Model risk of expected shortfall

Model Risk of Expected Shortfall *

Emese Lazar† and Ning Zhang‡

Abstract

In this paper we propose to measure the model risk of Expected Shortfall as the optimal correction needed to pass several ES backtests, and investigate the properties of our proposed measures of model risk from a regulatory perspective. Our 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.

Keywords: model risk, Expected Shortfall, backtesting.

JEL classification: C15, C22, C52, C53, G15.

*We would like to thank Radu Tunaru and the anonymous reviewers for their valuable comments, and the session participants at the 2018 Financial Management Association European Conference, the 2018 International Association for Applied Econometrics Annual Conference, the 2018 Financial Econometric Conference, Lancaster University and the 12th International Conference on Computational and Financial Econometrics.

†Correspondence to: Emese Lazar, ICMA Centre, Henley Business School, University of Reading, Whiteknights, Reading RG6 6BA, UK; e.lazar@icmacentre.ac.uk

‡ICMA Centre, Henley Business School, University of Reading, Whiteknights, Reading RG6 6BA, UK; N.Zhang3@pgr.reading.ac.uk

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 (Danielsson 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. Danielsson 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

[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

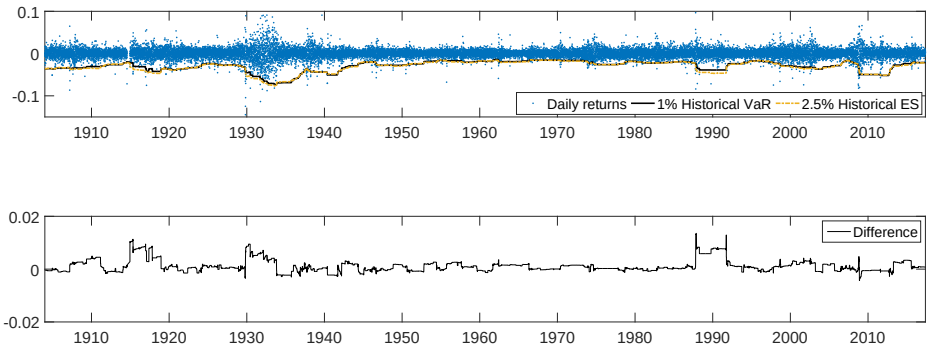


Figure 1: DJIA index daily returns, the daily historical VaR estimates ($\alpha = 1\%$) and the daily historical ES estimates ($\alpha = 2.5\%$) from 28/12/1903 to 23/05/2017, as well as the difference between the 2.5% historical ES and the 1% historical VaR are presented. We use a four-year rolling window to compute the risk estimates.

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. 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 6**, **Appendix B**), measuring the model risk of ES has become plausible.

Figure 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;

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:

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

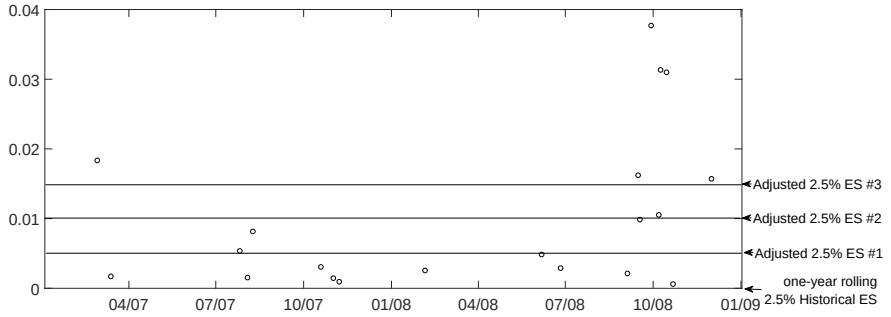


Figure 2: Peaked-over-ES and adjustments, based on the DJIA index from 01/01/2007 to 01/01/2009. One-year moving window is used to forecast daily historical ES ($\alpha = 2.5\%$).

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**, 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 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 paper, we mainly focus on several ES backtests with respect to the following properties[4] 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

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

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 the paper is as follows: section 2 analyzes the sources of ES model risk focusing on estimation and specification errors, and performs Monte Carlo simulations to quantify them; section 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 4 presents the empirical study and section 5 concludes.

2 Model risk of Expected Shortfall

2.1 Sources of model risk

We first establish a general scheme (see **Figure 3**) 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).

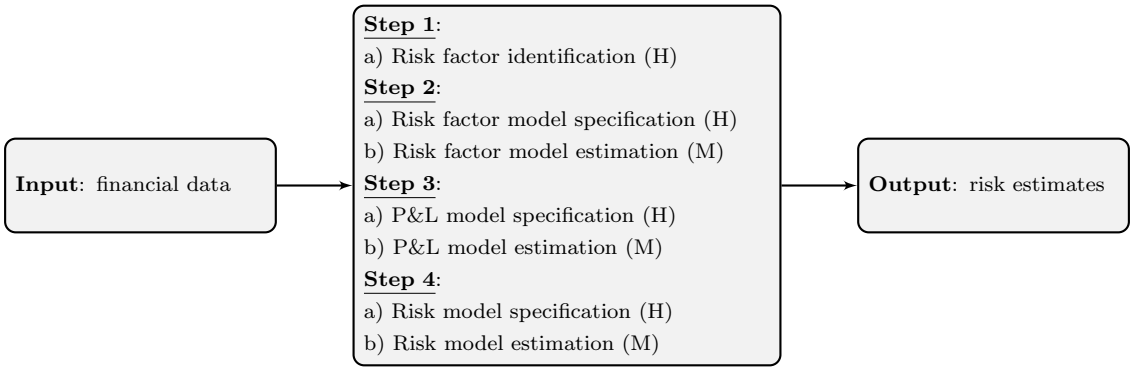


Figure 3: Risk estimation process

Notation: H and M represent high and medium impacts on risk estimates, respectively.

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 A**. In a similar vein, we investigate the impact of the earlier mentioned two errors on the ES estimates, deriving the theoretical formulae for estimation and specification errors, as well as correction of ES. VaR[5], for a given distribution function F and a given significance level α, is defined as:

$$VaR_t(\alpha) = -inf\{q : F(q) \geq \alpha\}, \tag{2.1}$$

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

[5]The values of VaR and ES are considered positive in this paper.

losses into account; it is given by:

$$ES_t(\alpha) = \frac{1}{\alpha}\int_0^{\alpha} VaR_t(u)du \tag{2.2}$$

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:

$$ThVaR(\theta_0, \alpha) = -q_{\alpha}^{F} = -F_{\alpha}^{-1} \tag{2.3}$$

$$ThES(\alpha) = \frac{1}{\alpha}\int_0^{\alpha} ThVaR(\theta_0, u)du \tag{2.4}$$

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:

$$ThVaR(\theta_0, \alpha) = VaR(\hat{\theta}_0, \alpha) + bias(\theta_0, \hat{\theta}_0, \alpha) \tag{2.5}$$

We also have that:

$$ThVaR(\theta_0, \alpha) - \mathbb{E}(VaR(\hat{\theta}_0, \alpha)) = \mathbb{E}(bias(\theta_0, \hat{\theta}_0, \alpha)) \tag{2.6}$$

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 ES($\hat{\theta}_0, \alpha$), and we have that

$$ThES(\theta_0, \alpha) - \mathbb{E}[ES(\hat{\theta}_0, \alpha)] = \frac{1}{\alpha}\int_0^{\alpha} \mathbb{E}[bias(\theta_0, \hat{\theta}_0, v)]dv, \tag{2.7}$$

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

$$ES^{E}(\hat{\theta}_0, \alpha) = ES(\hat{\theta}_0, \alpha) + \frac{1}{\alpha}\int_0^{\alpha} \mathbb{E}[bias(\theta_0, \hat{\theta}_0, v)]dv \tag{2.8}$$

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:

$$VaR(\hat{\theta}_1, \alpha) = -q_\alpha^{\hat{F}} = -\hat{F}_\alpha^{-1} \tag{2.9}$$

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

$$ThVaR(\theta_0, \alpha) = VaR(\hat{\theta}_1, \alpha) + bias(\theta_0, \theta_1, \hat{\theta}_1, \alpha) \tag{2.10}$$

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:

$$ThVaR(\theta_0, \alpha) - \mathbb{E}(VaR(\hat{\theta}_1, \alpha)) = \mathbb{E}(bias(\theta_0, \theta_1, \hat{\theta}_1, \alpha)) \tag{2.11}$$

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), the mean estimation and specification biases of ES can be formulated as below:

$$ThES(\theta_0, \alpha) - \mathbb{E}[ES(\hat{\theta}_1, \alpha)] = \frac{1}{\alpha}\int_0^\alpha \mathbb{E}[bias(\theta_0, \theta_1, \hat{\theta}_1, v)]dv \tag{2.12}$$

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

$$ES^{SE}(\hat{\theta}_1, \alpha) = ES(\hat{\theta}_1, \alpha) + \frac{1}{\alpha}\int_0^\alpha \mathbb{E}[bias(\theta_0, \theta_1, \hat{\theta}_1, v)]dv \tag{2.13}$$

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 C**, we give a review of risk forecasting models used in this paper.

2.3 Monte Carlo simulations

In this section, assume a simplified risk estimation process (**Figure 3**) 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 (2.7). We also forecast ES based on other models to examine the values of joint estimation and specification biases of ES, as in equation (2.12).

In our setup[6], a GARCH(1,1) model with normal disturbances (GARCH(1,1)-N) is assumed to be the 'true' data generating process, given by:

$$r_t = \mu + \varepsilon_t \tag{2.14}$$

$$\varepsilon_t = \sigma_t \cdot z_t,\ z_t \sim \mathcal{N}(0,1) \tag{2.15}$$

$$\sigma^2{}_t = \omega + \alpha\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2 \tag{2.16}$$

Using market data, we first estimate the parameters[7] 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 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**, 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.

[6]We also consider a different model, MS(2)-GARCH(1,1)-N, as the data generating process, and give simulated biases in **Table 10**, **Appendix E**.

[7]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 $\beta = 0.9017$.

Table 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

Note: The results are 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 $\alpha = 5\%$, 2.5% and 1%.

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.

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.

3 Measuring ES model risk

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 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 paper, 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:

$$ES^B_{i,j}(\hat{\theta}_1, \alpha) = ES_j(\hat{\theta}_1, \alpha) + C^*_{i,j} \tag{3.1}$$

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 our paper, $i \in \{1, 2, 3, 4\}$; $C_{1,j}$, $C_{2,j}$, 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 paper, we define model risk as $MR : \mathbb{R}^n \times V_M \to \mathbb{R}^+$, where $MR\left((X_{0,t}), M_j\right)$ 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 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 $MR : \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^+$:

$$MR^I\left((X_{0,t}), (v_{j,t}), (e_{j,t})\right) = \max_I(C^*_{i,j}). \tag{3.2}$$

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 test), the conditional coverage test (CC 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.

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/CC 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 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:

$$\hat{H}_t(\alpha) = \frac{1}{\alpha}(\alpha - \hat{u}_t)\mathbf{1}(\hat{u}_t \leqslant \alpha) \tag{3.3}$$

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 $t-1$.

$$\hat{u}_t = \hat{F}(r_t, \Omega_{t-1}, \hat{\theta}_1) \tag{3.4}$$

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

$$H_1 : \mathbb{E}\left[H_t(\alpha, \theta_0) - \frac{\alpha}{2}\right] = 0 \tag{3.5}$$

Hence, the simple t-test statistic[8] and its distribution is:

$$U_{ES} = \frac{\sqrt{n}\left(1/n \sum_{t=1}^{n} \hat{H}_t(\alpha) - \alpha/2\right)}{\sqrt{\alpha(1/3 - \alpha/4)}} \sim N(0,1) \tag{3.6}$$

Exception frequency and independence test

The conditional coverage test (CC 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:

$$H_2 : \mathbb{E}\left[H_t(\alpha, \theta_0) - \frac{\alpha}{2} | \Omega_{t-1}\right] = 0 \tag{3.7}$$

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[8], $C_{ES}(1)$, is written as:

$$C_{ES}(1) = \frac{n^3}{(n-1)^2} \cdot \frac{\left(\sum_{t=2}^{n} (\hat{H}_t(\alpha) - \alpha/2)(\hat{H}_{t-1}(\alpha) - \alpha/2)\right)^2}{\left(\sum_{t=1}^{n} (\hat{H}_t(\alpha) - \alpha/2)(\hat{H}_t(\alpha) - \alpha/2)\right)^2} \sim \chi^2(1) \tag{3.8}$$

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.

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

Exception frequency and magnitude test

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

$$\mathrm{Z}_2 = \sum_{t=1}^{T} \frac{r_t I_t}{T\alpha ES_{\alpha,t}} + 1 \tag{3.9}$$

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[9], 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:

$$er_t = (r_t + ES_{\alpha,t}) \cdot I_t \tag{3.10}$$

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.

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.

[9]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).

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 risk models that are not able to accommodate for bigger losses 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 datasets.

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 (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 and CC 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, CC 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 Danielsson et al. (2005), we consider two independent assets, X_1 and X_2, but with the same distribution, specified as:

$$X = \epsilon + \eta, \quad \epsilon \sim \text{IID}\mathcal{N}(0,1), \quad \eta = \begin{cases} 0 & \text{with a probability } 0.991 \\ -10 & \text{with a probability } 0.009 \end{cases} \tag{3.11}$$

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

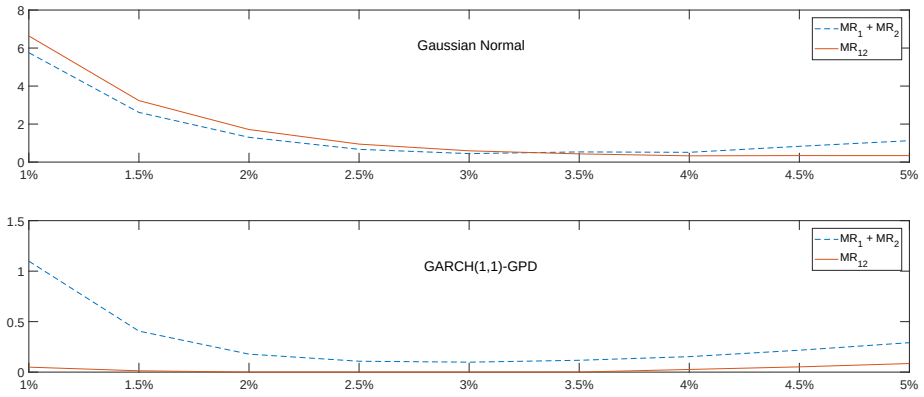


Figure 4: Average values of ES model risk of an equally weighted portfolio, $(X_1 + X_2)$, and the sum of ES model risks of X_1 and X_2, based on the Gaussian Normal ES and the GARCH(1,1)-GPD ES for a series of significance levels.

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 4**. 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 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 4**, subadditivity is satisfied when the model fits the data well.

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 paper. 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). 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)[10]. 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.

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:

$$r_t = \sigma_t Z_t, \qquad \sigma_t^2 = 2.18 \times 10^{-6} + 0.109 r_{t-1}^2 + 0.890 \sigma_{t-1}^2, \tag{3.12}$$

[10]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).

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 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** 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 (3.12). Thus, according to the last two rows in **Table 2**, 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.

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 01/01/1900 to 05/03/2017 (29,486 daily returns in total). Based on equation (3.1), we quantify the model risk of ES as the maximum of minimum corrections required to pass the ES backtests[11] and make comparisons among different models, where backtesting is performed over a year. Moreover, we examine this measure of model risk based

[11]The UC and CC 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 distirbutions (Broda and Paolella, 2009) can also be examined using these backtests.

Table 2: 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.

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	0.046	0.03	0.03	0.3%	0.3%
GARCH(1,1)-GPD	0.045	0.03	0.02	0.7%	0.6%
GARCH(1,1)-t	0.097	0.01	0.01	0.3%	0.3%
DGP	0.046	0.00	0.00	0.1%	0.1%

Note: 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.

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 5 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 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 6 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 test), the conditional test for ES (CC test) and the magnitude test

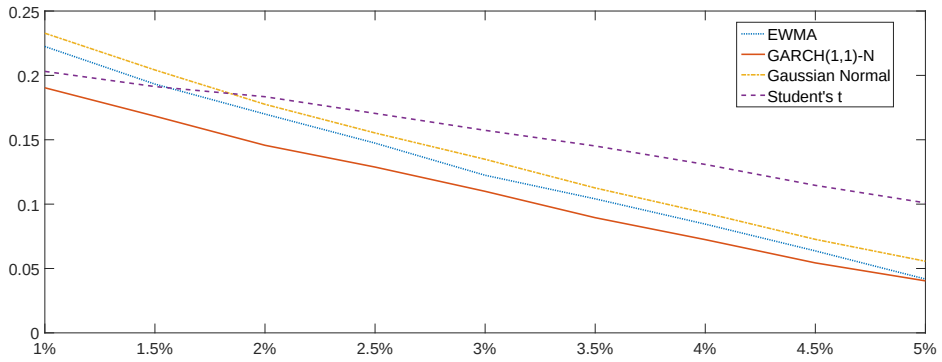
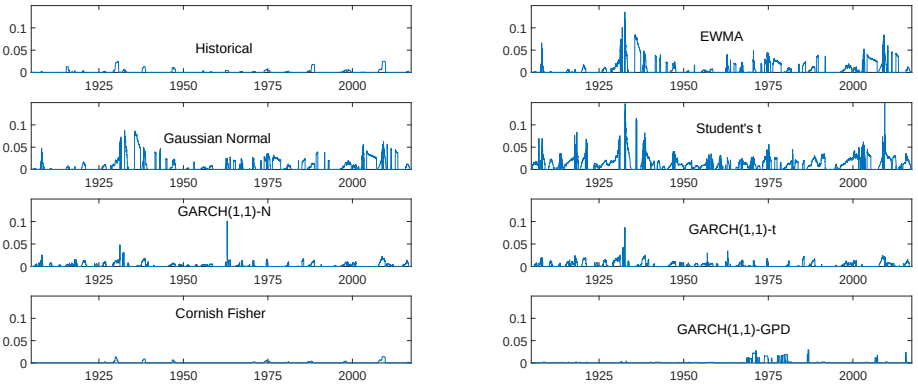


Figure 5: Relative corrections based on the UC test made to the daily ES associated with EWMA, GARCH(1,1)-N, Gaussian Normal, and Student's t along with a range of alpha levels, which is computed as the ratio of the absolute correction over the average daily ES.

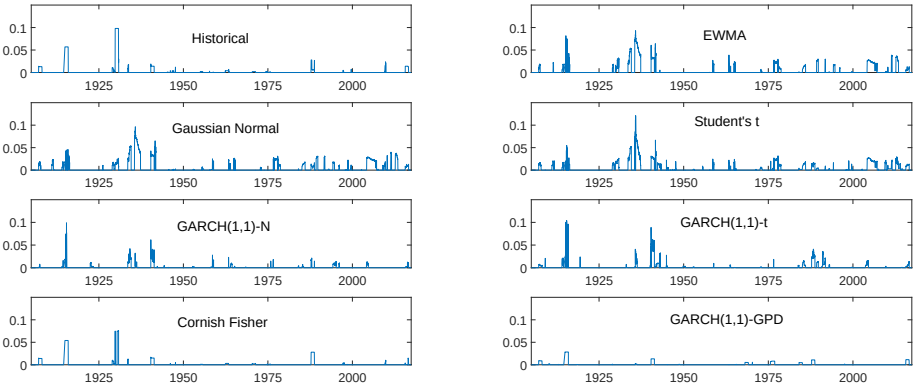
(Z_2 test), respectively, where ES is estimated at a 2.5% significance level using a four-year moving window[12] and the evaluation period for backtesting procedures is one year. This figure shows that a series of dynamic adjustments are needed for the daily ES ($\alpha = 2.5\%$) across all different 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, CC, 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 and CC 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 7**, for the UC, CC 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

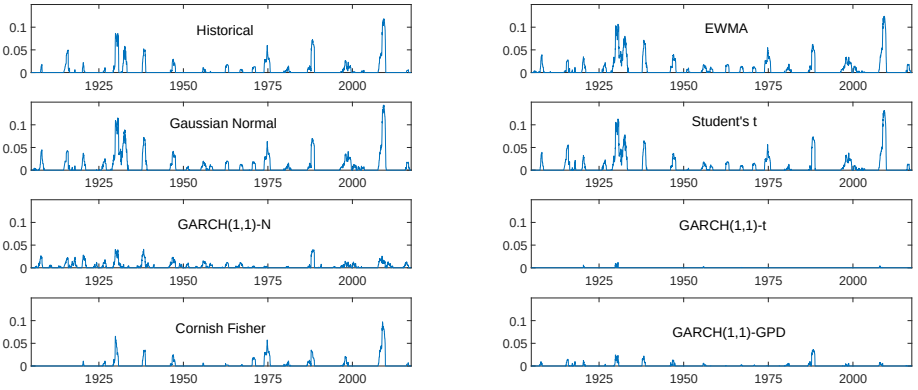
[12]The results computed using a five-year moving window and a three-year moving window are very similar to those required here (available from the authors on request).



(a) UC test for ES



(b) CC test for ES



(c) Z_2 test

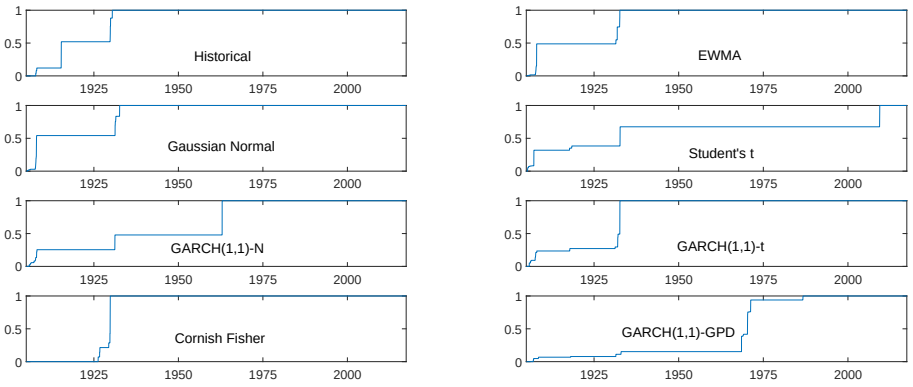
Figure 6: Dynamic optimal corrections made to the daily ES estimates ($\alpha = 2.5\%$) associated with various models for the DJIA index from 01/01/1900 to 23/05/2017, required to pass the UC , CC , and Z_2 tests, respectively. The parameters are re-estimated using a four-year moving window (1,000 daily returns) and the evaluation window for backtesting is one year.

Appendix D, Table 7 for the dates when the highest corrections are required). When considering the UC and CC tests, the highest values of the optimal corrections made to the daily ES of various models are achieved before the 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 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 D, Figure 9**, which presents extreme optimal corrections of ES forecasts based on different models, required to pass various backtests.

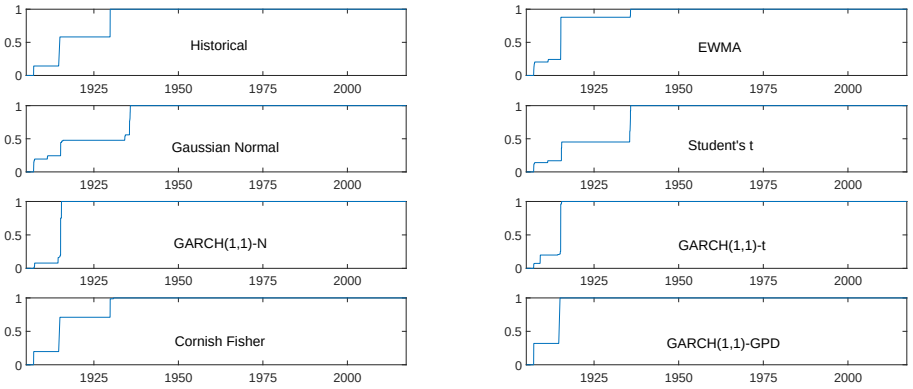
In **Table 3**, we measure the model risk of ES forecasts made by various risk models for the DJIA index, and compare the model risk of the 2.5% ES with that of the 1% VaR. Besides, we look into how ES model risk is affected by the model risk of VaR as discussed in section 3.4. **Panel A** and **Panel 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 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%, repectively, 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)[13], we compute the relative corrections made to one-step ahead

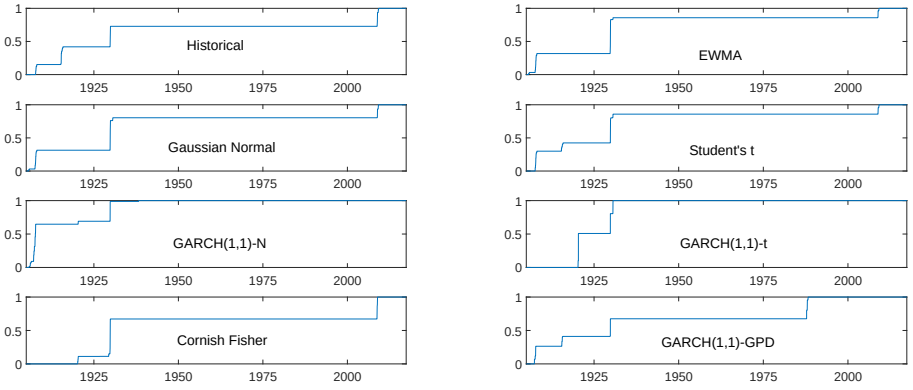
[13]Boucher et al. (2014) only present the results for the 5% VaR.



(a) UC test for ES



(b) CC test for ES



(c) Z_2 test

Figure 7: Required relative optimal adjustments made to the daily ES estimates by passing the UC, CC, Z_2 tests, which is expressed as the ratio of the corrections over the maximum of the optimal corrections over the entire period.

VaR forecasts by passing three VaR backtests[14], reported in **Panel C** of **Table 3**. 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 and CC 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 3**. 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 accomdating 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 9, Appendix 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.

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

Table 3: 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, based on different backtests across various models.

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%

Note: 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 test), the conditional coverage test (CC 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.

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 MSFT shares (adjusted or non-adjusted for dividends) from 01/01/1987 to 04/10/2017. **Panel A and B** of **Table 4** report the absolute and relative corrections required for the GARCH(1,1)-GPD ES ($\alpha = 2.5\%$) of various asset classes[15]. 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 8** of **Appendix 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 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 5** 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

[15]See the data source in the note to **Table 4**.

Table 4: Maximum and mean of the absolute and relative corrections made to the daily GARCH(1,1)-GPD ES ($\alpha = 2.5\%$), and the relative corrections made to the daily GARCH(1,1)-GPD VaR ($\alpha = 1\%$) for different asset classes based on different backtests.

	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%

Note: 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, CC 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).

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 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.

Table 5: Dollar exposures to the model risk of GARCH(1,1)-GPD ES ($\alpha = 2.5\%$) of the USD/GBP exchange rate and Microsoft equity, based on various ES backtests.

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

Note: 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, CC and Z_2 tests accordingly, when considering a position of 1 million dollars in the asset specified in the first column.

5 Conclusions

In this paper, 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 and CC 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. 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.

Appendix A. Theoretical analysis of estimation and specification errors of VaR[16]

Estimation bias and correction of VaR

Based on equation (2.5) and (2.6), correcting for the estimation error, the VaR estimate can be written as:

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

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 (2.10) and (2.11), correcting for these biases (specification and estimation biases), the VaR estimate can be written as:

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

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.

[16]The analysis is based on Boucher et al. (2014).

Appendix B. Backtesting measures of VaR and ES

Table 6: Selected backtesting methodologies for VaR and ES

VaR backtests	ES backtests
Exception Frequency Tests:	***Exception Frequency Tests:***
1)UC test - Kupiec (1995)	1)UC 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 (2017)
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 test- Christoffersen (1998)	1)CC test- Du and Escanciano (2016); Costanzino and Curran (2015, 2018)
2)dynamic quantile-Engle and Manganelli (2004);Patton et al. (2018)	2)dynamic quantile- Patton et al. (2018)
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)

Appendix 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 paper, 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

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

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

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

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

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 $\hat{r}_{t+\tau} = \hat{\mu}_{t+\tau} + \hat{\sigma}_{t+\tau}\Phi^{-1}_{\alpha}$, where $\hat{\mu}_{t+\tau}$ and $\hat{\sigma}^2_{t+\tau}$ 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 $VaR^{\alpha}_{t+\tau}$ as

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

ES can be derived as

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

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:

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

where $\Gamma(\cdot)$ is the gamma function and d is the degree of freedom larger than 2. The one-step ahead return is $\hat{r}_{t+\tau} = \hat{\mu}_{t+\tau} + \hat{\sigma}_{t+\tau}t^{-1}_{\alpha}(\hat{d})$, where $t^{-1}_{\alpha}(\hat{d})$ 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

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

ES is given by

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

where $\hat{\mu}_{t+\tau}$ and $\hat{\sigma}^2_{t+\tau}$ 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

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

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 (C.3) and (C.4). We can refer to equation (C.6) and (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

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

VaR and ES are computed as in equations (C.3) and (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 our paper, we use the GARCH(1,1) model with standardised t disturbances, combined with the EVT methodology (GARCH(1,1)-GPD). First, we obtain the standardised empirical losses via GARCH(1,1), assuming they are distributed as a standardised t distribution.

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

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

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

where

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

$\tilde{x}$ 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

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

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:

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

where

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

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

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

where

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

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.

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

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

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

ES is formulated as

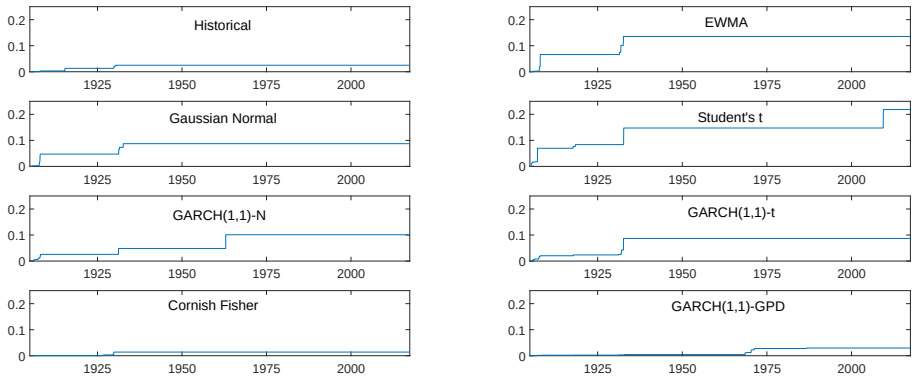
$$\widehat{ES}^{\alpha}_{t+\tau} = -\hat{\sigma}_{t+\tau} ES_{CF(\alpha)} \tag{C.20}$$

where

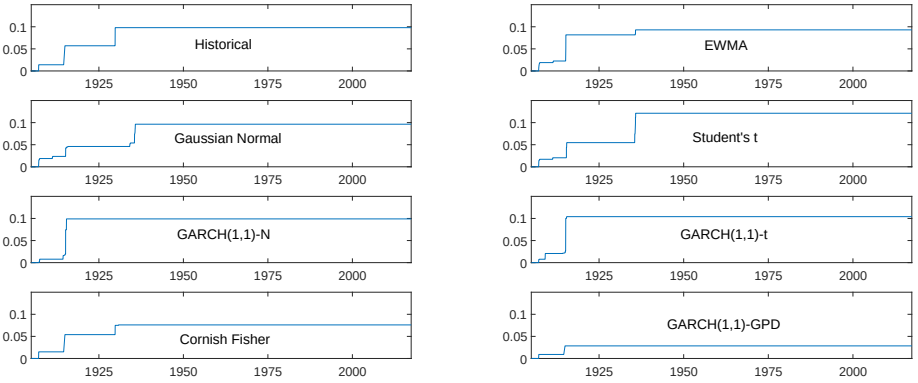
$$ES_{CF(\alpha)} = \frac{-\phi(CF_{\alpha}^{-1})}{\alpha}\left[1 + \frac{\hat{\zeta}_1}{6}(CF_{\alpha}^{-1})^3 + \frac{\hat{\zeta}_2}{24}\left[(CF_{\alpha}^{-1})^4 - 2(CF_{\alpha}^{-1})^2 - 1\right]\right] \tag{C.21}$$

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

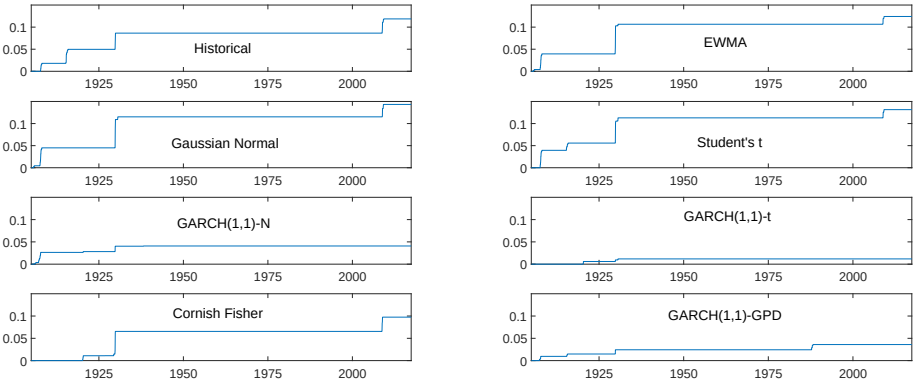
Appdendix D. Empirical results



(a) UC test for ES



(b) CC test for ES



(c) Z_2 test

Figure 8: Historical maximum of required optimal adjustments made to the daily ES estimates by passing the UC, CC and Z_2 tests, respectively.

Table 7: The highest values of the absolute minimum corrections made to the daily ES ($\alpha = 2.5\%$) based on several models and different backtests.

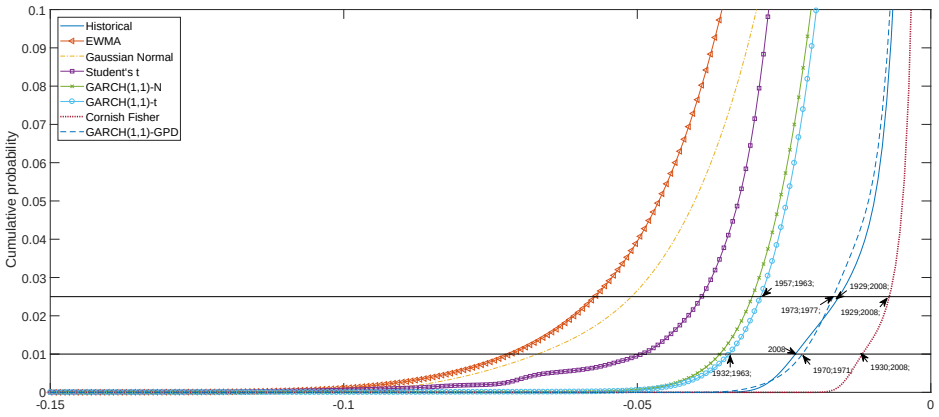
		UC test		CC test		Z_2 test	
Model		Date	C_1	Date	C_2	Date	C_3
Historical	1	16/06/1930	0.0250	29/10/1929	0.0980	20/04/2009	0.1186
	2	11/09/2009	0.0240	14/12/1914	0.0570	30/03/2009	0.1176
	3	20/11/2008	0.0230	30/10/1930	0.0300	05/03/2009	0.1172
	4	12/12/1929	0.0220	13/12/1915	0.0280	19/05/2009	0.1167
EWMA	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

Note: The results 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 test), the conditional coverage test (CC test) and the magnitude test (Z_2 test), respectively.

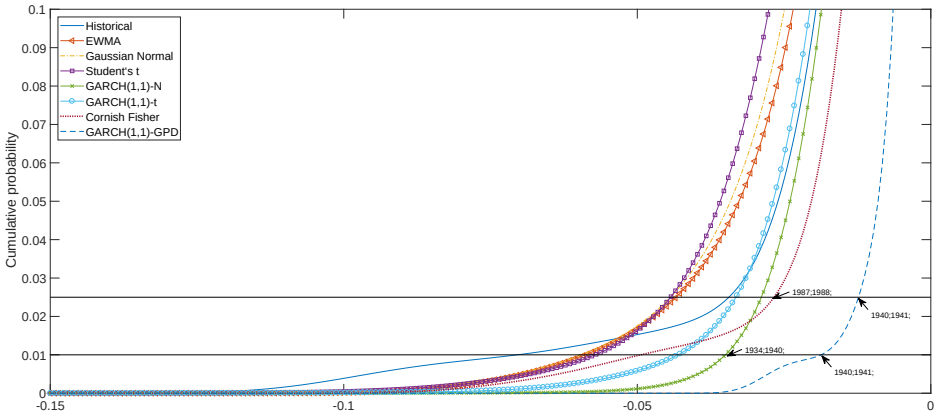
Table 8: The highest values of the absolute minimum corrections made to the GARCH(1,1)-GPD ES ($\alpha = 2.5\%$) for different assets by passing different backtests.

		UC test		CC 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

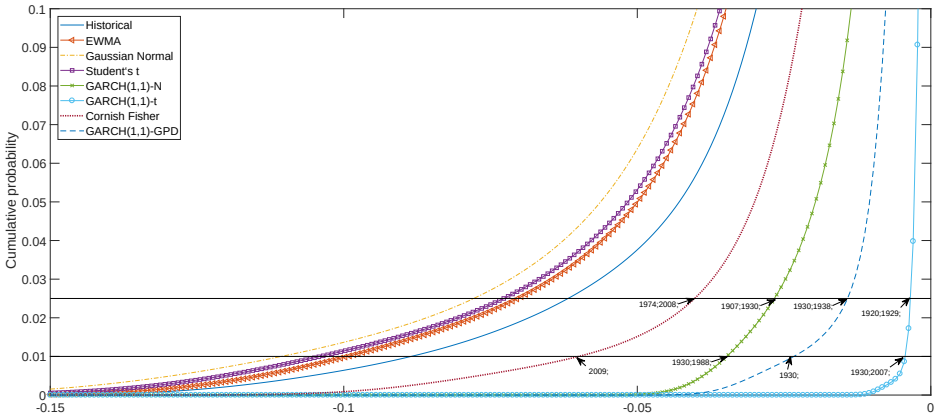
Note: Downloaded from DataStream. For the equity, we use a composite index with 95% "MSCI Europe Index" and 5% "MSCI World Index"; for the bond, we use the "Bank of America Merrill Lynch US Treasury & Agency Index"; for the commodity, we use the "CRB Spot Index", from 31/10/1986 to 07/07/2017. We compute 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 test), the conditional coverage test (CC test) and the magnitude test (Z_2 test), respectively.



(a) UC test for ES



(b) CC test for ES



(c) Z_2 test

Figure 9: The left tail of the cumulative distribution (using Gaussian Kernel smoothing) of the negative of required optimal adjustments made to the daily ES estimates by passing the UC, CC, and Z_2 tests, respectively.

Table 9: 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, CC and Z_2 backtests.

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

Note: Based on the DJIA index from 01/01/1900 to 23/05/2017, downloaded from DataStream.

Appendix E. Simulated Bias

Similar to **Table 1**, we conduct a simulation study to show the impacts of estimation and specification biases on the ES forecasts in the table below. 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:

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

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:

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

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 (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 10**.

Table 10: 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

Note: The results are based on the DJIA index from 03/01/2000 to 30/12/2011, downloaded from DataStream.

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.

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.

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

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

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.

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

Broda, S., Paolella, M., 2009. Calculating expected shortfall for distributions in finance. Tech. rep., Mimeo, Swiss Banking Institute, University of Zurich.

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.

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

Christoffersen, P., 2009. Backtesting. Encyclopedia of Quantitative Finance .

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. Available at SSRN 2514403 .

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

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

Danielsson, J., Jorgensen, B. N., Mandira, S., Samorodnitsky, G., De Vries, C. G., 2005. Subadditivity re-examined: the case for value-at-risk. Tech. rep., Cornell University Operations Research and Industrial Engineering.

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

Du, Z., Escanciano, J. C., 2016. Backtesting expected shortfall: accounting for tail risk. Management Science 63, 940–958.

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.

Engle, R. F., Manganelli, S., 2004. Caviar: Conditional autoregressive value at risk by regression quantiles. Journal of Business & Economic Statistics 22, 367–381.

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.

Farkas, W., Fringuellotti, F., Tunaru, R., 2016. Regulatory capital requirements: Saving too much for rainy days? EFMA annual meeting.

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.

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 .

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.

Kupiec, P. H., 1995. Techniques for verifying the accuracy of risk measurement models. Journal of Derivatives 3, 73–84.

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.

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.

Moldenhauer, F., Pitera, M., 2017. Backtesting expected shortfall: is it really that hard? arXiv preprint arXiv:1709.01337 .

Patton, A. J., Ziegel, J. F., Chen, R., 2018. Dynamic semiparametric models for expected shortfall (and value-at-risk). Journal of Econometrics, forthcoming .

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.

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.