B by replacing (β_0, β_1) with (γ_0, γ_1). The expected loss $\bar{S}_0$ (the same expression holds for $\bar{S}_1^A$) associated with model A is derived as:

$$\begin{aligned} \bar{S}_0^A(y, \hat{v}^a, \hat{e}^a; \alpha, G_1, G_2) = (1-\beta_1)\frac{G_2(e)}{\alpha}\mathbb{E}[X \cdot \mathbb{1}\{X \geq \frac{\beta_0}{1-\beta_1}\}] \\ - \beta_0\frac{G_2(e)}{\alpha}\mathbb{E}[\mathbb{1}\{X \geq \frac{\beta_0}{1-\beta_1}\}] \\ + \beta_0 G_2(e) + \beta_1 G_2(e)\mathbb{E}[X] + eG_2(e) - \mathcal{G}_2(e). \end{aligned}$$

In case of model B, the expected loss $\bar{S}_0^B$ (the same expression holds for $\bar{S}_1^B$) is:

$$\begin{aligned}\bar{S}_0^B(y, \hat{v}^b, \hat{e}^b; \alpha, G_1, G_2) = (1-\gamma_1)\frac{G_2(e)}{\alpha}\mathbb{E}[X \cdot \mathbb{1}\{X \leq \frac{\gamma_0}{1-\gamma_1}\}] \\ - \gamma_0\frac{G_2(e)}{\alpha}\mathbb{E}[\mathbb{1}\{X \leq \frac{\gamma_0}{1-\gamma_1}\}] \\ + \gamma_0 G_2(e) + \gamma_1 G_2(e)\mathbb{E}[X] + eG_2(e) - \mathcal{G}_2(e).\end{aligned}$$

Numerically, the results of this example conclude our proof:

$$\bar{S}_0^A = 1.883 \ < \ \bar{S}_0^B = 116.504; \qquad \bar{S}_1^A = -0.259 \ \ \bar{S}_1^B = -2410.$$

□

3.B Risk measurement models

In our investigations we use a set of widely known risk models considered in Nolde and Ziegel (2017a): the nonparametric method includes Historical Simulation (HS); the semi-parametric methods include the GARCH(1,1) models with the normal, standardised Student's t, and skewed t innovations, combined with Filtered Historical Simulation (NFHS, TFHS, and SKTFHS). The parametric methods include the GARCH(1,1) processes with the normal, standardised Student's t, and skewed t distributed innovations (NFP, TFP, and SKTFP), as well as the same distributions combined with the Extreme Value Theory methodol-

ogy (NEVT, TEVT, and SKTEVT). In addition, we include the newly proposed semiparametric models based on $FZ0$ minimization of Patton et al. (2019) (FZ1F, GFZ and Hybrid), and add the EWMA model to the set of parametric approaches. These risk estimation methods are used to compute the ex ante one step ahead VaR and ES measures, at a given significance level, using rolling windows of length $L = 1,000$.

3.B.1 Nonparametric approaches

Relying on the historical data series of returns $\{r\}$, we use the easy-to-implement Historical Simulation model (HS) to compute model-free the daily VaR and ES at α significance level at time t over the previous $L = 1,000$ observations:

$$\hat{v}_t^{HS} = percentile\left\{\{r\}_{t-L}^{t-1}, 100\alpha\right\}, \quad \hat{e}_t^{HS} = \frac{\sum_{i=t-L}^{t-1} r_i \cdot \mathbb{1}\{r_i \leq \hat{v}_t^{HS}\}}{\sum_{i=t-L}^{t-1} \mathbb{1}\{r_i \leq \hat{v}_t^{HS}\}}. \tag{3.B.1}$$

3.B.2 Semiparametric approaches

For the GARCH(1,1)-FHS models, we incorporate the GARCH(1,1) processes with the normal, standardised Student's t, and skewed t disturbances, keeping the non-parametric nature of Historical Simulation in these disturbances. The risk measures by NFHS, TFHS, and SKTFHS are estimated in section 2.3.

Patton et al. (2019) propose several semiparametric models in which the model parameters are estimated by minimizing the $FZ0$ scoring function. We use the

one-factor GAS model (denoted by FZ1F), the GARCH model via FZ minimization (denoted by GFZ) as well as the hybrid GAS/GARCH model (denoted by Hybrid) they propose.

3.B.3 Parametric approaches

The EWMA model (the RiskMetrics model) is a simple variance model that captures the persistence of a shock to the variance dynamics. The risk estimates under EWMA at time t are

$$\hat{v}_t^{EWMA} = \hat{\mu}_t + \hat{\sigma}_t \Phi^{-1}(\alpha), \qquad \hat{e}_t^{EWMA} = \hat{\mu}_t + \frac{\hat{\sigma}_t}{\alpha} \int_0^{\alpha} \Phi^{-1}(x) dx \tag{3.B.2}$$

where $\hat{\mu}_t$ is the average return within the estimation window and the conditional variance is estimated as: $\hat{\sigma}_t^2 = (1-\lambda) r_{t-1}^2 + \lambda \hat{\sigma}_{t-1}^2$, with $\lambda = 0.94$.

We also compute the VaR and ES measures for the GARCH(1,1) models:

$$\begin{aligned} r_t &= \hat{\sigma}_t Z_t, \quad Z_t \sim F \\ \hat{\sigma}_t^2 &= \hat{\omega} + \hat{\beta} r_{t-1}^2 + \hat{\gamma} \sigma_{t-1}^2, \quad where\ \hat{\beta} + \hat{\gamma} < 1, \end{aligned} \tag{3.B.3}$$

F denotes a cumulative Normal, Student's t or Skewed t distribution for the residuals. The parameters $\hat{\omega}$, $\hat{\beta}$ and $\hat{\gamma}$ are estimated via Maximum Likelihood Estimation in a moving window of $L = 1,000$. Subsequently, the VaR and ES

estimates are written as:

$$\hat{v}_t^{FP} = \hat{\sigma}_t F^{-1}(\alpha), \qquad \hat{e}_t^{FP} = \frac{\hat{\sigma}_t}{\alpha} \int_0^{\alpha} F^{-1}(x)dx. \tag{3.B.4}$$

For the extreme value approach, we fit the GPD distribution to the exceedances beyond the threshold in the standardised residuals obtained by the GARCH(1,1) process with various innovations. The risk forecasts are displayed in section 2.3.

Additionally, to allow for switching market states, the extended Markov Switching GARCH(1,1) model with normal disturbances (Klaassen, 2002) is also employed:

$$r_t = \sqrt{\hat{h}_{s_t}} Z_t, \quad Z_t \sim iid\ N(0,1), \quad s_t = \{1,2\}, \tag{3.B.5}$$

where $\hat{h}_{s_t} = \hat{\omega}_{s_t} + \hat{\beta}_{s_t} r_{t-1}^2 + \hat{\gamma}_{s_t} \cdot \sum_{i=1}^{2} p_{ij} \hat{h}_{i,t-1}, \hat{\omega}_{s_t}, \hat{\beta}_{s_t}$ and $\hat{\gamma}_{s_t} > 0$; $s_t = 1$ or 2, showing the possible market state at time t; p_{ij} denotes the probability of state j at time t conditional that the market is in state i at time $t-1$ and $\hat{h}_{i,t-1}$ denotes the conditional variance dynamics in state i at time $t-1$. In this case, we write the VaR and ES estimates as:

$$\hat{v}_t^{MS} = \sqrt{\hat{h}_{s_t}} \Phi^{-1}(\alpha), \qquad \hat{e}_t^{MS} = \frac{\sqrt{\hat{h}_{s_t}}}{\alpha} \int_0^{\alpha} \Phi^{-1}(x)dx. \tag{3.B.6}$$

Notes

[1]The primary focus of this chapter is on the standard regulatory VaR and ES measures, but other variants like expectile-based value-at-risk (expected shortfall) (Newey and Powell, 1987; Kuan et al., 2009) and mark to market value-at-risk (MMVaR) (Boudoukh et al., 2004; Chen et al., 2019) have been discussed in the academic literature. Detering and Packham (2016) propose a model risk measure applicable to derivatives contracts trading that improves over the price range measure introduced in Cont (2006) as a yardstick of model contingent claim pricing uncertainty. The latter is incompatible with regulatory capital charges while the former can be used for reserve buffer computations and it is based on value-at-risk or expected shortfall.

[2]See Embrechts et al. (2014) for a comprehensive discussion.

[3]See more properties of scoring functions in Davis (2016) and Gneiting (2011).

[4]In (3.A.2), $\mathbb{1}$ denotes the indicator function; the first summand is the GPL family and only depends on VaR; the second summand depends on VaR and ES. That is, ES is not elicitable per se, but jointly elicitable with VaR. On most occasions, $\mathcal{G}_2(r)$ is disregarded (see Nolde and Ziegel 2017a).

[5]The simulation study undertaken by Nolde and Ziegel (2017a) shows that the values of the constants are irrelevant.

[6]Here we used the Matlab codes (http://public.econ.duke.edu/ ap172/) provided by Patton et al. (2019), for which we are very grateful.

[7]The optimization to find a constrained minimum of a multivariate function shown in equation (3.3.2) is done in MATLAB by implementing the 'fmincon' function with the 'sqp' algorithm which guarantees the existence of a solution, as discussed in Nocedal and Wright (2006).

[8]In this chapter, we compute the optimized multipliers using a multiplier estimation window length of 2,000 in order to reduce the effect of data noise (Nolde and Ziegel, 2017a,b). That is, $\tau = 1,999$ throughout this chapter.

[9]We also compute the model risk over the course of a four-year backtesting period, around 1,000 trading days. The results are available upon request.

[10]The alternative measure replacing the formulation in (3.3.3) with the RMSE type produces similar results.

[11]We consider VaR and ES as negative risk measures throughout this chapter.

[12]An excellent discussion on the "square root of time" rule used to compute multi-day risk measures can be found in Diebold et al. (1997) and Daníelsson and Zhou (2017).

Chapter 4

Model Risk of Volatility Models

4.1 Introduction

Volatility forecasting often constitutes a significant impact in many applications, for example, in derivatives pricing, statistical risk measure estimation and investment decision-making. If the volatility forecast is wrong, then the implications can be widespread. The existing enormous volatility modeling literature covers the family of autoregressive conditional heteroscedasticity (ARCH) models, stochastic volatility models as well as volatility models based on realized data, in a univariate or multivariate setting (see an extensive overview of volatility models in Bauwens et al., 2012). This chapter contributes to the line of volatility modeling literature in measuring and managing model risk numerically.

The primary issue in evaluating the accuracy of volatility models is that the target variable (e.g., the true variance denoted by σ^2) is unobservable and latent

(Hansen and Lunde, 2006 and Patton, 2011). This is addressed by using a conditionally unbiased variance estimator of the true conditional variance (hereafter, also called the volatility proxy and denoted by $\hat{\sigma}^2$), namely the daily squared return, the realized variance, or the range-based variance to name the main ones (see Alizadeh et al., 2002; Barndorff-Nielsen and Shephard, 2002, and Andersen et al., 2003).

One strand of the volatility forecasting literature focuses on the accuracy of a single model. A simple and well-known approach to evaluate the accuracy of a single volatility model is the Mincer and Zarnowitz (1969) (MZ) regression. This method[1] regresses the conditionally unbiased proxy ($\hat{\sigma}_t^2$) for the true variance on the variance forecast (h_t) of a given model and estimates an intercept parameter (α) (indicating systematic over/under-estimation) and a coefficient (β), and it is written as $\hat{\sigma}_t^2 = \alpha + \beta h_t + e_t$. The null hypothesis of the forecast optimality is that $H^0 : \alpha = 0$ and $\beta = 1$. The R^2 of the regression equation is considered as a criterion for the accuracy (efficiency) of the volatility forecasting model. Instead of evaluating a single model, a second strand considers model comparisons based on scoring functions. The pairwise comparisons between two competing forecasts (see the tests of Diebold and Mariano, 1995 and West, 1996 as well as a general discussion in Giacomini and White, 2006) and the multiple comparisons among volatility models (e.g., Hansen and Lunde, 2005, and Hansen et al., 2011) have been well-documented.

The drawback of the MZ regression and the pairwise comparison tests is that

the noisy volatility proxy may distort the results, as argued by Hansen and Lunde (2006) and Patton (2011). To solve this problem, Patton (2011) proposes a class of robust and homogeneous scoring functions for the volatility, which leads to an invariant inference in the ranking of competing models to the choice of volatility proxy. Within the proposed family of scoring functions, the mean square error (MSE) and QLIKE scoring functions are widely accepted for the evaluation of volatility forecasting models as in Forsberg and Ghysels (2007), Bauwens et al. (2012), Engle and Siriwardane (2018) and others.

Although an extensive study of volatility forecast comparisons has been conducted around the average loss, or distance between the estimated variances of competing models over a forecasting period (e.g., Patton, 2011 and Hansen and Lunde, 2005), much less is known about the exact magnitude of model risk of the volatility forecast of a given model. Since the true volatility is never known in practice, any volatility model is considered to be exposed to unobserved and implicit model risk associated with the distance between the raw volatility estimates to the true volatility. We approximate this type of model risk based on the distance between the imperfect variance estimates and the improved variance estimates based on the MSE or QLIKE loss function, thus facilitating model risk management for volatility models.

We develop a model risk estimation methodology for volatility models, considering the choice of scoring function (MSE or QLIKE) and the effect of volatility proxy. We estimate the model risk as the average distance between the raw

and improved variance estimates over a model risk evaluation window (typically 250 trading days, similar to the backtesting period for statistical risk measures as suggested by the Basel Committee on Banking Supervision, 2019), in which an improvement is achieved by minimizing the expected score of a given robust scoring function (MSE or QLIKE) using a volatility proxy for a given univariate volatility model.

We study this methodology via Monte Carlo simulations by comparing different optimization strategies and different lengths of optimization windows and model risk evaluation windows. Then the simulation results show that the QLIKE-based model risk estimation method with additive adjustments made to the volatility estimates, which we propose in this chapter, is a good approximation of true model risk according to several measures of similarity, based on the set of volatility models considered. We mainly use different specifications within the broad GARCH class, and find that the proposed method at least has a correlation of 0.88 with the true model risk measure across various models considered.

Considering the desirable coherence properties (Artzner et al., 1999) of a manageable (from a regulatory perspective) measure of risk for our proposed QLIKE-based model risk estimation methodology, we find that all properties are satisfied except for the subadditivity. Despite this, the proposed measure of model risk can be effectively regulated as the monotonicity, positive homogeneity and translation invariance properties hold.

In an empirical study we apply the proposed model risk measure associated

with different volatility proxies (the squared return and the realized variance) to different asset classes, showing that the level of estimated model risk based on the QLIKE loss function is not sensitive to the choice of volatility proxy across various models in general. The model risk of volatility models adapts to market events, particularly increasing when the market becomes very volatile. The increase in the values of R^2 of the MZ regressions after adjusting variance estimates for model risk shows that model risk has a negative effect on the predictive accuracy of volatility models. We also disentangle the model risk of volatility models into parameter estimation risk and model misspecification risk, and conclude that model misspecification risk generally plays a more dominant role than parameter estimation risk.

The rest of Chapter 4 proceeds as follows: Section 4.2 introduces a model risk estimation methodology based on the MSE and QLIKE loss functions, as well as the definitions of model risk measures; Section 4.3 justifies the QLIKE-based model risk measure via simulations, and Section 4.4 examines the desirable coherence properties of this measure; Section 4.5 applies the proposed QLIKE-based model risk measure to different asset classes; Section 4.7 concludes.

4.2 Quantifying model risk

4.2.1 Evaluating volatility models using scoring functions

The discriminatory analysis between competing models is conducted based on scoring functions. A scoring function is defined as a function $S : \mathbb{R}_+ \times \mathcal{H} \to \mathbb{R}_+$ and $\mathcal{H}$ is a compact subset of $\mathbb{R}_{++}$, where $\mathbb{R}_+$ and $\mathbb{R}_{++}$ represent the non-negative and positive parts of the real line, respectively.

In terms of model comparisons, the evaluation of volatility forecasting models depends on the choice of volatility proxy $\hat{\sigma}^2$ and scoring function S. To compare two time series of competing volatility forecasts, $\{h^k\}$ and $\{h^j\}$, of model k and j over a period from t to $t+\tau$, we compare the expected scores $\mathbb{E}\left[S(\hat{\sigma}^2, h^k)\right] = \frac{1}{\tau+1} \cdot \sum_{i=t}^{t+\tau} S(\hat{\sigma}_i^2, h_i^k)$ and $\mathbb{E}\left[S(\hat{\sigma}^2, h^j)\right] = \frac{1}{\tau+1} \cdot \sum_{i=t}^{t+\tau} S(\hat{\sigma}_i^2, h_i^j)$, given a volatility proxy $\hat{\sigma}^2$ and scoring function S. A smaller expected score indicates the superior forecasting ability of a volatility model. For a given scoring function and volatility proxy, the optimal volatility forecast denoted by h_t^* can be obtained by minimizing the expected score and is defined as below, where $\mathcal{F}_{t-1}$ denotes the time $t-1$ information set (see Patton, 2011; this is further generalized for point forecasts of interest in Gneiting, 2011):

$$h_t^* \equiv \arg\min_{h \in \mathcal{H}} \mathbb{E}\left[S(\hat{\sigma}_t^2, h) | \mathcal{F}_{t-1}\right]. \tag{4.2.1}$$

We consider the MSE and QLIKE scoring (loss) functions (denoted by S_{mse} or

S_{qlike}) in this chapter. The robustness property of scoring functions distinguishes the MSE and QLIKE scoring functions from a number of widely used scoring functions in volatility forecast applications in that the ordering of any two (possibly imperfect) volatility forecasts by the expected score of MSE or QLIKE is the same whether the ordering is done using the true conditional variance or some conditionally unbiased variance (Patton, 2011). These two prominent robust scoring functions are listed below, when a conditionally unbiased volatility proxy $\hat{\sigma}^2$ is used:

$$\text{MSE} : S_{mse}(\hat{\sigma}^2, h) = (\hat{\sigma}^2 - h)^2; \quad \text{QLIKE} : S_{qlike}(\hat{\sigma}^2, h) = \log(h) + \frac{\hat{\sigma}^2}{h} \quad (4.2.2)$$

4.2.2 Measuring model risk of volatility models

In the following, we quantify the model risk of volatility model j for a time series of observed daily volatility proxy $\hat{\sigma}_t^2, ..., \hat{\sigma}_{t+T}^2$ and a time series of out-of-sample daily variance estimates $h_t^j, ..., h_{t+T}^j$ (computed in our case using rolling windows) at time t, $t+1$,..., $t+T$.

Definition 1. If the sequence of true variances $\{\sigma^2\}$ is known, and the volatility forecaster produces a time series of conditional variance forecasts $\{h^j\}$ by using volatility model j, then the true model risk of volatility model j over a model risk

evaluation window from t to $t+n$ is quantified by $p^j_{[t,t+n]}{}^2$:

$$p^j_{[t,t+n]} = \frac{1}{n+1} \cdot \sum_{i=t}^{t+n} \left| \sigma_i^2 - h_i^j \right| . \tag{4.2.3}$$

In practice, the true variance σ^2 is unobservable, which can be recovered by the observed volatility proxy $\hat{\sigma}^2$. Thus, the proxy $\hat{p}^j_{[t,t+n]}$ of the true model risk of model j is calculated as below:

$$\hat{p}^j_{[t,t+n]} = \frac{1}{n+1} \cdot \sum_{i=t}^{t+n} \left| \hat{\sigma}_i^2 - h_i^j \right| . \tag{4.2.4}$$

In order to approximate the true model risk of volatility models based on scoring functions, we consider two estimation methods related to different optimization strategies via making the additive or multiplicative improvements to variance forecasts, under (i) an *additive* structure or (ii) a *multiplicative* structure:

(i) Given a volatility model j, based on (4.2.5) we find an optimized constant[3] $c^{*,S,j}_{a,t+\tau+k}$ (added to a series of variance forecasts $\{h_i^j\}_{i=t+k}^{t+\tau+k}$) by minimizing the expected score of some scoring function S over an optimization window from $t+k$ to $t+\tau+k$ of length $\tau+1$, where $k = 0 : T-\tau$. Parameter c_a is restricted so that $h_i^j + c_a > 0$ is satisfied for all i in order to ensure the positivity of variance forecasts:

$$c^{*,S,j}_{a,t+\tau+k} = \arg\min_{c_a} \frac{1}{\tau+1} \cdot \sum_{i=t+k}^{t+\tau+k} S\left(\hat{\sigma}_i^2, h_i^j + c_a\right) . \tag{4.2.5}$$

As the optimization window of length $\tau+1$ is rolled forward at every step, a time

series of optimized increments $\{c_{a,i}^{*,S,j}\}_{i=t+\tau}^{t+\tau+n}$ is generated for variance estimates of model j. Subsequently, the estimated model risk of model j over a model risk evaluation window from $t+\tau$ to $t+\tau+n$ is given by $\rho_{1,[t+\tau,t+\tau+n]}^{S,j}$, under an additive structure:

$$\rho_{1,[t+\tau,t+\tau+n]}^{S,j} = \frac{1}{n+1} \cdot \sum_{i=t+\tau}^{t+\tau+n} \left| (h_i^j + c_{a,i}^{*,S,j}) - h_i^j \right|. \tag{4.2.6}$$

(ii) In an approach different from the one based on an optimized incremental component in (4.2.5), we calculate an optimized multiplier $c_{m,t+\tau+k}^{*,S,j}$ that is assigned to the conditional variance forecasts $\{h_i^j\}_{i=t+k}^{t+\tau+k}$ via minimizing the expected score over an optimization window from $t+k$ to $t+\tau+k$ with window length $\tau+1$, where $k = 0 : T-\tau$. Parameter c_m is constrained to satisfy $c > 0$:

$$c_{m,t+\tau+k}^{*,S,j} = \arg\min_{c_m} \frac{1}{\tau+1} \cdot \sum_{i=t+k}^{t+\tau+k} S\left(\hat{\sigma}_i^2, h_i^j \cdot c_m\right). \tag{4.2.7}$$

Then the model risk of volatility model j is estimated by $\rho_{2,[t+\tau,t+\tau+n]}^{S,j}$ under a multiplicative structure:

$$\rho_{2,[t+\tau,t+\tau+n]}^{S,j} = \frac{1}{n+1} \cdot \sum_{i=t+\tau}^{t+\tau+n} \left| (h_i^j \cdot c_{m,i}^{*,S,j}) - h_i^j \right|. \tag{4.2.8}$$

In the following, we will omit the subscripts for the time intervals of $p_{[t,t+n]}^j, \hat{p}_{[t,t+n]}^j$, $\rho_{1,[t+\tau,t+\tau+n]}^{S,j}$ and $\rho_{2,[t+\tau,t+\tau+n]}^{S,j}$ for brevity. In order to detect the similarity of model

risk estimation measures defined in (4.2.4), (4.2.6) and (4.2.8) to true model risk measure defined in (4.2.3), we first compute Pearson's linear correlation coefficient $\mathcal{C}^{\mathcal{M}} = Correl(p^{\mathcal{M}}, \hat{p}^{\mathcal{M}} \text{ or } \rho^{S,\mathcal{M}})$ between true model risk ($p^{\mathcal{M}}$) and model risk measure estimates across the set of volatility models $\mathcal{M}$ discussed in Table 4.3.1, in which $\rho^{S,\mathcal{M}} = \rho_1^{S,\mathcal{M}}$ or $\rho_2^{S,\mathcal{M}}$. This can only show a linear relationship between the two series, so we additionally consider the possibly nonlinear association between true model risk and model risk measure estimates by using the $\tau_x^{\mathcal{M}} = \tau_x(p^{\mathcal{M}}, \hat{p}^{\mathcal{M}}$ or $\rho^{S,\mathcal{M}})$ correlation coefficient from Emond and Mason (2002) that extends the nonparametric Kendall's τ_b measure. For a model j, the explanatory power of model risk estimation measure over true model risk measure is defined as $\psi^j = \rho^{S,j}/p^j$ or $\hat{p}^j/p^j$, where $\rho^{S,j}$ can be $\rho_1^{S,j}$ or $\rho_2^{S,j}$.

4.3 Simulation study

In this section, we verify via simulations whether the model risk estimation methodology is able to capture the size of true model risk of a given volatility model. Considering that the conditional distribution of financial time series is often fat-tailed and asymmetric, we use the GARCH(1,1) model with the skewed Student's t distributed innovations (SKTGARCH), allowing for kurtosis

and skewness, as the data generating process that is specified as:

$$\begin{aligned} r_t &= \sqrt{h_t} Z_t, \quad Z \sim \text{skewed Student's } t\ (\nu, \lambda), \\ h_t &= \hat{\omega} + \hat{\alpha} r_{t-1}^2 + \hat{\beta} h_{t-1}, \end{aligned} \tag{4.3.1}$$

where r_t denotes a realization of return and h_t denotes the one-step ahead conditional variance forecast at time t. The density function of the standardized returns Z is $f(z|\nu, \lambda)$ (see Appendix 4.A), in which ν is the degree of freedom parameter and λ is the skewness parameter. The model parameters[4] are estimated on the S&P500 Index daily returns from 2000/01/03 to 2010/12/31 (2869 observations): $\hat{\omega} = 7.8183e^{-07}, \hat{\alpha} = 0.0770, \hat{\beta} = 0.9205, \hat{\nu} = 7.1845$ and $\hat{\lambda} = -0.0848$. Using these values, we generate a time series of 10,000 daily returns.

Based on the simulated returns, we employ 19 volatility models[5] specified in Table 4.3.1 to make one-step ahead conditional variance estimates. More precisely, the models are: 1) historical volatility measures (RW250 and RW1000), which are non-parametric; 2) the RiskMetrics model with $\lambda = 0.94$; 3) the autoregressive conditional heteroscedasticity (ARCH(1)) models (Engle, 1982) with one lag, combined with four specifications[6] for the standardised errors following the normal, Student's t, skewed Student's t and generalized error distributions, respectively; and 4) specifications of the generalized autoregressive conditional heteroskedasticity models combined with the aforementioned four distributional assumptions for the standardised errors, including the symmetric GARCH(1,1)

models (Bollerslev, 1986), as well as the models of EGARCH(1,1) (Nelson, 1991) and GJR-GARCH(1,1) (Glosten et al., 1993) with leverage terms to consider asymmetry in volatility clustering.

Table 4.3.1: Volatility models for one-step ahead conditional variance forecasts

RW250:	$h_t = \frac{1}{249}\sum_{i=t-250}^{t-1}\left(r_i - \frac{1}{250}\sum_{i=t-250}^{t-1} r_i\right)^2$
RW1000:	$h_t = \frac{1}{999}\sum_{i=t-1000}^{t-1}\left(r_i - \frac{1}{1000}\sum_{i=t-1000}^{t-1} r_i\right)^2$
RiskMetrics:	$h_t = (1-\lambda)r_{t-1}^2 + \lambda h_{t-1}$, where $\lambda = 0.94$
ARCH(1):	$h_t = \omega + \alpha r_{t-1}^2$
GACRH(1,1):	$h_t = \omega + \alpha r_{t-1}^2 + \beta h_{t-1}$
EGARCH(1,1):	$\log(h_t) = \omega + \alpha\left[\frac{\lvert r_{t-1}\rvert}{\sqrt{h_{t-1}}} - \mathbb{E}\{\frac{\lvert r_{t-1}\rvert}{\sqrt{h_{t-1}}}\}\right] + \kappa(\frac{\lvert r_{t-1}\rvert}{\sqrt{h_{t-1}}}) + \beta\log(h_{t-1})$
GJR-GARCH(1,1):	$h_t = \omega + \alpha r_{t-1}^2 + \xi\mathbb{1}\{r_{t-1}<0\}r_{t-1}^2 + \beta h_{t-1}$

This table shows that for all (G)ARCH specifications $r_t = \sqrt{h_t}Z_t$, where Z_t denotes the standardized return and follows the normal, Student's t*, skewed Student's* t *and generalized error distributions.*

We first compute the daily variance estimates using rolling windows with length 1,000 (except for the RW250 method, for which we use the previous 250 observations to compute the historical variance in a rolling window scheme). Then for each model, we calculate the model risk of the daily volatility forecasts using the squared return as the volatility proxy for a given scoring function (S_{mse} or S_{qlike}), considering several optimization windows of length $\tau_1 = 250$, $\tau_2 = 500$, $\tau_3 = 1,000$ and $\tau_4 = 2,000$ with respect to the expected score and two model risk evaluation windows of length $n_1 = 250$ and $n_2 = 1,000$.

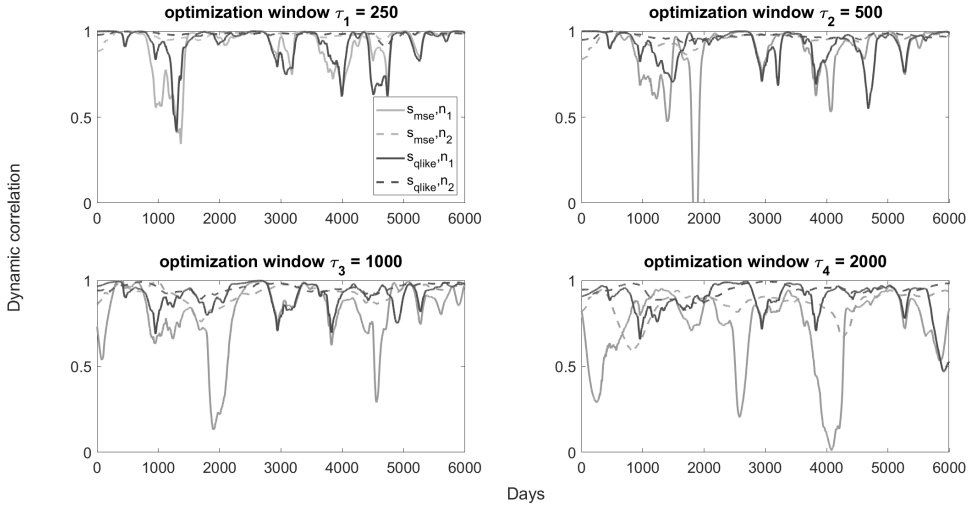
Panel (a) of Figure 4.3.1 presents the dynamic correlation[7] between true model risk in (4.2.3) and estimated model risk in (4.2.6) based on the MSE and QLIKE

loss functions under an additive structure across all the models considered in this chapter, based on data simulated by the SKTGARCH model, whilst panel (b) shows that the model risk estimation method that assumes a multiplicative structure in (4.2.8) leads to lower correlations between the true and estimated model risk. Thus, in the remaining part of this chapter, we only estimate model risk using the additive structure and present the corresponding results. Also, we find that the longer the model risk evaluation window, the higher the correlation between the true and estimated model risk. Moreover, Figure 4.3.2 illustrates the average percentage of true model risk explained by the QLIKE-based model risk measure estimates under an additive structure, calculated using an optimization window of length $\tau_2 = 500$ and model risk estimation windows of length $n_1 = 250$ and $n_2 = 1,000$. Across all the models considered, the model risk estimation measure computed over a shorter model risk evaluation window ($n_1 = 250$) can capture a larger part of true model risk than the measure computed over a longer window ($n_2 = 1,000$). Besides, this QLIKE-based model risk estimation method explains up to about 80% of true model risk.

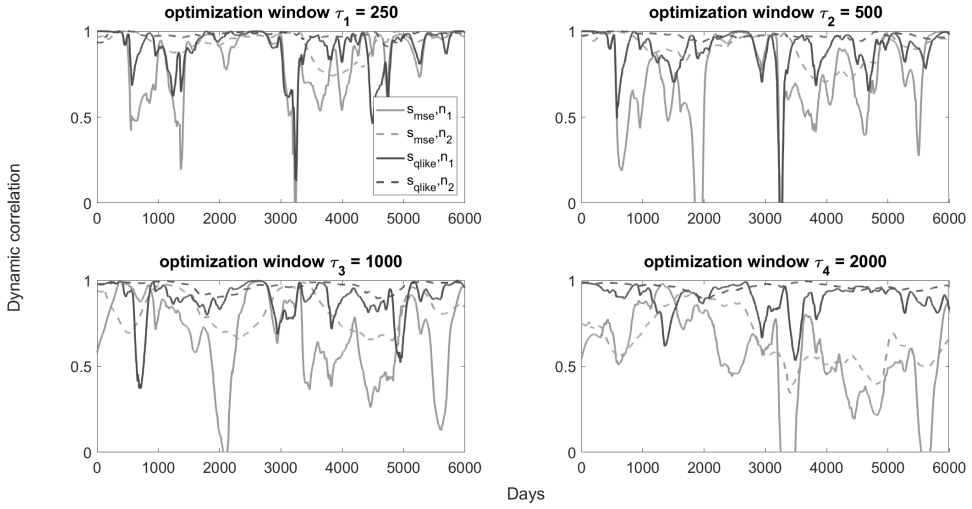
In order to measure the similarity of model risk measure estimates to true model risk, Table 4.3.2 reports average values of the correlation, τ_x correlation coefficient and explanatory power (denoted by $\bar{\mathcal{C}}^{\mathcal{M}}$, $\bar{\tau}_x^{\mathcal{M}}$ and $\bar{\psi}^{\mathcal{M}}$ respectively). In panel A, we report the results of model risk measures based on the MSE and QLIKE loss functions, considering the squared return as the volatility proxy. We find that the model risk estimation method based on the QLIKE loss function

Figure 4.3.1: Dynamic correlation between true model risk and estimated model risk

(a) Under an additive structure

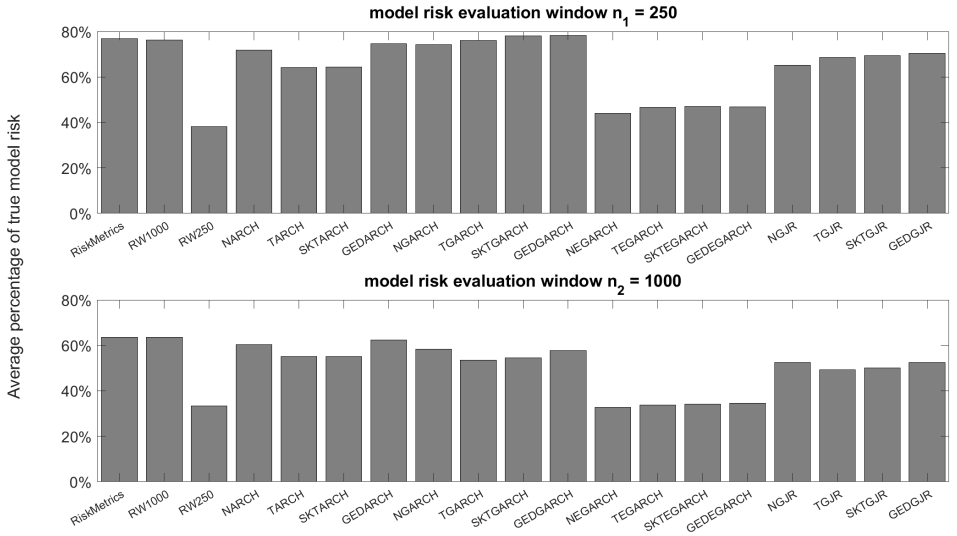


(b) Under a multiplicative structure



Panel (a) and panel (b) of this figure show the dynamic correlation between true model risk in (4.2.3) *and estimated model risk under the additive structure in* (4.2.6) *and under the multiplicative structure in* (4.2.8) *across various volatility models accordingly, based on data simulated by the SKTGARCH model. Model risk of daily volatility forecasts is estimated using scoring function S_{mse} or S_{qlike}, and the squared returns are used as the volatility proxy. We consider optimization windows $\tau_1 = 250$, $\tau_2 = 500$, $\tau_3 = 1,000$ and $\tau_4 = 2,000$ and model risk evaluation windows $n_1 = 250$ and $n_2 = 1,000$.*

Figure 4.3.2: Average percentage of true model risk explained



This figure shows the average percentage of true model risk explained by the QLIKE-based model risk defined in (4.2.6)*, using an optimization window of length* $\tau_2 = 500$*.*

outperforms the one based on the MSE loss function, as the former generally has a higher correlation (and τ_x coefficient) with the true model risk measure for a given optimization window and model risk evaluation window. The QLIKE-based technique is highly consistent with the true model risk measure with a correlation averaging from 0.88 to 0.98. In terms of the length of optimization windows, the QLIKE-based model risk estimation methodology using a window length of $\tau_3 = 1,000$ generally leads to the highest correlation with the true model risk measure for a given model risk window, which is followed by estimation using a window length of $\tau_2 = 500$ as shown in panel A. Nevertheless, the latter method is able to explain a higher proportion of true model risk.

In panel B of Table 4.3.2, we look at the similarity of the true model risk proxy

Table 4.3.2: Similarity of model risk measures to true model risk measure

Panel A: Similarity of the MSE or QLIKE-based model risk measure					
Model risk measure	optimization window length	model risk window length	$\bar{\mathcal{C}}^{\mathcal{M}}$	$\bar{\tau}_x^{\mathcal{M}}$	$\bar{\psi}^{\mathcal{M}}$
$\rho_1^{S_{mse}}$	$\tau_1 = 250$	$n_1 = 250$	0.91	0.68	115%
		$n_2 = 1,000$	0.96	0.67	88%
	$\tau_2 = 500$	$n_1 = 250$	0.87	0.73	92%
		$n_2 = 1,000$	0.94	0.88	60%
	$\tau_3 = 1,000$	$n_1 = 250$	0.82	0.79	66%
		$n_2 = 1,000$	0.91	0.87	44%
	$\tau_4 = 2,000$	$n_1 = 250$	0.73	0.88	41%
		$n_2 = 1,000$	0.86	0.94	29%
$\rho_1^{S_{qlike}}$	$\tau_1 = 250$	$n_1 = 250$	0.89	0.65	96%
		$n_2 = 1,000$	0.98	0.61	85%
	$\tau_2 = 500$	$n_1 = 250$	0.88	0.81	65%
		$n_2 = 1,000$	0.97	0.95	50%
	$\tau_3 = 1,000$	$n_1 = 250$	0.92	0.88	43%
		$n_2 = 1,000$	0.96	0.99	34%
	$\tau_4 = 2,000$	$n_1 = 250$	0.89	0.92	32%
		$n_2 = 1,000$	0.94	1.00	26%
Panel B: Similarity of the true model risk proxy measure					
$\hat{p}$		$n_1 = 250$	0.35	1.00	832%
		$n_2 = 1,000$	0.44	1.00	763%

This table presents several ways to measure the degree of similarity of model risk measures ($\rho_1^{S_{mse}}$, $\rho_1^{S_{qlike}}$, $\hat{p}$) to true model risk measure, based on daily returns simulated by the SKTGARCH model: $\bar{\mathcal{C}}^{\mathcal{M}}$ and $\bar{\tau}_x^{\mathcal{M}}$ represent average values of linear and nonlinear association between true and estimated model risk; $\bar{\psi}^{\mathcal{M}}$ shows the average explanatory power of model risk measures across the set of volatility models. We consider optimization windows $\tau_1 = 250$, $\tau_2 = 500$, $\tau_3 = 1,000$ and $\tau_4 = 2,000$ and model risk evaluation windows $n_1 = 250$ and $n_2 = 1,000$. The volatility proxy is the squared return.

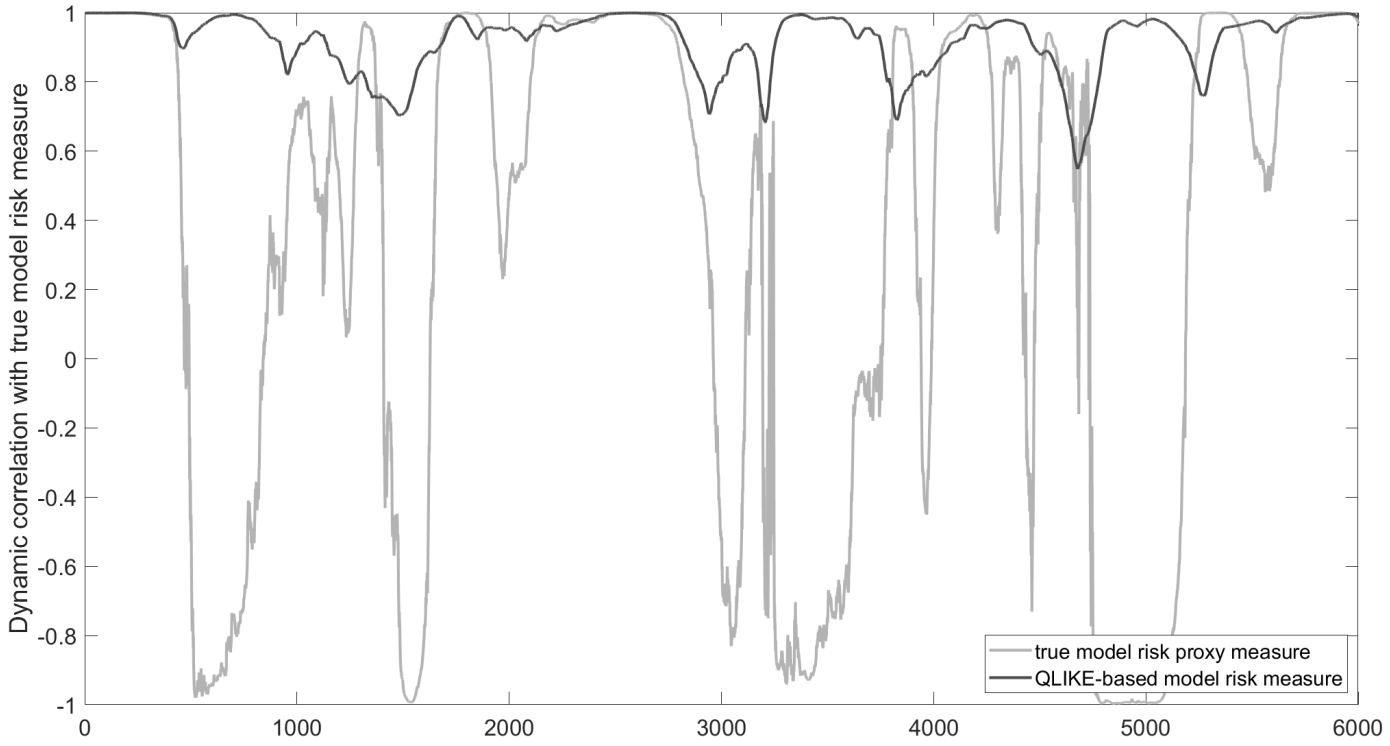
measure estimates to true model risk, and find the average correlations around 0.35 and 0.44 which are less than half of the corresponding values presented in panel A. Additionally, the true model risk proxy measure tends to over-estimate model risk that would be more than seven times of true model risk. From a dynamic perspective, Figure 4.3.3 compares the dynamic correlation of the true model risk proxy measure and the QLIKE-based model risk measure with the true model risk measure, where the squared return is used as the volatility proxy and model risk is computed over a model risk evaluation window $n_1 = 250$. Unlike the QLIKE-based model risk measure, the true model risk proxy measure is unable to give a reasonable approximation of true model risk in that its negative correlation with true model risk measure occurs frequently.

Generally, we can conclude based on the simulation analysis that the scoring function-based model risk estimation methodology using the additive structure defined in (4.2.6) can be a practical tool to provide a rational approximation of true model risk of volatility models, particularly when the QLIKE scoring function is used for optimization over windows $\tau_2 = 500$ and $\tau_3 = 1,000$.

4.4 Properties of model risk estimates

To facilitate model risk management from the regulators' perspective, a reasonable positive measure $\rho(\cdot)$ of risk should satisfy the coherence properties (McNeil et al., 2015): 1) Monotonicity: for returns r_1 and r_2 with $r_1 \leq r_2$, we have that

Figure 4.3.3: Dynamic correlation of true model risk proxy measure or QLIKE-based model risk measure with true model risk measure



This figure focuses on the true model risk proxy measure and the QLIKE-based model risk measure, and compares their dynamic correlations with true model risk measure. The squared return is used as the volatility proxy. Model risk is computed over a model risk evaluation window of length $n_1 = 250$, based on daily returns simulated by the SKTGARCH model.

$\rho(r_1) \geq \rho(r_2)$; 2) Positive homogeneity: for any positive number $k \in \mathbb{R}^+$, we have that $\rho(k \cdot r) = k \cdot \rho(r)$ where r denotes the returns; 3) Translation invariance: for any $a \in \mathbb{R}$, we have that $\rho(r + a) = \rho(r) - a$; and 4) Subadditivity: for any r_1 and r_2, we have that $\rho(r_1 + r_2) \leq \rho(r_1) + \rho(r_2)$.

In a similar vein, we focus on the properties of the QLIKE-based model risk measure denoted by $\rho^{S_{qlike},j}(r, h^j)$ using the squared returns r^2 as the volatility proxy in which r denotes the daily returns of a certain asset and h^j denotes one-step ahead variance forecasts of a model j. Consider the following properties that a reasonable measure of the model risk of volatility models should satisfy:

i) *Monotonicity*: If $\sigma^2 < h^i < h^j$ or $\sigma^2 > h^i > h^j$ for all t, then $\rho^{S_{qlike},i}(r, h^i) < \rho^{S_{qlike},j}(r, h^j)$, assuming that two different volatility models i and j produce variance estimates h^i and h^j respectively, when applied to the returns r of a certain asset.

This property states that if the variance estimates of a certain model are closer to the true variances σ^2, then this model will carry a lower level of model risk.

ii) *Positive homogeneity*: For $k \in \mathbb{R}^+$ and a model j, $\rho^{S_{qlike},j}(k \cdot r, k^2 \cdot h^j) = k^2 \cdot \rho^{S_{qlike},j}(r, h^j)$, given the returns r of a certain asset and the corresponding variance estimates h^j of volatility model j.

This states that if the return data is rescaled by a positive constant k and the variance estimates are rescaled by k^2, then the model risk will be resized by k^2 as well.

iii) *Translation invariance*: For a model j, if a constant a with $0 > a > \max(\sigma^2 - h^j)$ and $\sigma^2 < h^j$, or with $0 < a < \min(\sigma^2 - h^j)$ and $\sigma^2 > h^j$ for all t, $\rho^{S_{qlike},j}(r, h^j + a) = \rho^{S_{qlike},j}(r, h^j) + |a|$.

This property says that when the variance estimates are shifted by a constant a that satisfies the condition $0 > a > \max(\sigma^2 - h^j)$ with $\sigma^2 < h^j$, or $0 < a < \min(\sigma^2 - h^j)$ with $\sigma^2 > h^j$, then the model risk of model j will increase by the absolute value of a.

iv) *Subadditivity*: $\rho^{S_{qlike},j}(r_{(X+Y)}, h^j_{(X+Y)}) < \rho^{S_{qlike},j}(r_X, h^j_X) + \rho^{S_{qlike},j}(r_Y, h^j_Y)$, considering that a model j produces the variance estimates h^j_X, h^j_Y and $h^j_{(X+Y)}$ when applied to individual assets X and Y, and an equally weighted portfolio $(X + Y)$ consisting of these two assets.

This states that for a given volatility model, the model risk for an equally weighted portfolio comprised of assets X and Y is lower than the sum of model risk for the constituents. This property should not be required for measures of model risk of volatility models, as it does not follow the expected behavior of model risk measures.

Via Monte Carlo simulations, we find that the properties of monotonicity, positive homogeneity and translation invariance hold for the QLIKE-based model risk estimation method using the squared return as the volatility proxy, whilst the subadditivity property does not. In Figure 4.4.1, we revisit the subadditivity property of our proposed model risk measure in simulated cases as in Daníelsson

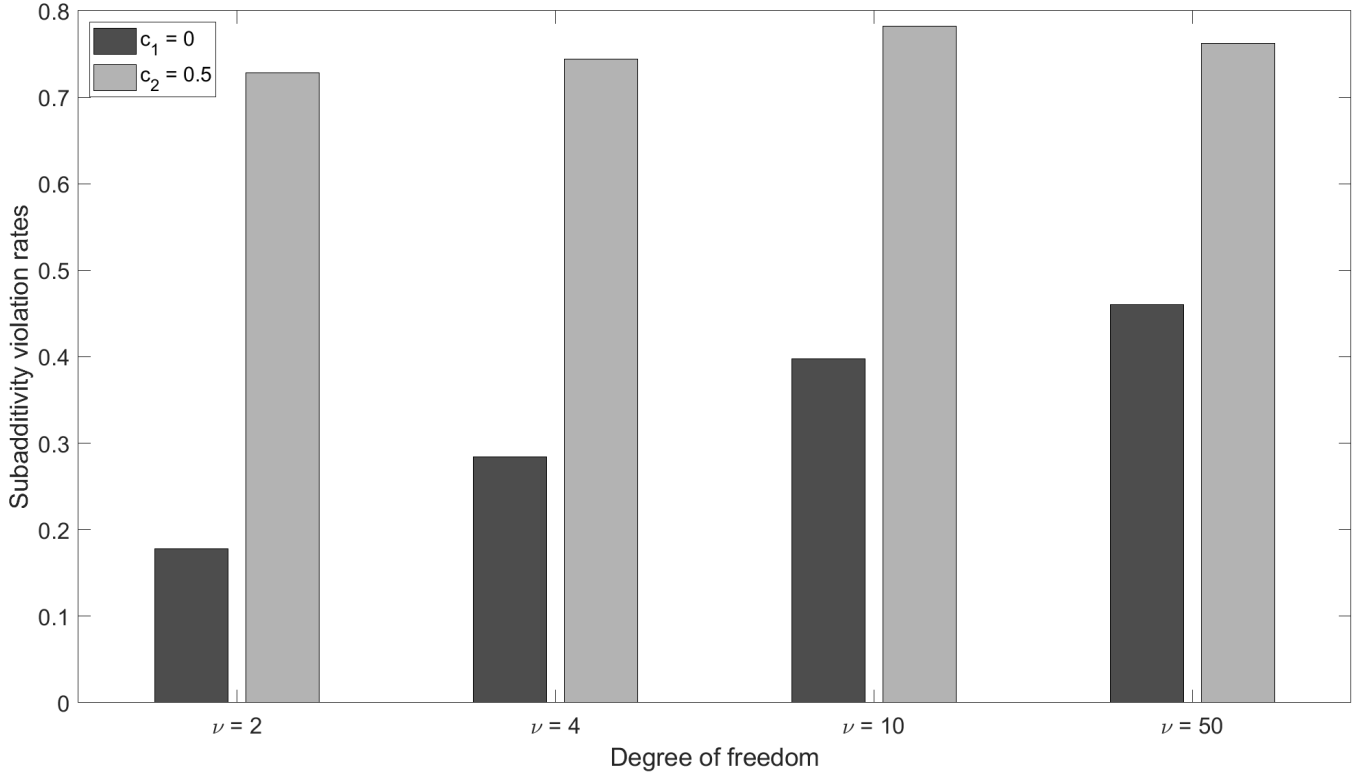
et al. (2013), and report the subadditivity violation rates.

Specifically, assuming that assets X and Z are independent but follow the same Student's t distribution with the degree of freedom $\nu = 2, 4, 10$, and 50, we construct asset Y defined as $Y = cX + \sqrt{1-c^2}Z$, thereby being correlated with asset X with a correlation coefficient c. We consider two cases: in the first one X and Y are independent ($c_1 = 0$); in the second case X and Y are correlated ($c_2 = 0.5$). We simulate 500 paths of 1750 returns for X and Y, and build an equally weighted portfolio $(X + Y)$. Subsequently, we make one-step ahead variance estimates by using the RW1000 model and compute the QLIKE-based model risk over an optimization window $\tau_2 = 500$ and a model risk evaluation window $n_1 = 250$ for the individual assets and the portfolio. If the model risk of the portfolio is larger than the sum of individual model risk of the component assets, the subadditivity property will be violated for this simulated path. As the results show, the subadditivity violations are very high.

4.5 Empirical application

In this section, we apply the QLIKE-based model risk measure under the additive structure in (4.2.6) using an optimization window of length $\tau_2 = 500$ and $\tau_3 = 1,000$ for empirical illustrations, as this measure shows high correlations with the true model risk measure as evidenced by Figure 4.3.1 and Table 4.3.2 in the simulation study. A shorter model risk evaluation period $n_1 = 250$ is used, since

Figure 4.4.1: Subadditivity violations for RW1000



This figure presents the subadditivity violation rates when we apply the RW1000 model to the simulated 500 paths of 1750 returns of assets X and Y and an equally weighted portfolio $(X+Y)$ to produce daily variance forecasts. Assuming that assets X and Z are independent but follow the same Student's t distribution with the degree of freedom $\nu = 2, 4, 10,$ and 50, asset Y is defined as $Y = cX + \sqrt{1-c^2}Z$ with $c_1 = 0$ and $c_2 = 0.5$. The QLIKE-based model risk is computed over an optimization window $\tau_2 = 500$ and a model risk evaluation window $n_1 = 250$, using the squared returns as the volatility proxy.

it is in line with the backtesting period of market risk models, and based on this shorter evaluation period, the proposed model risk estimation method captures a higher proportion of true model risk than the method based on a longer evaluation period $n_2 = 1{,}000$.

We illustrate the QLIKE-based model risk estimation method for several asset classes with daily data (30/12/1983 - 21/10/2019), downloaded from DataStream: 1) FTSE100 Index close prices (FTSE100); 2) JP Morgan Chase close prices (JPM); 3) Europe Brent spot prices (dollars per barrel) for Crude Oil (Crude Oil); and 4) Foreign exchange USD/GBP rates (USD/GBP). To consider an alternative volatility proxy for the conditionally unbiased variance estimator, we also download the daily close prices and the 5-min realized variances of the FTSE100 Index (04/01/2000 to 10/10/2019) from the realized library of Oxford-Man Institute of Quantitative Finance[8]. We compute daily log-returns of different assets and then produce out-of-sample one-step ahead variance forecasts in a rolling window scheme (all the models detailed in Table 4.3.1 use rolling windows of length 1,000 to build volatility forecasts, except for RW250 using windows of length 250).

Table 4.5.1 reports average ratios[9] of the QLIKE-based model risk estimates based on two proxies, namely the squared returns and 5-min realized variances, to estimated variances. Model risk estimates are calculated over two optimization windows $\tau_2 = 500$ and $\tau_3 = 1000$ and a model risk window $n_1 = 250$, using daily returns and 5-min realized variances of the FTSE100 Index from 04/01/2000 to

Table 4.5.1: Average ratios of the QLIKE-based model risk estimates, using different volatility proxies, to estimated variances

	The volatility proxy $\hat{\sigma}^2$ is			
	squared returns		5-min realized variances	
Models	$\tau_2 = 500$	$\tau_3 = 1000$	$\tau_2 = 500$	$\tau_3 = 1000$
RiskMetrics	51.6%	33.4%	47.1%	33.5%
RW1000	57.5%	36.9%	53.0%	36.9%
RW250	16.6%	12.4%	14.4%	8.3%
NARCH	36.9%	25.9%	36.2%	26.7%
TARCH	36.9%	27.8%	35.9%	27.5%
SKTARCH	36.9%	28.0%	36.0%	27.5%
GEDARCH	37.5%	27.3%	36.8%	27.4%
NGARCH	7.4%	4.8%	8.7%	8.1%
TGARCH	7.3%	5.1%	8.8%	8.3%
SKTGARCH	7.1%	5.0%	8.5%	8.0%
GEDGARCH	7.4%	5.2%	8.7%	8.1%
NEGARCH	8.0%	4.7%	8.3%	4.0%
TEGARCH	8.7%	4.9%	8.4%	4.2%
SKTEGARCH	9.0%	5.2%	8.6%	4.4%
GEDEGARCH	8.2%	4.6%	8.3%	4.0%
NGJR	8.0%	6.5%	9.0%	6.6%
TGJR	7.6%	6.3%	8.5%	6.2%
SKTGJR	7.6%	6.2%	8.5%	6.1%
GEDGJR	7.8%	6.4%	8.8%	6.4%

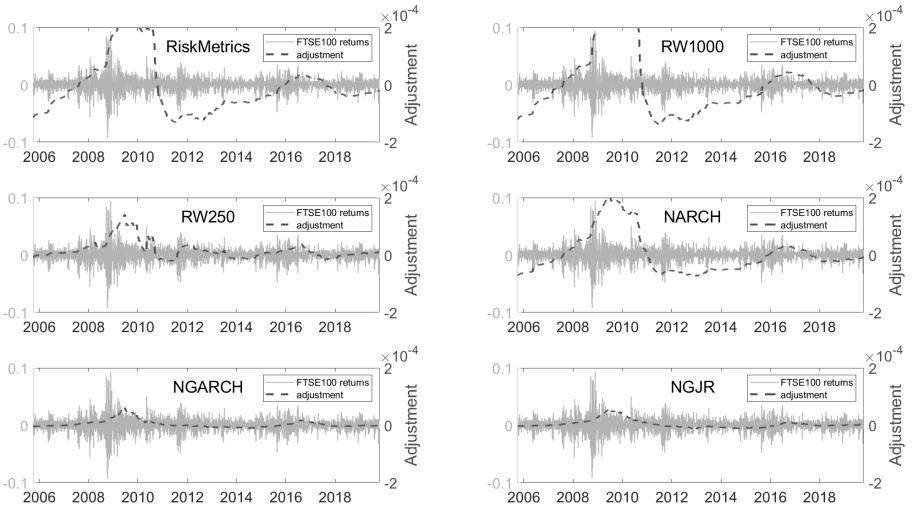
This table presents average ratios of the QLIKE-based model risk estimates, using different volatility proxies, to estimated variances. The daily prices and the 5-min realized variances of the FTSE100 Index range from 04/01/2000 to 10/10/2019. The volatility proxies used are squared returns and 5-min realized variances. The optimization window length is $\tau_2 = 500$ *and* $\tau_3 = 1000$*; the model risk evaluation window length is* $n_1 = 250$*.*

10/10/2019. We find that for the badly fitting models, the proposed technique, using squared returns as the volatility proxy, generally estimates a higher level of model risk than the one using realized data as the proxy. Here, the badly fitting models are defined as those affected by model risk amounting to more than 25% of estimated variances, and these include the RiskMetrics method, RW1000 as well as the ARCH models. Generally, the average ratio of estimated model risk to estimated variances is not sensitive to the use of the volatility proxy due to the similar values of the level of model risk estimated over the same optimization window.

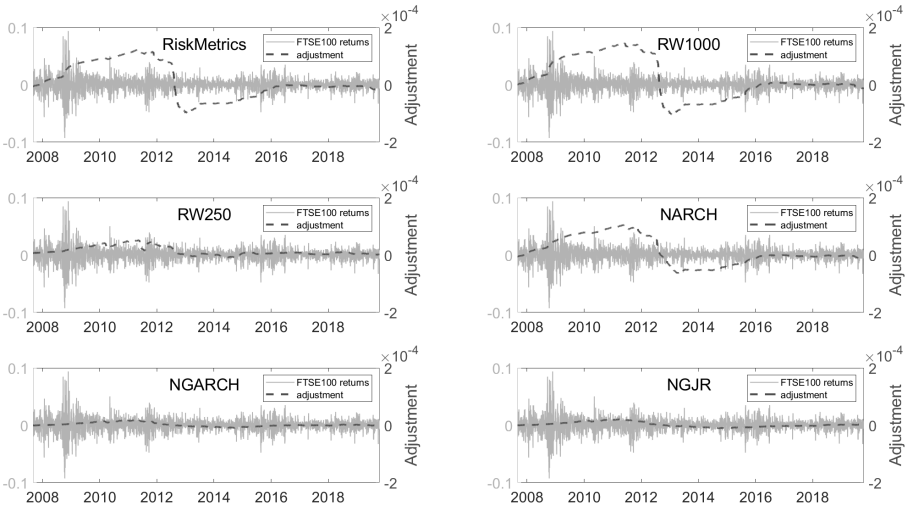
Regardless of the volatility proxy used for the computation of the QLIKE-based model risk estimates, it is interesting to notice in Table 4.5.1 that the model risk estimation method based on an optimization window of $\tau_2 = 500$ always gives higher ratios of model risk estimates than the method based on an optimization window of $\tau_3 = 1,000$ for the set of models. To get a better understanding of this phenomenon, in Figure 4.5.1 we compare the additive adjustments with respect to the optimization windows $\tau_2 = 500$ and $\tau_3 = 1,000$, and show the time series of adjustments, obtained based on the QLIKE loss function and the squared return used as the volatility proxy, made to volatility estimates of several selected models, using the FTSE100 Index returns from 04/01/2000 to 10/10/2019. Clearly, the QLIKE-based model risk measure computed over $\tau_2 = 500$ in panel (a) responds to market events in a more timely and effective manner. It allows a higher level of additive adjustments, which also supports its higher explanatory power in

Figure 4.5.1: Dynamic additive adjustments made to volatility estimates of selected models

(a) Optimization window $\tau_2 = 500$



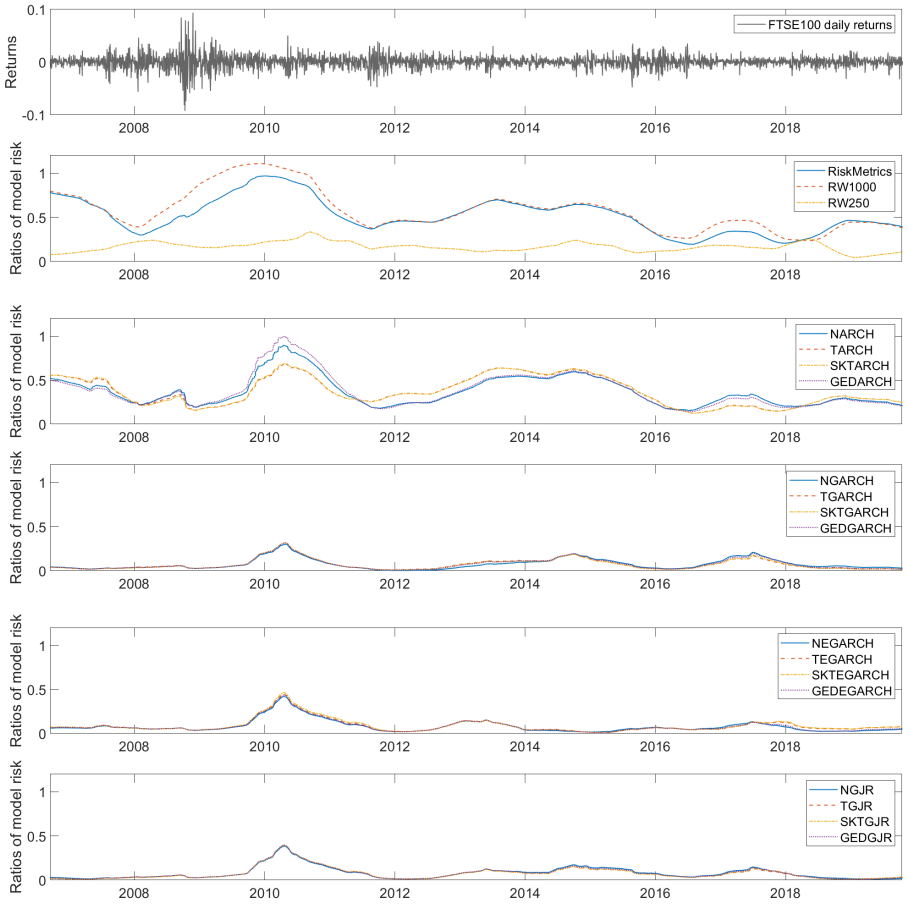
(b) Optimization window $\tau_2 = 1000$



The figure displays dynamic additive adjustments made to volatility estimates of selected models, based on the QLIKE loss function, for the FTSE100 Index from 04/01/2000 to 10/10/2019. The optimization windows $\tau_2 = 500$ and $\tau_3 = 1,000$ are considered and the volatility proxy is the squared return.

the simulation study, as compared with the measure computed over $\tau_3 = 1,000$ days presented in panel (b) of Figure 4.5.1. Therefore, in terms of the QLIKE-based model risk measures, an optimization window of $\tau_2 = 500$ is recommended to warrant effective adjustments for model risk and high consistency with true model risk at the same time.

Figure 4.5.2: Time-varying ratios of the QLIKE-based model risk estimates to estimated variances



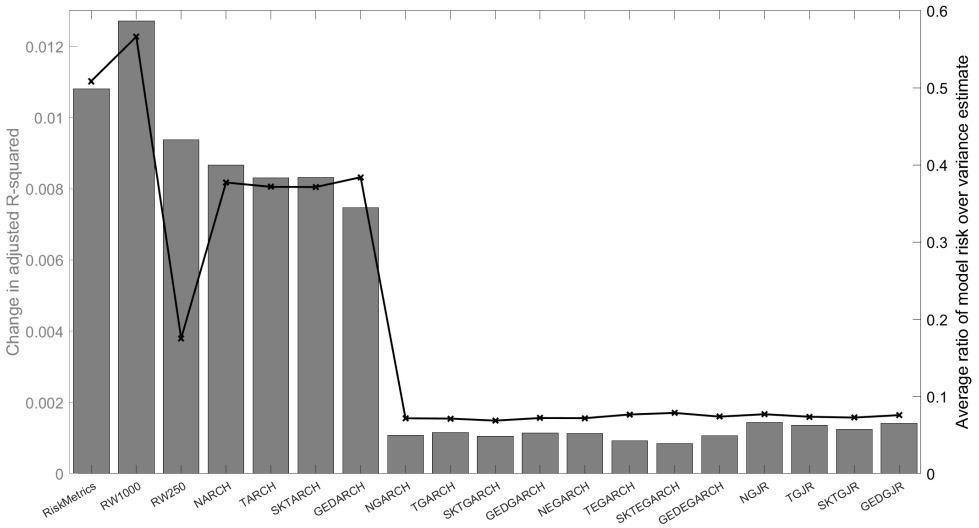
This figure shows time-varying ratios of the QLIKE-based model risk estimates to estimated variances. The squared return is used as the volatility proxy. Model risk is computed over $n_1 = 250$ trading days using an optimization window of length $\tau_2 = 500$, based on the FTSE100 Index from 04/01/2000 to 10/10/2019.

Figure 4.5.2 shows the time-varying ratios of the QLIKE-based model risk estimates of various models to variance estimates where the volatility proxy used is the squared return (this can be compared with Figure 4.B.1 of Appendix 4.B, in which the alternative proxy is the 5-min realized variance). Model risk is estimated over $n_1 = 250$ trading days with an optimization window $\tau_2 = 500$ for FTSE100. Within the sample period, the RiskMetrics method, RW1000 and the ARCH models are characterised by higher ratios of model risk over the variance forecasts, compared with the rest of the models considered. Noticeably, when the market is highly volatile, the model risk of volatility models increases in general. For example, the FTSE100 Index experiences its most uncertain period around 2009, following which the ratios of estimated model risk to estimated variances of various models reach the peak level around 2010 due to the evaluation period for model risk having a length of $n_1 = 250$ (about one year).

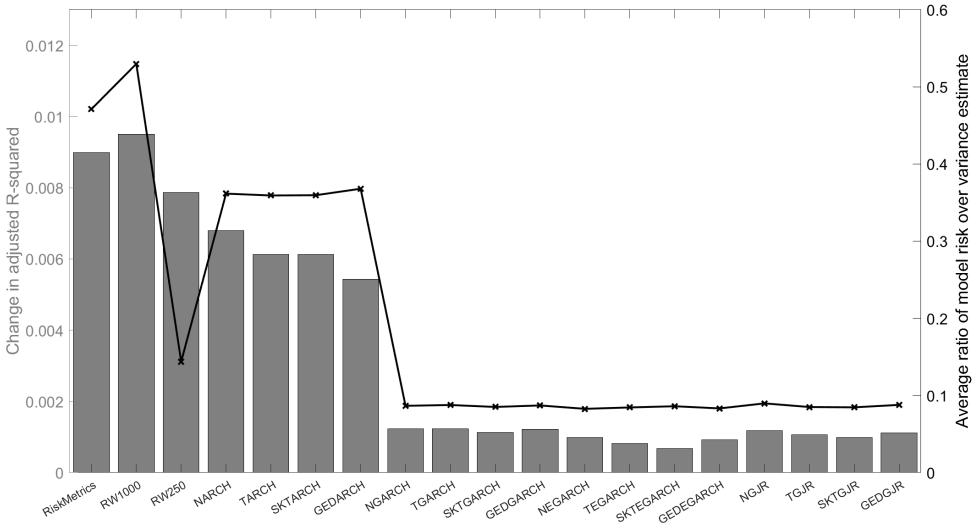
It is common practice to use the adjusted R^2 of the Mincer-Zarnowitz (MZ) regression to assess the degree of predictability of the volatility models. We use the 5-min realized variance as endogenous variable and the estimated variance, or rather the improved variance estimate for model risk as explanatory variable in the MZ regression. In order to analyze the performance of the forecasted volatility adjusted for model risk, in Figure 4.5.3 we compute the change in the adjusted R^2 of the MZ regressions (displayed in bars), which shows a very similar pattern to the average ratio (displayed in lines) of model risk estimate over estimated variance. The model risk estimates are computed over an optimization window

Figure 4.5.3: Change in the adjusted R-squared of the MZ regressions when adjusting for model risk

(a) Using the squared return as the volatility proxy



(b) Using the 5-min realized variance as the volatility proxy



This figure shows the change in the adjusted R^2 of the MZ regressions in which the 5-min realized variance and the variance forecast are dependent and independent variables, accordingly, based on the FTSE Index data from 04/01/2000 to 10/10/2019, after adjusting variance estimates for the QLIKE-based model risk of different volatility models. An optimization window $\tau_2 = 500$ and a model risk window $n_1 = 250$ are considered.

$\tau_2 = 500$ and model risk evaluation window $n_1 = 250$, based on the QLIKE loss function related to the squared return and the realized variance as volatility proxies, for the FTSE100 Index returns from 04/01/2000 to 10/10/2019. After taking model risk into account, the volatility models have more predictive ability as evidenced by an increase in the adjusted R^2 across the set of models considered. In general, the higher the model risk, the higher the increase in the adjusted R^2 of the MZ regressions when adjusting for model risk of various models.

A second application based on several asset classes from 30/12/1983 to 21/10/2019 is illustrated in Table 4.5.2 which presents average ratios of the QLIKE-based model risk to variance forecasts for the set of models given in Table 4.3.1. Here we use squared returns as the volatility proxy and compute the model risk based on optimization windows of length $\tau_2 = 500$ and $\tau_3 = 1,000$, and a model risk evaluation window of length $n_1 = 250$. For all assets considered, the RW1000 method carries the highest level of model risk among the set of volatility models, followed by the RiskMetrics method as well as the ARCH(1) models. Interestingly, volatility models have the highest model risk when applied to the JP Morgan Chase stock as compared with the other assets in general.

In Figure 4.5.4 we plot the time-varying ratios of the QLIKE-based model risk over variance estimates of two models (SKTGARCH(1,1) in panel a and RiskMetrics in panel b, respectively) applied to various assets, based on data from 30/12/1983 to 21/10/2019. We estimate model risk based on a model risk window $n_1 = 250$ and an optimization window $\tau_2 = 500$, using the squared return

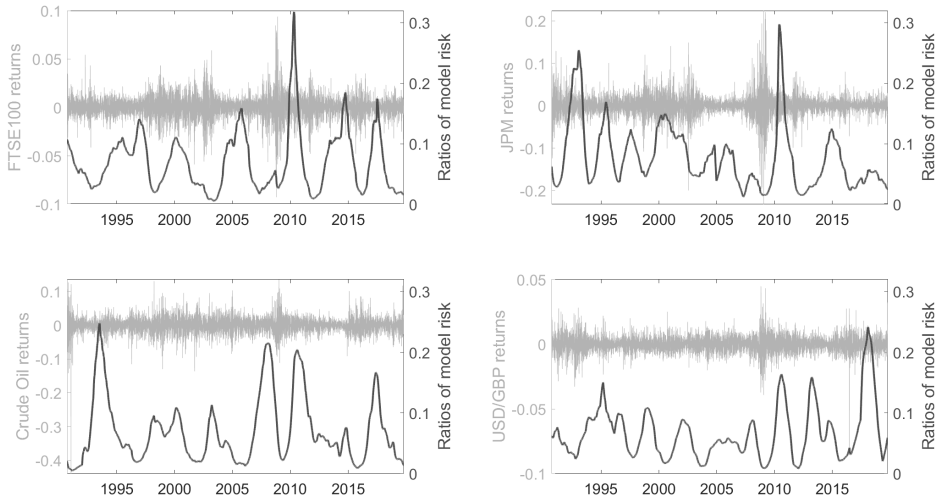
Table 4.5.2: Average ratios of the QLIKE-based model risk, with squared returns as the volatility proxy, to estimated variances of various models for different assets.

	FTSE100		JPM		Crude Oil		USD/GBP	
Models	$\tau_2 = 500$	$\tau_3 = 1000$	$\tau_2 = 500$	$\tau_3 = 1000$	$\tau_2 = 500$	$\tau_3 = 1000$	$\tau_2 = 500$	$\tau_3 = 1000$
RiskMetrics	42.2%	28.4%	44.8%	34.2%	40.5%	23.8%	30.0%	20.7%
RW1000	*46.4%*	*30.5%*	*50.2%*	*37.8%*	*44.6%*	*27.1%*	*33.0%*	*22.7%*
RW250	14.3%	10.1%	17.3%	11.8%	17.7%	13.0%	13.9%	10.5%
NARCH	32.8%	23.4%	36.7%	31.4%	33.2%	21.9%	27.6%	20.0%
TARCH	31.3%	23.3%	35.2%	29.6%	31.3%	21.8%	28.6%	20.5%
SKTARCH	31.4%	23.3%	35.3%	29.7%	31.4%	22.0%	28.6%	20.6%
GEDARCH	32.3%	23.5%	37.3%	33.4%	34.7%	23.3%	28.5%	20.3%
NGARCH	7.7%	5.5%	8.4%	6.0%	7.0%	**4.7%**	6.8%	4.3%
TGARCH	7.1%	5.2%	7.7%	5.3%	**6.9%**	5.2%	6.6%	4.0%
SKTGARCH	**7.0%**	5.1%	7.7%	**5.3%**	7.0%	5.3%	**6.6%**	**4.0%**
GEDGARCH	7.5%	5.4%	7.9%	5.5%	7.0%	4.8%	6.8%	4.3%
NEGARCH	7.7%	**4.6%**	8.8%	6.9%	9.5%	6.3%	8.5%	5.8%
TEGARCH	7.8%	4.7%	7.7%	6.3%	8.6%	6.3%	8.0%	4.6%
SKTEGARCH	8.1%	4.8%	**7.7%**	6.2%	8.6%	6.3%	7.9%	4.5%
GEDEGARCH	7.8%	4.6%	7.9%	6.5%	9.0%	5.8%	8.1%	5.4%
NGJR	7.7%	5.8%	9.6%	7.2%	8.0%	5.6%	8.2%	5.3%
TGJR	6.9%	5.3%	8.7%	6.3%	7.7%	5.7%	8.0%	4.6%
SKTGJR	6.9%	5.1%	8.8%	6.5%	7.7%	5.7%	7.9%	4.5%
GEDGJR	7.4%	5.6%	9.0%	6.6%	7.9%	5.5%	8.2%	5.1%

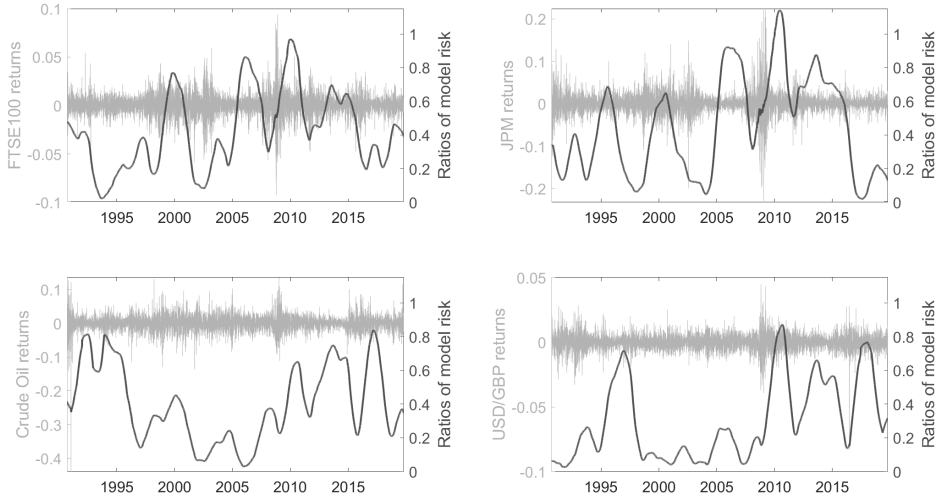
This table shows average ratios of the QLIKE-based model risk, with squared returns as the volatility proxy, to estimated variances of various models for different assets from 30/12/1983 to 21/10/2019. The model risk is computed over $n_1 = 250$ trading days using optimization windows of length $\tau_2 = 500$ and $\tau_3 = 1,000$. Numbers in bold (italics) are the lowest (highest) values per column.

Figure 4.5.4: Time-varying ratios of the QLIKE-based model risk to estimated variances, given a specific model applied to various assets

(a) SKTGARCH(1,1)



(b) RiskMetrics



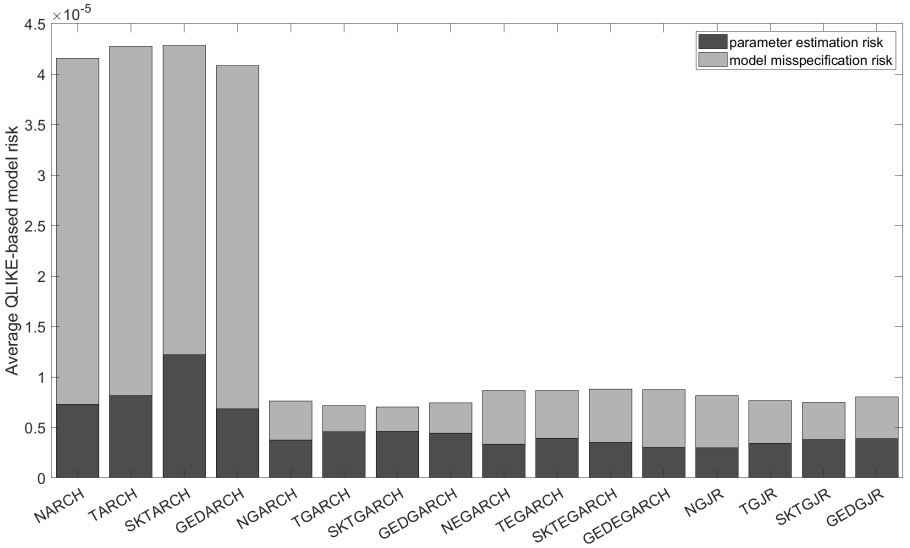
This figure shows time-varying ratios of the QLIKE-based model risk to estimated variances, given a specific model applied to various assets. The volatility proxy is squared return and model risk is computed over $n_1 = 250$ trading days using an optimization window of length $\tau_2 = 500$, based on data from 30/12/1983 to 21/10/2019.

as the volatility proxy. Considering the contrasting models, SKTGARCH(1,1) and RiskMetrics, we notice that the ratios of estimated model risk over the variance forecasts fluctuate dramatically between 1% and 115% for different assets. Particularly for the equity JP Morgan Chase, the peak value (about 115%) of the ratios of model risk estimates of the RiskMetrics model is around four times higher than for the SKTGARCH(1,1) model (about 30%). As such, the investors need to be conscious of the level of model risk of volatility models for different assets that increases in uncertain times.

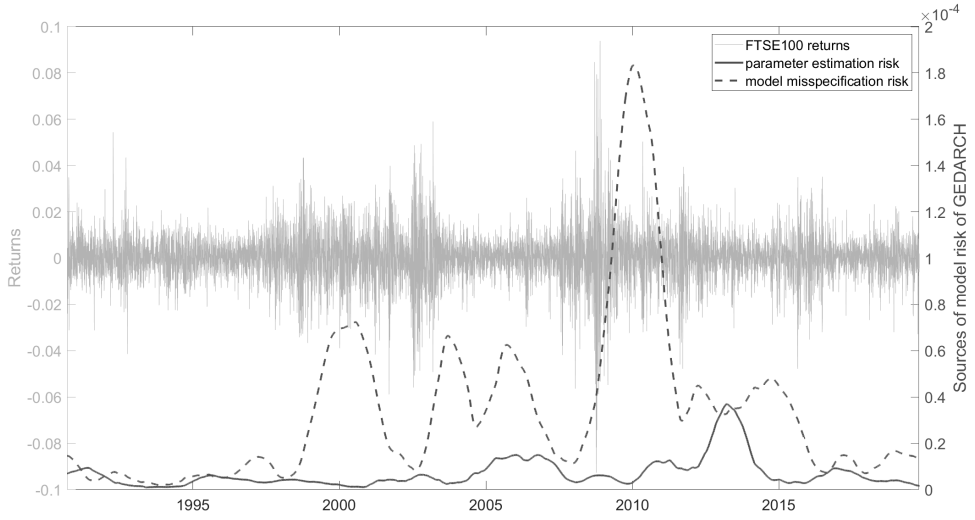
The components of model risk estimates are of much interest for the regulatory authority, practitioners and academics. The major sources of model risk are parameter estimation risk and model misspecification risk (Kerkhof et al., 2010). Figure 4.5.5 disentangles the QLIKE-based model risk estimates of volatility models into these two types of risk: panel (a) decomposes the model risk across various models; whilst panel (b) shows the time-varying values of the components of model risk for the GEDARCH model. The calculation of model risk is done over an optimization window $\tau_2 = 500$ and a model risk evaluation window $n_1 = 250$. The squared return is used as the volatility proxy. For a given volatility model, we compute estimation risk via simulations of this model with model parameters estimated on the FTSE100 Index from 30/12/1983 to 21/10/2019. Model misspecification risk generally contributes more to the total model risk than parameter estimation risk across various models and over time, as illustrated in panel (a) and panel (b) respectively, though a few exceptions appear when GARCH(1,1)

Figure 4.5.5: Decomposition of the QLIKE-based model risk of volatility models

(a) Average model risk across various models



(b) Time-varying model risk of GEDARCH



This figure shows components of the QLIKE-based model risk estimates for various models in panel (a) and for the GEDARCH model in panel (b), based on FTSE100 from 30/12/1983 to 21/10/2019. An optimization window $\tau_2 = 500$ and a model risk evaluation window $n_1 = 250$ are considered, and the volatility proxy is the squared return.

models are considered. When the market becomes volatile, model misspecification risk is aggravated. After the 2008 global financial crisis, we find that the estimate of model misspecification risk peaks, as can be seen in panel (b).

In an additional exercise, we investigate the relation between the constituents of model risk estimates and model dependent variance forecasts. Based on the daily prices and the 5-min realized variances of the FTSE100 Index from 04/01/2000 to 10/10/2019, the model risk of the set of volatility models is computed based on the QLIKE loss function over an optimization window $\tau_2 = 500$ and a model risk evaluation window $n_1 = 250$, using the squared return as the volatility proxy, and then it is decomposed into model misspecification risk and parameter estimation risk. Table 4.5.3 reports the coefficients, the associated t statistics with White (1980) standard errors robust to heteroscedasticity adjusted for clusters as well as the adjusted R^2 of the panel regressions based on a fixed-effects (within) estimation. More specifically, we regress model misspecification (estimation) risk estimates on the explanatory variables which are related to: RET is the average daily return; $RV5$ is the average 5-min realized variance and Var is the average variance estimate over the past 250 days. The results show an increase by 0.177 (0.196) in the values of adjusted R^2 after containing information on the models when model misspecification (estimation) risk is as endogenous variable. To this end, our proposed methodology can spot the inefficiency of volatility models in making volatility forecasts regarding major sources of model risk, which reinforces the reliability of this proposed technique.

Table 4.5.3: Panel regression results of misspecification risk and estimation risk

	Model misspecification risk		Parameter estimation risk	
RET	0.013***	0.022***	-0.001*	0.000
	(14.055)	(4.353)	(-1.926)	(0.037)
$RV5$	0.165***	0.149***	0.003*	-0.002
	(5.003)	(4.868)	(1.885)	(-1.222)
$Var - RV5$	0.351***		0.075***	
	(6.445)		(7.236)	
$RET \cdot RV5$	37.845**	8.245***	6.111**	2.004
	(2.330)	(4.381)	(2.798)	(1.237)
$RET \cdot (Var - RV5)$	221.177***		-2.213	
	(3.637)		(-0.388)	
$RV5 \cdot (Var - RV5)$	-447.549		-160.78***	
	(-1.6741)		(-5.331)	
$Adj.R^2$	0.516	0.339	0.221	0.025

*This table reports: 1) panel regression coefficients of model misspecification risk (or parameter estimation risk) on the variables shown in the first column where RET is the average daily return, RV5 is the average 5-min realized variance and Var is the average variance estimate over the past 250 days; 2) the associated t-statistics with White (1980) standard errors robust to heteroscedasticity adjusted for clusters, presented in parentheses; 3) adjusted R^2 of panel regressions in the last row. Calculations are based on the daily returns and the 5-min realized variances of the FTSE100 Index from 04/01/2000 to 10/10/2019. Model risk of the set of volatility models is estimated based on the QLIKE loss function over an optimization window $\tau_2 = 500$ and a model risk evaluation window $n_1 = 250$, using the squared return as the volatility proxy. *, **, *** indicate statistical significance at the 10%, 5% and 1% critical levels.*

4.6 Alternative measure of model risk

We consider an alternative definition of model risk measure, i.e. the RMSE formulation based on squared differences instead of the MAE formulation based on absolute differences in Section 4.2. For example, we can compute true model risk as below, rather than using the expression in (4.2.3):

$$p^j_{[t,t+n]} = \sqrt{\frac{1}{n+1} \cdot \sum_{i=t}^{t+n} \left(\sigma_i^2 - h_i^j\right)^2}. \tag{4.6.1}$$

In a similar manner, we can derive RMSE formulations for the true model risk proxy and the model risk estimates based on the MSE and QLIKE loss functions to replace (4.2.4), (4.2.6) and (4.2.8).

Table 4.6.1 reports the degree of similarity of the QLIKE-based model risk estimate to the true model risk, in which model risk is computed in RMSE formulations, based on simulated daily returns by the SKTGARCH(1,1) model. We consider different lengths of optimization windows and model risk windows and use the squared return as the variance proxy. Comparing with Panel A of Table 4.3.2, we find that the RMSE formulated model risk estimates in Table 4.6.1 derive similar values of correlations but tend to overestimate the magnitude of model risk, compared with the MAE formulated model risk estimates. Thus, our proposed model risk measure based on the MAE formulation are preferable over the alternative measure based on the RMSE formulation.

Table 4.6.1: Similarity of the QLIKE-based model risk estimate to the true model risk, using RMSE alternatives

Model risk estimate	optimization window length	model risk window length	$\bar{\mathcal{C}}^{\mathcal{M}}$	$\bar{\tau}_x^{\mathcal{M}}$	$\bar{\psi}^{\mathcal{M}}$
$\rho_1^{S_{qlike}}$	$\tau_1 = 250$	$n_1 = 250$	0.90	0.66	1.26
		$n_2 = 1,000$	0.97	0.61	1.15
	$\tau_2 = 500$	$n_1 = 250$	0.88	0.83	0.71
		$n_2 = 1,000$	0.95	0.92	0.56
	$\tau_3 = 1,000$	$n_1 = 250$	0.92	0.92	0.51
		$n_2 = 1,000$	0.94	1.00	0.42
	$\tau_4 = 2,000$	$n_1 = 250$	0.89	0.96	0.44
		$n_2 = 1,000$	0.93	1.00	0.36

This table shows similarity between the true model risk and QLIKE-based model risk estimates for the set of volatility models, computed using RMSE formulations based on data simulated by the SKTGARCH model. The squared return is used as the variance proxy. We consider optimization windows of length $\tau_1 = 250$, $\tau_2 = 500$, $\tau_3 = 1000$ *and* $\tau_4 = 2000$ *and model risk windows of length* $n_1 = 250$ *and* $n_2 = 1000$.

4.7 Conclusions

To assess the accuracy of volatility models which are of much importance in the financial world, we propose a new model risk estimation methodology based on scoring functions to measure the model risk of volatility models. We investigate this methodology considering the choice of volatility proxy (the squared return or the 5-min realized variance) and loss function (MSE or QLIKE) for a set of univariate models. It would be interesting to consider the model risk of volatility models in a multivariate setting in future research.

In a simulation analysis, we consider different optimization strategies to im-

prove on variance estimates, compare different lengths of optimization windows and model risk evaluation windows, and then recommend the QLIKE-based model risk estimation methodology with additive adjustments made to the volatility estimates, as we find that the proposed method leads to high correlations, averaging from 0.88 to 0.98, between the estimated and true model risk measures. Particularly the technique based on an optimization window of length $\tau_2 = 500$ and a model risk evaluation window of length $n_1 = 250$ is highly consistent with the true model risk measure, and can explain 65% of the true model risk on average across the models. We examine the desirable properties of a reasonable measure of model risk for our proposed technique, and find that the required properties are satisfied.

In an empirical study, we explore the effect of different volatility proxies on the proposed QLIKE-based model risk measure, concluding that the model risk measure using the squared return as volatility proxy generally produces a higher level of model risk for the badly fitting models (the RiskMetrics method, RW100 and the ARCH models), compared with the model risk measure that uses the realized variance. The level of estimated model risk based on the QLIKE loss function is not sensitive to the use of the volatility proxy across various models in general. After adjusting variance estimates for model risk, the degree of predictability of volatility models has been improved as evidenced by an increase in the values of adjusted R^2 of the MZ regressions.

In addition, applying our proposed methodology to several asset classes, we

find that the RiskMetrics method, the historical volatility measure RW1000 and the ARCH models are most affected by model risk, and that the volatility models applied to various assets carry a higher level of model risk during stressed market states than in normal market states, as expected. We also show that model misspecification risk generally contributes more to model risk than parameter estimation risk.

Appendices

4.A Density functions for error distributions

Normal density function

The probability density function of the normal distribution is:

$$f(z|\mu_z, \sigma_z) = \frac{1}{\sigma_z\sqrt{2\pi}} \exp^{-(z-\mu_z)^2/2\sigma_z^2},$$

where μ_z and σ_z denote the mean and standard deviation of z.

Student's t density function

The Student's t density function is written as:

$$f(z|\nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})} \frac{1}{\sqrt{\nu\pi}} \frac{1}{(1+\frac{z^2}{\nu})^{\frac{\nu+1}{2}}},$$

where ν denotes the degrees of freedom and $\Gamma(\cdot)$ denotes the Gamma function.

Skewed Student's *t* density function

Following Hansen (1994), the skewed Student's t density function is given as:

$$f(z|\nu,\lambda) = \begin{cases} bc\left(1+\frac{1}{\nu-2}\left(\frac{bz+a}{1-\lambda}\right)^2\right)^{-(\nu+1)/2}, & \text{if } z < -a/b, \\ bc\left(1+\frac{1}{\nu-2}\left(\frac{bz+a}{1+\lambda}\right)^2\right)^{-(\nu+1)/2}, & \text{if } z \geq -a/b, \end{cases}$$

where the degree of freedom parameter ν with $2 < \nu < \infty$ controls the kurtosis and the skewness parameter λ is $-1 < \lambda < 1$. The constants a, b and c are given by:

$$a = 4\lambda c\left(\frac{\nu-2}{\nu-1}\right), \quad b^2 = 1+3\lambda^2-a^2, \quad \text{and } c = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi(\nu-2)}\Gamma(\frac{\nu}{2})}.$$

Generalized error distribution (GED) density function

The probability density function of the generalized error distribution of the standardized residuals z beyond the threshold u is shown as below, where ξ and β are the shape and scale parameters with $\beta > 0$, respectively:

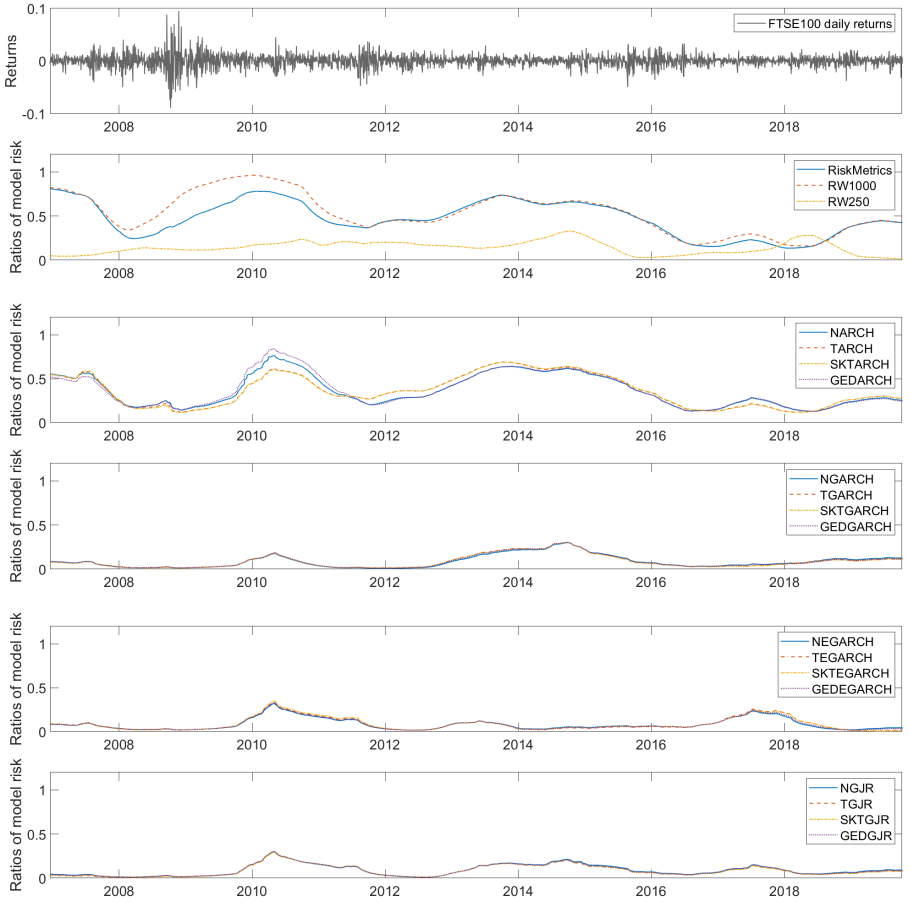
$$f(z|\xi,\beta) = \begin{cases} 1-(1+\xi z/\beta)^{-1/\xi}, & \text{if } \xi > 0, \\ 1-\exp(-z/\beta), & \text{if } \xi = 0, \end{cases} \qquad for\ all\ z \geq u.$$

4.B Additional results

Figure 4.B.1 reports the time-varying ratios of the QLIKE-based model risk estimates of various models to variance estimates where the volatility proxy used is the 5-min realized variance, which can be compared with Figure 4.5.2.

Table 4.B.1 reports the panel regression coefficients of misspecification (estimation) risk on the variables as shown in the first column, which can be compared with Table 4.5.3.

Figure 4.B.1: Time-varying ratios of the QLIKE-based model risk estimates to estimated variances



This figure shows time-varying ratios of the QLIKE-based model risk estimates to estimated variances. The 5-min realized variance is used as the volatility proxy. Model risk is computed over $n_1 = 250$ trading days using the optimization window length of $\tau_2 = 500$, based on the FTSE100 Index from 04/01/2000 to 10/10/2019.

Table 4.B.1: Panel regression results of misspecification risk and estimation risk

	Model misspecification risk		Parameter estimation risk	
RET	0.007***	0.023***	-0.002**	0.000
	(6.159)	(4.414)	(-2.759)	(0.014)
RET^2	0.153***	0.135***	0.002	-0.002
	(4.962)	(4.867)	(1.010)	(-1.559)
$Var - RET^2$	0.336***		0.070***	
	(6.674)		(7.250)	
$RET \cdot RET^2$	70.041***	9.307***	5.537***	1.659
	(3.319)	(4.715)	(3.268)	(1.241)
$RET \cdot (Var - RET^2)$	272.023***		3.656	
	(6.317)		(0.839)	
$RET^2 \cdot (Var - RET^2)$	-276.988		-97.477***	
	(-1.690)		(-5.073)	
$Adj.R^2$	0.534	0.335	0.241	0.027

*This table reports: 1) panel regression coefficients of model misspecification risk (or parameter estimation risk) on the variables shown in the first column where RET is the average daily return, RET^2 is the average squared return and Var is the average variance estimate over the past 250 days; 2) the associated t-statistics with White (1980) standard errors robust to heteroscedasticity adjusted for clusters, presented in parentheses; 3) adjusted R^2 of panel regressions in the last row. Calculations are based on the daily returns and the 5-min realised variances of the FTSE100 Index from 04/01/2000 to 10/10/2019. Model risk of the set of volatility models is computed based on the QLIKE loss function over an optimization window $\tau_2 = 500$ and a model risk evaluation window $n_1 = 250$, using the squared return as the volatility proxy. *, **, *** indicate statistical significance at the 10%, 5% and 1% critical levels.*

Notes

[1]Other transformations of the latent variables in similar regressions are discussed by Jorion (1995), Bollerslev and Wright (2001), and Hansen and Lunde (2006).

[2]This definition is based on absolute differences. We also use an alternative formulation based on squared differences with similar (albeit weaker) results which are available in Appendix 4.B.

[3]We use the default "interior-point" method in MATLAB for a constrained minimization problem.

[4]The model parameters are constrained to satisfy that $\hat{\omega}, \hat{\alpha}, \hat{\beta} > 0, \hat{\alpha} + \hat{\beta} < 1, 2 < \hat{\nu} < \infty$ and $-1 < \hat{\lambda} < 1$.

[5]See a comprehensive review of volatility models in Hansen and Lunde (2005).

[6]Density functions for standardised error distributions considered here are shown in Appendix 4.A.

[7]We also examine the model risk measures related to the formulation of RMSE as alternatives to (4.2.3), (4.2.4), (4.2.6) and (4.2.8), and then produce the similar dynamic correlation between true and estimated model risk as seen in Figure **??** of Appendix 4.B.

[8]Thanks to the data available from https://realized.oxford-man.ox.ac.uk/.

[9]The purpose of computing the ratio of model risk to the variance forecast of a given model is to make an easy comparison across various volatility models and assets.

Chapter 5

Conclusions and Further Research

5.1 Summary of the Findings and Contributions of the Thesis

This thesis makes significant contributions to the quantification of model risk of the widely used risk models and volatility models using several methods. It thus provides guidance on model risk management for the regulatory authorities, risk managers and financial decision-making for practitioners.

In Chapter 2, we quantify the model risk of ES as an optimal correction needed for ES to pass several ES backtests jointly, regarding the desirable backtesting criteria: 1) an appropriate frequency of exceptions; 2) the absence of autocor-

relations in exceptions; and 3) a suitable magnitude of exceptions. Considering whether the backtesting-based correction method for ES satisfies the coherence properties of a measure of risk from a regulatory point of view, in a simulation study we find that all the properties hold for our chosen measure of ES model risk with respect to the UC_{ES} and CC_{ES} tests, except that the subadditivity property is not guaranteed. We compare the 2.5% ES with the 1% VaR in terms of model risk across different models and based on different assets. We find that the 2.5% ES is less affected by model risk than the 1% VaR, needing a smaller correction to pass the three ES backtests jointly. Besides, commodity ES carries the highest model risk especially around 2008, compared to equity and bond ES. Moreover, we consider the impact of VaR model risk on ES model risk in terms of the ES calculations and the ES backtests. If VaR model risk is first accounted for, then ES model risk reduces by approximately 50%. The results are strengthened when the standard deviations of the corrections for model risk are considered: the GARCH(1,1) models not only require the smallest corrections for model risk, but the level of the corrections are the most stable, when compared to the other models considered in this study.

In Chapter 3, we develop a general scoring function-based model risk estimation methodology to quantify joint (VaR, ES) model risk and ES model risk, and disentangle the components of model risk of financial market risk models, based on strictly consistent FZ scoring functions applied to the risk functionals (VaR, ES). We show that, when model risk is present, the ordering of (VaR, ES) models

is sensitive to the FZ specification function, although in the simulated cases the model ranking is not sensitive to the choice of homogeneous FZ scoring function when the pair of (VaR, ES) is estimated at small critical levels (e.g., 2.5%). The proposed model risk estimation methodology is confirmed with a simulation study in which we use three specific FZ scoring functions which are 0, 0.5 and -1 positively homogenous. We find a high similarity between the true and estimated model risk of (VaR, ES) risk measures as well as for the ES model risk, across various risk models, with correlations varying from 0.8 to 0.987, with an explanatory power above 50%. In a simulation analysis, the newly proposed measures of joint model risk and ES model risk satisfy numerically all coherence properties of a measure of risk, except for the subadditivity property. The empirical results point out that, among all models considered the RiskMetrics model and Historical Simulation have a very high level of joint model risk and ES model risk, particularly during extreme events. In addition, the backtesting performance of these models is improved upon adjusting for model risk.

In Chapter 4, we introduce a model risk methodology for volatility estimates based on scoring functions. We study this methodology via simulations by comparing different optimization strategies and different lengths of optimization windows and model risk evaluation windows, and then recommend the QLIKE-based model risk estimation method with additive adjustments made to the volatility estimates, as we find that the proposed method leads to high correlations, averaging from 0.88 to 0.98, between the estimated and true model risk measures. Partic-

ularly the technique based on an optimization window of length $\tau_2 = 500$ and a model risk evaluation window of length $n_1 = 250$ is highly consistent with the true model risk measure, and can explain 65% of the true model risk on average across the models. We examine the desirable properties of a reasonable measure of model risk for our proposed technique, and find that the required properties are satisfied. In an empirical study, we explore the effect of different volatility proxies (squared returns and 5-min realized variances, respectively) on the proposed QLIKE-based model risk measures, concluding that the level of estimated model risk based on the QLIKE loss function is not sensitive to the choice of volatility proxy across various models in general. After adjusting the variance estimates for model risk, the efficiency of volatility models can be improved as evidenced by an increase in the values of adjusted R^2 of the MZ regressions. Additionally, applying our proposed methodology to several asset classes, we find that the RiskMetrics method, the historical volatility measure RW1000 and the ARCH-type models are most affected by model risk, and show that misspecification risk generally contributes more to model risk than estimation risk.

5.2 Suggestions for Future Research

Whilst we believe that this thesis makes significant contributions to model risk measurement of market risk models or of univariate volatility models, there are still many gaps that need to be filled in, which would expand our knowledge

about financial risk management. In the following, we discuss future research that builds on the findings of this thesis in several directions.

Financial Risk This study produces model risk estimates for the models of market risk measures and volatility forecasting in a univariate setting and identifies the major sources of model risk, namely, parameter estimation risk and model misspecification risk. First, the analysis may be extended to a multivariate setting. In the diagram of risk estimation process for risk estimates shown in Figure 2.2.1 of Chapter 2, model risk arises at step 2 that would specify and estimate the models describing the risk factors, which has been addressed in this study, whilst it would be interesting to estimate model risk occurring at step 3 that would model the P&L of a portfolio as a function of these risk factors.

Second, the scoring function-based model risk estimation methodology introduced by Chapter 3 may facilitate other extensions for measuring model risk of the predictive models. For example, one can investigate the individual VaR model risk measure by replacing the FZ class with the GPL class.

Finally, whilst the coherence properties of our proposed model risk measures based on scoring functions in Chapter 3 and Chapter 4 are examined via Monte Carlo simulations, these properties await consideration on a theoretical front.

Scoring Function This thesis significantly contributes to the implications of using scoring functions in measuring model risk. The current literature (e.g. Gneiting, 2011 and Patton et al., 2019) documents the implications of using scor-

ing functions in making model comparisons and estimating model parameters. However, it is not very clear what consequences the choice of scoring function will bring about. It may be worth connecting the analysis of model risk measures based on a single scoring function to measures based on a combination of scoring functions. Another extension would construct a model parameter estimation methodology drawing on the combined scoring functions using different weights.

Backtesting Chapter 2 relates model error to statistical backtesting, and finds a correction required for ES in order to pass ES backtests jointly. Another promising avenue for future research would derive some backtesting methodologies for the regulatory risk measures like VaR and ES which take into account the effect of model risk on risk estimates.

Volatility Forecasting Chapter 4 shows that the model risk of the broad (G)ARCH-type models increases following crisis periods, and that the forecasting ability of these models can be improved after accounting for model risk. This invites the research on improving on these models. Additionally, what is the relation between model risk and volatility clustering? To what extent can model risk be explained by economic/financial variables?

Relatedly, another interesting area of research lies in the stochastic volatility models as well as the volatility models based on high-frequency data. One may attempt to quantify the model risk and improve the forecasting accuracy of models in these categories.

References

Acerbi, C., Szekely, B., 2014. Backtesting expected shortfall. Risk 27, 76–81.

Acerbi, C., Tasche, D., 2002. On the coherence of expected shortfall. Journal of Banking & Finance 26, 1487–1503.

Alexander, C., Sarabia, J. M., 2012. Quantile uncertainty and value-at-risk model risk. Risk Analysis: An International Journal 32, 1293–1308.

Alizadeh, S., Brandt, M. W., Diebold, F. X., 2002. Range-based estimation of stochastic volatility models. Journal of Finance 57, 1047–1091.

Andersen, T. G., Bollerslev, T., Christoffersen, P. F., Diebold, F. X., 2013. Financial risk measurement for financial risk management. In: *Handbook of the Economics of Finance*, Elsevier, vol. 2, pp. 1127–1220.

Andersen, T. G., Bollerslev, T., Diebold, F. X., Labys, P., 2003. Modeling and forecasting realized volatility. Econometrica 71, 579–625.

Angelidis, T., Degiannakis, S. A., 2006. Backtesting var models: A two-stage procedure. Journal of Risk Model Validation 1, 27–48.

Artzner, P., Delbaen, F., Eber, J.-M., Heath, D., 1999. Coherent measures of risk. Mathematical Finance 9, 203–228.

Barndorff-Nielsen, O. E., Shephard, N., 2002. Estimating quadratic variation using realized variance. Journal of Applied Econometrics 17, 457–477.

Barrieu, P., Ravanelli, C., 2015. Robust capital requirements with model risk. Economic Notes 44, 1–28.

Barrieu, P., Scandolo, G., 2015. Assessing financial model risk. European Journal of Operational Research 242, 546–556.

Basel Committee on Banking Supervision, 2011. A global regulatory framework for more resilient banks and banking systems. Available online at: `http://www.bis.org/publ/bcbs189.pdf`.

Basel Committee on Banking Supervision, 2019. Minimum capital requirements for market risk. Available online at: `https://www.bis.org/bcbs/publ/d457.pdf`.

Bauwens, L., Hafner, C. M., Laurent, S., 2012. Handbook of volatility models and their applications, vol. 3. John Wiley & Sons.

Bellini, F., Bignozzi, V., 2015. On elicitable risk measures. Quantitative Finance 15, 725–733.

Berkowitz, J., 2001. Testing density forecasts, with applications to risk management. Journal of Business & Economic Statistics 19, 465–474.

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

Berkowitz, J., Obrien, J., 2002. How accurate are value-at-risk models at commercial banks? Journal of Finance 57, 1093–1111.

Bollerslev, T., 1986. Generalized autoregressive conditional heteroskedasticity. Journal of Econometrics 31, 307–327.

Bollerslev, T., Wright, J. H., 2001. High-frequency data, frequency domain inference, and volatility forecasting. Review of Economics and Statistics 83, 596–602.

Boucher, C. M., Daníelsson, J., Kouontchou, P. S., Maillet, B. B., 2014. Risk models-at-risk. Journal of Banking & Finance 44, 72–92.

Boudoukh, J., Richardson, M. P., Stanton, R., Whitelaw, R., 2004. Maxvar-long horizon value-at-risk in a mark-to-market environment. Journal of Investment Management 2, 1–6.

Brock, W. A., Durlauf, S. N., West, K. D., 2003. Policy evaluation in uncertain economic environments. Tech. rep., National Bureau of Economic Research.

Brock, W. A., Durlauf, S. N., West, K. D., 2007. Model uncertainty and policy evaluation: Some theory and empirics. Journal of Econometrics 136, 629–664.

Broda, S. A., Paolella, M. S., 2011. Expected shortfall for distributions in finance. In: *Statistical tools for finance and insurance*, Springer, pp. 57–99.

Campbell, S. D., 2006. A review of backtesting and backtesting procedures. Journal of Risk 9, 1–17.

Candelon, B., Colletaz, G., Hurlin, C., Tokpavi, S., 2010. Backtesting value-at-risk: a gmm duration-based test. Journal of Financial Econometrics 9, 314–343.

Chen, Y., Wang, Z., Zhang, Z., 2019. Mark to market value-at-risk. Journal of Econometrics 208, 299–321.

Christoffersen, P., 1998. Evaluating interval forecasts. International Economic Review 39, 841–62.

Christoffersen, P., 2009. Backtesting. Encyclopedia of Quantitative Finance Available online at: `https://onlinelibrary.wiley.com/doi/10.1002/9780470061602.eqf15018`.

Christoffersen, P., Gonçalves, S., 2005. Estimation risk in financial risk management. Journal of Risk 7, 1–28.

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

Christoffersen, P. F., 2012. Elements of financial risk management. Academic Press.

Clift, S. S., Costanzino, N., Curran, M., 2016. Empirical performance of backtesting methods for expected shortfall. Available at SSRN 2618345 .

Colletaz, G., Hurlin, C., Pérignon, C., 2013. The risk map: A new tool for validating risk models. Journal of Banking & Finance 37, 3843–3854.

Cont, R., 2006. Model uncertainty and its impact on the pricing of derivative instruments. Mathematical Finance 16, 519–547.

Cont, R., Deguest, R., Scandolo, G., 2010. Robustness and sensitivity analysis of risk measurement procedures. Quantitative Finance 10, 593–606.

Costanzino, N., Curran, M., 2015. Backtesting general spectral risk measures with application to expected shortfall. Journal of Risk Model Validation 9, 21–31.

Costanzino, N., Curran, M., 2018. A simple traffic light approach to backtesting expected shortfall. Risks 6, 2–8.

Daníelsson, J., James, K., Valenzuela, M., Zer, I., 2016. Model risk of risk models. Journal of Financial Stability 23, 79–91.

Daníelsson, J., Jorgensen, B. N., Samorodnitsky, G., Sarma, M., de Vries, C. G., 2013. Fat tails, var and subadditivity. Journal of Econometrics 172, 283–291.

Daníelsson, J., Zhou, C., 2017. Why risk is so hard to measure. Systemic Risk Centre discussion paper 36. London School of Economics.

Davis, M. H., 2016. Verification of internal risk measure estimates. Statistics & Risk Modeling 33, 67–93.

Detering, N., Packham, N., 2016. Model risk of contingent claims. Quantitative Finance 16, 1357–1374.

Diebold, F. X., Hickman, A., Inoue, A., Schuermann, T., 1997. Converting 1-day volatility to h-day volatility: scaling by $\sqrt{h}$ is worse than you think. Wharton Working Paper No. 97-34.

Diebold, F. X., Mariano, R. S., 1995. Comparing predictive accuracy. Journal of Business and Economic Statistics 13, 253–263.

Dimitriadis, T., Bayer, S., 2019. A joint quantile and expected shortfall regression framework. Electronic Journal of Statistics 13, 1823–1871.

Du, Z., Escanciano, J. C., 2016. Backtesting expected shortfall: accounting for tail risk. Management Science 63, 940–958.

Efron, B., 1991. Regression percentiles using asymmetric squared error loss. Statistica Sinica 1, 93–125.

Ehm, W., Gneiting, T., Jordan, A., Krüger, F., 2016. Of quantiles and expectiles: consistent scoring functions, choquet representations and forecast rankings.

Journal of the Royal Statistical Society: Series B (Statistical Methodology) 78, 505–562.

Embrechts, P., Puccetti, G., Rüschendorf, L., Wang, R., Beleraj, A., 2014. An academic response to Basel 3.5. Risks 2, 25–48.

Emmer, S., Kratz, M., Tasche, D., 2015. What is the best risk measure in practice? a comparison of standard measures. Journal of Risk 18, 31–60.

Emond, E. J., Mason, D. W., 2002. A new rank correlation coefficient with application to the consensus ranking problem. Journal of Multi-Criteria Decision Analysis 11, 17–28.

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

Engle, R. F., Manganelli, S., 2004. Caviar: Conditional autoregressive value at risk by regression quantiles. Journal of Business & Economic Statistics 22, 367–381.

Engle, R. F., Siriwardane, E. N., 2018. Structural garch: the volatility-leverage connection. Review of Financial Studies 31, 449–492.

Escanciano, J. C., Olmo, J., 2010a. Backtesting parametric value-at-risk with estimation risk. Journal of Business & Economic Statistics 28, 36–51.

Escanciano, J. C., Olmo, J., 2010b. Robust backtesting tests for value-at-risk models. Journal of Financial Econometrics 9, 132–161.

Escanciano, J. C., Pei, P., 2012. Pitfalls in backtesting historical simulation var models. Journal of Banking & Finance 36, 2233–2244.

European Banking Authority, 2014. Guidelines on common procedures and methodologies for the supervisory review and evaluation process (srep). Available online at: `https://eba.europa.eu/regulation-and-policy/supervisory-review-and-evaluation-srep-and-pillar-2/`.

Farkas, W., Fringuellotti, F., Tunaru, R., 2016. Regulatory capital requirements: Saving too much for rainy days? EFMA annual meeting.

Federal Reserve, 2011. Supervisory guidance on model risk management. Available online at: `https://www.federalreserve.gov/supervisionreg/srletters/sr1107.htm`.

Fissler, T., Fasciati-Ziegel, J., Gneiting, T., 2016. Expected shortfall is jointly elicitable with value-at-risk implications for backtesting. Risk Magazine pp. 58–61.

Fissler, T., Ziegel, J. F., 2016. Higher order elicitability and Osbands principle. Annals of Statistics 44, 1680–1707.

Fissler, T., Ziegel, J. F., et al., 2019. Order-sensitivity and equivariance of scoring functions. Electronic Journal of Statistics 13, 1166–1211.

Forsberg, L., Ghysels, E., 2007. Why do absolute returns predict volatility so well? Journal of Financial Econometrics 5, 31–67.

Gaglianone, W. P., Lima, L. R., Linton, O., Smith, D. R., 2011. Evaluating value-at-risk models via quantile regression. Journal of Business & Economic Statistics 29, 150–160.

Garcia, R., Renault, É., Tsafack, G., 2007. Proper conditioning for coherent VaR in portfolio management. Management Science 53, 483–494.

Giacomini, R., White, H., 2006. Tests of conditional predictive ability. Econometrica 74, 1545–1578.

Glasserman, P., Xu, X., 2014. Robust risk measurement and model risk. Quantitative Finance 14, 29–58.

Glosten, L. R., Jagannathan, R., Runkle, D. E., 1993. On the relation between the expected value and the volatility of the nominal excess return on stocks. Journal of Finance 48, 1779–1801.

Gneiting, T., 2011. Making and evaluating point forecasts. Journal of the American Statistical Association 106, 746–762.

Hansen, B. E., 1994. Autoregressive conditional density estimation. International Economic Review 35, 705–730.

Hansen, P. R., Lunde, A., 2005. A forecast comparison of volatility models: does anything beat a GARCH(1, 1)? Journal of Applied Econometrics 20, 873–889.

Hansen, P. R., Lunde, A., 2006. Consistent ranking of volatility models. Journal of Econometrics 131, 97–121.

Hansen, P. R., Lunde, A., Nason, J. M., 2011. The model confidence set. Econometrica 79, 453–497.

Hartz, C., Mittnik, S., Paolella, M., 2006. Accurate value-at-risk forecasting based on the normal-garch model. Computational Statistics & Data Analysis 51, 2295–2312.

Huggenberger, M., Zhang, C., Zhou, T., 2018. Forward-looking tail risk measures. Available at SSRN: https://ssrn.com/abstract=2909808 .

Jorion, P., 1995. Predicting volatility in the foreign exchange market. Journal of Finance 50, 507–528.

Jorion, P., 1996. Risk2: Measuring the risk in value at risk. Financial Analysts Journal 52, 47–56.

Kellner, R., Rösch, D., 2016. Quantifying market risk with value-at-risk or expected shortfall? Consequences for capital requirements and model risk. Journal of Economic Dynamics and Control 68, 45 – 63.

Kellner, R., Rösch, D., Scheule, H. H., 2016. The role of model risk in extreme value theory for capital adequacy. Journal of Risk 18, 39–70.

Kerkhof, J., Melenberg, B., Schumacher, H., 2010. Model risk and capital reserves. Journal of Banking & Finance 34, 267–279.

Klaassen, F., 2002. Improving garch volatility forecasts with regime-switching garch. In: *Advances in Markov-Switching Models*, Springer, pp. 223–254.

Krätschmer, V., Schied, A., Zähle, H., 2012. Qualitative and infinitesimal robustness of tail-dependent statistical functionals. Journal of Multivariate Analysis 103, 35–47.

Krätschmer, V., Schied, A., Zähle, H., 2014. Comparative and qualitative robustness for law-invariant risk measures. Finance and Stochastics 18, 271–295.

Krätschmer, V., Schied, A., Zähle, H., 2015. Quasi-hadamard differentiability of general risk functionals and its application. Statistics & Risk Modeling 32, 25–47.

Kratz, M., Lok, Y. H., McNeil, A. J., 2018. Multinomial var backtests: A simple implicit approach to backtesting expected shortfall. Journal of Banking & Finance 88, 393–407.

Kuan, C.-M., Yeh, J.-H., Hsu, Y.-C., 2009. Assessing value at risk with care, the conditional autoregressive expectile models. Journal of Econometrics 150, 261–270.

Kupiec, P. H., 1995. Techniques for verifying the accuracy of risk measurement models. Journal of Derivatives 3, 73–84.

Lazar, E., Zhang, N., 2019. Model risk of expected shortfall. Journal of Banking & Finance 105, 74 – 93.

Leccadito, A., Boffelli, S., Urga, G., 2014. Evaluating the accuracy of value-at-risk forecasts: New multilevel tests. International Journal of Forecasting 30, 206–216.

Lönnbark, C., 2013. On the role of the estimation error in prediction of expected shortfall. Journal of Banking & Finance 37, 847–853.

McNeil, A. J., Frey, R., 2000. Estimation of tail-related risk measures for heteroscedastic financial time series: an extreme value approach. Journal of Empirical Finance 7, 271–300.

McNeil, A. J., Frey, R., Embrechts, P., 2015. Quantitative risk management: Concepts, techniques and tools - Revised edition. Princeton university press.

Mincer, J. A., Zarnowitz, V., 1969. The evaluation of economic forecasts. In: *Economic forecasts and expectations: Analysis of forecasting behavior and performance*, NBER, pp. 3–46.

Moldenhauer, F., Pitera, M., 2019. Backtesting expected shortfall: a simple recipe? Journal of Risk, forthcoming .

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

Newey, W. K., Powell, J. L., 1987. Asymmetric least squares estimation and testing. Econometrica: Journal of the Econometric Society pp. 819–847.

Nieto, M. R., Ruiz, E., 2016. Frontiers in VaR forecasting and backtesting. International Journal of Forecasting 32, 475–501.

Nocedal, J., Wright, S., 2006. Numerical optimization. Springer Science & Business Media.

Nolde, N., Ziegel, J. F., 2017a. Elicitability and backtesting: Perspectives for banking regulation. Annals of Applied Statistics 11, 1833–1874.

Nolde, N., Ziegel, J. F., 2017b. Rejoinder:“ Elicitability and backtesting: Perspectives for banking regulation”. Annals of Applied Statistics 11, 1901–1911.

Patton, A. J., 2011. Volatility forecast comparison using imperfect volatility proxies. Journal of Econometrics 160, 246–256.

Patton, A. J., 2019. Comparing possibly misspecified forecasts. Journal of Business & Economic Statistics pp. 1–23, forthcoming.

Patton, A. J., Ziegel, J. F., Chen, R., 2019. Dynamic semiparametric models for expected shortfall (and value-at-risk). Journal of Econometrics 211, 388–413.

Pitera, M., Schmidt, T., 2018. Unbiased estimation of risk. Journal of Banking & Finance 91, 133–145.

Ruiz, E., Pascual, L., 2002. Bootstrapping financial time series. Journal of Economic Surveys 16, 271–300.

So, M. K., Wong, C.-M., 2012. Estimation of multiple period expected shortfall and median shortfall for risk management. Quantitative Finance 12, 739–754.

West, K. D., 1996. Asymptotic inference about predictive ability. Econometrica: Journal of the Econometric Society 64, 1067–1084.

White, H., 1980. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. Econometrica 48, 817–838.

Wong, W. K., 2008. Backtesting trading risk of commercial banks using expected shortfall. Journal of Banking & Finance 32, 1404–1415.

Wong, W. K., 2010. Backtesting value-at-risk based on tail losses. Journal of Empirical Finance 17, 526–538.

Ziegel, J. F., 2016. Coherence and elicitability. Mathematical Finance 26, 901–918.

Ziegel, J. F., Krüger, F., Jordan, A., Fasciati, F., 2017. Murphy diagrams: Forecast evaluation of expected shortfall. arXiv preprint arXiv:1705.04537 .

Ziegel, J. F., Krüger, F., Jordan, A., Fasciati, F., 2020. Robust forecast evaluation of expected shortfall. Journal of Financial Econometrics 18, 95–120.