Journal of Financial Econometrics, 2021, 1–41
doi: 10.1093/jjfinec/nbab008
Article

OXFORD

Backtesting Value-at-Risk and Expected Shortfall in the Presence of Estimation Error*

Sander Barendse [1], Erik Kole[2] and Dick van Dijk[3]

[1]Department of Economics, University of Oxford, and Nuffield College, University of Oxford, [2]Econometric Institute, Erasmus University Rotterdam, and Tinbergen Institute and [3]Econometric Institute, Erasmus University Rotterdam, Tinbergen Institute, and Erasmus Research Institute of Management

Address correspondence to Roetersstraat 11, 1018 WB Amsterdam, The Netherlands, or e-mail: s.c.barendse@uva.nl.

Received August 15, 2019; revised February 12, 2021; editorial decision March 1, 2021; accepted March 2, 2021

Abstract

We investigate the effect of estimation error on backtests of expected shortfall (ES) forecasts. These backtests are based on first-order conditions of a recently introduced family of jointly consistent loss functions for value-at-risk (VaR) and ES. For both single and multiperiod horizons, we provide explicit expressions for the additional terms in the asymptotic covariance matrix that result from estimation error, and propose robust tests that account for it. Monte Carlo experiments show that the tests that ignore these terms suffer from size distortions, which are more pronounced for higher ratios of out-of-sample to in-sample observations. Robust versions of the backtests perform well with power against common alternatives. We also introduce a novel standardization of the conditional joint test statistic that removes the need to estimate higher-order moments and significantly improves its performance. In an application to VaR and ES forecasts for daily FTSE 100 index returns as generated by (GJR-)GARCH and HEAVY models, we find that estimation error substantially impacts the outcome of the backtests, and is not bound to particular subperiods such as the credit crisis.

Key words: backtesting, expected shortfall, risk management, tail risk, value-at-risk

JEL classification: C12, C53, C58, G17

* We thank Fabio Trojani (the editor), three anonymous referees, Julie Schnaitmann and seminar participants at the CFE-CMStatistics 2018 conference in Pisa for valuable comments and feedback. Errors remain our own.

Research into financial risk management can nowadays be split into two strands. The first and oldest investigates risk measures from both a theoretical and practical perspective [see Emmer, Kratz, and Tasche (2015) for a recent example]. The second, more recent strand investigates what Cont, Deguest, and Scandolo (2010) have coined the risk measurement procedure, that is, the procedure with which the forecast of the risk measure is obtained. It encompasses the selection of the model, the method and data selection to estimate or calibrate the model parameters, and the forecasting method.[1] This procedure is typically evaluated by a backtest, that is, a formal statistical assessment of the quality of the risk forecasts by means of a loss function. Nolde and Ziegel (2017) propose a framework to jointly backtest expected shortfall (ES) and value-at-risk (VaR), motivated by the fact that ES is only elicitable in combination with VaR (see also Gneiting, 2011; Fissler and Ziegel, 2016). Because they are based on the elicitability property, the resulting backtests are not only suited to test the correctness of a given risk measurement procedure, but can also be used to rank competing alternative procedures. Furthermore, Nolde and Ziegel (2017) propose unconditional as well as conditional tests for correct specification.

In this paper, we focus on the effect that the estimation part of the risk measurement procedure has on backtesting. Because the parameters of the model with which the risk forecasts are generated are typically not known but estimated, the backtests are affected by estimation error. For a sound evaluation and comparison of risk measurement procedures, this effect should be taken into consideration. We quantify this effect for the tests of correct specification proposed by Nolde and Ziegel (2017) and propose robust versions of these tests that account for it.

The theoretical foundation of our analysis follows from West (1996) and McCracken (2000), who investigate the impact of estimation error on out-of-sample tests of predictive ability (see also West, 2006). Because the tests of Nolde and Ziegel (2017) are derived from an identification function implied by the family of jointly consistent loss functions for VaR and ES introduced in Fissler and Ziegel (2016), they fit into the general framework of McCracken (2000). We establish that estimation error results in additional terms in the asymptotic covariance matrix of the test statistics and find explicit expressions and consistent estimators for these terms. They are functions of the estimation scheme, that is, the choice of fixed, rolling or expanding window estimation, and the (asymptotic) ratio of in-sample and out-of-sample observations, and they apply to both single and multi-period forecasting. We then propose robust tests that use this consistent estimator of the asymptotic covariance matrix. We also introduce a novel conditional test statistic that standardizes the deviations from the forecasts by the forecast themselves. Because it removes the need to estimate higher-order moments, estimation of the covariance matrix becomes easier.

Our analysis of the effect of estimation error on joint backtests of VaR and ES and the robust test versions we propose complement earlier research in several respects, notably Escanciano and Olmo (2010) and Du and Escanciano (2016). Escanciano and Olmo (2010) examine the effect of estimation error in a similar way as we do, though they focus on one-period-ahead forecasts of VaR only, whereas our framework applies to VaR and ES jointly

1 See Giacomini and White (2006) for a general discussion of these issues in forecasting. For an investigation of choices of models and forecasting methods for VaR, see, for example, Kole et al. (2017).

and to single and multi-period forecasting. Du and Escanciano (2016) propose conditional tests of ES also limited to one-period-ahead forecasts, and robust versions of these and the corresponding unconditional test. These tests do not meet the conditions of Nolde and Ziegel (2017), and their derivation of the impact of estimation error additionally requires an estimate of the complete conditional distribution of the realization. We include the tests of Escanciano and Olmo (2010) and Du and Escanciano (2016) in our analysis for comparison.

We conduct several Monte Carlo experiments to examine the effect of estimation error on the size and power properties of the standard backtests, and to examine the extent to which the proposed robustification corrects this. We evaluate conditional and unconditional joint tests for VaR and ES forecasts with coverage levels of both 95% and 97.5% for 1- and 10-day-ahead forecasts. The size experiments are based on a standard GARCH data generating process (DGP) with Student's t distributed errors. In the power experiments we consider deviations in the specification of the mean, the volatility, or the error distribution. We estimate the parameters of the model of the null hypothesis with a rolling window scheme, and analyze the tests for several choices of the lengths of the in-sample and out-of-sample windows.

Our results show that estimation error leads to size distortions. The empirical rejection rates at the 5% nominal significance level are typically between 5% and 15%. We find the worst performance for one of the conditional tests (highest rejection rate of 0.53), but the novel standardization we propose largely eliminates its size distortion (highest rejection rate of 0.09). Effects of estimation error are largest when the ratio of out-of-sample to in-sample observations is large. The size distortions are somewhat smaller for 10-step-ahead forecasts. It does not vary much over the coverage levels of the VaR and ES forecasts, or the error distribution. These results hold equally for the joint tests of VaR and ES of Nolde and Ziegel (2017) as the separate tests of VaR and ES of Escanciano and Olmo (2010) and Du and Escanciano (2016).

The robust versions of the tests correct well for estimation error, and have empirical rejection rates that fluctuate around the nominal 5%. They work well for both coverage levels and both forecast horizons. We do not observe large differences between conditional and unconditional tests anymore. The size distortions for the VaR tests of Escanciano and Olmo (2010) and the ES tests of Du and Escanciano (2016) are a bit smaller than for the joint tests, but this may be related to the number of test conditions.

Our analysis of the power of the different tests, in which we correct the tests for size distortions, shows that together they offer reasonable power against various forms of misspecification. The power curves of the standard and robust versions are generally close to each other. The empirical rejection rates of both versions generally increase when the model used to obtain the risk forecasts deviates more from the true DGP. The joint tests of Nolde and Ziegel (2017) have some power against joint VaR and ES forecasts that result from misspecification of the error distribution and the mean, but perform less when the volatility is misspecified. Though the tests of Du and Escanciano (2016) have less power for these first two cases, their test that uses higher-order autocorrelation of scaled ES distances has some power when the volatility is misspecified.

In our empirical application we evaluate VaR and ES forecasts for the FTSE 100 index returns, generated by (GJR-)GARCH and HEAVY models estimated with rolling windows of 500 and 2,500 observations with backtests based on the period from December 8, 2009

to April 17, 2019. Estimation error generally has a non-negligible effect on the conclusions that would likely be drawn based on the test outcome, in the sense that the p-values often increase from values below 5% or 10% for the standard tests to values well above these levels for their robust counterparts. For the standard GARCH model, such a decisive increase in p-values is mostly present for unconditional tests, but less so for the GJR-GARCH model, where also robust tests reject correct specification. For the HEAVY model estimated with 2,500 observations and backtested with 1-day-ahead forecasts, the increase is such that in their standard version six out of nine tests have p-values below 5%, while in the robust version only one has. An analysis of the evolution of p-values over time shows that for all combinations of tests, models, and forecasts we find periods where the standard and robust versions lead to different decisions, though these periods are not the same, and are hard to trace back to particular aspects of the estimation or forecasting period.

Based on our analyses, we conclude that estimation error substantially affects backtests of VaR and ES both in simulated and real-life settings. This conclusion concurs with Escanciano and Olmo (2010) for VaR and Du and Escanciano (2016) for ES, both based on single-period forecasts. We extend their results by proposing robust versions for jointly testing VaR and ES in the framework of McCracken (2000), which enables us to also handle multi-period forecasts and different estimation schemes. As a second extension, we propose a new conditional test that is easier to estimate. We extend Nolde and Ziegel (2017) by showing how estimation error can formally be included in their testing framework, and by our results for multi-step-ahead forecasting.

Our results also carry practical relevance, because ES is gaining popularity at the expense of VaR as the risk measure that the financial industry uses to assess market positions (BCBS, 2016). ES has better theoretical properties than VaR, because it is coherent (see Artzner et al., 1999, 1997; Acerbi and Tasche, 2002, for elaborations). While tests for the correctness of ES forecasts have been proposed before,[2] the discussion whether ES could theoretically be backtested (see, e.g., Gneiting, 2011; Acerbi and Székely, 2014) has only recently been solved by the work of Fissler and Ziegel (2016) and Nolde and Ziegel (2017). It means that financial institutions can now design sound backtests to evaluate their risk measurement procedure with ES. Our results show that estimation error in this procedure should be taken into account when constructing backtests.

We discuss the methodology in Section 1, and the test specifications in Section 2. We set up and study the results of our Monte Carlo experiments in Section 3, and an empirical application in Section 4. We conclude in Section 5. The Supplemental Materials to this article with additional theoretical and simulation results can be found in the online version at the journal's website.

1 Theory

1.1 Environment

Let $Y_{t+\tau}$ denote the return generated by holding an asset from period t to $t+\tau$, with horizon $\tau \geq 1$. Let $W_t = \left(Y_t, Z_t', Y_{t-1}, Z_{t-1}', \ldots\right)'$ denote the information set at time t, with Z_t

2 See McNeil and Frey (2000), Berkowitz (2001), Kerkhof and Melenberg (2004), and Wong (2008, 2010).

denoting a vector of other relevant variables. Let $\mathcal{F}_t = \sigma(W_t)$ denote the σ-algebra generated by W_t and define $E_t[\cdot] = E[\cdot|W_t]$ as the expectation conditional on W_t.

The τ-step-ahead VaR, denoted $\text{VaR}_{t,\tau}$ and ES, denoted $\text{ES}_{t,\tau}$, at coverage level $1-\alpha \in (0,1)$, are (implicitly) defined as

$$P\left(Y_{t+\tau} \leq \text{VaR}_{t,\tau}|W_t\right) = \alpha, \tag{1}$$

and

$$\text{ES}_{t,\tau} = \frac{1}{\alpha}E_t\left[Y_{t+\tau}\mathbb{1}\left\{Y_{t+\tau} < \text{VaR}_{t,\tau}\right\}\right] = E_t\left[Y_{t+\tau}|Y_{t+\tau} < \text{VaR}_{t,\tau}\right], \tag{2}$$

almost surely (a.s.), for all t, and where $\mathbb{1}\{\cdot\}$ denotes the indicator function that takes the value 1 if the event within curly brackets is true and zero otherwise.

We consider a model of VaR and ES as given by the parametric family of functions $\mathcal{M} = \{m(\cdot,\theta) : \theta \in \Theta \subset \mathbb{R}^p\}$, with parameter space Θ and $p < \infty$, and where $m(\cdot,\theta) = (m_1(\cdot,\theta), m_2(\cdot,\theta))'$ is some 2×1 vector. We let $m_1(\cdot,\theta)$ and $m_2(\cdot,\theta)$ be the VaR and ES forecasts generated by the model at parameter vector θ, respectively, and write $m_t(\theta) = \left(m_{t,1}(\theta), m_{t,2}(\theta)\right)' = (m_1(W_t,\theta), m_2(W_t,\theta))'$. For notational convenience we ignore dependence on α generally, and ignore the dependence on τ in the model notation specifically.

1.2 Correct Specification Hypothesis

We hypothesize the existence of some true parameter vector $\theta_0 \in \Theta$, with Θ some compact parameter space, at which the model is correctly specified, that is,

$$m_t(\theta_0) = (\text{VaR}_{t,\tau}, \text{ES}_{t,\tau})\text{`a.s., for all } t. \tag{3}$$

If the hypothesis is valid and we can estimate θ_0 consistently, we may expect our model to perform well when implemented.

To formally test this hypothesis we consider an *identification function*, $g_{t,\tau}(\theta)$, and a *test matrix*, $H_t(\theta)$, that contains conditioning information, and design moment conditions based on their product $k_{t,\tau}(\theta) = H_t(\theta)g_t(\theta)$. This gives us the actionable null hypothesis

$$\mathcal{H}_{0,b} : E\left[k_{t,\tau}(\theta_0)\right] = 0 \text{ for all } t, \tag{4}$$

as it allows for the application of the out-of-sample testing framework of McCracken (2000).

The 2×1 identification function, $g_{t,\tau}(\theta)$, introduced in Nolde and Ziegel (2017), provides a transformation of the asset return $Y_{t+\tau}$ and the VaR and ES forecasts, $m_{t,1}(\theta)$ and $m_{t,2}(\theta)$:

$$\begin{aligned} g_{t,\tau}(\theta) &= \begin{bmatrix} g_{t,1,\tau}(\theta) \\ g_{t,2,\tau}(\theta) \end{bmatrix} \\ &= \begin{bmatrix} \mathbb{1}\{Y_{t+\tau} - m_{t,1}(\theta) < 0\} - \alpha \\ m_{t,2}(\theta) - m_{t,1}(\theta) - \frac{1}{\alpha}\mathbb{1}\left\{aY_{t+\tau} - m_{t,1}(\theta) < 0\right\}\left(Y_{t+\tau} - m_{t,1}(\theta)\right) \end{bmatrix}, \end{aligned} \tag{5}$$

where the first element gives a centered VaR violation, and the second element provides a

measure of the distance between the ES forecast and the actual return. Importantly, the identification function has a conditional mean uniquely equal to zero at θ_0, that is, $E_t\left[g_{t,\tau}(\theta_0)\right] = 0$, given unique quantiles. Moreover, the first element forms the basis for the coverage tests of Kupiec (1995) and Christoffersen (1998). Nolde and Ziegel (2017) also show that $E_t\left[g_{t,\tau}(\theta_0)\right]$ is proportional, up to some $\mathcal{F}_t$-measurable proportionality constant, to the gradient of the conditional mean of any member of the family of joint consistent scoring function for VaR and ES introduced in Fissler and Ziegel (2016). Barendse (2020) proposes a joint semi-parametric estimator of ES using Equation (5).

The specification of the test matrix $H_t(\theta_0)$ should consist of (transformations of) variables in the information set at time t, but is otherwise discretionary. We obtain an unconditional test by setting $H_t(\theta_0) = I_2$, the 2×2 identity matrix. For conditional tests, $H_t(\theta_0)$ should contain variables that are correlated with the risk in the economy or particular to the asset under consideration. An intuitive specification includes lags of $g_{t,\tau}(\theta_0)$, to mimic the conditional VaR testing, which frequently relies on lags of $g_{t,1,\tau}(\theta_0)$ (e.g., Christoffersen, 1998; Escanciano and Olmo, 2010). On the other hand, Nolde and Ziegel (2017) argue that conditioning on lags of $g_{t,\tau}(\theta_0)$ leads to bad size properties and argue for a $H_t(\theta_0)$ specification that is smoother and divides by the conditional volatility to reduce sensitivity to outliers. In the following we show that the poor size properties of the conditional tests arise due to estimation of fourth moments in the asymptotic covariance matrix of the test. Once we use model implications to circumvent the estimation of fourth-order sample moments, the conditional test performs well, see Section 2.2. Section 2 contains several examples of $H_t(\theta)$ informed by prior research.

Finally, it follows from two steps that a test of Equation (4) implies a test of correct model specification (Equation 3). First, it is easy to see that correct model specification (Equation 3) and unique quantiles together imply $E_t\left[g_{t,\tau}(\theta)\right] = 0$ uniquely at $\theta = \theta_0$. A test of correct model specification can therefore be based on the following null hypothesis:

$$\mathcal{H}_0 : E_t\left[g_{t,\tau}(\theta_0)\right] = 0, \text{ a.s. for all } t. \tag{6}$$

Second, we use the equivalence statement (see Giacomini and White, 2006):

$$E_t\left[g_{t,\tau}(\theta_0)\right] = 0, \text{a.s.} \iff E\left[g_{t,\tau}(\theta_0)\tilde{h}_t\right] = 0,$$

for all $\mathcal{F}_t$-measurable functions $\tilde{h}_t$ and for all t. Setting $\tilde{h}_t = H_t$, it follows that evidence against Equation (4) provides evidence against Equation (6) and thus Equation (3).

1.3 Out-of-Sample Testing

To formally test the null hypothesis in Equation (6) we consider backtests based on out-of-sample VaR and ES forecasts. Our test framework is based on the theory in McCracken (2000) for general out-of-sample tests derived from moment conditions on non-differentiable functions, such as our identification function $g_{t,\tau}(\theta)$. This framework allows us to estimate and correct for the effect of estimation error on our tests. This is important given that (i) the general theory establishes that this effect can be pronounced once the number of in-sample and out-of-sample observations are proportional, which is often the case in risk-management practice and (ii) previous research has shown that risk measure estimation is imprecise and affects, for instance, unconditional VaR tests (see, e.g., Escanciano and Olmo, 2010).

To implement out-of-sample testing we first define subsamples and estimation strategies. Consider the sample $\{Y_t, Z_t'\}_{t=1}^{T}$, with total sample size $T \geq 1$. We let the first R observations denote the first in-sample period, and the subsequent $P = T - R - \tau + 1$ observations denote the out-of-sample period minus the forecast horizon. We will consider fixed, rolling, and recursive forecasting schemes, as in West and McCracken (1998), McCracken (2000), and Escanciano and Olmo (2010). The schemes differ in the observations that are used to estimate the unknown parameters. Consider an estimator $\widehat{\theta}_t$ of θ_0 at time t. The fixed scheme uses the first R observations to compute $\widehat{\theta}_t$, for all t. The rolling scheme uses the R most recently observed observations, that is, the observations at times $t - R + 1, \ldots, t$. The recursive scheme uses all observations up to time t, that is, the observations at times $1, \ldots, t$.

Returning to the testing problem, since the true parameter θ_0 is generally unknown, we must base our tests on the normalized sample moment of our test condition

$$S_P = S(R, P) = \frac{1}{\sqrt{P}} \sum_{t=R}^{T-\tau} k_{t,\tau}(\widehat{\theta}_t), \tag{7}$$

with $k_{t,\tau}(\theta)$ evaluated at the estimated $\widehat{\theta}_t$ instead of the true θ_0. It is well known that the presence of the estimators $\widehat{\theta}_t$ in S_P implies that standard Wald test results do not apply (see, e.g., West, 2006). In the following section we derive the non-standard asymptotic distribution, which relies on an alternative asymptotic covariance matrix that includes additional terms related to estimation error.

1.4 Asymptotic Theory

We obtain the asymptotic distribution of S_P under the following assumptions, showing that it converges to a normal vector zero mean and asymptotic covariance matrix $\mathbf{\Omega}$.

To obtain the non-standard asymptotic distribution of S_P we utilize an asymptotic expansion that many estimators in literature admit, including maximum-likelihood estimators as well as a range of GMM estimators [see, e.g., West (1996) and McCracken (2000) for elaborations].

Assumption 1. *The estimate $\widehat{\theta}_t$ satisfies the expansion $\widehat{\theta}_t - \theta_0 = B(t)M(t) + o_P(t^{-1/2})$, with $B(t)$ a $p \times q$ matrix of rank p, and $M(t)$ a $q \times 1$ vector, with (a) $B(t) \xrightarrow{a.s} B$, B a matrix of rank p, (b) $M(t) = R^{-1} \sum_{s=1}^{R} l_s(\theta_0)$, $M(t) = R^{-1} \sum_{s=t-R+1}^{t} l_s(\theta_0)$, and $M(t) = t^{-1} \sum_{s=1}^{t} l_s(\theta_0)$, for the fixed, rolling, and recursive forecasting schemes, respectively, and (c) $E_{t-\tau}[l_t(\theta_0)] = 0$ (a.s.) and $l_t(\theta_0)$ $\mathcal{F}_t$−measurable, for all $t = R, \ldots, T$.*

Assumption 2. *$R, P \to \infty$ as $T \to \infty$, and $\lim_{T\to\infty} \frac{P}{R} \to \pi$, $0 \leq \pi < \infty$.*

Assumption 2 is equivalent to Assumption 2 in McCracken (2000) and allows the out-of-sample period to grow proportionally to the in-sample period. Notice that both the in-sample and out-of-sample sizes must diverge simultaneously.

Define

$$v_t(\theta) = \left[k_{t,\tau}(\theta)', l_t(\theta)'\right]'. \tag{8}$$

Assumption 3. *For some $r > 1$, (a) $\left(Y_t, Z_t', m_t(\theta)'\right)'$ is strong mixing with coefficients of size $-2r/(r-1)$ uniformly over Θ; (b) $v_t(\theta_0)$ is covariance stationary, and (c) Ω is positive definite.*

Assumption 3 is equivalent to Assumption 3 in McCracken (2000). Patton, Ziegel, and Chen (2019) derive the parameter subset Θ for which condition (a) holds for VaR and ES models that are based on a GARCH(1,1) model. For VaR and ES models that are based on conditional location-scale models the conditional mean specification should be uniformly mixing too, and Pham and Tran (1985) establish such results for ARMA models. Covariance stationarity of $v_t(\theta_0)$ is primarily assumed for simplification of the algebra in the proofs establishing the consistency of an estimator of the asymptotic covariance matrix Ω. It is weaker than strict stationary of (Y_t, Z_t'), but due to the nonlinear nature of $v_t(\theta_0)$ it is close to it.

We also impose several smoothness and moment conditions. Let $F_{t,\tau}(y) = P(Y_{t+\tau} < y|W_t)$ denote the conditional distribution of $Y_{t+\tau}$ for some $\tau \geq 1$, and let $f_{t,\tau}(y)$ denote the associated density. Moreover, for any matrix B, let $|B|$ denote the max norm of B, let $\|\cdot\|_Q$ denote the L^Q norm for $Q \in [0,\infty)$ and the essential supremum if $Q = \infty$, and let $\sup_t$ denote $\sup_{R \leq t \leq T}$.

Assumption 4. *For each t, (a) let there exist finite constants C, $\phi > 0$, and $Q \geq 2r$ such that for all $\Theta(\varepsilon) \subset \Theta_0$, with Θ_0 some open neighborhood of θ_0 and $\Theta(\varepsilon) = \Theta(\theta_0, \varepsilon) = \{\theta \in \mathbb{R}^p : |\theta - \theta_0| < \varepsilon\}$, $H_t(\theta)$ satisfies the weak Lipschitz condition $\sup_t \| \sup_{\theta \in \Theta(\varepsilon)} H_{t,i,j}(\theta) - H_{t,i,j}(\theta_0)\|_Q \leq C\varepsilon^\phi$, for all $i = 1, \ldots, l$, and $j = 1, 2$; (b) let $m_t(\theta)$ be (a.s.) continuously differentiable on Θ_0 with Jacobian matrix $J_t(\theta) = \nabla m_t(\theta)$, respectively; (c) let $E[l_t(\theta)]$ be continuously differentiable on Θ_0 with Jacobian matrix $\nabla E[l_t(\theta)]$, and let $l_t(\theta)$ satisfies the weak Lipschitz condition $\sup_t \| \sup_{\theta \in \Theta(\varepsilon)} l_{t,j}(\theta) - l_{t,j}(\theta_0)\|_Q \leq C\varepsilon^\phi$, for $j = 1, \ldots, m$; and (d) let $G = G_t = \nabla E[l_t(\theta)]|_{\theta=\theta_0}$ and $A = A_t$, with*

$$A_t = E\left(H_t(\theta_0)\begin{bmatrix} f_{t,\tau}(m_{t,1}(\theta_0)) & 0 \\ 0 & 1 \end{bmatrix} J_t(\theta_0)\right).$$

Assumption 4 imposes (a.s.) differentiability on $m_t(\theta)$ and $E[l_t(\theta)]$ on some neighborhood Θ of θ_0. Moreover, a weak Lipschitz condition is imposed on $H_t(\theta)$ and $l_t(\theta)$, as required in the theory of McCracken (2000).

Assumption 5. *For each t, let the conditional cdf $F_{t,\tau}(\cdot)$ have (a.s.) continuous conditional density $f_{t,\tau}(\cdot)$. Moreover, for the neighborhood $N_t(\Theta_0) = \{y : y \in \{m_{t,1}(\theta)\}_{\theta \in \Theta_0}\}$ let $f_{t,\tau}(\cdot)$ be bounded on $N_t(\Theta_0)$, that is, $\sup_{y \in N_t(\Theta_0)} f_{t,\tau}(y) < C_f < \infty$ (a.s.), and $|f_{t,\tau}(y) - f_{t,\tau}(y')| \leq L|y - y'|$ (a.s.) for all $y, y' \in N_t(\Theta_0)$ and each t, and some constant $L < \infty$.*

Assumption 5 imposes a boundedness condition and Lipschitz continuity on the conditional distribution of $Y_{t+\tau}$ in some neighborhood of the α-quantile of $F_{t,\tau}(\cdot)$, given by $m_{t,1}(\theta_0)$, that are similar to those imposed in Escanciano and Olmo (2010).

Assumption 6. *For constants $c_H \in [1,\infty]$ and $c_g \in [1,\infty]$, such that $1/c_H + 1/c_g = 1$, we impose the following moment conditions, for some arbitrary constant $C < \infty$: (a) $\sup_t \|Y_t\|_{c_g Q} < C$; (b) $\sup_t \| \sup_{\theta \in \Theta_0} |m_t(\theta)|\|_{c_g Q} < C$; (c) $\sup_t \| \sup_{\theta \in \Theta_0} |J_t(\theta)|\|_{c_g Q} < C$;*

(d) $\sup_t \| \sup_{\theta \in \Theta_0} |H_t(\theta)| \|_{c_H Q} < C$; *(e)* $\sup_t \| \sup_{\theta \in \Theta_0} |l_t(\theta)| \|_Q < C$; *and (f)* $\sup_t \sup_{\theta \in \Theta_0} |\nabla E[l_t(\theta)]| < C$.

Assumption 6 imposes moment conditions on the relevant quantities. Notice that if $|H_t(\theta)|$ is uniformly bounded on Θ_0 (e.g., when $H_t = I_2$), we can set $c_H = \infty$ and $c_g = 1$, such that we effectively impose $2r$-moment conditions on Y_t, $m_t(\theta)$, $J_t(\theta)$, and $l_t(\theta)$ when we set $Q = 2r$, with $r > 1$.

Under these conditions we establish that S_P converges to a multivariate normal limit with zero mean and the non-standard $l \times l$ asymptotic covariance matrix

$$\Omega = \Sigma + \lambda_{hl}(AB\rho + \rho' B' A') + \lambda_{ll} ABVB'A', \tag{9}$$

which includes additional terms found for out-of-sample tests (see, e.g., West, 2006), consisting of a $l \times l$ matrix

$$\Sigma = \Sigma(0) + \sum_{j=1}^{\tau-1} (\Sigma(j) + \Sigma(j)'),$$
$$\Sigma(j) = E\left[k_{t-j,\tau}(\theta_0) k_{t,\tau}(\theta_0)'\right], j = 0, 1, \ldots,$$

a $q \times l$ matrix

$$\rho = \sum_{j=-(\tau-1)}^{\tau-1} E\left[l_{t+\tau-j}(\theta_0) k_{t,\tau}(\theta_0)'\right],$$

that describes the correlation between the estimator and the test conditions, a $q \times q$ matrix

$$V = E[l_t(\theta_0) l_t(\theta_0)'] + \sum_{j=1}^{\tau-1} \left(E\left[l_t(\theta_0) l_{t-j}(\theta_0)'\right] + E\left[l_{t-j}(\theta_0) l_t(\theta_0)'\right] \right).$$

Finally, the matrix A is defined in Assumption 4, whereas the matrix B is used in the expansion of the estimator θ_t discussed above and defined in Assumption 1.

Finally, $\lambda_{hl}(\pi)$ and $\lambda_{ll}(\pi)$ are defined as

	$\lambda_{hl}(\pi)$	$\lambda_{ll}(\pi)$
fixed :	0	π
rolling($\pi \leq 1$) :	$\pi/2$	$\pi - \pi^2/3$
rolling($\pi > 1$) :	$1 - (2\pi)^{-1}$	$1 - (3\pi)^{-1}$
recursive :	$1 - \pi^{-1}\log(1+\pi)$	$2[1 - \pi^{-1}\log(1+\pi)]$.

(10)

The following theorem formalizes the convergence in distribution of S_P to a normal limit.

Theorem 1. *Let Assumptions 1–6 be satisfied. It follows under* $\mathcal{H}_0$ *that* $S_P \xrightarrow{d} N(0, \Omega)$ *with* Ω *defined in* Equation (9). *Moreover, if* $\pi = 0$, *then* $S_P \xrightarrow{d} N(0, \Sigma)$.

Proof. See Appendix A. □

1.5 Asymptotic Covariance Matrix Estimation

Theorem 1 shows that using estimated parameters introduces additional terms in the asymptotic covariance matrix of S_P in case the in-sample size P grows proportionally to the out-of-sample size R.

We propose the following consistent estimator of the asymptotic covariance matrix Ω:

$$\widehat{\Omega}_P = \widehat{\Sigma}_P + \widehat{\lambda}_{bl}(\widehat{\pi}_P)\left(\widehat{A}_P\widehat{B}_P\widehat{\rho}_P + \widehat{\rho}_P'\widehat{B}_P'\widehat{A}_P'\right) + \widehat{\lambda}_{ll}(\widehat{\pi}_P)\widehat{A}_P\widehat{B}_P\widehat{V}_P\widehat{B}_P'\widehat{A}_P'. \tag{11}$$

with the consistent estimators of Σ, A, ρ, V, and π:

$$\begin{aligned}\widehat{\Sigma}_P &= \widehat{\Sigma}_P(0) + \sum_{j=1}^{\tau-1} w_{P,j}\left(\widehat{\Sigma}_P(j) + \widehat{\Sigma}_P(j)'\right),\\ \widehat{\Sigma}_P(j) &= \frac{1}{P}\sum_{t=R+j}^{T-\tau} k_{t-j,\tau}\left(\widehat{\theta}_{t-j}\right)k_{t,\tau}(\widehat{\theta}_t)', j = 0, 1, \ldots,\end{aligned} \tag{12}$$

$$\widehat{A}_P = \frac{1}{P}\sum_{t=R}^{T-\tau}\left(H_t(\widehat{\theta}_t)\begin{bmatrix}(2\widehat{c}_P)^{-1}\mathbb{1}\left(|Y_{t+\tau} - m_{t,1}(\widehat{\theta}_t)| < \widehat{c}_P\right) & 0\\ 0 & 1\end{bmatrix}J_t(\widehat{\theta}_t)\right), \tag{13}$$

$$\begin{aligned}\widehat{\rho}_P = &\frac{1}{P}\sum_{t=R}^{T-\tau}\left[l_{t+\tau}(\widehat{\theta}_t)k_{t,\tau}(\widehat{\theta}_t)'\right]\\ &+\frac{1}{P}\sum_{j=1}^{\tau-1}\left(\sum_{t=R+j}^{T-\tau} l_{t+\tau-j}\left(\widehat{\theta}_{t-j}\right)k_{t,\tau}(\widehat{\theta}_t)' + \sum_{t=R}^{T-\tau-j} l_{t+\tau}(\widehat{\theta}_t)k_{t+j,\tau}\left(\widehat{\theta}_{t+j}\right)'\right),\end{aligned} \tag{14}$$

$$\widehat{V}_P = \frac{1}{P}\sum_{t=R}^{T-\tau} l_t(\widehat{\theta}_t)l_t(\widehat{\theta}_t) + \frac{1}{P}\sum_{j=1}^{\tau-1}\nu_{P,j}\sum_{t=R+j}^{T-\tau}\left(l_{t-j}\left(\widehat{\theta}_{t-j}\right)l_t(\widehat{\theta}_t)' + l_t(\widehat{\theta}_t)l_{t-j}\left(\widehat{\theta}_{t-j}\right)'\right), \tag{15}$$

$$\widehat{\pi}_P = P/R, \tag{16}$$

where we include weight functions $w_{P,j}$ and $\nu_{P,j}$ to ensure positive definiteness of the covariance estimators $\widehat{V}_P$ and $\widehat{\Sigma}_P$, with $w_{P,j}\to^{\mathrm{p}}1$ and $w_{P,j}\nu_{P,j}\to^{\mathrm{p}}1$ for all $j = 1, \ldots, \tau - 1$. White (2001, p. 153-154) provides a convenient practical choice for $w_{P,j}$ (and $\nu_{P,j}$ similarly). Set $w_{P,j} = 1$, for all $j = 1, \ldots, \tau$, if $\widehat{V}_P$ (or $\widehat{\Sigma}_P$) has all positive eigenvalues at $w_{P,j} = 1$ (which ensures $w_{P,j}\xrightarrow{p}1$). Otherwise, set $w_{P,j} = \max(0, 1 - \kappa_P\tau)$, with κ_P the smallest positive number that ensures $\widehat{V}_P$ (or $\widehat{\Sigma}_P$) has all positive eigenvalues. In practice, a grid of potential values of κ_P can be used. Alternatively, one can use a HAC estimator of V_0 and Σ_0, such as the estimator in Newey and West (1987) under conditions similar to Assumption 7 in McCracken (2000). HAC estimation does not use the zero autocorrelation condition for lags beyond $\tau - 1$ as implied under the null hypothesis, which can result in increased variance in the estimator. It will also require stronger moment and mixing conditions.

These estimators, except for $\widehat{A}_P$, are similar to those used in Escanciano and Olmo (2010). $\widehat{A}_P$ is based on Powell (1986), and similar estimators are used in Engle and Manganelli (2004) and Patton, Ziegel, and Chen (2019). In the definition of $\widehat{A}_P$, the sequence $\widehat{c}_P$ is a potentially stochastic sequence that converges in probability to zero at a slower rate than $P^{-1/2}$. A typical choice is $\widehat{c}_P = P^{-1/3}$. Instead of $\widehat{V}_P$ above, we can also opt for an estimator of V provided by a statistical computing package for the end-of-sample estimation of $\widehat{\theta}_{T-\tau}$, as long as we know this estimator is consistent. A strongly consistent estimator $\widehat{B}_P$ of B can usually be obtained similarly.

The following result states that $\widehat{\Omega}_P$ is consistent, and that $\widehat{\Omega}_P^{-1/2} S_P$ converges to a standard normal random vector. Hence, we can derive tests that have standard critical values.

Corollary 1. *Let Assumptions 1–6 be satisfied, let* $\widehat{B}_P \xrightarrow{a.s.} B$, *and let* $\widehat{c}_P / c_P \xrightarrow{p} 1$, *where the non-stochastic* c_P *satisfies* $c_P = o(1)$ *and* $c_P^{-1} = o(P^{1/2})$. *Under* $\mathcal{H}_0$ *it follows that* $\widehat{\Omega}_P \xrightarrow{p} \Omega$ *and* $\widehat{\Omega}_P^{-1/2} S_P \xrightarrow{d} N(0, I)$.

Proof. See Appendix A. □

2 Tests

2.1 Test Specifications

The above results imply we can construct tests using the Wald test statistic $T_P = S_P' \widehat{\Omega}_P^{-1} S_P$, where we reject the null hypothesis H_0 at a $100 \cdot q\%$ significance level if T_P exceeds $\chi^2_{l,1-q}$, with $\chi^2_{l,1-q}$ denoting the $100 \cdot (1-q)\%$ quantile of the χ^2-distribution with l degrees of freedom. We refer to *robust* versions of the tests when we use $\widehat{\pi}_P$ in $\widehat{\Omega}_P$, such that we account for the effect of estimation error. We refer to *standard* versions of the tests when we impose $\pi = 0$, that is, $\widehat{\Omega}_P = \widehat{\Sigma}_P$, such that we ignore estimation effects. We study nine different tests in total. The first two are the classical tests for VaR that are also considered in Escanciano and Olmo (2010), and are special cases of our framework. Next, we consider four joint tests for VaR and ES that differ in the specification of $H_t(\theta_0)$. The final three tests are the ES statistics introduced in Du and Escanciano (2016), and considered benchmarks as they do not fit our framework.

We follow the standard nomenclature in the literature, see, for example, Christoffersen (1998), regarding unconditional and conditional tests. Unconditional tests are tasked with finding large forecasts errors on average, and therefore assess whether $E\left[g_{t,\tau}(\theta_0)\right]$ is far away from zero in our terminology. Conditional tests search for clusters of large forecast errors and are therefore concerned with interactions that involve lags, that is, whether $E\left[g_{t-\tau,i,\tau}(\theta_0) g_{t,j,\tau}(\theta_0)\right]$ is far away from zero, for all $i, j = 1, 2$.

Escanciano and Olmo (2010) consider unconditional and conditional tests for VaR, and introduce both standard and robust test statistics. In our framework, the statistic of the unconditional test $EO_P^{(1)}$ follows from using

$$EO_P^{(1)} : H_t(\theta) = \begin{bmatrix} 1 & 0 \end{bmatrix}.$$

It corresponds with the test for correct unconditional coverage proposed by Kupiec (1995) and Christoffersen (1998). The conditional VaR test is based on test statistic $EO_P^{(2)}$ which follows in our framework from

$$EO_P^{(2)} : H_t(\theta) = \begin{bmatrix} g_{t-\tau,1,\tau}(\theta) & 0 \end{bmatrix}.$$

It corresponds with the independence test proposed by Christoffersen (1998). Berkowitz, Christoffersen, and Pelletier (2011) base an alternative conditional test on the same sequence $k_{t,\tau}(\theta) = H_t(\theta)g_{t,\tau}(\theta)$. Finally, we follow convention and replace $\widehat{\Sigma}_P(0)$, the first term in the definition of $\widehat{\Sigma}_P$, with an alternative estimator that is common in the literature, $(\frac{1}{P}\sum_{t=R}^{T-\tau} g_{t,1,\tau}(\widehat{\theta}_t)^2)^2$, since $E[g_{t,\tau}(\theta_0)g_{t-\tau,1,\tau}(\theta_0)] = E[g_{t,1,\tau}(\theta_0)^2]^2$, see, for example, Escanciano and Olmo (2010).

The first joint (VaR, ES) test statistics $T_P^{(1)}$ is a straightforward generalization of the unconditional VaR test $EO_P^{(1)}$,

$$T_P^{(1)} : H_t(\theta) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

In the second test statistic $T_P^{(2)}$, we standardize the element related to $g_{t,2,\tau}(\theta)$ by the difference between the ES and the VaR forecast, $\big(m_{t,2}(\theta) - m_{t,1}(\theta)\big)$,

$$T_P^{(2)} : H_t(\theta) = \begin{bmatrix} 1 & 0 \\ 0 & 1/\big(m_{t,2}(\theta) - m_{t,1}(\theta)\big) \end{bmatrix}.$$

The rationale for this standardization is discussed in detail below.

The next joint tests are conditional ones, in that the conditioning information is given by lags of $g_{t,\tau}(\theta)$, with $T_P^{(3)}$ the generalization of $EO_P^{(2)}$, and $T_P^{(4)}$ the standardized version

$$T_P^{(3)} : H_t(\theta) = \begin{bmatrix} g_{t-\tau,1,\tau}(\theta) & 0 \\ 0 & g_{t-\tau,2,\tau}(\theta) \end{bmatrix}$$

$$T_P^{(4)} : H_t(\theta) = \begin{bmatrix} g_{t-\tau,1,\tau}(\theta) & 0 \\ 0 & \dfrac{g_{t-\tau,2,\tau}(\theta)}{\big(m_{t,2}(\theta) - m_{t,1}(\theta)\big)\big(m_{t-\tau,2}(\theta) - m_{t-\tau,1}(\theta)\big)} \end{bmatrix}.$$

The standardization in $T_P^{(4)}$ allows an alternative strategy to estimate the asymptotic covariance matrix, as shown in Theorem 2 in Section 2.2. This strategy circumvents the estimation of the second moment of the function $g_{t,2,\tau}(\theta_0)g_{t-s,2,\tau}(\theta_0)$, $s \geq \tau$, which is imprecise and leads to bad size properties (see Nolde and Ziegel, 2017). Instead, we can take the square of the second moment of $g_{t,2,\tau}(\theta_0)$, which is well estimable.

Finally, we consider the conditional test statistic $T_P^{(5)}$ based on

$$T_P^{(5)} : H_t(\theta) = \frac{1}{\sigma_{t+\tau}(\theta)} \begin{bmatrix} \dfrac{m_{t,2}(\theta) - m_{t,1}(\theta)}{\alpha} & 1 \end{bmatrix},$$

where $\sigma_{t+\tau}(\theta)$ denotes the conditional volatility of $Y_{t+\tau}$, which should be consistently estimated in practice. This particular choice of $H_t(\theta_0)$ follows Nolde and Ziegel (2017), who show that this choice results in an approximation of the test of McNeil and Frey (2000), and who prefer it in their application.

For comparison, we consider the unconditional and conditional ES backtests of Du and Escanciano (2016). Note that the theoretical properties of these tests have only been derived for the fixed window estimation scheme and $\tau = 1$. These tests fall outside our framework, because their testing condition cannot be written as in Equation (4). Let $\widehat{D}_t = \frac{1}{\alpha}\mathbb{1}(\alpha \leq \widehat{u}_t)(\alpha - \widehat{u}_t)$, where $\widehat{u}_t = G_t\Big(Y_{t+\tau}, \widehat{\theta}_t\Big)$, with $G_t\Big(\cdot, \widehat{\theta}_t\Big)$ an estimator of the

conditional cdf of $Y_{t+\tau}$. The unconditional ES test of Du and Escanciano (2016) is based on test statistic

$$DE_P^{(1)} = \left(\frac{1}{\sqrt{P}\sqrt{\alpha(1/3 - \alpha/4) + d_U}} \sum_{t=R}^{T-\tau} \left(\widehat{D}_t - \alpha/2 \right) \right)^2.$$

The conditional ES test uses test statistic

$$DE_P^{(2)} = (P-1)(1+d_C)^{-1} \tilde{\xi}_{P,1}^2,$$

which includes a measure for autocorrelation in the $\widehat{D}_t$ sequence $\tilde{\xi}_{P,j} = \tilde{\gamma}_{P,j}/\tilde{\gamma}_{P,0}$ with $\tilde{\gamma}_{P,j} = (P-j)^{-1} \sum_{t=R+j}^{T-\tau} \left(\widehat{D}_t - \alpha/2 \right) \left(\widehat{D}_{t-j} - \alpha/2 \right)$. The terms d_U and d_C are additional terms in the asymptotic covariance matrix due to estimation error further elaborated on in Du and Escanciano (2016). Both test statistics converge in distribution to a χ_1^2-distributed random variable. Du and Escanciano (2016) also consider conditional tests that include autocorrelations $\tilde{\xi}_{P,j}$ with more distant lags, for example, $j = 5$.

We therefore also consider

$$DE_P^{(3)} = (P-5) \xi_{P,1:5}' \left(I_j + D_C \right)^{-1} \xi_{P,1:5},$$

with $\xi_{P,1:j} = \left(\xi_{P,1}, \ldots, \xi_{P,j} \right)'$, I_j the $j \times j$ identity matrix, and where D_C is given in Du and Escanciano (2016).

2.2 Asymptotic Covariance Matrix Estimation for Conditional Standardized Tests

We provide a novel result for the estimation of the asymptotic covariance matrix of standardized conditional tests, like $T_P^{(4)}$. This result provides an alternative estimator of the asymptotic covariance matrix estimator, $\tilde{\Omega}_P$ (we emphasize the notational difference to the standard estimator $\widehat{\Omega}_P$), that exploits a property that is similar to the one that improves the estimation of the asymptotic covariance matrix of the conditional VaR test, $EO_P^{(2)}$, and the conditional ES test, $DE_P^{(2)}$.

Consider the following assumption.

Assumption 7. *Let the following conditions hold:*

i. $E_t \left[\{ g_{t,2,\tau}(\theta_0) / (\mathrm{ES}_{t,\tau} - \mathrm{VaR}_{t,\tau}) \}^2 \right] = \overline{c}^{(g,2)}$ *with* $\overline{c}^{(g,2)}$ *some constant, for all t;*
ii. $E_t \left[\left(g_{t,1,\tau}(\theta_0) g_{t,2,\tau}(\theta_0) \right) / (\mathrm{ES}_{t,\tau} - \mathrm{VaR}_{t,\tau}) \right] = \overline{c}^{(g,3)}$ *with* $\overline{c}^{(g,3)}$ *some constant, for all t.*

Assumption 7 imposes a condition on the (VaR, ES)-tests that is similar to the one exploited for conditional VaR tests. For these tests, we often use the fact that $E_t[g_{t,1,\tau}(\theta_0)^2] = E_t[(\mathbb{1}\{Y_{t+\tau} - m_{t,1}(\theta_0) < 0\} - \alpha)^2] = \alpha(1-\alpha)$, a constant, such that it can be shown that the covariance matrix term $E[g_{t,1,\tau}(\theta_0)^2 g_{t-s,1,\tau}(\theta_0)^2] = E[g_{t,1,\tau}(\theta_0)^2]^2$, $s \geq \tau$. As a result, we can approximate this term with the square of the second sample moment of $g_{t,1,\tau}$, instead of the sample moment of $g_{t,1,\tau}^2 g_{t-s,1,\tau}^2$ if we do not exploit the property.[3] The

3 In case of the conditional VaR test the value of $E[g_{t,1,\tau}(\theta_0)^2] = \alpha(1-\alpha)$ independent of the conditional distribution of $Y_{t+\tau}$, so we could of course use this value instead of the sample estimators. This does not hold for the joint (VaR, ES)-test.

latter estimation provides a much worse approximation in practice, most likely due to it involving a fourth-order multiplication of relatively nonsmooth terms.

We show in the Supplemental Materials that this assumption holds for VaR and ES forecasts derived from location-scale models of $Y_{t+\tau}$, that is,

$$\begin{aligned} Y_{t+1} &= \mu(W_t) + \upsilon_{t+1} \\ \upsilon_{t+1} &= \sigma(W_t)\varepsilon_{t+1}, \\ \varepsilon_{t+1} &\sim iidF, \end{aligned}$$

where $\mu(W_t)$ denotes the conditional mean and $\sigma(W_t)$ the conditional volatility of Y_{t+1} given the information set W_t. Since many risk models in the literature are derived from location-scale models, this assumption will usually be satisfied in practice.

Theorem 2. *Let the assumptions of Corollary 1 and Assumption 7 be satisfied.*

Consider the estimator

$$\tilde{\Sigma}_P^{(0)} := \begin{bmatrix} \left(\frac{1}{P}\sum_{t=R}^{T-\tau} g_{t,1,\tau}^2(\widehat{\theta}_t)\right)^2 & \left(\frac{1}{P}\sum_{t=R}^{T-\tau} \frac{g_{t,1,\tau}(\widehat{\theta}_t)g_{t,2,\tau}(\widehat{\theta}_t)}{m_{t,2}(\widehat{\theta}_t) - m_{t,2}(\widehat{\theta}_t)}\right)^2 \\ \left(\frac{1}{P}\sum_{t=R}^{T-\tau} \frac{g_{t,1,\tau}(\widehat{\theta}_t)g_{t,2,\tau}(\widehat{\theta}_t)}{m_{t,2}(\widehat{\theta}_t) - m_{t,2}(\widehat{\theta}_t)}\right)^2 & \left(\frac{1}{P}\sum_{t=R}^{T-\tau} \frac{g_{t,2,\tau}^2(\widehat{\theta}_t)}{\left(m_{t,2}(\widehat{\theta}_t) - m_{t,2}(\widehat{\theta}_t)\right)^2}\right)^2 \end{bmatrix}. \tag{17}$$

If

$$H_t = \begin{bmatrix} g_{t-s,1,\tau}(\theta) & 0 \\ 0 & \frac{g_{t-s,2,\tau}(\theta)}{(m_{t,2}(\theta) - m_{t,1}(\theta))(m_{2,t-s}(\theta) - m_{1,t-s}(\theta))} \end{bmatrix},$$

for $s \geq \tau$, then under $\mathcal{H}_0$ it follows that $\tilde{\Omega}_P \xrightarrow{p} \Omega$, and $\tilde{\Omega}_P^{-1/2} S_P \xrightarrow{d} N(0, I)$, for

$$\tilde{\Omega}_P = \tilde{\Sigma}_P + \widehat{\lambda}_{hl}(\widehat{\pi})\left(\widehat{A}_P\widehat{B}_P\widehat{\rho}_P + \widehat{\rho}_P'\widehat{B}_P'\widehat{A}_P'\right) + \widehat{\lambda}_{ll}(\widehat{\pi})\widehat{A}_P\widehat{B}_P\widehat{V}_P\widehat{B}_P'\widehat{A}_P', \tag{18}$$

with $\tilde{\Sigma}_P = \tilde{\Sigma}_P(0) + \sum_{j=1}^{\tau-1} w_{P,j}\widehat{\Sigma}_P(j)$.

Proof. See the Supplemental Materials. □

The novelty of the asymptotic covariance matrix introduced in Theorem 2, $\tilde{\Omega}_P$, is that it exchanges $\widehat{\Sigma}_P(0)$ for $\tilde{\Sigma}_P(0)$. It is therefore not harder to estimate than $\widehat{\Sigma}_P$.

3 Simulation Study

3.1 Design

We investigate the finite sample performance of the VaR and ES tests by means of Monte Carlo experiments with rolling window estimation. We report empirical rejection rates in 1000 Monte Carlo samples for the tests at the 5% significance level. Both the lengths of the in-sample window R and of the out-of-sample window P can either be 500 or 2,500. These are relevant sample sizes corresponding to, respectively, approximately 2 and 10 years of daily return data. The resulting four combinations account for scenarios that include short or long

out-of-sample periods, as well as scenarios that are substantially or only slightly impacted by estimation error. We consider $\tau = 1$ and 10-step-ahead VaR and ES forecasts at coverage levels $1 - \alpha = 97.5\%$ and 95%. The Basel committee requires evaluation of ES at the 97.5% coverage level. We include 95% to assess how size and power are affected by the coverage level.

Our simulation setup is similar to Du and Escanciano (2016). In all experiments, the GARCH(1,1) model as in Bollerslev (1986) forms the null model H_0 to construct the forecasts. Its specification is given by

$$\begin{aligned} Y_{t+1} &= \sigma_{t+1}\varepsilon_{t+1}, \\ \sigma^2_{t+1} &= \omega_0 + \alpha_0 Y_t^2 + \beta_0 \sigma_t^2, \end{aligned}$$

such that the risk measure forecasts are equal to

$$\begin{aligned} \mathrm{VaR}_{t,\tau}(\alpha) &= -\sqrt{E_t\left[\sigma^2_{t,\tau}\right]}F^{-1}(\alpha), \\ \mathrm{ES}_{t,\tau}(\alpha) &= -\sqrt{E_t\left[\sigma^2_{t,\tau}\right]}E\left[\varepsilon_t | \varepsilon_t \leq F^{-1}(\alpha)\right] \end{aligned} \tag{19}$$

with $E_t\left[\sigma^2_{t,\tau}\right] = (\alpha_0 + \beta_0)^{\tau-1}\sigma^2_{t+1} + \sum_{i=1}^{\tau-1}\omega_0(\alpha_0 + \beta_0)^{i-1}$ (see the Supplemental Materials), and where ε_t follows a standardized distribution, of which $F^{-1}(\alpha)$ denotes its α-quantile. In particular, we consider the standardized Student's t distribution with $\nu_0 = 5$ and 30 degrees of freedom, to account for both fat and approximately normal tails.[4] We set the GARCH(1,1) parameters $(\omega_0, \alpha_0, \beta_0) = (0.05, 0.1, 0.85)$.

We need several quantities to determine the effects of estimation error. Given the continuous differentiability of $m_t(\theta)$ we use numerical approximation to find $J_t(\widehat{\theta}_t)$ for each $t = R+1, \ldots, T-\tau+1$. Moreover, in our setting $l_t(\widehat{\theta}_t) = \partial \log\left[f_{\widehat{\nu}_t}\left(Y_t/\sigma_t(\widehat{\theta}_t)\right)/\sigma_t(\widehat{\theta}_t)\right]/\partial\theta$, with f_ν denoting the standardized Student's t pdf with ν degrees of freedom. $\widehat{B}_P$ denotes a strongly consistent estimator of the asymptotic covariance matrix of $\sqrt{t}(\widehat{\theta}_t - \theta_0)$, for any $t = R+1, \ldots, T$, which can, for instance, be obtained as the negative of the inverse of the sample information matrix evaluated at $\widehat{\theta}_T$, that is, $\widehat{B}_P = \left[-\frac{1}{T}\sum_{t=2}^{T}\partial^2 \log\left[f_{\widehat{\nu}_T}\left(Y_t/\sigma_t(\widehat{\theta}_T)\right)/\sigma_t(\widehat{\theta}_T)\right]/(\partial\theta\partial\theta\prime)\right]^{-1}$. We employ numerical approximation to obtain these quantities, and provide details in the Supplemental Materials.

In the power study we consider three DGPs that differ from the GARCH(1,1) model with regard to the specification of the error distribution, the volatility, or the mean process. Similar DGPs have been studied in Du and Escanciano (2016) and Escanciano and Olmo (2010).

- A_1. GARCH(1,1) with mixed-normal innovations:

4 When ε_t follows a standardized Student's t distribution with ν degrees of freedom, $E\left[\varepsilon_t | \varepsilon_t \leq F^{-1}(\alpha)\right] = \sqrt{\frac{\nu-2}{\nu}}\frac{\nu + \left(G_\nu^{-1}(\alpha)\right)^2}{\nu-1}\frac{g_\nu\left(G_\nu^{-1}(\alpha)\right)}{\alpha}$, with $G_\nu^{-1}(\alpha)$ the α-quantile of the standard t-distribution with ν degrees of freedom, and $g_\nu(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi\nu}}\left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$.

$$
\begin{aligned}
&Y_{t+1} = \sigma_{t+1}\varepsilon_{t+1}\\
&\sigma^2_{t+1} = \omega_0 + \alpha_0 Y_t^2 + \beta_0\sigma_t^2,\\
&\varepsilon_t \sim \left[\frac{1}{1+q}N(0,q) + \frac{q}{1+q}N(0,1/q)\right], \quad q = 1+\frac{3}{2}c;
\end{aligned}
$$

- A_2. GJR-GARCH(1,1):

$$
\begin{aligned}
&Y_{t+1} = \sigma_{t+1}\varepsilon_{t+1}\\
&\sigma^2_{t+1} = \omega_0 + \alpha_0(1-c)Y_t^2 + 2\alpha_0 c\mathbb{1}(\varepsilon_t < 0)Y_t^2 + \beta_0\sigma_t^2,\\
&\varepsilon_t \sim t;
\end{aligned}
$$

- A_3. GARCH(1,1)-in-mean:

$$
\begin{aligned}
&Y_{t+1} = 2.5c\left(\sigma^2_{t+1} - 1\right) + \nu_{t+1},\\
&\nu_t = \sigma_t\varepsilon_t,\\
&\sigma^2_{t+1} = \omega_0 + \alpha_0\nu_t^2 + \beta_0\sigma_t^2\\
&\varepsilon_t \sim t.
\end{aligned}
$$

We let the hyperparameter c vary over $[0, 1]$. When $c = 0$ all DGPs actually correspond with the GARCH(1,1) model used to obtain the VaR and ES forecasts, whereas larger values of c indicate larger deviations from the null model. As c increases in specification A_1, the error distribution becomes leptokurtic in a way that is not captured by the Student's t distribution.[5] Having $c = 1$ results in a kurtosis of $3 + 27/10 = 5.7$, which corresponds to values found in the empirical literature. In specification A_2 the volatility equation differs from the null model for $c \neq 0$ by introducing a leverage effect, corresponding with the GJR-GARCH model of Glosten, Jagannathan, and Runkle (1993). The parameterization ensures stationarity. Specification A_3 corresponds with a GARCH-in-mean model for $c \neq 0$. Setting $c = 1$ results in a GARCH-in-mean model with coefficient 2.5 which is similar in magnitude to Du and Escanciano (2016), whereas smaller values of c around 0.10 are more in line with estimates found in, for example, Christensen, Dahl, and Iglesias (2012). We subtract the unconditional value of the variance, which is equal to 1, from σ^2_{t-1} in the mean equation to impose that the unconditional mean of Y_t remains zero, because the forecasting model does not have an intercept term. If the tests detect deviations from the null, it indicates that the tests can pick up deviations in the conditional mean equation. By studying these DGPs we believe to cover the most important types of misspecification from the null model.

3.2 Results

Table 1 provides size properties of the tests for forecasts with a horizon of one period and different combinations of R and P. The results show a varying performance of the different standard tests. The VaR tests $EO_P^{(1)}$ and $EO_P^{(2)}$, the joint tests $T_P^{(4)}$ and $T_P^{(5)}$, and the ES tests $DE_P^{(1)}$ and $DE_P^{(2)}$ score reasonably well with empirical rejection rates that are generally

5 The kurtosis of ε_t in specification A_1 equals $3(q-1)^2/q + 3 = 27c^2/(6c+4) + 3$.

Table 1 Empirical rejection rates in the size experiment for one-step-ahead forecasts

Panel A: Standard tests

	VaR		Joint (VaR, ES)					ES (D&E, 2016)		
	Unc.	Cond.	Unc.		Cond.			Unc.	Cond.	
R/P	$EO_P^{(1)}$	$EO_P^{(2)}$	$T_P^{(1)}$	$T_P^{(2)}$	$T_P^{(3)}$	$T_P^{(4)}$	$T_P^{(5)}$	$DE_P^{(1)}$	$DE_P^{(2)}$	$DE_P^{(3)}$
Coverage level $1-\alpha = 97.5\%$										
500/500	0.08	0.09	0.12	0.11	0.48	0.09	0.07	0.08	0.06	0.13
2500/500	0.07	0.08	0.18	0.16	0.51	0.07	0.09	0.10	0.06	0.22
500/2500	0.05	0.04	0.06	0.06	0.53	0.07	0.06	0.04	0.07	0.26
2500/2500	0.04	0.06	0.08	0.06	0.52	0.05	0.08	0.01	0.04	0.27
Coverage level $1-\alpha = 95\%$										
500/500	0.07	0.06	0.11	0.11	0.33	0.08	0.09	0.06	0.08	0.13
2500/500	0.05	0.05	0.12	0.12	0.36	0.04	0.10	0.04	0.06	0.16
500/2500	0.04	0.07	0.03	0.03	0.26	0.09	0.04	0.04	0.05	0.22
2500/2500	0.05	0.09	0.05	0.06	0.25	0.07	0.04	0.03	0.05	0.27

Panel B: Robust tests

Coverage level $1-\alpha = 97.5\%$										
500/500	0.03	0.05	0.10	0.09	0.35	0.06	0.05	0.04	0.04	0.09
2500/500	0.04	0.07	0.17	0.14	0.40	0.06	0.06	0.08	0.05	0.19
500/2500	0.01	0.03	0.05	0.04	0.36	0.06	0.04	0.01	0.05	0.24
2500/2500	0.01	0.05	0.06	0.04	0.41	0.03	0.07	0.01	0.04	0.26
Coverage level $1-\alpha = 95\%$										
500/500	0.03	0.03	0.07	0.07	0.23	0.06	0.07	0.02	0.03	0.09
2500/500	0.02	0.04	0.09	0.10	0.29	0.03	0.10	0.02	0.05	0.16
500/2500	0.00	0.06	0.02	0.02	0.20	0.07	0.02	0.01	0.04	0.17
2500/2500	0.01	0.09	0.02	0.03	0.22	0.07	0.04	0.01	0.05	0.26

Notes: This table presents empirical rejection rates of the tests introduced in Section 2. The forecasts are constructed with the GARCH model H_0 with Student's t distributed errors with $\nu = 5$ degrees of freedom. The GARCH(1,1) parameters are $(\omega_0, \alpha_0, \beta_0) = (0.05, 0.1, 0.85)$. The forecast horizon τ is 1 period, and the coverage levels are $1-\alpha = 97.5\%$ and 95%. We present rejection based on 1000 Monte Carlo experiments with a rolling window with different combinations of 500 and 2500 periods for the in-sample size R and out-of-sample size P.

somewhat larger than the nominal size of 5% but never exceed 10%. Tests $T_P^{(1)}$ and $T_P^{(2)}$ perform slightly worse with rates as high as 18% and 16% for $R = 2500$ and $P = 500$. Tests $T_P^{(3)}$ and $DE_P^{(3)}$ show large size deviations with rejection rates in the range of 25–53% and 13–27%. The bad result for $T_P^{(3)}$ is in line with the findings of Nolde and Ziegel (2017), indicating that using the lagged value of the identification function $g_{t-\tau,2,\tau}(\theta)$ in the test matrix does not work well, even though it corresponds with a straightforward generalization of the conditional test $EO_P^{(2)}$. The good performance of $T_P^{(4)}$ shows how beneficial standardization of the test matrix is in this case. The deterioration of $DE_P^{(3)}$ compared with $DE_P^{(2)}$ may be related to the required estimation of additional autocorrelation terms. Overall, we find similar size distortions as Escanciano and Olmo (2010) and Du and Escanciano (2016).

Panel B of Table 1 shows that the robust versions of the tests generally have better size properties. Reductions in rejection rates vary between 0% and 5%-points for all tests except $T_P^{(3)}$, where reductions are larger. A closer comparison per test shows that the unconditional tests $EO_P^{(1)}$ and $DE_P^{(1)}$ tend to become undersized. The differences between standard and robust versions are small for $EO_P^{(2)}$, $T_P^{(4)}$, and $DE_P^{(2)}$ confirming that their standardization reduces estimation error. The robust versions of $T_P^{(3)}$ and $DE_P^{(3)}$ score better than their standard versions, but the empirical rejection rates are still too high.

Comparing the different combinations R/P of in-sample and out-of-sample window lengths gives some further insights. The improvement in the rejection rates tends to be larger for $R = P = 500$ (with $\widehat{\pi} = 1$) than for $R = 2500$ and $P = 500$ (with $\widehat{\pi} = 1/5$). Comparing $R = 500$ with 2500 when $P = 2500$ does not show a clear pattern. When P is large, the rejection rates of the robust tests with $R = 500$ are close to those with $R = 2500$. The robust tests still show large distortions for the combination of $R = 2500$, $P = 500$, indicating that the statistics may not have completely converged to the normal distribution when the out-of-sample window is short. We do not observe a systematic effect of the coverage level.

We next compare the standard and robust versions of the tests for 10-step-ahead forecasts in Table 2. We do not include the tests of Du and Escanciano (2016) as their robust versions have not been extended to multi-step-ahead forecasts. Overall, the tests perform a bit better than for the one-step-ahead forecast, with empirical rejection rates closer to the nominal 5%. Because 10-step-ahead forecasts more strongly reflect the steady-state distribution, they may be less affected by estimation error than one-step-ahead forecasts. As before, the rejection rates for the standard versions of $EO_P^{(1)}$, $EO_P^{(2)}$, $T_P^{(4)}$, and $T_P^{(5)}$ are below 10%. The highest rates for the standard versions of $T_P^{(1)}$ and $T_P^{(2)}$ come down to 15%. The performance of $T_P^{(3)}$ has improved a bit, but generally it remains strongly oversized.

Also for 10-step-ahead forecasts, robust versions of the tests generally have better size. The empirical rejection rates for $EO_P^{(2)}$ and all T_p tests except $T_P^{(3)}$ are closer to 5% and sometimes below. The improvement is largest for the unconditional tests $T_P^{(1)}$ and $T_P^{(2)}$. The robust version of $EO_P^{(1)}$ seems to become undersized, in particular for the coverage level of 95%. Test $T_P^{(3)}$ remains oversized, though its performance with rates between 12% and 40% is considerably better than for the standard version.

The effect of accounting for estimation error for the different combinations of in-sample size R and out-of-sample size P lines up more clearly with the theory now. The effect is largest for $R = 500$ and $P = 2500$ (with $\widehat{\pi} = 5$), then $R = P = 500$ and $R = P = 2500$ (with both $\widehat{\pi} = 1$), and smallest for $R = 2500$ and $P = 500$ ($\widehat{\pi} = 1/5$). Size distortions for the robust versions of the tests are also largest for this combination. The performance of the tests seems a bit better for the 95% coverage level, though differences with the 97.5% level are small.

Overall, we conclude that accounting for estimation error by using robust test versions leads to sizes that are closer to the nominal 5%. The improvement is smaller for the conditional tests, in particular when well-chosen standardization in the test statistic reduces the impact of estimation error. As expected, the improvement is largest when the in-sample size is large compared with the out-of-sample size. Finally, we see that the tests $T_P^{(3)}$ and $DE_P^{(3)}$ perform badly, whether we correct for estimation error or not.

A repetition of the size experiment in which we set the degrees of freedom of the Student's t distribution for the errors equal to $\nu = 30$ largely confirms these conclusions (see

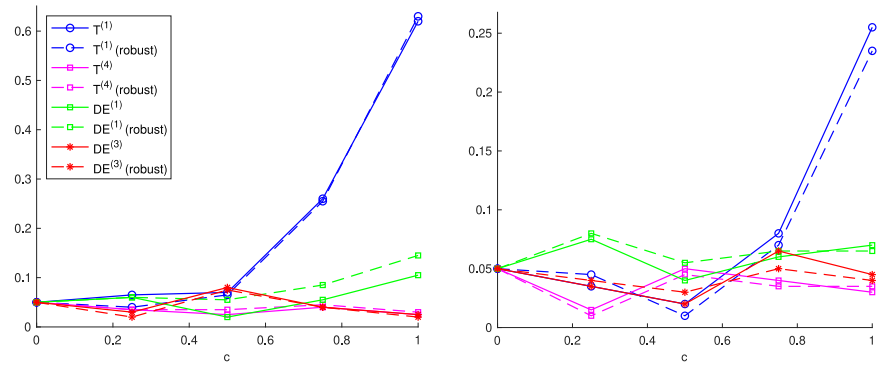
Table 2 Empirical rejection rates in the size experiment for 10-step-ahead forecasts

Panel A: Standard tests							
	VaR		Joint (VaR, ES)				
	Unc.	Cond.	Unc.		Cond.		
R/P	$EO_P^{(1)}$	$EO_P^{(2)}$	$T_P^{(1)}$	$T_P^{(2)}$	$T_P^{(3)}$	$T_P^{(4)}$	$T_P^{(5)}$
Coverage level $1-\alpha = 97.5\%$							
500/500	0.07	0.06	0.09	0.12	0.34	0.06	0.06
2500/500	0.08	0.08	0.15	0.15	0.26	0.10	0.07
500/2500	0.01	0.07	0.01	0.03	0.37	0.06	0.04
2500/2500	0.01	0.05	0.05	0.07	0.51	0.07	0.04
Coverage level $1-\alpha = 95\%$							
500/500	0.04	0.06	0.09	0.10	0.25	0.06	0.04
2500/500	0.08	0.06	0.12	0.11	0.28	0.08	0.07
500/2500	0.02	0.04	0.04	0.10	0.22	0.05	0.06
2500/2500	0.03	0.06	0.06	0.08	0.24	0.07	0.04
Panel B: Robust tests							
Coverage level $1-\alpha = 97.5\%$							
500/500	0.02	0.04	0.05	0.05	0.17	0.04	0.03
2500/500	0.05	0.07	0.11	0.10	0.18	0.08	0.05
500/2500	0.00	0.04	0.00	0.01	0.25	0.04	0.01
2500/2500	0.01	0.04	0.03	0.04	0.40	0.06	0.03
Coverage level $1-\alpha = 95\%$							
500/500	0.01	0.04	0.05	0.04	0.12	0.04	0.02
2500/500	0.03	0.05	0.09	0.09	0.19	0.07	0.04
500/2500	0.01	0.03	0.02	0.06	0.15	0.05	0.05
2500/2500	0.01	0.05	0.04	0.06	0.21	0.07	0.03

Notes: This table presents empirical rejection rates of the tests introduced in Section 2. The forecasts are constructed with the GARCH model H_0 with Student's t distributed errors with $\nu = 5$ degrees of freedom. The GARCH(1,1) parameters are $(\omega_0, \alpha_0, \beta_0) = (0.05, 0.1, 0.85)$. The forecast horizon τ is 10 periods, and the coverage levels are $1-\alpha = 97.5\%$ and 95%. We present rejection based on 1000 Monte Carlo experiments with a rolling window with different combinations of 500 and 2500 periods for the in-sample size R and out-of-sample size P.

the Supplemental Materials). For both 1- and 10-step-ahead forecasts, the robust versions of the tests have better size properties, and tests $T_P^{(3)}$ and $DE_P^{(3)}$ perform badly. The empirical rejection rates for the conditional tests do not differ much from those in Tables 1 and 2. For the unconditional tests, the sizes of the standard versions applied to one-period forecasts are better than in Table 1 and the robust versions tend to become undersized, whereas the performance of the standard versions of $T_P^{(1)}$ and $T_P^{(2)}$ with 10-step-ahead forecasts is slightly worse compared with Table 2. The robust versions of these tests show rejection rates closer to the nominal 5%. Overall, differences between the forecasting horizons seem to be smaller.

To compare the power of the different standard and robust tests, we calculate empirical rejection rates for the three DGPs A_1 to A_3, as we let c take values on an equally spaced



(a) horizon $\tau = 1$ (b) $\tau = 10$

Figure 1 Empirical (size-corrected) power in case of misspecifaction of the error distribution. This figure plots the empirical rejection rates of the standard and robust versions of tests $T_P^{(1)}$, $T_P^{(4)}$, $DE_P^{(1)}$, and $DE_P^{(3)}$ as a function of c which determines the deviation from the null model. The DGP is A_1 with $(\omega_0, \alpha_0, \beta_0) = (0.05, 0.1, 0.85)$. We estimate the model H_0 with a rolling window of $R = 2500$ observations, and generate forecasts of VaR and ES with a coverage level of $1 - \alpha = 97.5\%$ for horizons of $\tau = 1$ and 10 periods. The backtests use an out-of-sample window of $P = 2500$ observations. We evaluate the test statistics with critical values such that all tests have sizes of 5% and use 1000 simulations.

five-point grid of $[0, 1]$. We focus on the results for $R/P = 2500/2500$, because the size experiments suggest that the distribution of the robust statistics is close to normal when $P = 2500$. Our choice for $R = 2500$ is guided by the size distortions of the standard tests that are not too large to make a comparison with the robust tests irrelevant. When risk is measured on a daily basis, 2500 observations roughly correspond with 10 years, which is easily attainable for many financial time series. We cover the tests $T_P^{(1)}$, $T_P^{(4)}$, $DE_P^{(1)}$, and $DE_P^{(3)}$, because we are mostly interested in the power properties of the joint (VaR, ES) tests, and their robust versions. We include $DE_P^{(1)}$ and $DE_P^{(3)}$ as benchmarks. We correct the critical values such that all tests have correct size.

We present the power curves of the different tests for the different DGPs in Figures 1–3. To make for a fair comparison, we determine the critical value for each test such that they have correct size. The pictures show that the power curves of the standard and robust versions of each test lie close to each other. If lower, the power of the robust version is generally at most 5%-point lower. The performance of the tests shows quite some variation over the different DGPs, meaning that there is not a clear (robust) test that works well in all the directions of misspecification that we consider.

The power curves in Figure 1 show that both the standard and robust versions of $T_P^{(1)}$ detect the mixed-normal alternative A_1 well for values of $c > 0.5$, which corresponds with kurtosis exceeding $27/28 + 3 \approx 4$. Apparently, mixtures with small values of c can be reasonably well approximated by a Student's t distribution. The difference between the power of the standard and robust version of $T_P^{(1)}$ is quite small. Since the misspecification is related to the error distribution and not to the dynamics of the H_0 model, the conditional tests $T_P^{(4)}$ and $DE_P^{(3)}$ have no power against it. It's remarkable that the unconditional test $DE_P^{(1)}$ also seems to have low power against this alternative even for large c.

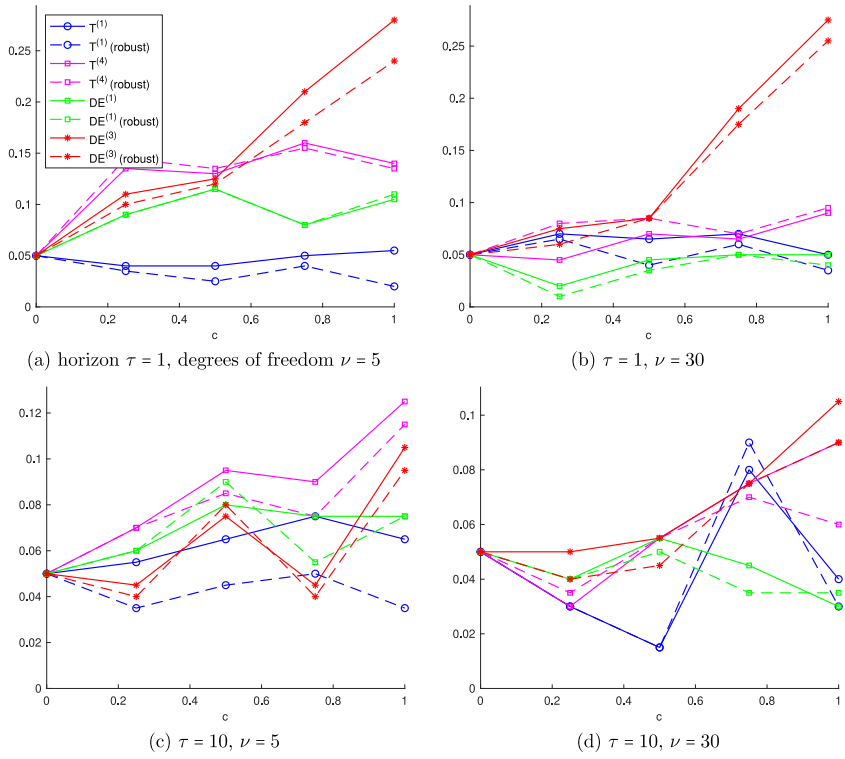


Figure 2 Empirical (size-corrected) power in case of misspecification of the volatility process. This figure plots the empirical rejection rates of the standard and robust versions of tests $T_P^{(1)}$, $T_P^{(4)}$, $DE_P^{(1)}$, and $DE_P^{(3)}$ as a function of c which determines the deviation from the null model. The DGP is A_1 with $(\omega_0, \alpha_0, \beta_0) = (0.05, 0.1, 0.85)$ and degrees of freedom $\nu = 5$ or 30. We estimate the model H_0 with a rolling window of $R = 2500$ observations, and generate forecasts of VaR and ES with a coverage level of $1 - \alpha = 97.5\%$ for horizons of $\tau = 1$ and 10 periods. The backtests use an out-of-sample window of $P = 2500$ observations. We evaluate the test statistics with critical values such that all tests have sizes of 5% and use 1000 simulations.

The results for the GJR-GARCH alternative A_2 in Figure 2 present a more mixed picture. For the horizon of $\tau = 1$ period, the conditional tests $T_P^{(4)}$ and in particular $DE_P^{(3)}$ have some power against this alternative. However, even for this test power is limited. For a value of $c = 1$, in which only negative innovations affect the volatility in the next period, the rejection rates do not exceed 30% (standard) and 25% (robust). The power of $T_P^{(4)}$ does not exceed 15 for $\nu = 5\%$ or 10% for $\nu = 30\%$. The low power for the unconditional tests is in line with Du and Escanciano (2016), who note that the conditional tests have more power against ARCH(2) and EGARCH alternatives, which also only differ in terms of the volatility equation. The misspecification does not affect much forecasts for longer horizons and consequently none of the tests shows much power for the horizon of $\tau = 10$ periods.

The analysis of the GARCH-in-mean alternative A_3 in Figure 3 shows that predominantly $T_P^{(1)}$ has power in detecting the conditional mean misspecification. The rejection rate

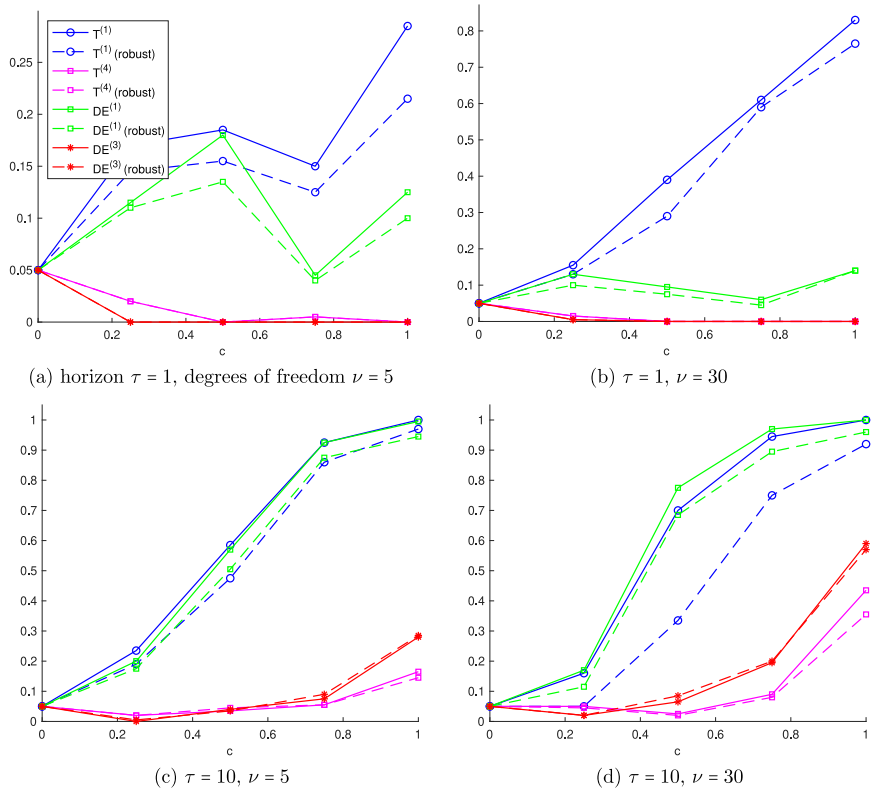


Figure 3 Empirical (size-corrected) power in case of misspecification of the mean process. This figure plots the empirical rejection rates of the standard and robust versions of tests $T_P^{(1)}$, $T_P^{(4)}$, $DE_P^{(1)}$, and $DE_P^{(3)}$ as a function of c which determines the deviation from the null model. The DGP is A_1 with $(\omega_0, \alpha_0, \beta_0) = (0.05, 0.1, 0.85)$ and degrees of freedom $\nu = 5$ or 30. We estimate the model H_0 with a rolling window of $R = 2500$ observations, and generate forecasts of VaR and ES with a coverage level of $1 - \alpha = 97.5\%$ for horizons of $\tau = 1$ and 10 periods. The backtests use an out-of-sample window of $P = 2500$ observations. We evaluate the test statistics with critical values such that all tests have sizes of 5% and use 1000 simulations.

of the standard and robust versions rise to around 25% for one-step-ahead forecasts when $\nu = 5$, and to around 80% when $\nu = 30$. The power is larger when 10-step-ahead forecasts are backtested. The power of both $T_P^{(1)}$ and $DE_P^{(1)}$ approximate 1 when the misspecification is large. The loss in power of using robust tests is bit larger here, in particular for the robust version of $T_P^{(1)}$ for $\nu = 30$. The remarkable increase in power for the longer forecasting horizon may be related to the persistence of the volatility process which translates into a shift of the whole distribution. The conditional tests $T_P^{(4)}$ and $DE_P^{(3)}$ do not have any power against A_3 for $\tau = 1$. For $\tau = 10$, power is larger, in particular when $\nu = 30$, but is far less compared with the unconditional tests.

We conclude that together the tests have reasonable power in detecting misspecification. For two out of the three cases of misspecification, the unconditional test $T_P^{(1)}$ performs well. We find that tests $T_P^{(4)}$ and $DE_P^{(3)}$ work well in one of the three cases, whereas $DE_P^{(1)}$

only performs well in one case with horizon $\tau = 10$. Similar as in Du and Escanciano (2016) it means that the tests should be employed together.

Our results for the other combinations of R and P, included in the Supplemental Materials, confirm these conclusions. These figures generally show the same patterns. As expected, an increase of P while keeping π fixed at one, moving from $R/P = 500/500$ to $R/P = 2500/2500$ improves power, indicative of test consistency. A reduction of R leads to a larger loss of power for the robust tests.

4 Empirical Analysis

4.1 Data and Models

In our empirical application, we backtest VaR and ES forecasts for daily returns of the FTSE 100 index as generated by three different models: GARCH, GJR-GARCH, and HEAVY. The models all prescribe a conditional mean and volatility model of the form

$$Y_{t+1} = a_0 Y_t + \nu_{t+1} \tag{20}$$

$$\nu_t = \sigma_t \varepsilon_t, \tag{21}$$

where the innovations ε_t follow a standardized Student's t distribution with degrees of freedom parameter ν which is also estimated. We include an AR(1)-term to accommodate the moderate level of autocorrelation in the daily index returns.

The models differ in their specification of the conditional volatility σ_t. For the GJR-GARCH model, the conditional volatility is given by

$$\sigma_{t+1}^2 = \omega_0 + \alpha_0 \nu_t^2 + \gamma_0 \nu_t^2 \, \mathbb{1}\{\nu_t < 0\} + \beta_0 \sigma_t^2. \tag{22}$$

The GARCH model follows as the special case with $\gamma_0 = 0$. The HEAVY model is a GARCH-type model that includes a realized measure RM_t in the conditional volatility specification and specifies ARMA-type dynamics for the realized measure itself

$$\sigma_{t+1}^2 = \omega_0 + \delta_0 RM_t + \beta_0 \sigma_t^2, \tag{23}$$

$$E_t[RM_{t+1}] = \omega_{RM,0} + \delta_{RM,0} RM_t + \beta_{RM,0} E_{t-1}[RM_t]. \tag{24}$$

We follow Shephard and Sheppard (2010) and use the realized kernel of Barndorff-Nielsen et al. (2008) to construct the realized measure. The τ-day-ahead VaR and ES forecasts are generated as described in the Supplemental Materials.

We obtain the returns on the FTSE 100 index as well as the realized measure from the Realized Library of the Oxford-Man Institute (Heber et al., 2009), and consider the sample period January 5, 2000 to April 17, 2019, adding up to 4865 observations. The starting date corresponds to the earliest available observation in the data set. We present summary statistics in the Supplementary Material. We first perform the tests with $R = 2500$ in-sample observations and $P = 2365$ out-of-sample observations, with estimation in a rolling window. This set-up aligns closely with the simulation scenario that is shown to have good

Table 3 Summary statistics of parameter estimates

	GARCH			GJR-GARCH			HEAVY		
	Mean	Median	SD	Mean	Median	SD	Mean	Median	SD
a_0	−0.027	−0.028	0.026	−0.031	−0.031	0.027	−0.014	−0.018	0.017
ω_0	0.012	0.010	0.004	0.025	0.021	0.008	0.022	0.022	0.009
α_0	0.097	0.099	0.009	0.039	0.036	0.008	–	–	–
δ_0	–	–	–	–	–	–	0.317	0.312	0.051
γ_0	–	–	–	0.082	0.075	0.021	–	–	–
β_0	0.893	0.893	0.011	0.879	0.888	0.015	0.693	0.694	0.051
$\omega_{RM,0}$							0.024	0.024	0.002
$\delta_{RM,0}$	–	–	–	–	–	–	0.326	0.318	0.024
$\beta_{RM,0}$	–	–	–	–	–	–	0.663	0.664	0.021
ν	10.813	11.223	2.278	13.863	13.904	2.484	14.644	15.450	3.970

Notes: This table provides summary statistics for the parameter estimates of the GARCH, GJR-GARCH, and HEAVY models, estimated using rolling windows of $R = 2500$ observations. The sample runs from January 5, 2000 to April 17, 2019, with the first estimation window running until December 7, 2009, resulting in a total of 2365 parameter estimations.

properties in Section 3. We also consider an in sample window of $R = 500$ observations. We then perform a subsample analysis in which we set $R = P = 1500$ to inspect the evolution of the tests. We focus again on 1- and 10-day-ahead forecasts.

Table 3 provides summary statistics of the parameter estimates for the different models using a rolling window of 2500 observations. There is substantial variation in the parameter estimates over this period, as evinced by the relatively large standard deviations. The small estimates for a_0 indicate weak autocorrelation of the daily returns in all specifications.[6] The parameters of the GARCH components fluctuate around their typical values. The introduction of the leverage parameter in the GJR-GARCH model shows that negative return shocks have a larger effect on conditional volatility, since γ_0 is generally positive. The HEAVY model has smaller β_0 estimates, consistent with Shephard and Sheppard (2010) who note that a value of β_0 about 0.6 is common in empirical applications. The degrees of freedom parameter ν_0 is generally quite large, suggesting that the returns are not very fat-tailed. The correction for estimation effects should therefore work well, as suggested by the simulation results.

4.2 Main Analysis

Figure 4 plots the 1- and 10-day-ahead ES forecasts at coverage level $1 - \alpha = 97.5\%$ as generated by the three models, estimated over a rolling window of $R = 2500$ observations. The forecasts lie close together, but the HEAVY model reacts more forcefully to large shocks and reverts more quickly, as explained by its reliance on the realized measure. Compared with the 1-day-ahead forecasts, the 10-day-ahead forecasts response to shocks is delayed.

6 We have repeated analysis imposing $a_0 = 0$ and results are qualitatively similar. Results can be obtained from the authors upon request.

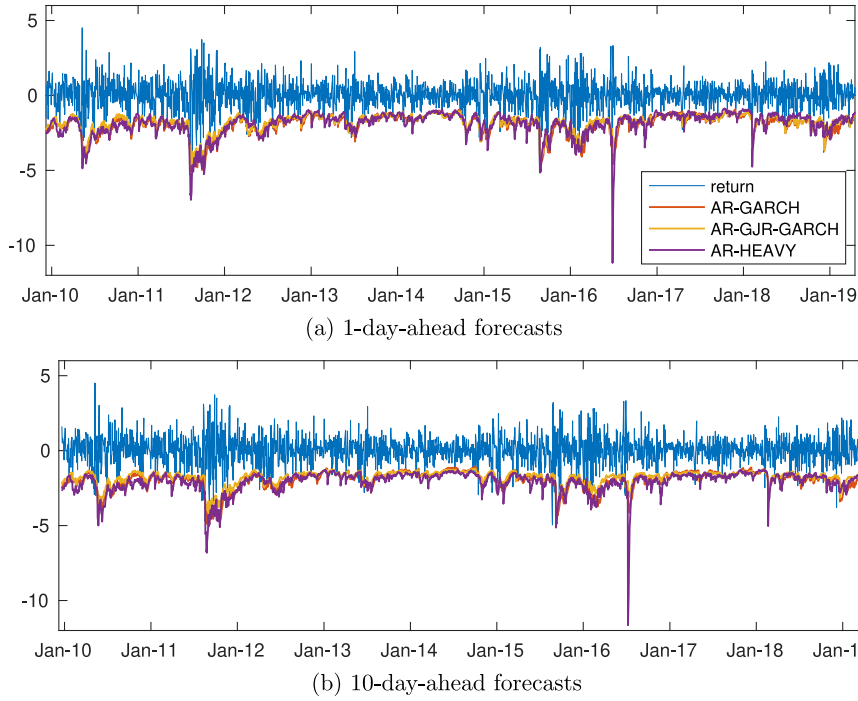


Figure 4 ES forecasts generated by GARCH, GJR-GARCH, and HEAVY models. This figure plots $\tau = 1$ and 10-day-ahead ES forecasts with a coverage level of $1 - \alpha = 97.5\%$ generated by the GARCH, GJR-GARCH, and HEAVY models over the out-of-sample period December 8, 2009 to April 17, 2019. The models are estimated with the rolling window scheme with $R = 2500$ in-sample observations.

The largest shock in the sample period is caused by the Brexit referendum on June 24, 2016.

Panel A of Table 4 shows the realized fraction of daily VaR violations

$$\widehat{\alpha} = \frac{1}{P}\sum_{t=R}^{T-\tau} \mathbb{1}\left[Y_{t+\tau} < \widehat{\mathrm{VaR}}_{t,\tau}\right], \tag{25}$$

with $\widehat{\mathrm{VaR}}_{t,\tau} = m_t(\widehat{\theta}_t)$ for $R = 500$ and 2500 in-sample observations. All models produce too optimistic VaR forecasts, since all fractions of VaR violations exceed the nominal 2.5%, ranging from 3.9% to 9.2% for 1-day-ahead, and from 10.9% to 19.5% for 10-day-ahead forecasts. The simple GARCH model yields fractions that are closest to 2.5%. One-day-ahead forecasts typically have closer to nominal VaR violations than 10-day-ahead forecasts, consistent with an earlier response to shocks. We do not find that either $R = 500$ or $R = 2500$ works best.

To evaluate the ES forecasts, we calculate the mean ES error (MESE) conditional on a VaR violation,

Table 4 Sample fraction of VaR violations and MESEs for a coverage level of 97.5%

Panel A: Sample ratio of violations of VaR(2.5%)			
Model	R	1-day-ahead	10-day-ahead
GARCH	500	0.039	0.109
	2500	0.052	0.123
GJR-GARCH	500	0.057	0.195
	2500	0.092	0.180
HEAVY	500	0.075	0.141
	2500	0.071	0.121
Panel B: Sample mean of ES(2.5%) error of violations			
Model	R	1-day-ahead	10-day-ahead
GARCH	500	0.055	0.052
	2500	0.063	0.059
GJR-GARCH	500	0.060	0.057
	2500	0.065	0.068
HEAVY	500	0.061	0.055
	2500	0.060	0.051

Notes: The models are estimated over a rolling window of R observations. Panel A reports the sample fraction of VaR violations defined in Equation (25). Panel B reports the MESE given a VaR violation, which is defined in Equation (26). We consider the coverage level $1-\alpha = 97.5\%$. VaR and ES forecasts are generated for a total 2365 out-of-sample observations. The sample runs from January 5, 2000 to April 17, 2019, with the first estimation window running until December 7, 2009.

$$\text{MESE} = \frac{1}{\widehat{\alpha} P} \sum_{t=R}^{T-\tau} \left(Y_{t+\tau} - \widehat{\text{ES}}_{t,\tau} \right) \mathbb{1}\left[Y_{t+\tau} < \widehat{\text{VaR}}_{t,\tau} \right], \tag{26}$$

with $\widehat{\alpha}$ as in Equation (25) and $\widehat{\text{ES}}_{t,\tau} = m_{t,2}(\widehat{\theta}_t)$. If a model is correctly specified, its MESE should be close to zero. Positive (negative) values for MESE indicate that the ES forecast is too pessimistic (optimistic). Panel B of Table 4 shows similar positive values for all models, so all generate too pessimistic ES forecasts. The MESE values are similar for both forecast horizons, and both number of in-sample observations.

Next we evaluate the quality of the forecasts by means of the conditional and unconditional tests of correct specification of Section 2. Table 5 shows the p-values of the standard and robust versions of the tests applied to the forecasts generated by the GARCH model. We omit $T_P^{(3)}$ because of its bad size properties. We first consider one-day-ahead forecasts. The standard versions of the tests generally indicate that the null of correct VaR and ES specification should be rejected. For $R = 500$ all unconditional tests reject the null, but the conditional tests $EO_P^{(2)}$, $T_P^{(4)}$, $T_P^{(5)}$, and $DE_P^{(2)}$ do not. For $R = 2500$ all tests reject except $T_P^{(5)}$, and the p-value of test $EO_P^{(2)}$ is 0.06. The robust versions of the tests generate larger p-values, resulting in changes in test outcomes depending on the chosen significance level (e.g., the p-values change from 6% to 22% for $EO_P^{(2)}$ and from 2% to 5% for $T_P^{(1)}$). The unconditional tests still mostly reject the null. The conditional tests are less affected because their standardization reduces estimation error.

The tests of the 10-day-ahead forecasts show a different picture. As before, we do not show Du and Escanciano (2016) test results, since formal inference results have not been

Table 5 *p*-values of VaR and ES tests applied to forecasts generated by the GARCH model

In-sample size R	VaR tests		Joint (VaR, ES) tests				ES tests		
	Unc.	Cond.	Unc.		Cond.		Unc.	Cond.	
	$EO_P^{(1)}$	$EO_P^{(2)}$	$T_P^{(1)}$	$T_P^{(2)}$	$T_P^{(4)}$	$T_P^{(5)}$	$DE_P^{(1)}$	$DE_P^{(2)}$	$DE_P^{(3)}$
Panel A: 1-day-ahead forecasts									
Standard tests									
500	0.00	0.86	0.01	0.00	0.68	0.21	0.00	0.64	0.00
2500	0.00	0.06	0.02	0.02	0.00	0.49	0.00	0.01	0.01
Robust tests									
500	0.01	0.86	0.03	0.02	0.68	0.21	0.01	0.64	0.01
2500	0.02	0.22	0.05	0.05	0.00	0.52	0.02	0.04	0.02
Panel B: 10-day-ahead forecasts									
Standard tests									
500	0.02	0.42	0.03	0.01	0.05	0.01	–	–	–
2500	0.08	0.31	0.21	0.15	0.57	0.07	–	–	–
Robust tests									
500	0.13	0.49	0.06	0.02	0.05	0.01	–	–	–
2500	0.15	0.31	0.36	0.22	0.57	0.09	–	–	–

Notes: This table presents *p*-values for the standard and robust versions of the tests introduced in Section 2. The VaR and ES forecasts are generate by the GARCH model estimated over rolling windows of $R = 500$ and 2500 observations. We consider coverage level $1 - \alpha = 97.5\%$, and forecast horizons $\tau = 1$ and 10 days. The sample runs from January 5, 2000 to April 17, 2019, with the first estimation window running until December 7, 2009, resulting in a total of 2365 parameter estimations.

derived for multi-period forecasts for these tests. For $R = 500$ the standard versions of all tests except $EO_P^{(2)}$ reject, but the robust versions lead to *p*-values that are often considerably higher. The *p*-value of test $EO_P^{(1)}$ increases from 0.02 to 0.13, and the one of $T_P^{(1)}$ from 0.03 to 0.21. So in these cases, estimation error has a profound effect and leads to different conclusions. The conditional tests are less affected. For $R = 2500$ the robust tests yield higher *p*-values but the standard versions did already not lead to rejection.

The results for the forecasts based on the GJR-GARCH specification in Table 6 are a bit more extreme compared with the GARCH specification. For one-day-ahead forecasts, both versions of the different unconditional tests reject the null, whereas both versions of the different conditional tests do not. The differences between the *p*-values of the standard and the robust versions of the tests are small. The same applies to 10-day-ahead forecasts with $R = 500$. For $R = 2500$, the outcome of the unconditional tests can change as the *p*-values increase from 0.01 to 0.02 for the standard tests to 0.05 to 0.07 for the robust tests. In all, the forecasting quality of this specification is less impacted by estimation error, and should generally be rejected.

The differences between the results of the standard and robust versions of the tests based on the forecasts from the HEAVY specification in Table 7 clearly show the impact of estimation error. For one-day-ahead forecasts and $R = 500$ all four unconditional tests reject in

Table 6 *p*-values of VaR and ES tests applied to forecasts generated by the GJR-GARCH model

In-sample size R	VaR tests		Joint (VaR, ES) tests				ES tests		
	Unc.	Cond.	Unc.		Cond.		Unc.	Cond.	
	$EO_P^{(1)}$	$EO_P^{(2)}$	$T_P^{(1)}$	$T_P^{(2)}$	$T_P^{(4)}$	$T_P^{(5)}$	$DE_P^{(1)}$	$DE_P^{(2)}$	$DE_P^{(3)}$
Panel A: 1-day-ahead forecasts									
Standard tests									
500	0.00	0.65	0.00	0.00	0.63	0.16	0.00	0.85	0.28
2500	0.00	0.63	0.00	0.00	0.15	0.10	0.00	0.19	0.15
Robust tests									
500	0.00	0.67	0.00	0.00	0.64	0.17	0.00	0.86	0.33
2500	0.00	0.70	0.01	0.01	0.16	0.18	0.00	0.45	0.27
Panel B: 10-day-ahead forecasts									
Standard tests									
500	0.00	0.48	0.00	0.00	0.49	0.00	–	–	–
2500	0.01	0.87	0.02	0.02	0.61	0.02	–	–	–
Robust tests									
500	0.01	0.55	0.00	0.00	0.51	0.00	–	–	–
2500	0.05	0.87	0.07	0.06	0.63	0.02	–	–	–

Notes: This table presents *p*-values for the standard and robust versions of the tests introduced in Section 2. The VaR and ES forecasts are generate by the GJR-GARCH model estimated over rolling windows of $R = 500$ and 2500 observations. We consider coverage level $1 - \alpha = 97.5\%$, and forecast horizons $\tau = 1$ and 10. The sample runs from January 5, 2000 to April 17, 2019, with the first estimation window running until December 7, 2009, resulting in a total of 2365 parameter estimations.

their standard and robust versions. However, for the larger estimation window of $R = 2500$, *p*-values increase from 0.00–0.03 to 0.03–0.10. In this case, we also see that the *p*-values of the conditionals tests $T_P^{(4)}$, $DE_P^{(2)}$, and $DE_P^{(5)}$ show increases from 0.02, 0.04 and 0.07 to 0.06, 0.09 and 0.12, which may lead to a different decision. The *p*-values for the short estimation window increase as well, but are less likely to change a decision. The picture for the 10-day-ahead forecasts is similar. Also, here *p*-values increase and for some tests this can lead to a different decision, most notably for $EO_P^{(1)}$, $T_P^{(1)}$, $T_P^{(2)}$, and $T_P^{(5)}$. Overall, the standard versions of the tests would indicate that the HEAVY model does not lead to correct VaR and ES forecasts, but accounting for estimation uncertainty, this conclusion does not survive when the HEAVY model is estimated with 2500 observations.

4.3 Subsample Analysis

To determine whether the impact of estimation error shows time-variation and can be related to the economic or financial cycle or events, we analyze the evolution of *p*-values for the $T_P^{(1)}$ and $T_P^{(4)}$ tests over consecutive subsamples of $R = P = 1500$ in-sample and out-of-sample observations (with rolling window estimation). We choose these two tests, because we can apply them to both one and 10-day-ahead forecasts, $T_P^{(1)}$ is an unconditional and $T_P^{(4)}$ a conditional test, and they have good theoretical and empirical properties.

Table 7 *p*-values of VaR and ES tests applied to forecasts generated by the HEAVY model

In-sample size R	VaR tests		Joint (VaR, ES) tests				ES tests		
	Unc.	Cond.	Unc.		Cond.		Unc.	Cond.	
	$EO_P^{(1)}$	$EO_P^{(2)}$	$T_P^{(1)}$	$T_P^{(2)}$	$T_P^{(4)}$	$T_P^{(5)}$	$DE_P^{(1)}$	$DE_P^{(2)}$	$DE_P^{(3)}$
Panel A: 1-day-ahead forecasts									
Standard tests									
500	0.00	0.55	0.00	0.00	0.58	0.16	0.00	0.89	0.02
2500	0.01	0.16	0.03	0.02	0.02	0.21	0.00	0.04	0.07
Robust tests									
500	0.00	0.56	0.00	0.00	0.59	0.19	0.00	0.89	0.04
2500	0.03	0.18	0.10	0.05	0.06	0.29	0.08	0.09	0.12
Panel B: 10-day-ahead forecasts									
Standard tests									
500	0.07	0.12	0.02	0.01	0.15	0.00	–	–	–
2500	0.58	0.52	0.21	0.04	0.53	0.03	–	–	–
Robust tests									
500	0.30	0.13	0.09	0.00	0.16	0.00	–	–	–
2500	0.74	0.54	0.63	0.12	0.66	0.10	–	–	–

Notes: This table presents *p*-values for the standard and robust versions of the tests introduced in Section 2. The VaR and ES forecasts are generate by the HEAVY model estimated over rolling windows of $R = 500$ and 2500 observations. We consider coverage level $1 - \alpha = 97.5\%$, and forecast horizons $\tau = 1$ and 10 days. The sample runs from January 5, 2000 to April 17, 2019, with the first estimation window running until December 7, 2009, resulting in a total of 2365 parameter estimations.

Because our sample consists of 4865 observations, the tests give 1865 *p*-values, the first one calculated with observations up to December 1, 2011. We choose these values for R and P because the previous analyses show that the effect of estimation error is not concentrated at windows of either 500 or 2500 observations, and in order to create a series of *p*-values that is long enough, and does not have too much overlap. Figures 5–7 plot the *p*-values with the end of the out-of-sample period on the horizontal axis, which we call the test date. The *p*-values with test date April 17, 2019, differ of course from those in Tables 5–7 because R and P are different. Of particular interest are the *p*-values reported from July 2014 to July 2015, because the observations belonging to the global credit crisis (July 2008 to July 2009) then move from the out-of-sample period to the in-sample period.

The *p*-values in Figure 5, based on forecasts from the GARCH model, show quite some variation over time, but the differences between those from the standard and from the robust tests seem quite constant. The *p*-values resulting from $T_P^{(1)}$ for both forecast horizons show a hump for test dates between July 2014 and July 2015, and from January 2017 onwards. The first hump coincides with the observations of the credit crisis being split over the in- and out-of-sample period. The second increase corresponds with in-sample (out-of-sample) periods from January 2005 (January 2011) onward. It may be explained by the volatility cluster around January 2016 in Figure 5 rolling into the out-of-sample window. If we choose the conventional significance level of 5%, the standard version of $T_P^{(1)}$ for one-

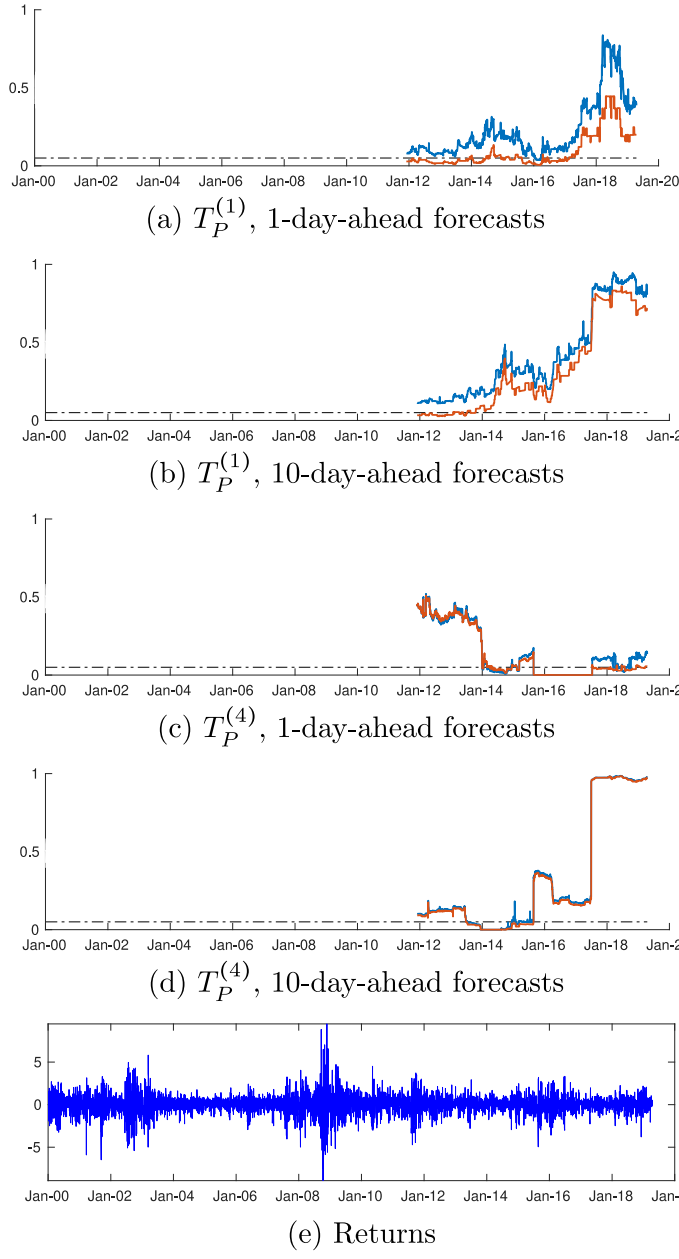


Figure 5 *p*-values over time based on forecasts from the GARCH model. This figure plots standard (red) and robust (blue) *p*-values of the unconditional test $T_P^{(1)}$ and the standardized conditional test $T_P^{(4)}$ through time, for the GARCH model with VaR and ES for a coverage level of $1-\alpha = 97.5\%$. Given a total sample of 4865 days of FTSE returns, and performing the tests using in- and out-sample sizes $R = P = 1500$, we obtain 1865 *p*-values for each test. We plot *p*-values at their test dates, which is the first working day after the end of the out-of-sample period. They run from December 1, 2011 to April 17, 2019. We use the rolling window scheme, and forecast horizons $\tau = 1$ and 10.

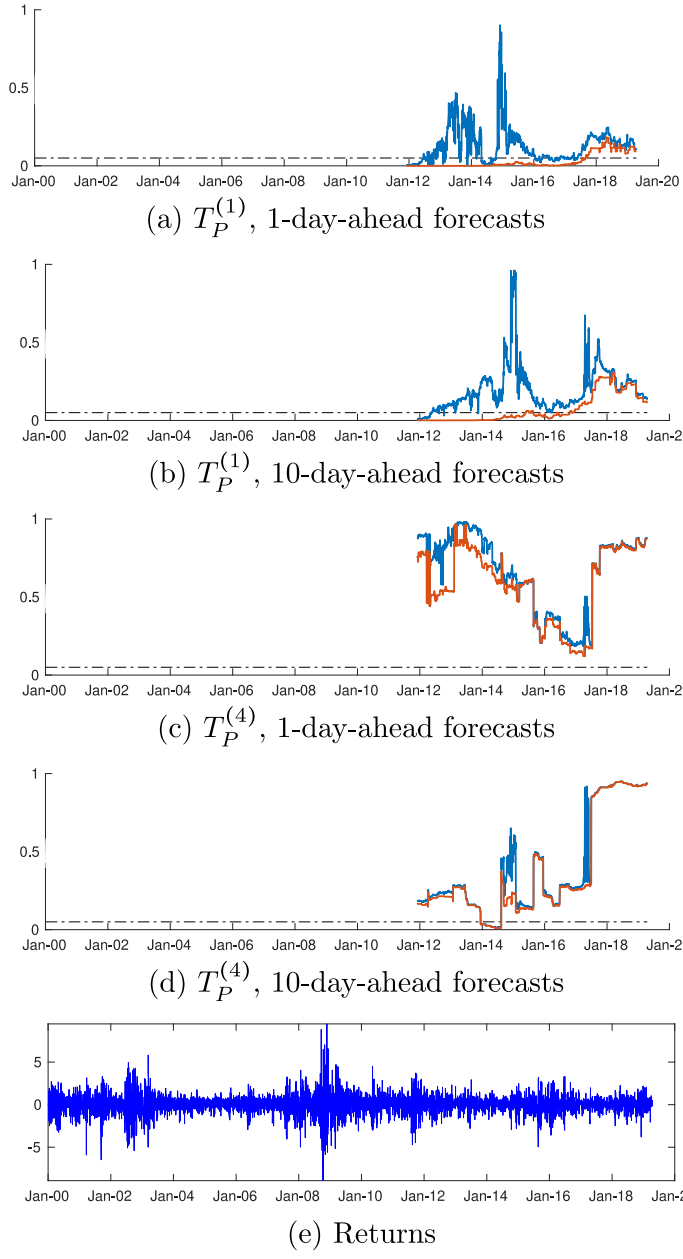
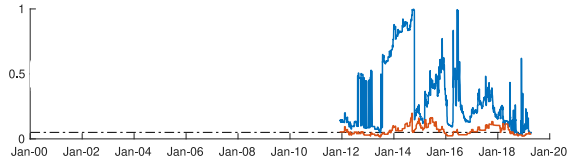
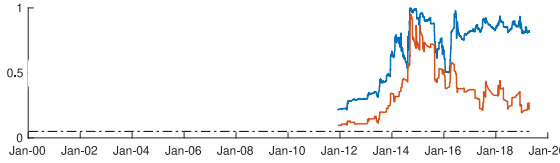


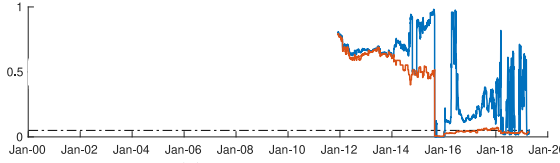
Figure 6 *p*-values over time based on forecasts from the GARCH model. This figure plots standard (red) and robust (blue) *p*-values of the unconditional test $T_P^{(1)}$ and the standardized conditional test $T_P^{(4)}$ through time, for the GARCH model with VaR and ES for a coverage level of $1 - \alpha = 97.5\%$. Given a total sample of 4865 days of FTSE returns, and performing the tests using in- and out-sample sizes $R = P = 1500$, we obtain 1865 *p*-values for each test. We plot *p*-values at their test dates, which is the first working day after the end of the out-of-sample period. They run from December 1, 2011 to April 17, 2019. We use the rolling window scheme, and forecast horizons $\tau = 1$ and 10.



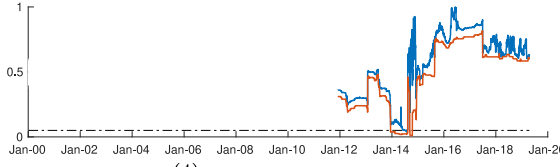
(a) $T_P^{(1)}$, 1-day-ahead forecasts



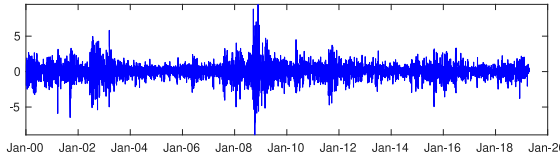
(b) $T_P^{(1)}$, 10-day-ahead forecasts



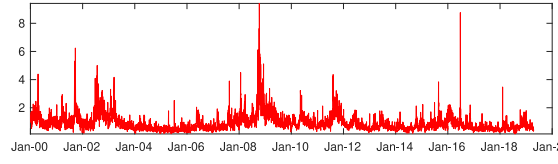
(c) $T_P^{(4)}$, 1-day-ahead forecasts



(d) $T_P^{(4)}$, 10-day-ahead forecasts



(e) Returns



(f) Realized volatility

Figure 7 *p*-values over time based on forecasts from the GARCH model. This figure plots standard (red) and robust (blue) *p*-values of the unconditional test $T_P^{(1)}$ and the standardized conditional test $T_P^{(4)}$ through time, for the GARCH model with VaR and ES for a coverage level of $1 - \alpha = 97.5\%$. Given a total sample of 4865 days of FTSE returns, and performing the tests using in- and out-sample sizes $R = P = 1500$, we obtain 1865 *p*-values for each test. We plot *p*-values at their test dates, which is the first working day after the end of the out-of-sample period. They run from December 1, 2011 to April 17, 2019. We use the rolling window scheme, and forecast horizons $\tau = 1$ and 10.

day-ahead forecasts leads to rejections for almost all test dates up to January 2017, while the robust version does not. Apparently, the effect of estimation error is large enough to reduce the evidence against correct specification. For 10 days ahead forecasts, the same applies to test dates up to July 2013. As before, test $T_P^{(4)}$ is less impacted by estimation error, so the p-values of the standard and the robust version are very close together. However, for the tests on one-day-ahead forecasts taken after July 2017, the standard version rejects while the robust version does not. For 10-day-ahead forecasts there are only a few number of dates where the test outcome differs.

The GJR-GARCH model has a richer specification than the GARCH model, and indeed, the differences between the p-values of the standard and the robust versions of the tests in Figure 6 are larger. The p-values from the robust version are also more jittery, indicating that the additional terms in the test statistic are less stable for this model. As a consequence, there is not a long period of consecutive tests dates for which the standard test rejects, while the robust ones do not, but we can conclude that the evidence against this specification for the period up to July 2017 is not as strong as the standard version of $T_P^{(1)}$ indicates. For the conditional test $T_P^{(4)}$, the differences between the standard and the robust versions are consequential for neither of the forecast horizons.

Figure 7 shows that for the case of the HEAVY model, the differences between the two series of p-values can also get large. Similar as for the GJR-GARCH model, the p-values of the robust test get jittery from time to time. We also see some upward and downward jumps in the p-values that persist for some time around 2015, that may be related to the high values for the realized volatility in Figure 7f in 2009, which move into the in-sample period for test dates in 2015. A closer inspection shows that estimation error reduces the evidences against the forecasts of the HEAVY model from time to time. This holds in particular for the test $T_P^{(1)}$ with dates between January 2012 and July 2013, the year 2016 and from July 18 onward, and for $T_P^{(4)}$ from March 16 onward, both based on one-day-ahead forecasts. For 10-day-ahead forecasts, both versions of $T_P^{(1)}$ never reject, whereas the outcomes from $T_P^{(4)}$ only differ for 2014.

Overall we conclude that there are no easy to spot patterns in the differences between the standard and robust versions of the tests. For the simple GARCH model, the differences are relatively constant and small. For the more extensive GJR-GARCH and HEAVY models, the differences are from time to time larger, and the robust versions of the tests lead to more jittery series of p-values. For all models, there are lengthy periods during which the standard tests reject the null of correct specification, while the robust tests do not, though these periods are not the same across the different models, tests, or forecast horizons. Though we suspect some effect of the observations of the credit crisis rolling from the out-of-sample into the in-sample period, the evidence for it is not clear-cut.

5 Concluding Remarks

In this paper, we examine the impact of estimation error on joint backtests for VaR and ES forecasts as proposed by Nolde and Ziegel (2017). Building on the general framework of McCracken (2000), we demonstrate that estimation error leads to additional terms in the asymptotic covariance matrix, which depend on the estimation scheme, the forecast horizon, and the ratio of in-sample to out-of-sample observations. We formulate robust tests that account for estimation error.

Using Monte Carlo simulations we show that standard tests may suffer from size distortions due to estimation error, with empirical rejection frequencies exceeding nominal significance levels. Robustifying the backtests generally corrects this issue quite successfully. We also show that standardization in the construction of the test statistic can help to improve the size of the test. We find that the robust tests have somewhat less power than the standard tests, but the reduction is quite modest, not exceeding 5–10%-point. The empirical application to daily VaR and ES forecasts for the FTSE 100 index illustrates that the effect of estimation error is not only a theoretical issue but also bears practical relevance. We find that estimation error has a substantial impact on the outcomes of the backtests, with p-values often increasing from below 0.05 to above when we switch from standard to robust versions of the backtests. This effect is not limited to particular periods or event in financial markets. The impact that estimation error potentially has on backtests means that financial institutions should take this into account when developing and evaluating risk measurement procedures for ES. Because it has better theoretical properties, ES is about to replace VaR. Ignoring the estimation uncertainty in the risk measurement procedure may actually lead to false rejections and may hinder the further development of these procedures.

Supplemental Data

Supplemental data are available at https://www.datahostingsite.com.

Appendix A: Proofs

A.1 Proof of Theorem 1

The result follows from Lemma 1 in Appendix B once we have shown that Assumptions C3–C5 in Appendix B are satisfied. We will do so in the following:

Assumption C3(c):
Notice that

$$\begin{aligned}
&\sup_t \| \sup_{\theta\in\Theta_0} v_t(\theta)\|_{2r} \\
&\le \sup_t \Big\| \sum_{i=1}^{l} \sup_{\theta\in\Theta_0} k_{t,i,\tau}(\theta) + \sum_{i=1}^{l} \sup_{\theta\in\Theta_0} E\big[k_{t,i,\tau}(\theta)\big] + \sum_{j=1}^{p} l_{t,j}(\theta)\Big\|_{2r} \\
&\le \sum_{i=1}^{l} \sup_t \| \sup_{\theta\in\Theta_0} k_{t,i,\tau}(\theta)\|_{2r} + \sum_{i=1}^{l} \sup_t \| \sup_{\theta\in\Theta_0} E\big[k_{t,i,\tau}(\theta)\big]\|_{2r} + \sum_{j=1}^{p} \sup_t \| \sup_{\theta\in\Theta_0} l_{t,j}(\theta)\|_{2r} \\
&\le 2\sum_{i=1}^{l} \sup_t \Big\| \sup_{\theta\in\Theta_0} k_{t,i,\tau}(\theta)\|_{2r} + \sum_{j=1}^{p} \sup_t \| \sup_{\theta\in\Theta_0} l_{t,j}(\theta)\|_{2r},
\end{aligned}$$

where the second inequality follows from the Triangle Inequality, and the third inequality follows from Hölder's Inequality (specifically $E|\cdot| \le \|\cdot\|_{2r}$, $r > 1/2$, for scalar random variables). We can thus establish (c) elementwise.

Notice that $g_{t,1,\tau}(\theta) \le 1$, and $\sup_{\theta\in\Theta_0} |g_{t,2,\tau}(\theta)| \le C\big(\sup_{\theta\in\Theta_0} |m(W_{t-1},\theta)| + |Y_t|\big)$. Moreover,

$$\begin{aligned}
\| \sup_{\theta\in\Theta_0} k_{t,i,\tau}(\theta)\|_{2r} &\le \| \sup_{\theta\in\Theta_0} H_{t,i,1}(\theta) \sup_{\theta\in\Theta_0} g_{t,1,\tau}(\theta)\|_{2r} + \| \sup_{\theta\in\Theta_0} H_{t,i,2}(\theta) \sup_{\theta\in\Theta_0} g_{t,1,\tau}(\theta)\|_{2r} \\
&\le \| \sup_{\theta\in\Theta_0} |H_{t,i,1}(\theta)|\|_{2c_H r} \times \| \sup_{\theta\in\Theta_0} |g_{t,1,\tau}(\theta)|\|_{2c_g r} + \| \sup_{\theta\in\Theta_0} |H_{t,i,2}(\theta)|\|_{2c_H r} \\
&\quad \times \| \sup_{\theta\in\Theta_0} |g_{t,2,\tau}(\theta)|\|_{2c_g r} \\
&\le C\left(1 + \| \sup_{\theta\in\Theta_0} |m_t(\theta)|\|_{2c_g r} + \|Y_t\|_{2c_g r}\right) < \infty,
\end{aligned}$$

where the first inequality follows from Minkowski's Inequality, the second inequality follows from Hölder's Inequality, since under Assumption 6 $c_H \in [1,\infty]$ and $c_g \in [1,\infty)$, and $1/c_H + 1/c_g = 1$. Moreover, by Hölder's Inequality we have under Assumption 6 that $\sup_t \| \sup_{\theta\in\Theta_0} H_{t,ij}(\theta)\|_{2c_H r} < C$, $\|Y_t\|_{2c_g r} < C$, $\sup_t \| \sup_{\theta\in\Theta_0} |m_t(\theta)|\|_{2c_g r} < C$, and $\sup_t \| \sup_{\theta\in\Theta_0} l_{t,j}(\theta)\|_{2r} < C$. The result follows.

Assumption C4:

Again we can work elementwise in terms of $k_{t,\tau}(\theta)$ and $l_t(\theta)$. By Assumption 4, $E[l_t(\theta)]$ is continuously differentiable on Θ_0. By the Mean Value Theorem, we obtain the expansion $E[l_{t,i}(\theta)] = E[l_{t,i}(\theta_0)] + \nabla E\left[l_{t,i}(\tilde{\theta})\right](\theta - \theta_0)$, for some $\tilde{\theta}$ between θ and θ_0 (elementwise). Additionally, under Assumption 6 we have $\sup_t \sup_{\theta\in\Theta_0} |\nabla E\left[l_{t,i}(\tilde{\theta})\right]| < C$, and $G = G_t$.

Now notice

$$\begin{aligned}
&E[k_{t,i,\tau}(\theta)] - E[k_{t,i,\tau}(\theta_0)] \\
&\quad = E\left[H_{t,i,j}(\theta)E_t[g_{t,j,\tau}(\theta)] - H_{t,i,j}(\theta_0)E_t[g_{t,j,\tau}(\theta_0)]\right] \\
&\quad = E\left[H_{t,i,j}(\theta)E_t[g_{t,j,\tau}(\theta) - g_{t,j,\tau}(\theta_0)] - \big(H_{t,i,j}(\theta_0) - H_{t,i,j}(\theta)\big)E_t[g_{t,j,\tau}(\theta_0)]\right] \\
&\quad = E[H_{t,i,j}(\theta)E_t[g_{t,j,\tau}(\theta) - g_{t,j,\tau}(\theta_0)]],
\end{aligned}$$

since $E_t[g_{t,j,\tau}(\theta_0)] = 0$ (a.s.).

Under the conditions on $F_t(\cdot)$ in Assumption 5 $E_t[g_{t,j,\tau}(\theta)]$ is (a.s.) continuously differentiable on Θ_0 with derivative

$$\begin{aligned}
\nabla E_t[g_{t,1,\tau}(\theta)] &= f_{t,\tau}\big(m_{t,1}(\theta)\big)\nabla m_{t,1}(\theta), \\
\nabla E_t[g_{t,2,\tau}(\theta)] &= \nabla m_{t,2}(\theta).
\end{aligned}$$

Hence, we obtain the expansion

$$E[k_{t,i,\tau}(\theta)] = E[k_{t,i,\tau}(\theta_0)] + E\left[H_{t,i,j}(\theta)\nabla E_t\left[g_{t,j,\tau}(\tilde{\theta})\right]\right](\theta - \theta_0),$$

for some $\tilde{\theta}$ between θ and θ_0 (elementwise).

That $\sup_t E\left[H_{t,i,j}(\theta)\nabla E_t\left[g_{t,j,\tau}(\tilde{\theta})\right]\right] < C$ follows if $\sup_t \| \sup_{\theta\in\Theta_0} |H_{t,i,j}(\theta)|\|_{c_H} < C$, and $\sup_t \| \sup_{\theta\in\Theta_0} |\nabla m_{t,2}(\theta)|\|_{c_g} < C$, as imposed in Assumption 6, and the boundedness of $f_{t,\tau}$ imposed in Assumption 5. That $A = A_t$ follows straightforwardly.

Assumption C5:

We establish C5 by showing (i) $\sup_t \| \sup_{\theta\in\Theta(\varepsilon)} k_{t,i,\tau}(\theta) - k_{t,i,\tau}(\theta_0)\|_Q \le C\varepsilon^{\phi}$, for all $i = 1,\dots,l$, and (ii) $\sup_t \| \sup_{\theta\in\Theta(\varepsilon)} l_{t,j}(\theta) - l_{t,j}(\theta_0)\|_Q \le C\varepsilon^{\phi}$, for all $j = 1,\dots,p$. Condition (ii) is imposed under Assumption 4.

To establish (ii) notice that

$$\begin{aligned}
&\| \sup_{\theta\in\Theta(\varepsilon)} g_{t,j,\tau}(\theta) - g_{t,j,\tau}(\theta_0)\|_Q \\
&\quad \le \| \sup_{\theta\in\Theta(\varepsilon)} H_{t,i,j}(\theta)\big(g_{t,j,\tau}(\theta) - g_{t,j,\tau}(\theta_0)\big) + g_{t,j,\tau}(\theta_0)\big(H_{t,i,j}(\theta) - H_{t,i,j}(\theta_0)\big)\|_Q \\
&\le \| \sup_{\theta\in\Theta(\varepsilon)} H_{t,i,j}(\theta)\big(g_{t,j,\tau}(\theta) - g_{t,j,\tau}(\theta_0)\big)\|_Q + \| \sup_{\theta\in\Theta(\varepsilon)} g_{t,j,\tau}(\theta_0)\big(H_{t,i,j}(\theta) - H_{t,i,j}(\theta_0)\big)\|_Q \\
&\quad \le \| \sup_{\theta\in\Theta} H_{t,i,j}(\theta)\|_{c_H Q} \times \| \sup_{\theta\in\Theta(\varepsilon)} g_{t,j,\tau}(\theta) - g_{t,j,\tau}(\theta_0)\|_{c_g Q} \\
&\qquad + \| \sup_{\theta\in\Theta(\varepsilon)} H_{t,i,j}(\theta) - H_{t,i,j}(\theta_0)\|_{c_H Q} \times \| \sup_{\theta\in\Theta} g_{t,j,\tau}(\theta)\|_{c_g Q}
\end{aligned}$$

That $\sup_t \| \sup_{\theta\in\Theta(\varepsilon)} H_{t,i,j}(\theta)\|_{c_H Q} < C$ and $\sup_t \| \sup_{\theta\in\Theta(\varepsilon)} g_{t,j,\tau}(\theta)\|_{c_g Q} < C$ follows from Assumption 6 and applying steps as in the preceding. Assumption 4 imposes $\| \sup_{\theta\in\Theta(\varepsilon)} H_{t,i,j}(\theta) - H_{t,i,j}(\theta_0)\|_{c_H Q} < C\varepsilon^{\phi}$.

Now notice that, for any $\xi > 1$, $|g_{t,1,\tau}(\theta) - g_{t,1,\tau}(\theta_0)|^{\xi} \le |g_{t,1,\tau}(\theta) - g_{t,1,\tau}(\theta_0)|$, since $|g_{t,1,\tau}(\theta)| \le 1$. Hence, $\|g_{t,1,\tau}(\theta) - g_{t,1,\tau}(\theta_0)\|_{c_g Q} \le \big(E|g_{t,1,\tau}(\theta) - g_{t,1,\tau}(\theta_0)|\big)^{1/(c_g Q)}$.

By monotonicity of the indicator function, it follows (a.s.)

$$\sup_{\theta,\theta'\in\Theta(\varepsilon)} |g_{t,1,\tau}(\theta) - g_{t,1,\tau}(\theta')| = g_{t,1,\tau}(\theta_{\max}(W_t)) - g_{t,1,\tau}(\theta_{\min}(W_t)),$$

where $\theta_{\max}(W_t)$ maximizes $m_1(W_{t-1},\cdot)$ and $\theta_{\min}(W_t)$ minimizes $m_1(W_{t-1},\cdot)$ over $\Theta(\varepsilon)$, for a given W_t. Moreover, we have $E_t[g_{t,1,\tau}(\theta_{\max}(W_t)) - g_{t,1,\tau}(\theta_{\min}(W_t))] = F_{t,\tau}(\theta_{\max}(W_t)) - F_{t,\tau}(\theta_{\min}(W_t)) \le C_f \sup_{\theta\in\Theta} |J_t(\theta)| \times 2\varepsilon$, with $C_f < \infty$ the upperbound imposed in Assumption 5.

Hence,

$$\begin{aligned}
\| \sup_{\theta\in\Theta(\varepsilon)} g_{t,1,\tau}(\theta) - g_{t,1,\tau}(\theta_0)\|_{c_g Q} &\le \left(E \sup_{\theta\in\Theta(\varepsilon)} |g_{t,1,\tau}(\theta) - g_{t,1,\tau}(\theta_0)|\right)^{1/(c_g Q)} \\
&\le \left(E[E_{t-1}[g_{t,1,\tau}((\theta_{\max}(W_t)) - g_{t,1,\tau}((\theta_{\min}(W_t))]]\right)^{1/(c_g Q)} \\
&\le (2C_f)^{1/(c_g Q)} E[\sup_{\theta\in\Theta_0} |J_t(\theta)|]^{1/(c_g Q)} \times \varepsilon^{1/(c_g Q)}
\end{aligned}$$

That $\sup_t E[\sup_{\theta\in\Theta_0} |J_t(\theta)|] < \infty$ is implied by $\sup_t \| \sup_{\theta\in\Theta_0} |J_t(\theta)|\|_{c_g Q}$, as imposed in Assumption 6.

Finally, it is easy to see that for all $\theta \in \Theta(\varepsilon)$ $g_{t,2,\tau}(\theta)$ satisfies the Lipschitz condition $|g_{t,2,\tau}(\theta) - g_{t,2,\tau}(\theta_0)| \le |m_t(\theta) - m_t(\theta_0)| \le \sup_{\theta\in\Theta_0} |J_t(\theta)\|\theta - \theta_0|$ (a.s.), with the second inequality following from the mean value theorem.

Hence, $\sup_t \| \sup_{\theta\in\Theta(\varepsilon)} |g_{t,2,\tau}(\theta) - g_{t,2,\tau}(\theta_0)|\|_{c_g Q} \le \sup_t \| \sup_{\theta\in\Theta_0} |J_t(\theta)|\|_{c_g Q} \times \varepsilon$, which is bounded under Assumption 6. The result follows. □

A.2 Proof of Corollary 1

The consistency of $\widehat{\Sigma}_P$ and $\widehat{V}_P$ follows from noting that the summands over j converge in probability by McCracken (2000)'s Lemma A.3. Since $w_{P,j}$ and $v_{P,j} \xrightarrow{p} 1$ the result follows straightforwardly. Also note that $\widehat{\rho}_P$ converges in probability by McCracken (2000)'s Lemma A.3.

We now consider $\widehat{A}_P$. The result follows from similar steps as in the proof of Theorem 3 in Engle and Manganelli (2004). For completeness we include the proof for a specific term in the definition of $\widehat{A}_P$:

$$\widehat{a}_P = \frac{1}{P}\sum_{t=R}^{T-\tau+1} H_{t,i,j}(\widehat{\theta}_t)(2\widehat{c}_P)^{-1}\mathbb{1}\Big(|Y_{t+\tau} - m_{t,1}(\widehat{\theta}_t)| < \widehat{c}_P\Big)J_{k,l,t}(\widehat{\theta}_t).$$

The proof for other terms contained in $\widehat{A}_P$ follow along similar lines, since conditions imposed on all elements of $H_t(\cdot)$ and $J_t(\cdot)$ are equivalent.

Define

$$\tilde{a}_P = \frac{1}{P}\sum_{t=R}^{T-\tau} H_{t,i,j}(\theta_0)(2c_P)^{-1}\mathbb{1}\big(|Y_{t+\tau} - m_{t,1}(\theta_0)| < c_P\big)J_{k,l,t}(\theta_0), \text{ and}$$

$$a_P = E\left[\frac{1}{P}\sum_{t=R}^{T-\tau} H_{t,i,j}(\theta_0)f_{t,\tau}\big(m_{t,1}(\theta_0)\big)J_{k,l,t}(\theta_0)\right].$$

We will first establish $\widehat{a}_P = \tilde{a}_P + o_P(1)$, and subsequently $\tilde{a}_P = a_P + o_P(1)$.

Also define $\widehat{\varepsilon}_t = Y_{t+\tau} - m_{t,1}(\widehat{\theta}_t)$, $\varepsilon_{0,t} = Y_{t+\tau} - m_{t,1}(\theta_0)$, and $\delta_t(\theta) = m_{t,1}(\theta) - m_{t,1}(\theta_0)$. Then,

$$\begin{aligned}
|\widehat{a}_P - \tilde{a}_P| \le{}& \frac{c_P}{\widehat{c}_P}\Big|(2Pc_P)^{-1} \times \sum_{t=R}^{T-\tau}\Big\{[\mathbb{1}(|\widehat{\varepsilon}_t| < \widehat{c}_P) - \mathbb{1}(|\varepsilon_{0,t}| < c_P)]H_{t,i,j}(\widehat{\theta}_t)J_{k,l,t}(\widehat{\theta}_t)\\
&+ \mathbb{1}(|\varepsilon_{0,t}| < c_P)(H_{t,i,j}(\widehat{\theta}_t) - H_{t,i,j}(\theta_0))J_{k,l,t}(\widehat{\theta}_t))\\
&+ \mathbb{1}(|\varepsilon_{0,t}| < c_P)(J_{k,l,t}(\widehat{\theta}_t) - J_{k,l,t}(\theta_0))H_{t,i,j}(\theta_0)\\
&+ \frac{c_P - \widehat{c}_P}{c_P}\mathbb{1}(|\varepsilon_{0,t}| < c_P)H_{t,i,j}(\theta_0)J_{k,l,t}(\theta_0)\}\Big|,
\end{aligned}$$

such that we have, for P sufficiently large, (a.s.) bounds

$$\begin{aligned}
|\widehat{a}_P - \tilde{a}_P| \le{}& \frac{c_P}{\widehat{c}_P}(2Pc_P)^{-1} \times \sum_{t=R}^{T-\tau}\Bigg\{\{\mathbb{1}(|\varepsilon_{0,t} - c_P| < |\delta_t(\widehat{\theta}_t)| + |\widehat{c}_P - c_P|)\\
&+ \mathbb{1}(|\varepsilon_{0,t} + c_P| < |\delta_t(\widehat{\theta}_t)| + |\widehat{c}_P - c_P|)\} \sup_{\theta\in\Theta_0}|H_{t,i,j}(\theta)| \sup_{\theta\in\Theta_0}|J_{k,l,t}(\theta)|)\\
&+ \mathbb{1}(|\varepsilon_{0,t}| < c_P)|H_{t,i,j}(\widehat{\theta}_t) - H_{t,i,j}(\theta_0)| \sup_{\theta\in\Theta_0}|J_t(\theta)|\\
&+ \mathbb{1}(|\varepsilon_{0,t}| < c_P)|H_{t,i,j}(\widehat{\theta}_t)\\
&- H_{t,i,j}(\theta_0)| \sup_{\theta\in\Theta_0}|H_t(\theta)| + \frac{c_P - \widehat{c}_P}{c_P}\mathbb{1}(|\varepsilon_{0,t}| < c_P) \sup_{\theta\in\Theta_0}|H_t(\theta)| \sup_{\theta\in\Theta_0}|J_t(\theta)|\Bigg\}\\
\equiv{}& \frac{c_P}{\widehat{c}_P}(A_1 + A_2 + A_3 + A_4),
\end{aligned} \tag{27}$$

by the mean value theorem, and Assumptions 4 and 6 [see Engle and Manganelli (2004) for elaboration].

We can pick any $d > 0$, such that eventually (P large) $|c_P - \widehat{c}_P|/c_P < d$, and $c_P^{-1}\sup_t|\widehat{\theta}_t - \theta_0| < d$, where the latter inequality follows from Lemma A.1 in McCracken (2000), which holds under the assumptions imposed in our Theorem 1. Given the inequality in Equation (27), we can show $E[A_i] = O(d)$, for $i = 1, \ldots, 4$, such that we obtain $|\widehat{a}_P - \tilde{a}_P| = o_P(1)$ from Markov's inequality.

We will show $E(A_1) = O(d)$, with the other equalities following from similar steps. Notice,

$$
\begin{aligned}
E[A_1] \leq\ & (2Pc_P)^{-1} \times \sum_{t=R}^{T-\tau} E\Bigg\{ \Big\{ \mathbb{1}\Big(|\varepsilon_{0,t} - c_P| < \sup_{\theta\in\Theta_0} |J_t(\theta)| \times |\widehat{\theta}_t - \theta_0| + |\widehat{c}_P - c_P| \Big) \\
& + \mathbb{1}\Big(|\varepsilon_{0,t} + c_P| < \sup_{\theta\in\Theta_0} |J_t(\theta)| \times |\widehat{\theta}_t - \theta_0| + |\widehat{c}_P - c_P| \Big) \Big\} \\
& \times \sup_{\theta\in\Theta_0} |H_{t,i,j}(\theta)| \sup_{\theta\in\Theta_0} |J_{k,l,t}(\theta)| \Bigg\} \\
\leq\ & (2Pc_P)^{-1} \\
& \times \sum_{t=R}^{T-\tau} E\{ 4dC_f c_P \Big(\sup_{\theta\in\Theta_0} |J_t(\theta)| + 1 \Big) \times \sup_{\theta\in\Theta} |H_t(\theta)| \times \sup_{\theta\in\Theta_0} |J_t(\theta)| \} \\
\leq\ & P^{-1} \sum_{t=R}^{T-\tau} 4dC_f \| \Big(\sup_{\theta\in\Theta_0} |J_t(\theta)| + 1 \Big) \times \sup_{\theta\in\Theta_0} |H_t(\theta)| \times \sup_{\theta\in\Theta_0} |J_t(\theta)| \|_1 \\
\leq\ & 4dC_f \sup_t \| \Big(\sup_{\theta\in\Theta_0} |J_t(\theta)| + 1 \Big) \times \sup_{\theta\in\Theta_0} |H_t(\theta)| \times \sup_{\theta\in\Theta_0} |J_t(\theta)| \|_1 \quad \leq 4dC_f K,
\end{aligned}
$$

with K some finite constant, and where the first inequality follows from noting that, for example, $E_t\big[\mathbb{1}\big(|\varepsilon_{0,t} - c_P| < y\big)\big] = F_{t,\tau}(y + c_P) - F_{t,\tau}(-y + -c_P) \leq C_f|y + c_P|$, for all $y \in \mathbb{R}$, and $H_t(\theta)$ and $J_t(\theta)$ are $\mathcal{F}_t$-measurable functions, and the last inequality follows from the bounds in Assumption 6 and Hölder's Inequality. For instance, notice

$$
\begin{aligned}
& \| \sup_{\theta\in\Theta} |H_t(\theta)| \times \sup_{\theta\in\Theta} |J_t(\theta)|^2 \|_1 \\
& \leq \| \sup_{\theta\in\Theta} |J_t(\theta)|^2 \|_{c_g} \times \| \sup_{\theta\in\Theta} |H_t(\theta)| \|_{c_H} \\
& \leq \Big(\| \sup_{\theta\in\Theta} |J_t(\theta)| \|_{c_g Q} \Big)^2 \times \| \sup_{\theta\in\Theta} |H_t(\theta)| \|_{c_H} < \infty,
\end{aligned}
$$

where the first inequality follows from $1/c_H + 1/c_g = 1$, the second inequality from $Q > 2$ as imposed in Assumption 4, and the third inequality from Assumption 6.

That $E[A_i] = O(d)$, for $i = 2, 3, 4$, follows from the bounds in Assumption 6.

Now we establish $\tilde{a}_P = a_P + o_P(1)$.

Rewrite

$$
\begin{aligned}
|\tilde{a}_P - a_P| = (2Pc_P)^{-1} & \sum_{t=R}^{T-\tau+1} \{ \big[\mathbb{1}\big(|\varepsilon_{0,t}| < c_P \big) - E_t\big[\mathbb{1}\big(|\varepsilon_{0,t}| < c_P \big) \big] \big] H_{t,i,j}(\theta_0) J_{k,l,t}(\theta_0) \} \\
+ P^{-1} & \sum_{t=R}^{T-\tau+1} \{ (2c_P)^{-1} \Big[E_t\big[\mathbb{1}\big(|\varepsilon_{0,t}| < c_P \big) \big] - E_t\big[f_{t,\tau}\big(m_{t,1}(\theta_0) \big) \big] \Big] H_{t,i,j}(\theta_0) J_{k,l,t}(\theta_0) \}
\end{aligned}
$$

The first term has zero mean by a martingale difference sequence property, and has variance equal to

$$(2Pc_P)^{-2}E\left\{\sum_{t=R}^{T-\tau+1}\left[\mathbb{1}\left(|\varepsilon_{0,t}| < c_P\right) - E_t\left[\mathbb{1}\left(|\varepsilon_{0,t}| < c_P\right)\right]\right]H_{t,i,j}(\theta_0)J_{k,l,t}(\theta_0)\right\}^2$$
$$\leq \left(4Pc_P^2\right)^{-1}\sup_t E\left[\sup_{\theta\in\Theta_0}|H_t(\theta)|^2\sup_{\theta\in\Theta_0}|J_t(\theta)|^2\right] = o(1),$$

where the first inequality follows from the cross-terms being zero by the martingale difference sequence property, and the equality in the last display follows from Hölder's Inequality and the moment bounds in Assumption 6. Hence, the first term converges to zero in mean-square, and therefore in probability.

To show convergence in probability to zero of the second term note that

$$\begin{aligned}&\left|(2c_P)^{-1}\left[E_t\left[\mathbb{1}\left(|\varepsilon_{0,t}| < c_P\right)\right] - E_t\left[f_{t,\tau}\left(m_{t,1}(\theta_0)\right)\right]\right]\right|\\ &\leq \left|(2c_P)^{-1}\int_{-c_P}^{c_P} f_{t,\tau}(y)dy - f_{t,\tau}\left(m_{t,1}(\theta_0)\right)\right|\\ &\leq \left|(2c_P)^{-1}2c_P f_{t,\tau}(y^*) - f_{t,\tau}\left(m_{t,1}(\theta_0)\right)\right|\\ &\leq L|c_P| = o_P(1),\end{aligned}$$

where $y^* = \text{argmax}_{y\in[-c_P,c_P]}f_{t,\tau}(y)$, and the third inequality follows from Assumption 5. By substituting, and noting that $P^{-1}\sum_{t=R}^{T-\tau+1}H_{t,i,j}(\theta_0)J_{k,l,t}(\theta_0)$ converges in probability to a finite limit by a LLN for mixing sequences [see White (2001, Cor. 3.48)] under Assumptions 3 and 6. □

Appendix B: Extended McCracken (2000) Results

Assumption C3. *For some $r > 1$, (a) (Y_t, Z_t') is strong mixing with coefficients of size $-2r/(r-1)$, (b) $v_t(\theta_0)$ is covariance stationary, and (c) for neighborhood Θ_0 of θ_0, $\sup_t \|\sup_{\theta\in\Theta} v_t(\theta)\|_{2r} < \infty$, (d) Ω is positive definite.*

Assumption C4. *For each $i \in \{1,\ldots,l+q\}$: (a) $E[v_{i,t}(\theta)]$ admits an expansion $E[v_{i,t}(\theta)] = E[v_{i,t}(\theta_0)] + \left(\partial E\left[v_{i,t}(\tilde{\theta})\right]/\partial\theta\right)(\theta-\theta_0)$, with $\partial E[v_{i,t}(\theta)]/\partial\theta$ a continuous function, for all $\theta\in\Theta_0$, with $v_{i,t}(\tilde{\theta})$ a scalar, θ a $p\times 1$ vector, and $\tilde{\theta}$ on the line between θ and θ_0, (b) there exists a finite constant D such that $\sup_t\sup_{\theta\in\Theta}|\partial E[v_{i,t}(\theta)]/\partial\theta| < D$, and (c) for all t, $G = G_t \equiv \partial E[l_t(\theta)]/\partial\theta|_{\theta=\theta_0}$ and $A = A_t \equiv \partial E[k_{t,\tau}(\theta)]/\partial\theta|_{\theta=\theta_0}$.*

Assumption C5. *Let $\Theta(\varepsilon) = \Theta(\theta_0,\varepsilon) \equiv \{\theta\in\mathbb{R}^p : |\theta-\theta_0| < \varepsilon\}$. There exist finite constants C, $\phi > 0$, and $Q \geq 2r$ such that for all $\Theta(\varepsilon)\subset\Theta_0$, $\sup_t\|\sup_{\theta\in\Theta(\varepsilon)}(v_t(\theta)-v_t(\theta_0))\|_Q \leq C\varepsilon^\phi$.*

The following lemma is an extension of McCracken (2000)'s Theorem 3.2.1, which extends their result to (i) estimators adhering to $\widehat{\theta}_t - \theta_0 = B(t)M(t) + o_P(t^{-1/2})$ instead of $\widehat{\theta}_t - \theta_0 = B(t)M(t)$, and (ii) alternative expansions of $E[v_{i,t}(\theta)]$ around θ_0, that is, $\partial E\left[v_{i,t}(\tilde{\theta})\right]/\partial\theta$ and $\partial E[k_{t,\tau}(\theta)]/\partial\theta$ do not have to be the partial derivatives and Jacobian matrix of $E\left[v_{i,t}(\tilde{\theta})\right]$ and $E[k_{t,\tau}(\theta)]$, respectively, although they still have to be continuous functions.

The first expansion is important to allow for a larger class of estimators (see the comment below Assumption 1). The second expansion is specifically important to our case, as

it allows the application of Theorem 3.2.1 for non-smooth H_t. Both extensions do not alter the steps in the proof materially.

Lemma 1. *Let Assumptions 1, 2, C3, C4, and C5 be satisfied. It follows under* $\mathcal{H}_0$ *that* $S_P \xrightarrow{d} N(0, \Omega)$. *Moreover, if* $\pi = 0$, *then* $S_P \xrightarrow{d} N(0, \Sigma)$.

Proof. The result follows straightforwardly by application of Theorem 2.3.1 in McCracken (2000), with two modifications. These modifications allow for a larger class of estimators $\widehat{\theta}_t$, and non-smooth H_t functions, but they do not require changes to the steps in the proof of Theorem 2.3.1.

We establish that the conditions are sufficient for Theorem 2.3.1, once we consider alterations to the proof. We will denote the five assumptions imposed in Theorem 2.3.1 as MC1–MC5. Assumptions 1(a, b) are identical to Assumption MC1(a, b). MC1(c) follows from Assumption 1(c) by the Law of Iterated Expectation. Finally, notice the first modification imposing the equality $\widehat{\theta}_t - \theta_0 = B(t)M(t) + o_P(t^{-1/2})$ where McCracken (2000) imposes $\widehat{\theta}_t - \theta_0 = B(t)M(t)$. This modification does not change the proof materially, since it only results in additional $o_P(1)$ terms in the (in)equalities in the proofs of McCracken (2000)'s Lemma A.1 and Lemma 2.3.2.

Assumption C4 is equivalent to Assumption MC4, except for the second modification that we do not impose the expansion in (a) to be a mean-value expansion. This does not change the proofs in McCracken (2000). Assumptions 2, C3, and C5 are identical to Assumptions MC2, MC3, and MC5, respectively. □

References

Acerbi, C., and B. Székely. 2014. Back-Testing Expected Shortfall. *Risk* 76–81.

Acerbi, C., and D. Tasche. 2002. On the Coherence of Expected Shortfall. *Journal of Banking & Finance* 26: 1487–1503.

Artzner, P., F. Delbaen, J.-M. Eber, and D. Heath. 1997. Thinking Coherently. *Risk* 68–71.

Artzner, P., F. Delbaen, J.-M. Eber, and D. Heath. 1999. Coherent Measures of Risk. *Mathematical Finance* 9: 203–228.

Barendse, S. 2020. "Efficiently Weighted Estimation of Tail and Interquantile Expectations." Tinbergen Institute Discussion Paper 2017-034/III.

Barndorff-Nielsen, O. E., P. R. Hansen, A. Lunde, and N. Shephard. 2008. Designing Realized Kernels to Measure the Ex Post Variation of Equity Prices in the Presence of Noise. *Econometrica* 76: 1481–1536.

BCBS. 2016. "Minimum Capital Requirements for Market Risk." Technical Report, Basel Committee on Banking Supervision.

Berkowitz, J. 2001. Testing Density Forecasts, with Applications to Risk Management. *Journal of Business & Economic Statistics* 19: 465–474.

Berkowitz, J., P. Christoffersen, and D. Pelletier. 2011. Evaluating Value-at-Risk Models with Desk-Level Data. *Management Science* 57: 2213–2227.

Bollerslev, T. 1986. Generalized Autoregressive Conditional Heteroskedasticity. *Journal of Econometrics* 31: 307–327.

Christensen, B. J., C. M. Dahl, and E. M. Iglesias. 2012. Semiparametric Inference in a GARCH-in-Mean Model. *Journal of Econometrics* 167: 458–472.

Christoffersen, P. F. 1998. Evaluating Interval Forecasts. *International Economic Review* 39: 841–862.

Cont, R., R. Deguest, and G. Scandolo. 2010. Robustness and Sensitivity Analysis of Risk Measurement Procedures. *Quantitative Finance* 10: 593–606.

Du, Z., and J. C. Escanciano. 2016. Backtesting Expected Shortfall: Accounting for Tail Risk. *Management Science* 63: 940–958.

Emmer, S., M. Kratz, and D. Tasche. 2015. What is the Best Risk Measure in Practice? Comparison of Standard Measures. *The Journal of Risk* 18: 31–60.

Engle, R. F., and S. Manganelli. 2004. CAViaR: Conditional Autoregressive Value at Risk by Regression Quantiles. *Journal of Business & Economic Statistics* 22: 367–381.

Escanciano, J. C., and J. Olmo. 2010. Backtesting Parametric Value-at-Risk with Estimation Risk. *Journal of Business & Economic Statistics* 28: 36–51.

Fissler, T., and J. F. Ziegel. 2016. Higher Order Elicitability and Osband's Principle. *The Annals of Statistics* 44: 1680–1707.

Giacomini, R., and H. White. 2006. Tests of Conditional Predictive Ability. *Econometrica* 74: 1545–1578.

Glosten, L. R., R. Jagannathan, and D. E. Runkle. 1993. On the Relation between the Expected Value and the Volatility of the Nominal Excess Return on Stocks. *The Journal of Finance* 48: 1779–1801.

Gneiting, T. 2011. Making and Evaluating Point Forecasts. *Journal of the American Statistical Association* 106: 746–762.

Heber, G., A. Lunde, N. Shephard, and K. Sheppard. 2009. Oxford-Man Institute's Realized Library, Version 0.3. https://realized.oxford-man.ox.ac.uk/ (last accessed 17 April 2019).

Kerkhof, J., and B. Melenberg. 2004. Backtesting for Risk-Based Regulatory Capital. *Journal of Banking & Finance* 28: 1845–1865.

Kole, E., T. Markwat, A. Opschoor, and D. van Dijk. 2017. Forecasting Value-at-Risk under Temporal and Portfolio Aggregation. *Journal of Financial Econometrics* 15: 649–677.

Kupiec, P. H. 1995. Techniques for Verifying the Accuracy of Risk Measurement Models. *The Journal of Derivatives* 3: 73–84.

McCracken, M. W. 2000. Robust Out-of-Sample Inference. *Journal of Econometrics* 99: 195–223.

McNeil, A. J., and R. Frey. 2000. Estimation of Tail-Related Risk Measures for Heteroscedastic Financial Time Series: An Extreme Value Approach. *Journal of Empirical Finance* 7: 271–300.

Newey, W. K., and K. D. West. 1987. A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica* 55: 703–708.

Nolde, N., and J. F. Ziegel. 2017. Elicitability and Backtesting: Perspectives for Banking Regulation. *The Annals of Applied Statistics* 11: 1833–1874.

Patton, A. J., J. F. Ziegel, and R. Chen. 2019. Dynamic Semiparametric Models for Expected Shortfall (and Value-at-Risk). *Journal of Econometrics* 211: 388–413.

Pham, T. D., and L. T. Tran. 1985. Some Mixing Properties of Time Series Models. *Stochastic Processes and Their Applications* 19: 297–303.

Powell, J. L. 1986. Censored Regression Quantiles. *Journal of Econometrics* 32: 143–155.

Shephard, N., and K. Sheppard. 2010. Realising the Future: Forecasting with High-Frequency-Based Volatility (HEAVY) Models. *Journal of Applied Econometrics* 25: 197–231.

West, K. D. 1996. Asymptotic Inference about Predictive Ability. *Econometrica* 64: 1067–1084.

West, K. D. 2006. "Forecast Evaluation." In G., ElliottC., Grangerand A., Timmermann (eds.), *Handbook of Economic Forecasting*, vol. 1, Chapter 3, pp. 99–134. Elsevier.

West, K. D., and M. W. McCracken. 1998. Regression-Based Tests of Predictive Ability. *International Economic Review* 39: 817–840.

White, H. 2001. *Asymptotic Theory for Econometricians*. Cambridge, MA: Academic Press.

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.