TI 2019-058/III
Tinbergen Institute Discussion Paper

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

Sander Barendse[1]
Erik Kole[2,3]
Dick van Dijk[2,3,4]

[1] University of Oxford
[2] Econometric Institute, Erasmus University Rotterdam
[3] Tinbergen Institute
[4] Erasmus Research Institute of Management

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

Sander Barendse†[1], Erik Kole[2,3], and Dick van Dijk[2,3,4]

[1] *University of Oxford*

[2] *Econometric Institute, Erasmus University Rotterdam*

[3] *Tinbergen Institute*

[4] *Erasmus Research Institute of Management.*

August 2019

Abstract

We investigate the effect of estimation error on backtests of (multi-period) 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. 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, although this also depends on the choice of conditioning variables. In an application to VaR and ES forecasts for daily FTSE 100 index returns as generated by AR-GARCH, AR-GJR-GARCH, and AR-HEAVY models, we find that estimation error substantially impacts the outcome of the backtests.

Keywords: expected shortfall; backtesting; risk management; tail risk; Value-at-Risk.

JEL codes: C12, C53, C58, G17

*We thank Julie Schnaitmann and seminar participants at CFE-CMStatistics 2018 conference in Pisa for valuable comments and feedback. Errors remain our own.

†Corresponding Author: Nuffield College, New Road, Oxford, OX1 1NF, UK E-mail addresses are Barendse: sander.barendse[at]economics.ox.ac.uk Kole: kole[at]ese.eur.nl, Van Dijk: djvandijk[at]ese.eur.nl.

1 Introduction

Research into financial risk management can nowadays be split into two strands. The first and oldest strand investigates risk measures from both a theoretical and practical perspective (see Emmer et al., 2015, for a recent example). The second, more recent strand investigates what Cont et al. (2010) have coined the risk measurement procedure, i.e., the procedure with which the forecast of the risk measure is generated. This 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 et al., 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 using a finite sample of historical observations, 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 et al. (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. The extra terms are functions of the estimation scheme, i.e. the choice of fixed, rolling, expanding window

[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 Value-at-Risk, see, for example, Kole et al. (2017).

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.

Our analysis of the effect of estimation error on joint backtests of VaR and ES complements earlier analyses for backtests of VaR and of ES in isolation. Escanciano and Olmo (2010) examine the effect of estimation error on VaR backtests in a similar way as we do, though their approach cannot deal with multi-period ahead forecasts. Du and Escanciano (2016) propose new conditional tests for ES, 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%. The size experiments are based on a standard AR-GARCH data generating process (DGP) with normal or Student's t distributed errors. In the power experiments, we consider deviations in terms of the specification of the mean, the volatility or the error distribution.

Our results show that estimation error leads to considerable size distortions, in particular of unconditional tests. The empirical rejection rate can become as high 0.36 when using critical values that should theoretically lead to a rate of 0.05. The effect on conditional tests is smaller, although as also pointed out by Nolde and Ziegel (2017) the choice of conditioning variables can lead to tests for which both the standard and robust versions perform badly. Effects of estimation error are largest when the ratio of out-of-sample to in-sample observations is large. 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 are much closer to the theoretical 0.05. The largest rejection rate we observe is now 0.15. The robust versions work well for both coverage levels, and only slightly worse when the

errors stem from the fat-tailed Student's t distribution instead of the normal one. We do not observe 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.

We find that the robust versions of the tests have less power than the standard tests, but that the reduction is limited. The differences between the rejection rates of the two versions are not larger than 0.10–0.15. The empirical rejection rates of both versions generally increase when the DGP deviates more from the model used to obtain the risk forecasts. 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 an empirical application we evaluate VaR and ES forecasts for the FTSE 100 index returns, generated by AR-GARCH, AR-GJR-GARCH and AR-HEAVY models estimated on rolling windows of 500, 1000 and, if possible, 2,500 observations. We conduct backtests for the financial crisis period (June 2007 to June 2009) and a longer period (November 2009 to April 2019). We generally observe that estimation error has a non-negligible effect on the conclusions that would likely be drawn based on the test outcome, in the sense that p-values often increase from values below 0.05 or 0.10 to values well above these levels. This effect is largest when the forecasts of the two GARCH models are evaluated over the longer period, and the correction for estimation error leads to non-rejection of the hypothesis of correct forecasts. When the crisis period is considered, increases in p-values are smaller, providing stronger evidence of misspecification. In line with our simulation results, the effect of estimation error becomes smaller for larger in-sample periods, keeping the out-of-sample period fixed. The effect of estimation error is also smaller for the AR-HEAVY model, which may due to reduced estimation error related to the higher precision of the realized volatility measure that this model uses.

Based on the results of our simulation study and empirical analysis, we conclude that backtests of VaR and ES should not ignore the effect of estimation error. Estimation error leads to size distortions. We complement Escanciano and Olmo (2010) who analyze and propose corrections for the effect of estimation error on VaR backtests, and Du and Escanciano (2016) who do the same for ES in isolation. Because we use the framework of McCracken (2000), our results

apply to single and multi-period ahead forecasts of ES. 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-period 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, for example Gneiting, 2011; Acerbi and Székely, 2014) has only recently been solved by the work of Fissler et al. (2016); 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 2, and the test specifications in Section 3. We set up and study the results of our Monte Carlo experiments in Section 4, and an empirical application in Section 5. Section 6 concludes.

2 Theory

Let $Y_{t+\tau}$ denote the return generated by holding an asset from period t to $t+\tau$, $\tau \geq 1$. Let $W_t = (Y_t, Z_t', Y_{t-1}, Z_{t-1}', \ldots)'$ denote the agent's information set at time t, with Z_t denoting a vector of other relevant variables. Let $\mathcal{F}_t = \sigma(W_t)$ denote the σ-algebra generated by W_t, and define $E_t[\cdot] = E[\cdot|W_t]$.

Now (implicitly) define the $\mathcal{F}_t$-measurable, τ-step ahead Value-at-Risk (VaR), denoted $\mathrm{VaR}_{t,\tau}$ and Expected Shortfall (ES), denoted $\mathrm{ES}_{t,\tau}$, at coverage level $1-\alpha \in (0,1)$ as

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

and

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

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

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

We consider a model of VaR and ES given by the parametric family of functions $\mathcal{M} = \{m(\cdot,\theta) : \theta \in \Theta \subset \mathbb{R}^p\}$, with $p < \infty$, 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, respectively, and write $m_t(\theta) = (m_{t,1}(\theta), m_{t,2}(\theta))' = (m_1(W_t,\theta), m_2(W_t,\theta))'$. Throughout we ignore dependence on α for notational convenience.

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

$$m_t(\theta_0) = \big(m_{t,1}(\theta_0), m_{t,1}(\theta_0)\big)' = \big(\mathrm{VaR}_{t,\tau}, \mathrm{ES}_{t,\tau}\big)' \text{ a.s., for all } t. \tag{3}$$

We set up our testing framework to fit the out-of-sample testing theory in McCracken (2000) and base the tests on the following 2×1 *identification function* proposed in Nolde and Ziegel (2017):

$$g_{t,\tau}(\theta) = \begin{bmatrix} g_{t,1,\tau}(\theta) \\ g_{t,2,\tau}(\theta) \end{bmatrix} = \begin{bmatrix} \mathbb{1}\{Y_{t+\tau} - m_{t,1}(\theta) < 0\} - \alpha \\ m_{t,2}(\theta) - m_{t,1}(\theta) - \frac{1}{\alpha}\mathbb{1}\{Y_{t+\tau} - m_{t,1}(\theta) < 0\}\big(Y_{t+\tau} - m_{t,1}(\theta)\big) \end{bmatrix}, \tag{4}$$

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 return. The first element forms the basis for the coverage tests of Kupiec (1995) and Christoffersen (1998). Nolde and Ziegel (2017) also show that $E_t[g_{t,\tau}(\theta)]$ 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 et al. (2016). Barendse (2017) proposes a joint semi-parametric estimator of expected shortfall using Eq. (4).

It is easy to see that correct model specification (Eq. (3)) and unique quantiles together imply that $E_t[g_{t,\tau}(\theta)]$ is uniquely equal to zero at $\theta = \theta_0$. A test of correct model specification can therefore be based on the following null hypothesis:

$$\mathcal{H}_0 : E_t[g_{t,\tau}(\theta_0)] = 0, \text{ a.s. for all } t. \tag{5}$$

Under this null hypothesis $\{g_{t,\tau}(\theta_0), \mathcal{F}_t\}$ is a martingale difference sequence, such that we

can employ the equivalence statement

$$E_t[g_{t,\tau}(\theta_0)] = 0, \text{ a.s. } \iff E[g_{t,\tau}(\theta_0)\tilde{h}_t] = 0, \tag{6}$$

for all $\mathcal{F}_t$-measurable functions $\tilde{h}_t$ and for all t (see, e.g. Giacomini and White (2006)). Like Nolde and Ziegel (2017) we restrict our attention to a subset of these functions, namely those that are differentiable (a.s.) in a neighbourhood Θ_0 of θ_0. More specifically, we employ the $\mathcal{F}_t$-measurable, $l \times 2$ *test matrix* $H_t(\theta_0)$, with elements referred to as $H_{t,i,j}(\theta_0)$.

The smoothness assumption on H_t precludes usage of lags of (elements of) $g_{t,\tau}(\theta_0)$ as conditioning variables. This may ich seem restrictive because $g_{t-1,1,\tau}(\theta_0)$ is commonly used in conditional VaR tests, (e.g. Christoffersen, 1998; Escanciano and Olmo, 2010). However, Nolde and Ziegel (2017) show convincingly that conditioning on lags in the joint (VaR,ES) test leads to bad size properties, and advocate using smoother conditioning variables. We find similarly bad size properties in simulation experiments during earlier stages of this research. Based on these results, we argue that our framework is sufficiently general, even though we restrict $H_t(\theta_0)$ to be smooth.

We can test the following null hypothesis based on the $l \times 1$ vector $k_{t,\tau}(\theta_0) = H_t(\theta_0)g_t(\theta_0)$:

$$\mathcal{H}_{0,h} : E[k_{t,\tau}(\theta_0)] = 0 \text{ for all } t, \tag{7}$$

and Eq. (6) implies that a test of $\mathcal{H}_{0,h}$ provides a test of $\mathcal{H}_0$. Under suitable regularity conditions, at the true parameter vector θ_0,

$$S^0(R,P) \equiv S_P^0 = \frac{1}{\sqrt{P}} \sum_{t=R}^{T-\tau} k_{t,\tau}(\theta_0), \tag{8}$$

converges to a multivariate normal random variable with zero mean and $l \times l$ asymptotic covariance matrix

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

and testing can proceed using a standard Wald test. Our Assumptions 2, 3, and 6(a,b,d) are sufficient, for instance. The additional terms in Σ are due to overlapping data, implying that only for $j \geq \tau$ we have $E[k_{t-j,\tau}(\theta_0)k_{t,\tau}(\theta_0)'] = E\big[k_{t-j,\tau}(\theta_0) \times E_t[k_{t,\tau}(\theta_0)']\big] = 0$ under $\mathcal{H}_0$.

Usually we do not know θ_0, and must estimate it instead. Let us consider a sample $\{Y_t, Z_t'\}_{t=1}^{T}$, $T \geq 1$. 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 $\hat{\theta}_t$ of θ_0 at time t. The fixed scheme uses the first R observations for $\hat{\theta}_t$ for all t, i.e. the observations at times $1, \ldots, R$. The rolling scheme uses the R most recently observed observations, i.e. the observations at times $t - R + 1, \ldots, t$. The recursive scheme uses all observations up to time t, i.e. the observations at times $1, \ldots, t$.

Returning to the testing problem, since the true parameter θ_0 is generally unknown we must instead of Eq. (8) consider

$$S(R, P) \equiv S_P = \frac{1}{\sqrt{P}} \sum_{t=R}^{T-\tau} k_{t,\tau}(\hat{\theta}_t). \tag{9}$$

To obtain the 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). This expansion, which in the fixed window case is given by $\hat{\theta}_t - \theta_0 = B(t)t^{-1} \sum_{s=1}^{t} l_s(\theta_0) + o_p(t^{-1/2})$, $B(t) \xrightarrow{a.s.} B$, depends on estimator-specific definitions of $p \times q$ matrices $B(t)$ and B, and the $q \times 1$ vector $l_t(\theta)$, and is defined fully in Assumption 1 in the Appendix. This approach is standard in the analysis of estimation effects on statistical testing.

Under appropriate conditions we can then show that S_P converges to a multivariate normal distribution with zero mean and $l \times l$ asymptotic covariance matrix

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

which has the particular structure of Ω as dictated by the theory of West (1996), McCracken (2000), Escanciano and Olmo (2010), and where the different elements in the definition of Ω are defined as the $q \times l$ matrix

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

$q \times q$ matrix $V = E[l_t(\theta_0) l_t(\theta_0)']$, and $l \times p$ matrix

$$A = A_t \equiv \nabla E[g_{t,\tau}] \equiv E\left(H_t(\theta_0) \begin{bmatrix} f_{t,\tau}(\theta_0) & 0 \\ 0 & 1 \end{bmatrix} J_t(\theta_0) \right), \tag{10}$$

with $J_t(\theta_0) = \nabla_\theta m_t(\theta)|_{\theta=\theta_0}$ the (a.s.) continuous Jacobian matrix of $m_t(\theta)$ in a neighborhood of θ_0. The matrix B is used in the expansion discussed above and defined formally in Assumption 1 in the Appendix. Finally, let $\pi = \lim_{T\to\infty} P/R$, and $\lambda_{hl}(\pi)$ and $\lambda_{ll}(\pi)$ be defined as:

$$\begin{array}{lll} & \lambda_{hl}(\pi) & \lambda_{ll}(\pi) \\ \textit{fixed} : & 0 & \pi \\ \textit{rolling } (\pi \leq 1): & \pi/2 & \pi - \pi^2/3 \\ \textit{rolling } (\pi > 1): & 1 - (2\pi)^{-1} & 1 - (3\pi)^{-1} \\ \textit{recursive} : & 1 - \pi^{-1}\log(1+\pi) & 2\big[1 - \pi^{-1}\log(1+\pi)\big]. \end{array} \tag{11}$$

The following theorem formalizes the convergence in distribution of S_P to a normal limit under the assumptions provided in Appendix A.

Theorem 1. *Let Assumptions 1 to 6 in the Appendix be satisfied. It follows under $\mathcal{H}_{0,h}$ that $S_P \xrightarrow{d} N(0, \Omega)$. Moreover, if $\pi = 0$, then $S_P \xrightarrow{d} N(0, \Sigma)$.*

Proof: see Appendix B.

It is clear that using estimated parameters introduces additional terms in the asymptotic covariance matrix of S_P in case the in-sample size P grows proportionally with the out-of-sample size R.

We can obtain a consistent estimator of the asymptotic covariance matrix Ω as

$$\hat{\Omega}_P = \hat{\Sigma}_P + \hat{\lambda}_{hl}(\hat{\pi})(\hat{A}_P \hat{B}_P \hat{\rho}_P + \hat{\rho}_P' \hat{B}_P' \hat{A}_P') + \hat{\lambda}_{ll}(\hat{\pi}) \hat{A}_P \hat{B}_P \hat{V}_P \hat{B}_P' \hat{A}_P'. \tag{12}$$

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

$$\begin{aligned}\hat{\Sigma}_P =& \frac{1}{P}\sum_{t=R}^{T-\tau} k_{t,\tau}(\hat{\theta}_t)k_{t,\tau}(\hat{\theta}_t)' + \\ & \frac{1}{P}\sum_{j=1}^{\tau-1}\sum_{t=R+j}^{T-\tau}\left(k_{t-j,\tau}(\hat{\theta}_t)k_{t,\tau}(\hat{\theta}_t)' + k_{t,\tau}(\hat{\theta}_t)k_{t-j,\tau}(\hat{\theta}_t)'\right),\end{aligned} \tag{13}$$

$$\hat{A}_P = \frac{1}{P}\sum_{t=R}^{T-\tau}\left(H_t(\hat{\theta}_t)\begin{bmatrix}(2\hat{c}_P)^{-1}\mathbb{1}(|Y_{t+\tau} - m_{t,1}(\hat{\theta}_t)| < c_P) & 0 \\ 0 & 1\end{bmatrix}J_t(\hat{\theta}_t)\right), \tag{14}$$

$$\hat{\rho}_P = \frac{1}{P}\sum_{j=1}^{\tau}\sum_{t=R}^{T-\tau+1}[l_{t-j}(\hat{\theta}_t)k_{t,\tau}(\hat{\theta}_t)'], \tag{15}$$

$$\hat{V}_P = \frac{1}{P}\sum_{t=R}^{T-\tau+1} l_t(\hat{\theta}_t)l_t(\hat{\theta}_t)', \tag{16}$$

$$\hat{\pi}_P = P/R. \tag{17}$$

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

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

Corollary 1. *Let Assumptions 1 to 6 in the Appendix be satisfied, let* $\hat{B}_P \xrightarrow{a.s.} B$*, and let* $\hat{c}_P/c_P \xrightarrow{p} 1$*, where the nonstochastic* c_P *satisfies* $c_P = o(1)$ *and* $c_P^{-1} = o(P^{1/2})$*. Under Eq.* (5) *it follows that* $\hat{\Omega}_P \xrightarrow{p} \Omega$*, and* $\hat{\Omega}_P^{-1/2}S_P \xrightarrow{d} N(0, I)$*.*

Proof: see Appendix B.

3 Tests

We can now construct tests using the statistic $T_P = S_P'\hat{\Omega}_P^{-1}S_P$, 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 $\hat{\pi}_P$ in $\hat{\Omega}_P$, and to standard versions of the tests when we assume $\pi = 0$, such

that $\hat{\Omega}_P = \hat{\Sigma}_P$. 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). Next, we consider four joint tests for VaR and ES that differ in the specification of $H_t(\theta_0)$. The final three use a different test specification, and are analyzed by Du and Escanciano (2016).

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

$$EO_P^{(1)} : H_t(\theta_0) = \begin{bmatrix} 1 & 0 \end{bmatrix}.$$

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

$$EO_P^{(2)} : H_t(\theta_0) = \begin{bmatrix} g_{t-\tau,1,\tau}(\theta_0) & 0 \end{bmatrix}.$$

It corresponds with the independence test proposed by Christoffersen (1998). Berkowitz et al. (2011) base an alternative conditional test on the same sequence $k_{t,\tau}(\theta_0) = H_t(\theta_0) g_{t,\tau}(\theta_0)$. We use the robust tests proposed in Escanciano and Olmo (2010).[3]

The first two joint test statistics are straightforward generalizations of the two VaR tests. We denote the unconditional test statistic by $T_P^{(1)}$, which uses

$$T_P^{(1)} : H_t(\theta_0) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

The conditional test statistic $T_P^{(2)}$ uses

$$T_P^{(2)} : H_t(\theta_0) = \begin{bmatrix} g_{t-\tau,1,\tau}(\theta_0) & 0 \\ 0 & g_{t-\tau,2,\tau}(\theta_0) \end{bmatrix}.$$

Since $H_t(\theta)$ is not a.s. differentiable over a neighborhood of θ_0 in this specification, it violates Assumption 4, such that we cannot obtain a robust version of this test based on our theoretical framework. We include this test however to show this *intuitive* specification suffers from considerable size distortions. Nolde and Ziegel (2017) also note that using lagged elements of

[3] Alternatively, we could employ the bootstrap-based robust tests of Escanciano and Olmo (2011).

$g_{t,\tau}(\theta_0)$ can result in bad performance in small samples.

The next conditional test statistic, $T_P^{(3)}$, uses

$$T_P^{(3)} : H_t(\theta_0) = \begin{bmatrix} \sigma_t(\theta_0) & 0 \\ 0 & \sigma_t(\theta_0) \end{bmatrix},$$

with $\sigma_t(\theta_0)$ denoting the conditional volatility of Y_t. We use $\sigma_t(\theta_0)$ as conditioning variable because it is a good proxy for financial risk at time t and because it is smoother than lagged elements of $g_{t,\tau}(\theta_0)$. Assumption 4 may therefore hold for this specification depending on the conditional volatility model under consideration. It holds, for example for the (exponentially weighted) moving average (EWMA) estimator of the conditional variance.

Finally, we consider the conditional test statistic $T_P^{(4)}$ based on

$$T_P^{(4)} : H_t(\theta_0) = \frac{1}{\sigma_{t+\tau}} \begin{bmatrix} \dfrac{m_{2,t}(\theta_0) - m_{1,t}(\theta_0)}{\alpha} & 1 \end{bmatrix}.$$

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 this specification in their application. In practice we only use the information set up to and including time t to estimate $\sigma_{t+\tau}(\theta_0)$.

For comparison, we consider the unconditional and conditional ES backtests of Du and Escanciano (2016). These tests fall outside our framework, because their testing condition cannot be written as in Eq. (7). Let $\hat{D}_t = \frac{1}{\alpha}\mathbb{1}(\alpha \leq \hat{u}_t)(\alpha - \hat{u}_t)$, where $\hat{u}_t = G_{t-1}(Y_t, \hat{\theta}_t)$, with $G_{t-1}(\cdot, \hat{\theta}_t)$ an estimator of the conditional cdf of Y_t. The unconditional ES test of Du and Escanciano (2016) is based on test statistic

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

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 $\hat{D}_t$ sequence $\tilde{\xi}_{P,j} = \tilde{\gamma}_{P,j}/\tilde{\gamma}_{P,0}$ with $\tilde{\gamma}_{P,j} = (P-j)^{-1} \sum_{t=R+j}^{T-\tau+1} (\hat{D}_t - \alpha/2)(\hat{D}_{t-j} - \alpha/2)$. 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 Escan-

ciano (2016). Both test statistics converge in distribution to a χ^2_1-distributed random variable. Du and Escanciano (2016) also consider conditional tests that include autocorrelations $\tilde{\xi}_{P,j}$ with more distant lags, e.g. $j = 5$. We therefore also consider

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

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

4 Simulation study

4.1 Design

We investigate the finite sample performance of the VaR and ES tests by means of Monte Carlo experiments. We report empirical rejection rates in 1,000 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 2500. 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 one-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 used.

Our simulation setup is similar to Du and Escanciano (2016). In all experiments, we consider an AR(1)-GARCH(1,1) null model for Y_t, (see Bollerslev, 1987) given by,

$$\begin{aligned} Y_t &= -a_0 Y_{t-1} + v_t, \\ v_t &= \sigma_t \varepsilon_t, \\ \sigma_t^2 &= \omega_0 + \alpha_0 v_{t-1}^2 + \beta_0 \sigma_{t-1}^2, \end{aligned}$$

such that our risk measure forecasts are equal to

$$\begin{aligned} \mathrm{VaR}_t(\alpha) &= -a_0 Y_{t-1} - \sigma_t F^{-1}(\alpha), \\ \mathrm{ES}_t(\alpha) &= -a_0 Y_{t-1} - \sigma_t E[\varepsilon_t | \varepsilon_t \leq F^{-1}(\alpha)] \end{aligned}$$

where ε_t follows a standardized distribution, of which $F^{-1}(\alpha)$ denotes its α-quantile. In particular, we consider the normal and the student's t distribution with $\nu = 5$ degrees of freedom.[4]

In the simulation study we focus on fixed window estimation, because rolling or expanding window designs are too computationally costly. In each simulation we estimate θ_0 by the one-stage MLE estimator $\hat{\theta}_R$ over the first R observations.

We need several quantities to estimate the estimation error effects. Given the continuous differentiability of $m_t(\theta)$ we use numerical approximation to find $J_t(\hat{\theta}_R)$ for each $t = R+1, \ldots, T$. Moreover, in our setting $l_t(\hat{\theta}_R) = \partial \log[f\left((Y_t - \hat{a}_{t-1}Y_{t-1})/\sigma_t(\hat{\theta}_R)\right) - \sigma_t(\hat{\theta}_R)]/\partial\theta$, with f denoting the pdf of ε_t, and $\hat{B}_P$ denotes a strongly consistent estimator of the asymptotic covariance matrix of $\sqrt{R}(\hat{\theta}_R - \theta_0)$, for any $t = R+1, \ldots, T$, which can be obtained as the negative of the inverted Hessian matrix, i.e., $\hat{B}_P = \Big[-\frac{1}{R}\sum_{t=1}^{R} \partial^2 \log[f_{\hat{\nu}_{t-1}}\left((Y_t - \hat{a}_{t-1}Y_{t-1})/\sigma_t(\hat{\theta}_R)\right) - \sigma_t(\hat{\theta}_R)]/(\partial\theta\partial\theta') \Big]^{-1}$. We employ numerical approximation to obtain them.

For the power analysis we consider three data generating processes.

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

$$
\begin{aligned}
Y_t &= 0.05Y_{t-1} + v_t, \\
v_t &= \sigma_t \varepsilon_t \\
\sigma_t^2 &= 0.05 + 0.1v_{t-1}^2 + 0.85\sigma_{t-1}^2, \\
\varepsilon_t &\sim \left[\frac{1}{1+q}N(0,q) + \frac{q}{1+q}N(0,1/q)\right], \quad q = 1 + \frac{3}{2}c;
\end{aligned}
$$

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

$$
\begin{aligned}
Y_t &= 0.05Y_{t-1} + v_t, \\
v_t &= \sigma_t \varepsilon_t \\
\sigma_t^2 &= 0.05 + (0.1 - 0.1c)v_{t-1}^2 + 0.2c\mathbb{1}(\varepsilon_{t-1} < 0)v_{t-1}^2 + 0.85\sigma_{t-1}^2, \\
\varepsilon_t &\sim t_5;
\end{aligned}
$$

[4] Specifically, $E[\varepsilon_t | \varepsilon_t \leq F^{-1}(\alpha)] = \phi(\Phi^{-1}(\alpha))/\alpha$ when $\varepsilon_t \sim N(0,1)$ with ϕ and Φ denoting its pdf and cdf. When ε_t follows a standardized Student's t distribution with ν degrees of freedom, $E[\varepsilon_t | \varepsilon_t \leq F^{-1}(\alpha)] = \sqrt{\frac{\nu-2}{\nu}}\frac{\nu + \left(G_\nu^{-1}(\alpha)\right)^2}{\nu-1}\frac{g_\nu\left(G_\nu^{-1}(\alpha)\right)}{\alpha}$, with $G_\nu^{-1}(\alpha)$ the α-quantile of the standard t-distribution with ν degrees of freedom, and $g_\nu(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi\nu}}\left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$.

- A_3. AR(1)-GARCH(1,1)-in-Mean:

$$
\begin{aligned}
Y_t &= 0.05Y_{t-1} + 2.5c(\sigma_{t-1}^2 - 1) + v_t, \\
v_t &= \sigma_t \varepsilon_t, \\
\sigma_t^2 &= 0.05 + 0.1v_{t-1}^2 + 0.85\sigma_{t-1}^2 \\
\varepsilon_t &\sim t_5.
\end{aligned}
$$

Similar DGPs have been studied in Du and Escanciano (2016) and Escanciano and Olmo (2010). In the simulation study we let the hyperparameter c vary over $[0, 1]$. When $c = 0$ the DGP actually corresponds with the AR(1)-GARCH model used to obtain the VaR and ES forecasts, whereas larger values of c indicate larger deviations from the null model. In specification A_1 the error distribution differs from the null model. As c increases 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 et al. (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, e.g., Christensen et al. (2012). We subtract the unconditional value of the variance, which is equal to 1, from σ_{t-1}^2 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 we cover the most important types of misspecification from the null model.

4.2 Results

Table 1 provides size properties of the tests for the null model with standard normal errors and the different combinations of R and P. The standard versions of the tests generally lead to rejection rates that exceed the nominal size of 5%. This result corresponds with findings in Escanciano and Olmo (2010) and Du and Escanciano (2016). The rejection rates for the

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

conditional VaR test $EO_P^{(2)}$ and conditional ES test $DE_P^{(2)}$ are only marginally larger than 5%. They vary between 0.07 and 0.31 for the VaR test $EO_P^{(1)}$, the joint VaR and ES tests $T_P^{(1)}$ and $T_P^{(4)}$, and the ES tests $DE_P^{(1)}$ and $DE_P^{(3)}$. The test $T_P^{(2)}$ scores particularly bad, in line with the findings of Nolde and Ziegel (2017). Test $T_P^{(3)}$ also shows bad size properties, indicating that the conditional volatility may not be a good instrument in testing.

[Table 1 about here.]

The robust tests that we propose have better size properties, with rejection rates that are closer to 5%, in particular for $EO_P^{(1)}$, $T_P^{(1)}$, $T_P^{(4)}$, $DE_P^{(1)}$ and $DE_P^{(3)}$. For these tests, rejection rates are now all below 0.12, with many rates around 0.07, whereas we find rates as high as 0.31 and many around 0.15 for the standard versions. We cannot include a robust version of $T_P^{(2)}$, because our framework does not allow for non-smooth H_t. The robust version of $T_P^{(3)}$ still suffers from substantial size distortions. It means that the robust versions of the tests are also sensitive to the choice of conditioning variables, as the robust version of $T_P^{(4)}$ has good size properties.

Comparing the different combinations R/P of in-sample and out-of-sample window gives some further insights in the performance of the tests. The difference in rejection rates of the standard and robust tests show that the impact of estimation error is largest for $R, P = (500, 2500)$, as suggested by the theory. Rejection rates are closest to 5% when both the in-sample and ouf-of-sample windows are long. The robust tests still show large distortions for the combination of $R, P = (2500, 500)$, which indicates 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.

Table 2 shows that the using the Student's distribution with 5 degrees of freedom for the errors leads to somewhat different results. The size properties of the standard VaR-tests are a bit better. The joint tests for VaR and ES perform a bit worse, in particular for the $(R, P) = (500, 2500)$ combination. The changes in rejections rates for the DE-tests are mixed. However, differences are small. The size distortions are again a lot smaller when the robust versions of the tests are used. However, the size distortions for the $T_P^{(1)}$ and $T_P^{(4)}$ tests are larger than when the error distribution is normal. The robust versions of the DE-tests perform equally well for the normal as the Student's t distribution for the errors. As before $T_P^{(3)}$ does not work well. The conditional ES and VaR tests are less affected by fat-tailedness of the standardized errors.

Given that we only occasionally find standardized errors to be t-distributed with as low as five degrees of freedom, these results suggest that the correction for estimation errors should work fairly well in practice for the joint (VaR,ES) tests $T_P^{(1)}$ and $T_P^{(4)}$.

[Table 2 about here.]

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, five-point grid of $[0, 1]$. As c increases the deviation from the null model increases as well. 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 $P = 2,500$. Our choice for $R = 2,500$ 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, 2,500 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 the robust versions we propose. We include $DE_P^{(1)}$ and $DE_P^{(3)}$ as benchmarks.

We present the power curves of the different tests for the different DGPs in Fig. 4. As can be expected, the robust versions of the test have less power than the standard versions, but the difference is generally limited with rejection rates of robust tests being 0.10-0.15 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 cases of misspecification that we consider.

[Figure 1 about here.]

The power curves in Fig. 1a show that both the standard and robust versions of $T_P^{(1)}$ and $T_P^{(4)}$ detect the mixed-normal alternative A_1 well. This result holds in paricular for values of $c > 0.5$, which corresponds with kurtosis exceeding $27/28 + 3 \approx 4$. The rejection rates of both versions of $DE_P^{(1)}$ increase only slightly for $c > 0.5$, but stay below 0.2, whereas the rejection rates of $DE_P^{(3)}$ do not seem to increase at all.

The results for the AR-GJR-GARCH alternative A_2 in Fig. 1b show that the only test with some power against it is $DE_P^{(3)}$, both in the standard and robust version. However, for a value of $c = 1$, in which only negative innovations affect the volatility in the next period, the rejection

rates do not exceed 0.35 (standard) and 0.30 (robust). Apparently, the misspecification of the volatility equation in the null model leads to autocorrelation in the scaled ES distances on which test $DE_P^{(3)}$ is based. The unconditional tests $T_P^{(1)}$ and $DE_P^{(1)}$ do not have much power against this specification, and neither has test $T_P^{(4)}$. These results are in line with Du and Escanciano (2016), who note that the conditional tests have more power against AR-ARCH(2) and AR-EGARCH alternatives, which also only differ in terms of the volatility equation.

The analysis of the GARCH-in-Mean alternative A_3 in Fig. 1c shows that predominantly $T_P^{(1)}$ has some power in detecting the conditional mean misspecification. The rejection rate of the standard version rises from 0.13 at $c = 0$ (so it is oversized) to 0.47, and the robust version shows an increase from 0.07 to 0.3. Perhaps surprisingly, the conditional tests $DE_P^{(3)}$ and $T_P^{(4)}$ are not able to detect the misspecification in the mean equation well. Under this specification, an increase in the variance at $t-1$ leads to a wider density that is shifted to the right. In the left tail, these effects cancel to some extent, which may explain the low power of these tests. The rejection rates of both versions of $DE_P^{(1)}$ also do not rise beyond 0.2. Our result that $DE_P^{(1)}$ and $DE_P^{(3)}$ do not have much power against the GARCH-in-Mean alternative A_3 differs from the findings in Du and Escanciano (2016). This might be caused by our parameterization that keeps the unconditional mean at zero, whereas the previous authors do not impose this restriction. In separate simulations we indeed find that without imposing this restriction the tests do have some power.

We conclude that the robust versions of the tests perform quite well compared to the standard versions, and that their better size properties come at the expense of only limited reductions in power. 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 $DE_P^{(1)}$ does not seem to detect any of the misspecifications. Our results for the other combinations of R and P in Appendix D confirm these conclusions. Both test versions have less power for out-of-sample window size $P = 500$. Conform the theoretical result that the effect of estimation error is larger when the in-sample window R is shorter, the difference in power is larger for the combination $R/P = 500/2500$.

5 Empirical analysis

In our empirical application we evaluate VaR and ES forecasts for daily returns on the FTSE 100 index as produced by three different models: AR-GARCH, AR-GJR-GARCH, and AR-HEAVY. All three models make use of an AR(1) specification for the conditional mean,

$$Y_t = a_0 Y_{t-1} + v_t \tag{18}$$

$$v_t = \sigma_t \varepsilon_t, \tag{19}$$

where the innovations ε_t follow a standardized Student's t distribution with degrees of freedom parameter ν which is also estimated. The models differ in the specification of the conditional volatility σ_t. For the AR-GJR-GARCH model, this is given by

$$\sigma_t^2 = \omega_0 + \alpha_0 v_{t-1}^2 + \gamma_0 v_{t-1}^2 \mathbb{1}\{v_{t-1} < 0\} + \beta_0 \sigma_{t-1}^2. \tag{20}$$

When $\gamma_0 = 0$ is imposed, the AR-GARCH model results. The HEAVY model is a GARCH-type model that incorporates high-frequency estimates for the volatility in the conditional volatility specification,

$$\sigma_t^2 = \omega_0 + \delta_0 RM_{t-1} + \beta_0 \sigma_{t-1}^2, \tag{21}$$

where RM_{t-1} is the realized measure calculated for the previous period. We follow Shephard and Sheppard (2010) and use the realized kernel of Barndorff-Nielsen et al. (2008).

We consider two distinct sample periods. The first sample is defined by its out-of-sample period that runs from July 30, 2007 to July, 30, 2009, so it contains the financial crisis. We use 500 and 1,000 observations prior to July 30, 2007 as in-sample period (corresponding with starting dates July 7, 2005, and July 11, 2003). The second sample runs from January 5, 2000 to April 17, 2019 (4865 observations). We split the sample at November 8, 2009, creating in- and out-of-sample periods of similar length of approximately 2,500 observations. We refer to it as the long sample.

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). Table 3 contains descriptive statistics for the crisis and the long samples, which shows that they differ substantially. The crisis sample has negative mean and median returns, higher volatility, and different skewness

and kurtosis properties.

[Table 3 about here.]

We estimate the models by maximum likelihood using rolling windows of different sizes and construct one-period-ahead forecasts, i.e. $\tau = 1$. Table 4 provides summary statistics of the parameter estimates for the AR-GARCH, AR-GJR-GARCH, and AR-HEAVY models using a rolling window of 1,000 observations. The first rolling window ends at November 7, 2009, the last at April 16, 2019. We notice that there is substantial variation in the parameter estimates over this period, as shown by the relatively large standard deviation. The small estimates for a_0 indicate weak autocorrelation of the daily returns in all specifications. The other parameters in the AR-GARCH model fluctuate around their typical values. The introduction of the leverage parameter in the AR-GJR-GARCH model shows that negative returns have a larger effect on conditional volatility, as γ_0 is positive. The AR-HEAVY model has smaller β_0 estimates than the GARCH models, consistent with Shephard and Sheppard (2010) who note that β_0 about 0.6 is common in empirical applications. The degrees of freedom parameter ν_0 is generally quite large, suggesting that the FTSE 100 index returns are not very fat-tailed. The correction for estimation effects should therefore work well, as suggested by the simulation study results in Table 1.

[Table 4 about here.]

Panel A of Table 5 shows the sample fraction of daily VaR for the coverage level $1-\alpha = 0.975$,

$$\hat{\alpha} = \frac{1}{P} \sum_{t=R}^{T-1} \mathbb{1}[Y_{t+1} < \widehat{\mathrm{VaR}}_{t,1}], \tag{22}$$

with $\widehat{\mathrm{VaR}}_{t,1} = m_{t,1}(\hat{\theta}_t)$.

All models produce too optimistic VaR forecasts, as all fractions of VaR violations exceed the nominal 2.5%. In particular during the crisis period, the ratio of VaR violations ranges from 4.5 to 8.1%. The shorter estimation window of $R = 500$ leads to better violation frequencies than $R = 1,000$. The AR-GJR-GARCH specification performs worst. Over the long sample period, the violation ratios are still too large, but closer to 2.5%. The simple AR-GARCH model has the lowest ratios, but differences are small. The differences for the different window lengths are also small.

[Table 5 about here.]

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

$$\text{MESE} = \frac{1}{\hat{\alpha}P} \sum_{t=R+1}^{T-1} (Y_t - \widehat{\text{ES}}_{t,1}) \mathbb{1}[Y_t < \widehat{\text{VaR}}_{t,1})], \tag{23}$$

with $\hat{\alpha}$ as in Eq. (22) and $\widehat{\text{ES}}_{t,1} = m_{t,2}(\hat{\theta}_t)$. If a model is correctly specified, its MESE should be close to zero. Positive values for MESE indicate that the ES forecast is too negative. Panel B of Table 5 shows large deviations from zero for both GARCH models, but reasonable performance for the AR-HEAVY model. When the estimation window consists of 500 observations, the mean error is only 0.003% for the AR-HEAVY model, whereas the AR-GARCH model has a mean error of 0.330%. Both AR-GARCH models generally provide too liberal ES forecasts on average. Over the long sample the models perform much better, but without a clear winner. Neither do we see a clear pattern related to the estimation window.

Table 6 shows the results of the standard and robust versions of the tests of Section 3 applied to the forecasts of the AR-GARCH model. We do not apply test $T_P^{(2)}$ because we cannot construct a robust version. For the crisis sample, the standard versions of the tests indicate that the null of correct VaR and ES forecasts should be rejected. Only the conditional VaR test $EO_P^{(2)}$ and ES test $DE_P^{(2)}$ do not reject the null hypothesis. However, this conclusion changes when we account for estimation error, in particular when an in-sample window of $R = 500$ observations is used and unconditional tests are applied. The p-value for $EO_P^{(1)}$ increases from 0.02 to 0.12, for $T_P^{(1)}$ from 0.06 to 0.19 and for $DE_P^{(1)}$ from 0.01 to 0.04. Hence, correct specification of VaR forecasts is no longer rejected when accounting for estimation error. This is different for ES based on tests $T_P^{(3)}$, $T_P^{(4)}$, $DE_P^{(1)}$ and $DE_P^{(3)}$. When $R = 1,000$, p-values increase a bit for the robust tests compared to the standard versions, but here the conclusion remains unchanged.

[Table 6 about here.]

For the long sample period, the quality of the forecasts is better. The deviations from the theoretical values in Table 5 are much smaller than for the crisis period. Still, estimation uncertainty has a large impact on the evaluation of these forecasts. The standard versions of the tests give reasonable evidence to reject correct specification, but the robust versions much less so. For $R = 500$, standard versions of tests $EO_P^{(1)}$, $T_P^{(3)}$, $DE_P^{(1)}$ and $DE_P^{(3)}$ have p-values below 0.05. For $R = 1,000$, this is only the case for $DE_P^{(3)}$, with $EO_P^{(1)}$, $DE_P^{(1)}$ and $DE_P^{(2)}$ below

0.10. For $R = 2,500$, only $T_P^{(3)}$ and $T_P^{(4)}$ lead to p-values above 0.10, $T_P^{(1)}$ has $p = 0.06$ and all others are below 0.05. The robust versions lead again to sizable increases of the p-values, with no tests having p-values below 0.05 when $R = 500$, and only the p-value for $DE_P^{(3)}$ below 0.05 when $R = 1,000$ or $R = 2,500$. When $R = 2,500$ tests $EO_P^{(1)}$, $DE_P^{(1)}$ and $DE_P^{(2)}$ now only provide mild evidence against the null hypothesis.

Table 7 shows the results for the tests of the AR-GJR-GARCH forecasts. For the crisis period, we see a similar picture emerging from the standard version of the tests, as they mostly indicate rejection. However, in this case, estimation uncertainty has only a small effect on the EO_P and T_P tests, but a large effect on $DE_P^{(1)}$ and to a lesser extent $DE_P^{(3)}$, both when $R = 500$ and $R = 1,000$. For the long sample period, the findings also are similar to the AR-GARCH model. The evidence against correct specification becomes weaker when the robust tests are used. With standard tests, we generally reject based on $EO_P^{(1)}$, $T_P^{(1)}$, $DE_P^{(1)}$ and $DE_P^{(3)}$. Moving to robust tests, only the p-value for $EO_P^{(1)}$ when $R = 2,500$ remains below 0.05. The increase in p-values is again particularly large for the unconditional $DE_P^{(1)}$ but not so much for the conditional $DE_P^{(3)}$. Overall, only $EO_P^{(1)}$ and $DE_P^{(3)}$ provide (mild) evidence against correct specification.

[Table 7 about here.]

The results for the AR-HEAVY model also show a milder effect of estimation uncertainty. For the crisis period, the largest increase of a p-value is from 0.05 to 0.16 for $DE_P^{(1)}$. The rejections by $EO_P^{(1)}$, $T_P^{(1)}$, $T_P^{(3)}$, $DE_P^{(1)}$ and $DE_P^{(3)}$ mostly remain when we replace the standard versions by the robust versions. For the longer period, the largest increase in p-value is from 0.3 to 0.43 for $DE_P^{(2)}$. While some of the p-values increase from below typical thresholds of 0.05 or 0.10 to above, the conclusions derived from the tests are less impacted by estimation error in this setting.

[Table 8 about here.]

Overall, we conclude that estimation error has a non-trivial impact on the test outcomes. For many tests, we observe that accounting for estimation error increases p-values to such an extent that the null hypothesis is no longer rejected. This effect occurs for long and short estimation and evaluation windows, and is not restricted to the crisis period. Though estimation uncertainty affects all tests that we consider, the effect seems smallest for the $DE_P^{(3)}$ test, which

is in line with the simulation result that this test has good power against misspecification of the volatility dynamics. Finally, our finding that tests suffer less from estimation uncertainty when forecasts by the AR-HEAVY model are evaluated may be related to the fact that this model uses a realized volatility measure, which is more precise than the squared return used by the two GARCH models.

6 Concluding remarks

In this paper we examine the impact of estimation error on joint backtests for Value-at-Risk (VaR) and Expected Shortfall (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 substantial size distortions due to estimation error, with empirical rejection frequencies exceeding nominal significance levels by a large margin. Robustifying the backtests generally corrects this issue quite successfully, with the caveat that the size properties of conditional tests also depends on the conditioning variables used. We find that the robust tests have somewhat less power than the standard tests, but the reduction is quite modest, not exceeding 10-15%. 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. 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.

References

Acerbi, C. and Székely, B. (2014). Back-testing Expected Shortfall. *Risk*, pages 76–81.

Acerbi, C. and Tasche, D. (2002). On the Coherence of Expected Shortfall. *Journal of Banking & Finance*, 26(7):1487–1503.

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

Artzner, P., Delbaen, F., Eber, J.-M., and Heath, D. (1999). Coherent Measures of Risk. *Mathematical Finance*, 9(3):203–228.

Barendse, S. (2017). Interquantile Expectation Regression. *Working paper*.

Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A., and Shephard, N. (2008). Designing Realized Kernels to Measure the ex post Variation of Equity Prices in the Presence of Noise. *Econometrica*, 76(6):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(4):465–474.

Berkowitz, J., Christoffersen, P., and Pelletier, D. (2011). Evaluating Value-at-Risk Models with Desk-Level Data. *Management Science*, 57(12):2213–2227.

Bollerslev, T. (1987). A Conditionally Heteroskedastic Time Series Model for Speculative Prices and Rates of Return. *The Review of Economics and Statistics*, 69(3):542–547.

Christensen, B. J., Dahl, C. M., and Iglesias, E. M. (2012). Semiparametric Inference in a GARCH-in-Mean Model. *Journal of Econometrics*, 167(2):458–472.

Christoffersen, P. F. (1998). Evaluating Interval Forecasts. *International Economic Review*, 39(4):841–862.

Cont, R., Deguest, R., and Scandolo, G. (2010). Robustness and Sensitivity Analysis of Risk Measurement Procedures. *Quantitative Finance*, 10(6):593–606.

Du, Z. and Escanciano, J. C. (2016). Backtesting Expected Shortfall: Accounting for Tail Risk. *Management Science*, 63(4):940–958.

Emmer, S., Kratz, M., and Tasche, D. (2015). What is the Best Risk Measure in Practice? Comparison of Standard Measures. *Journal of Risk*, 18(2):31–60.

Engle, R. F. and Manganelli, S. (2004). CAViaR: Conditional Autoregressive Value at Risk by Regression Quantiles. *Journal of Business & Economic Statistics*, 22(4):367–381.

Escanciano, J. C. and Olmo, J. (2010). Backtesting Parametric Value-at-Risk With Estimation Risk. *Journal of Business & Economic Statistics*, 28(1):36–51.

Escanciano, J. C. and Olmo, J. (2011). Robust Backtesting Tests for Value-at-Risk Models. *Journal of Financial Econometrics*, 9(1):132–161.

Fissler, T., Ziegel, J. F., et al. (2016). Higher Order Elicitability and Osband's Principle. *The Annals of Statistics*, 44(4):1680–1707.

Giacomini, R. and White, H. (2006). Tests of Conditional Predictive Ability. *Econometrica*, 74(6):1545–1578.

Glosten, L. R., Jagannathan, R., and Runkle, D. E. (1993). On the Relation Between the Expected Value and the Volatility of the Nominal Excess Return on Stocks. *The Journal of Finance*, 48(5):1779–1801.

Gneiting, T. (2011). Making and Evaluating Point Forecasts. *Journal of the American Statistical Association*, 106(494):746–762.

Heber, G., Lunde, A., Shephard, N., and Sheppard, K. (2009). Oxford-Man Institute's Realized Library, Version 0.3.

Kerkhof, J. and Melenberg, B. (2004). Backtesting for Risk-Based Regulatory Capital. *Journal of Banking & Finance*, 28(8):1845–1865.

Kole, E., Markwat, T., Opschoor, A., and van Dijk, D. (2017). Forecasting Value-at-Risk under Temporal and Portfolio Aggregation. *Journal of Financial Econometrics*, 15(4):649–677.

Kupiec, P. H. (1995). Techniques for Verifying the Accuracy of Risk Measurement Models. *The Journal of Derivatives*, 3(2):73–84.

McCracken, M. W. (2000). Robust Out-of-Sample Inference. *Journal of Econometrics*, 99(2):195–223.

McNeil, A. J. and Frey, R. (2000). Estimation of Tail-Related Risk Measures for Heteroscedastic Financial Time Series: an Extreme Value Approach. *Journal of Empirical Finance*, 7(3-4):271–300.

Newey, W. K. and McFadden, D. L. (1994). Large sample estimation and hypothesis testing. In Engle, R. F. and McFadden, D. L., editors, *Handbook of Econometrics*, volume 4, chapter 36, pages 2111–2245. Elsevier.

Nolde, N. and Ziegel, J. F. (2017). Elicitability and Backtesting: Perspectives for Banking Regulation. *The Annals of Applied Statistics*, 11(4):1833–1874.

Patton, A. J., Ziegel, J. F., and Chen, R. (2019). Dynamic Semiparametric Models for Expected Shortfall (and Value-at-Risk). *Journal of Econometrics*, 211(2):388–413.

Powell, J. L. (1986). Censored Regression Quantiles. *Journal of Econometrics*, 32(1):143–155.

Ruppert, D. and Carroll, R. J. (1980). Trimmed Least Squares Estimation in the Linear Model. *Journal of the American Statistical Association*, 75(372):828–838.

Shephard, N. and Sheppard, K. (2010). Realising the Future: Forecasting with High-Frequency-Based Volatility (HEAVY) Models. *Journal of Applied Econometrics*, 25(2):197–231.

West, K. D. (1996). Asymptotic Inference about Predictive Ability. *Econometrica*, 64(5):1067–1084.

West, K. D. (2006). Forecast evaluation. In Elliott, G., Granger, C., and Timmermann, A., editors, *Handbook of Economic Forecasting*, volume 1, chapter 3, pages 99–134. Elsevier.

West, K. D. and McCracken, M. W. (1998). Regression-Based Tests of Predictive Ability. *International Economic Review*, 39(4):817–840.

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

Wong, W. K. (2008). Backtesting Trading Risk of Commercial Banks Using Expected Shortfall. *Journal of Banking & Finance*, 32(7):1404–1415.

Wong, W. K. (2010). Backtesting Value-at-Risk Based on Tail Losses. *Journal of Empirical Finance*, 17(3):526–538.

A Assumptions

Before we state the assumptions we need to introduce some additional notation. Let $F_{t,\tau}(y) = P(Y_{t+\tau} < y|W_t)$ denote the conditional distributions, for some $\tau \geq 1$, and let $f_{t,\tau}(y)$ denote the associated densities. 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 $\sup_{R\leq t\leq T}$. Finally, define

$$v_t(\theta) = [(k_{t,\tau}(\theta) - Ek_{t,\tau}(\theta))', l_t(\theta)]'. \tag{24}$$

Assumption 1. *The estimate $\hat{\theta}_t$ satisfies the expansion $\hat{\theta}_t - \theta_0 = B(t)M(t) + o_P(t^{-1/2})$, 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) $M(t) = t^{-1}\sum_{s=1}^{t} l_s(\theta_0)$, $M(t) = R^{-1}\sum_{s=t-R+1}^{t} l_s(\theta_0)$, and $M(t) = R^{-1}\sum_{s=1}^{R} l_s(\theta_0)$, for the recursive, rolling, and fixed forecasting schemes, respectively, and (c) $E_{t-1}\left[l_t(\theta_0)\right] = 0$, almost surely, for all $t = R,\ldots,T$.*

Many estimators in the literature satisfy Assumption 1, including maximum likelihood estimators and a range of GMM estimators. For instance, if $\hat{\theta}_t$ is the OLS estimator of a regression of Y_t on Z_t in a fixed forecasting scheme, then

$$\hat{\theta}_t - \theta_0 = B(t)\frac{1}{R}\sum_{s=1}^{R} Z_s Y_s,$$

with $B(t) = \left(\frac{1}{R}\sum_{t=1}^{R} Z_t Z_t'\right)^{-1} \xrightarrow{a.s} \left(E[Z_t Z_t']\right)^{-1} = B$.

Chapter 3 of Newey and McFadden (1994) discusses similar expansions for M-estimators and GMM estimators. For some estimators such an exact expansion does not exist. However, it can often be shown that the expansion is accurate up to a $o_P(t^{-1/2})$ term. This is the case for many non-smooth estimators. For, instance, the quantile regression estimator admit such an approximate expansion (see, e.g. Theorem 4 of Ruppert and Carroll (1980)). Chapter 7 of Newey and McFadden (1994) provides an elaboration on such approximate expansions for non-smooth M-estimators and GMM estimators.

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.

Assumption 3. *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) Ω is positive definite.*

Assumption 3 is equivalent to Assumption 3 in McCracken (2000). Covariance stationarity is primarily assumed for simplification of the algebra in the proofs establishing the consistency of estimators of A, V, and ρ.

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

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

Assumption 4 imposes (a.s.) differentiability on $H_t(\theta)$, $m_t(\theta)$, and $E[l_t(\theta)]$ on some neighbourhood Θ of θ_0, and a Lipschitz condition on $l_t(\theta)$, 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, let $f_{t,\tau}(\cdot)$ be uniformly bounded, i.e. $\sup_t \sup_{y \in \mathbb{R}} f_{t,\tau}(y) < C_f < \infty$, and $|f_{t,\tau}(y) - f_{t,\tau}(y')| \leq L|y - y'|$, for all $y, y' \in \mathbb{R}$ and each t, and some constant $L < \infty$.*

Assumption 5 imposes conditions on the conditional distribution of $Y_{t+\tau}$ 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) $\sup_t \| \sup_{\theta \in \Theta_0} |\partial H_{i,j,k,t}(\theta)| \|_{c_H Q} < C$, for all $i = 1, \ldots, l$, $j = 1, 2$, and $k = 1, \ldots, p$; (f) $\sup_t \| \sup_{\theta \in \Theta_0} |l_t(\theta)| \|_Q < C$; and (g) $\sup_t \sup_{\theta \in \Theta_0} |\nabla E[l_t(\theta)]| < C$.*

Assumption 6 imposes moment conditions on the relevant quantities. Notice that if $|H_t(\theta)|$ and $|\partial H_{i,j,k,t}(\theta)|$ are 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 $Q = 2r$.

B Proofs

B.1 Proof of Theorem 1

The result follows by application of Theorem 2.3.1 in McCracken (2000). We need to establish that our framework satisfies the five assumptions required for Theorem 2.3.1, which we denote MC1-MC5. MC3-MC5 are included in section C for completeness.

Conditions (a) and (b) of Assumption 1 are identical to conditions (a) and (b) of Assumption MC1. Meeting MC1(c) follows from Assumption 1(c) by the Law of Iterated Expectation. Finally, notice that we impose the equality $\hat{\theta}_t - \theta_0 = B(t)M(t) + o_P(t^{-1/2})$, whereas McCracken (2000) impose $\hat{\theta}_t - \theta_0 = B(t)M(t)$. This does not change the theory, as it only results in additional $o_P(1)$ term in the proofs of McCracken (2000)'s Lemma A.1 and Lemma 2.3.2.

Assumption 2 is identical to Assumption MC2.

Conditions (a), (b), and (d) of Assumption MC3 are imposed as conditions (a), (b), and (c) of Assumption 3(a,b,c). We will establish that MC3(c), MC4, and MC5 hold in the following.

MC3(c):

We establish MC3(c). Notice that

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

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

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

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

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 $\sup_t \| \sup_{\theta\in\Theta_0} H_{t,i,j}(\theta)\|_{2c_H r} < C$, $\|Y_t\|_{2c_g r} < C$, $\sup_t \| \sup_{\theta\in\Theta_0} |m_t(\theta)|\|_{2c_g r} < C$, and $\sup_t \| \sup_{\theta\in\Theta_0} l_{j,t}(\theta)\|_{2r} < C$. The result follows.

MC4:

We establish Assumption MC4. 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 $E[l_{i,t}(\theta)] = E[l_{i,t}(\theta_0)] + \big(\partial E[l_{i,t}(\tilde{\theta})]/\partial\theta\big)(\theta - \theta_0)$, for some $\tilde{\theta}$ between θ and θ_0 (elementwise). Additionally, under Assumption 6 we have $\sup_t \sup_{\theta\in\Theta_0} |\partial E[l_{i,t}(\theta)]/\partial\theta| < C$, and $G = G_t$.

Now notice

$$E[k_{t,i,\tau}(\theta)] = E\big[H_{t,i,j}(\theta) E_t[g_{t,j,\tau}(\theta)]\big].$$

Under Assumption 4 $H_{i,j,t-1}(\theta)$ is (a.s) differentiable on Θ_0, and under the conditions on $F_t(\cdot)$ in Assumption 5 $E_t[g_{t,j,\tau}(\theta)]$ is (a.s.) differentiable on Θ_0 with derivative

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

Hence, employing Leibniz' Integral Rule, we can obtain the mean value expansions

$$E[k_{t,i,\tau}(\theta)] = E[k_{t,i,\tau}(\theta_0)] + \big(\partial E[k_{t,i,\tau}(\tilde{\theta})]/\partial\theta\big)(\theta - \theta_0),$$

with

$$\begin{aligned}\partial E[k_{j,t,\tau}(\theta)]/\partial\theta &= E\big[\partial k_{j,t,\tau}(\theta)/\partial\theta\big] \\ &= E\big[H_{t,i,j}(\theta)\times\partial E_t[g_{t,j,\tau}(\theta)]/\partial\theta + \partial H_{t,i,j}(\theta)/\partial\theta\times E_t[g_{t,j,\tau}(\theta)]\big].\end{aligned}$$

That $\sup_t E\big[\sup_{\theta\in\Theta_0}|\partial k_{t,i,\tau}(\theta)/\partial\theta|\big] < C$ follows if $\sup_t\|\sup_{\theta\in\Theta_0}|\partial H_{t-1,i,j}(\theta)/\partial\theta|\|_{c_H} < C$, and $\sup_t\|\sup_{\theta\in\Theta_0}|\partial J_{k,l,t,\tau}(\theta)/\partial\theta\|_{c_g} < C$, and these conditions are imposed in Assumption 6.

Finally, notice that $E_t[g_{t,\tau}(\theta)]\big|_{\theta=\theta_0} = 0$ (a.s) under H_0, such that we find the specification of A_t in (10). That $A = A_t$ follows under Assumption 4. The result follows.

MC5:

We establish MC5, by showing (i) $\sup_t\big\|\sup_{\theta\in\Theta(\varepsilon)}k_{t,i,\tau}(\theta) - k_{t,i,\tau}(\theta_0)\big\|_Q \le C\varepsilon^{\phi}$, for all $i = 1,\dots,l$, and (ii) $\sup_t\big\|\sup_{\theta\in\Theta(\varepsilon)}l_{j,t}(\theta) - l_{j,t}(\theta_0)\big\|_Q \le C\varepsilon^{\phi}$, for all $j = 1,\dots,p$. Condition (ii) is imposed under Assumption 4.

To establish (ii) notice that

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

That $\sup_t\|\sup_{\theta\in\Theta(\varepsilon)}H_{t,i,j}(\theta)\|_{c_HQ} < C$ and $\sup_t\|\sup_{\theta\in\Theta(\varepsilon)}g_{t,j,\tau}(\theta)\|_{c_gQ} < C$ follows from Assumption 6 and applying steps as in the preceding.

By (a.s.) differentiability of $H_{t,i,j}(\theta)$ on Θ_0 under Assumption 4 we have $\|\sup_{\theta\in\Theta(\varepsilon)}H_{t,i,j}(\theta) - H_{t,i,j}(\theta_0)\|_{c_HQ} < \|\sup_{\theta\in\Theta_0}|\partial H_{t,i,j}(\theta)/\partial\theta|\|_{c_HQ}\sup_{\theta\in\Theta(\varepsilon)}|\theta-\theta| < C\varepsilon$, where the last inequality follows under Assumption 6.

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 $\|g_{t,1,\tau}(\theta) - g_{t,1,\tau}(\theta_0)\|_{c_gQ} \le \big(E|g_{t,1,\tau}(\theta) - g_{t,1,\tau}(\theta_0)|\big)^{1/(c_gQ)}$.

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

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

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 $E_t[g_{t,1,\tau}\left(\theta_{\max}(W_t)\right) - g_{t,1,\tau}\left(\theta_{\min}(W_t)\right)] = F_{t,\tau}(\theta_{\max}(W_t)) - F_{t,\tau}(\theta_{\min}(W_t)) \leq C_f \sup_{\theta\in\Theta} |J_t(\theta)| \times 2\varepsilon$, with $C_f < \infty$ the upperbound imposed in Assumption 5.

Hence,

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

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 $\theta \in \Theta(\varepsilon)$ $g_{t,2,\tau}(\theta)$ safisfies the Lipschitz condition $|g_{t,2,\tau}(\theta) - g_{t,2,\tau}(\theta_0)| \leq |m_t(\theta) - m_t(\theta_0)| \leq \sup_{\theta\in\Theta_0} |J_t(\theta)| |\theta - \theta_0|$ (a.s.), with the second inequality following from the Mean Value Theorem.

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

B.2 Proof of Corollary 1

We first consider $\hat{A}_P$. 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 $\hat{A}_P$:

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

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

Define

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

and

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

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

Also define

$$\hat{\varepsilon}_t = Y_{t+\tau} - m_{t,1}(\hat{\theta}_t),$$
$$\varepsilon_{0,t} = Y_{t+\tau} - m_{t,1}(\theta_0),$$

and

$$\delta_t(\theta) = m_{t,1}(\theta) - m_{t,1}(\theta_0).$$

Then,

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

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

$$
\begin{aligned}
&|\hat{a}_P - \tilde{a}_P| \\
&\quad \leq \frac{c_P}{\hat{c}_P}(2Pc_P)^{-1} \times \sum_{t=R}^{T-\tau} \Bigg\{ \Big\{ \mathbb{1}(|\varepsilon_{0,t} - c_P| < |\delta_t(\hat{\theta}_t)| + |\hat{c}_P + c_P|) \\
&\qquad\qquad + \mathbb{1}(|\varepsilon_{0,t} + c_P| < |\delta_t(\hat{\theta}_t)| + |\hat{c}_P + c_P|) \Big\} \sup_{\theta \in \Theta_0} |H_{t,i,j}(\theta)| \sup_{\theta \in \Theta_0} |J_{k,l,t}(\theta)|) \\
&\qquad\qquad + \mathbb{1}(|\varepsilon_{0,t}| < c_P) \sup_{\theta \in \Theta_0} |\partial H_{t,i,j}(\theta)/\partial\theta| \times |\hat{\theta}_t - \theta_0| \times \sup_{\theta \in \Theta_0} |J_t(\theta)| \\
&\qquad\qquad + \mathbb{1}(|\varepsilon_{0,t}| < c_P) \sup_{\theta \in \Theta_0} |\partial J_{k,l,t}(\theta)/\partial\theta| \times |\hat{\theta}_t - \theta_0| \times \sup_{\theta \in \Theta_0} \times |H_t(\theta)| \\
&\qquad\qquad + \frac{c_P - \hat{c}_P}{c_P} \mathbb{1}(|\varepsilon_{0,t}| < c_P) \sup_{\theta \in \Theta_0} |H_t(\theta)| \times \sup_{\theta \in \Theta_0} |J_t(\theta)| \Bigg\} \\
&\quad \equiv \frac{c_P}{\hat{c}_P}(A_1 + A_2 + A_3 + A_4),
\end{aligned}
\tag{25}
$$

by the Mean Value Theorem, and Assumptions 4 and 6 (see Engle and Manganelli (2004) for elaboration).

For P sufficiently large we can pick any $d > 0$, such that eventually $|c_P - \hat{c}_P|/c_P < d$, and $c_P^{-1} \sup_t |\hat{\theta}_t - \theta_0| < d$, 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 (25), we can show $E[A_i] = O(d)$, for $i = 1, \ldots, 4$, such that we obtain $|\hat{a}_P - \tilde{a}_P| = o_P(1)$ from Markov's inequality.

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

Notice,

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

with K some finite constant, and where the first inequality follows from noting that, e.g., $E_t[\mathbb{1}(|\varepsilon_{0,t} - c_P| < y)] = F_{t,\tau}(y + c_P) - F_{t,\tau}(-y + -c_P) \leq C_f|y + c_P|$, 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

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

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

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

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

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

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

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

where $y^* = \operatorname{argmax}_{y \in [-c_P, c_P]} f_{t,\tau}(y)$, 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.

□

C McCracken (2000) assumptions

Assumption MC3. *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 an open neighborhood Θ of θ_0, $\sup_t \| \sup_{\theta \in \Theta} v_t(\theta) \|_{2r} < \infty$, (d) Ω is positive definite.*

Assumption MC4. *For each $i \in \{1, \ldots, l+q\}$: (a) $E[v_{i,t}(\theta)]$ is continuously differentiable in the neighborhood Θ of θ_0 admitting a mean value expansion $E[v_{i,t}(\theta)] = E[v_{i,t}(\theta_0) + \big(\partial E[v_{i,t}(\tilde{\theta})] / \partial \theta\big)(\theta - \theta_0)]$, 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 $\sup_t \sup_{\theta \in \Theta} \left| \partial E[v_{i,t}(\theta)] / \partial \theta \right| < D$, and (c) for all t, $G = G_t \equiv \partial E[l_t(\theta)] / \partial \theta \big|_{\theta = \theta_0}$ and $F = F_t \equiv \partial E[k_{t,\tau}(\theta)] / \partial \theta \big|_{\theta = \theta_0}$.*

Assumption MC5. *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 $\Theta(\varepsilon) \subset \Theta$, $\sup_t \left\| \sup_{\theta \in \Theta(\varepsilon)} (v_t(\theta) - v_t(\theta_0)) \right\|_Q \leq C \varepsilon^{\phi}$.*

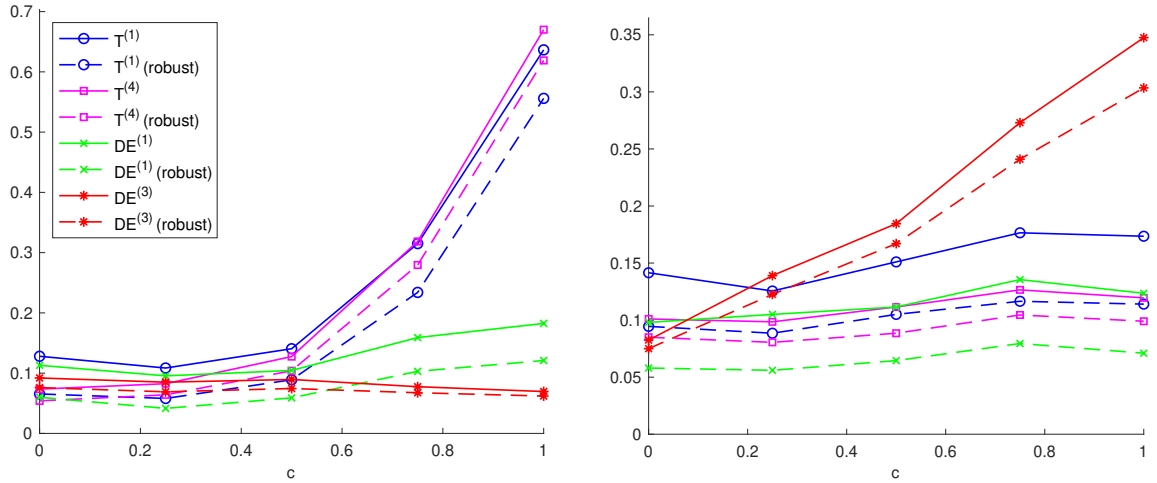
D Additional power plots

[Figure 2 about here.]

[Figure 3 about here.]

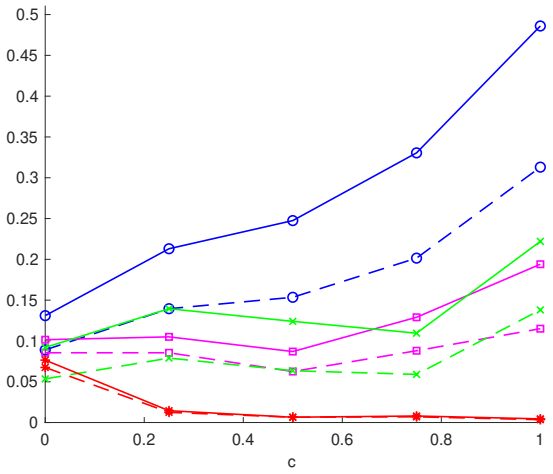
[Figure 4 about here.]

Figure 1: Empirical rejection rates for DGPs A_1, A_2, and A_3. $R/P = 2500/2500$.



(a) DGP A_1

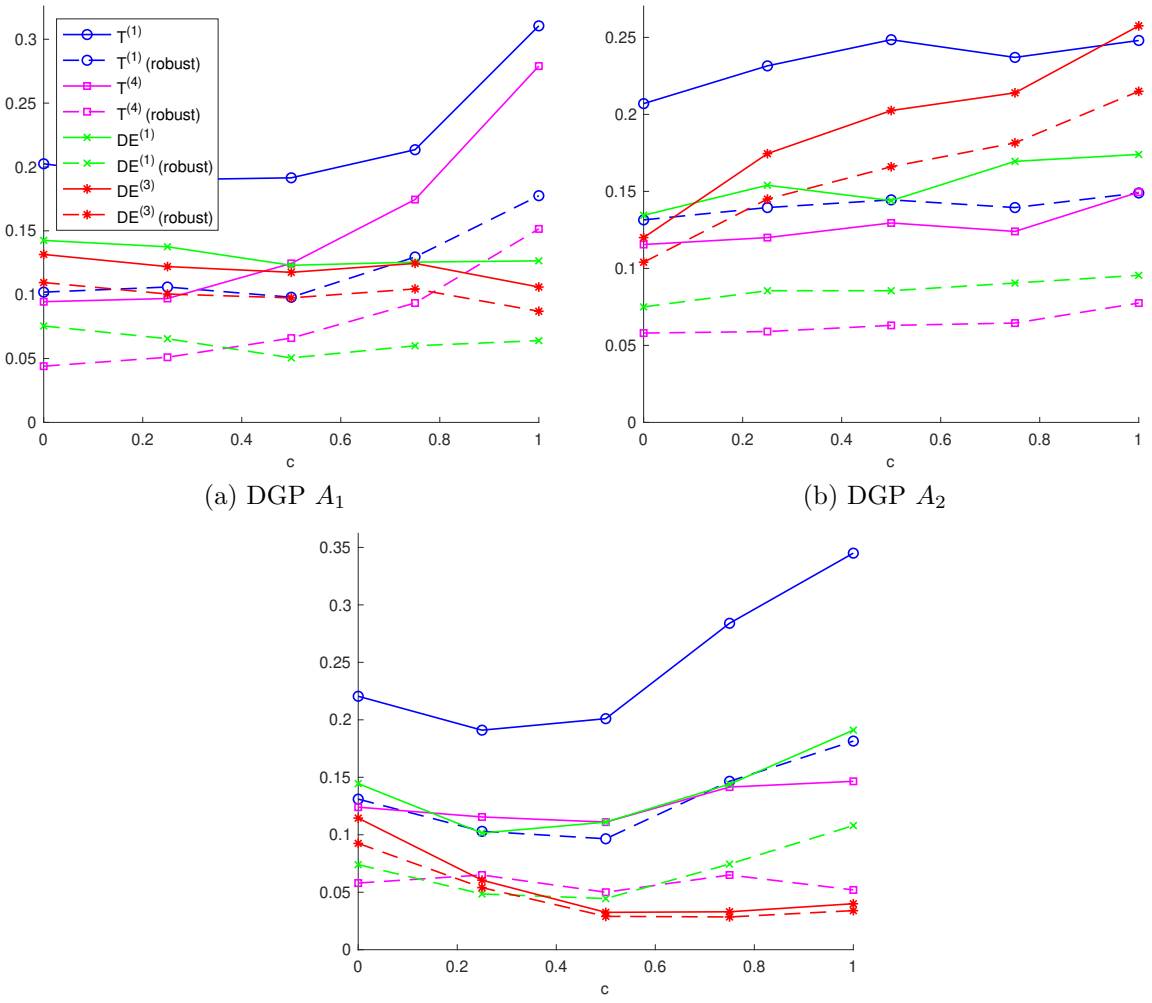
(b) DGP A_2



(c) DGP A_3

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 alternative models are specified as A_1, A_2 and A_3. We calculate VaR and ES for a coverage level of $1-\alpha = 97.5\%$, and use an in-sample and out-of-sample window of $R = P = 2,500$. We evaluate the test statistic with a 5% significance level and use 1,000 simulations.

Figure 2: Empirical rejection rates for DGPs A_1, A_2, and A_3. $R/P = 500/500$.



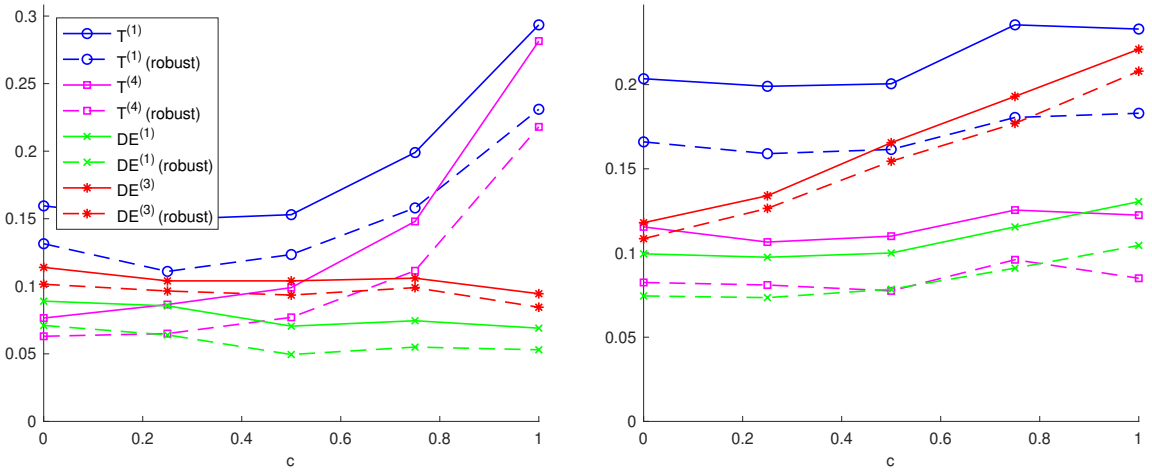
(a) DGP A_1

(b) DGP A_2

(c) DGP A_3

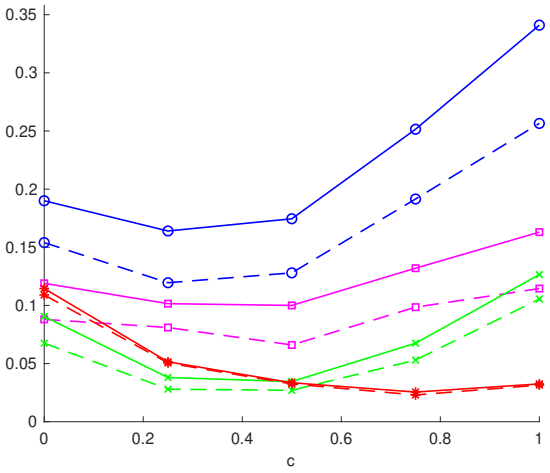
This figure plots the empirical rejection rates of the standard and robust versions of tests $T_P^{(1)}$, $T_P^{(4)}$, $DE_P^{(1)}$ and $DE_P^{(3)}$ as a function of c which determines the deviation from the null model. The alternative models are specified as A_1, A_2 and A_3. We calculate VaR and ES for a coverage level of $1 - \alpha = 97.5\%$, and use an in-sample and out-of-sample window of $R = P = 500$. We evaluate the test statistic with a 5% significance level and use 1,000 simulations.

Figure 3: Empirical rejection rates for DGPs A_1, A_2, and A_3. $R/P = 2,500\,/\,500$.



(a) DGP A_1

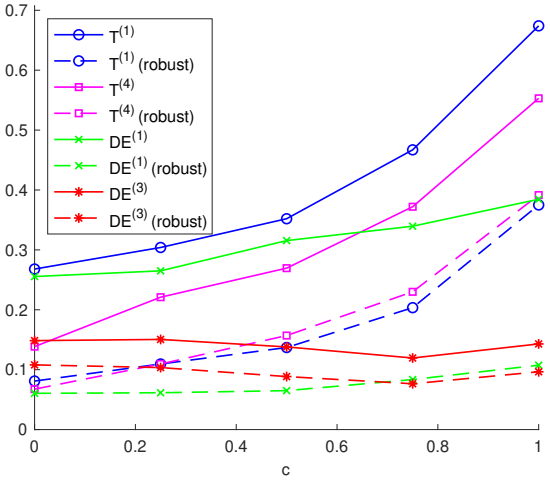
(b) DGP A_2



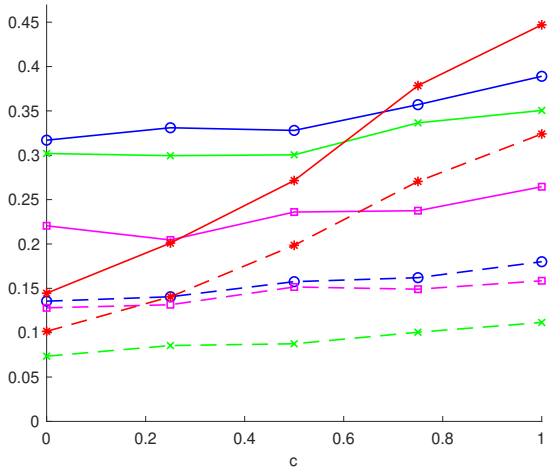
(c) DGP A_3

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 alternative models are specified as A_1, A_2 and A_3. We calculate VaR and ES for a coverage level of $1-\alpha = 97.5\%$, and use an in-sample and out-of-sample window of $R = 2,500$ and $P = 500$, respectively. We evaluate the test statistic with a 5% significance level and use 1,000 simulations.

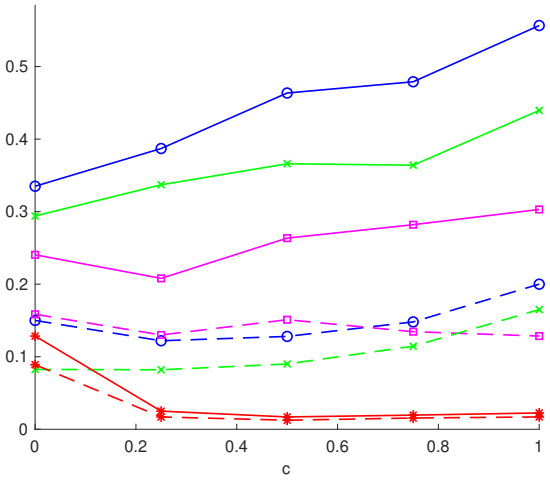
Figure 4: Empirical rejection rates for DGPs A_1, A_2, and A_3. $R/P = 500\,/\,2,500$.



(a) DGP A_1



(b) DGP A_2



(c) DGP A_3

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 alternative models are specified as A_1, A_2 and A_3. We calculate VaR and ES for a coverage level of $1 - \alpha = 97.5\%$, and use an in-sample and out-of-sample window of $R = 500$ and $P = 2,500$, respectively. We evaluate the test statistic with a 5% significance level and use 1,000 simulations.

Table 1: Empirical rejection rates in the size experiments with normal errors

Panel A: Standard Tests									
	VaR tests		Joint (VaR,ES) tests				ES tests		
	Unc.	Cond.	Unc.	Cond.			Unc.	Cond.	
R/P	$EO_P^{(1)}$	$EO_P^{(2)}$	$T_P^{(1)}$	$T_P^{(2)}$	$T_P^{(3)}$	$T_P^{(4)}$	$DE_P^{(1)}$	$DE_P^{(2)}$	$DE_P^{(3)}$
Coverage level $1-\alpha = 97.5\%$									
500/500	0.13	0.06	0.20	0.45	0.41	0.09	0.14	0.07	0.14
2500/500	0.07	0.06	0.15	0.44	0.36	0.08	0.09	0.06	0.12
500/2500	0.25	0.07	0.27	0.52	0.35	0.13	0.25	0.08	0.14
2500/2500	0.10	0.05	0.13	0.56	0.20	0.07	0.11	0.05	0.09
Coverage level $1-\alpha = 95\%$									
500/500	0.13	0.05	0.16	0.42	0.27	0.09	0.13	0.06	0.09
2500/500	0.08	0.04	0.13	0.40	0.26	0.08	0.08	0.06	0.09
500/2500	0.26	0.11	0.31	0.26	0.36	0.18	0.29	0.09	0.15
2500/2500	0.10	0.07	0.12	0.28	0.16	0.08	0.11	0.05	0.07
Panel B: Robust tests									
Coverage level $1-\alpha = 97.5\%$									
500/500	0.06	0.05	0.11	-	0.21	0.04	0.07	0.05	0.11
2500/500	0.06	0.05	0.12	-	0.26	0.06	0.07	0.05	0.11
500/2500	0.05	0.04	0.08	-	0.12	0.06	0.06	0.04	0.10
2500/2500	0.05	0.03	0.07	-	0.10	0.05	0.06	0.04	0.08
Coverage level $1-\alpha = 95\%$									
500/500	0.06	0.03	0.07	-	0.12	0.05	0.06	0.03	0.07
2500/500	0.04	0.03	0.10	-	0.18	0.07	0.06	0.05	0.08
500/2500	0.05	0.05	0.09	-	0.11	0.08	0.06	0.03	0.09
2500/2500	0.05	0.04	0.06	-	0.08	0.06	0.05	0.04	0.05

This table presents empirical rejection rates of the tests introduced in Section 3 for the AR-GARCH model with standard normal errors as in specification A_1 with $c = 0$. The rejection rates are calculated using 1,000 Monte Carlo experiments. Results are presented for coverage levels $1-\alpha = 97.5\%$ and 95%, as well as combinations of in-sample size R and out-of-sample size P.

Table 2: Empirical rejection rates in the size experiments with Students t-errors

Panel A: Standard Tests									
	VaR tests		Joint (VaR,ES) tests				ES tests		
	Unc.	Cond.	Unc.	Cond.			Unc.	Cond.	
R/P	$EO_P^{(1)}$	$EO_P^{(2)}$	$T_P^{(1)}$	$T_P^{(2)}$	$T_P^{(3)}$	$T_P^{(4)}$	$DE_P^{(1)}$	$DE_P^{(2)}$	$DE_P^{(3)}$
Coverage level $1-\alpha = 97.5\%$									
500/500	0.11	0.06	0.21	0.49	0.43	0.13	0.12	0.06	0.13
2500/500	0.08	0.05	0.19	0.45	0.45	0.10	0.09	0.06	0.11
500/2500	0.24	0.06	0.31	0.49	0.42	0.21	0.29	0.07	0.15
2500/2500	0.09	0.05	0.14	0.51	0.28	0.10	0.10	0.06	0.08
Coverage level $1-\alpha = 95\%$									
500/500	0.12	0.05	0.18	0.37	0.34	0.12	0.12	0.06	0.09
2500/500	0.07	0.05	0.15	0.38	0.33	0.11	0.07	0.06	0.09
500/2500	0.26	0.09	0.36	0.30	0.45	0.27	0.29	0.08	0.12
2500/2500	0.09	0.04	0.11	0.31	0.22	0.10	0.10	0.04	0.07
Panel B: Robust tests									
Coverage level $1-\alpha = 97.5\%$									
500/500	0.05	0.04	0.13	-	0.23	0.07	0.07	0.04	0.11
2500/500	0.06	0.04	0.15	-	0.36	0.08	0.07	0.04	0.11
500/2500	0.06	0.04	0.13	-	0.18	0.12	0.08	0.04	0.10
2500/2500	0.05	0.04	0.09	-	0.19	0.09	0.06	0.05	0.08
Coverage level $1-\alpha = 95\%$									
500/500	0.04	0.03	0.10	-	0.17	0.08	0.05	0.04	0.07
2500/500	0.04	0.04	0.11	-	0.25	0.09	0.05	0.05	0.08
500/2500	0.04	0.04	0.15	-	0.17	0.18	0.05	0.03	0.08
2500/2500	0.03	0.04	0.06	-	0.13	0.08	0.04	0.03	0.06

This table presents empirical rejection rates of the tests introduced in Section 3 for the AR-GARCH models with Student's t distributed errors with $\nu = 5$ degrees of freedom as in specification A_2 with $c = 0$. The rejection rates are calculated using 1,000 Monte Carlo experiments. Results are presented for coverage levels $1-\alpha = 97.5\%$ and 95%, as well as several combinations of in-sample size R and out-of-sample size P.

Table 3: Descriptive statistics of FTSE 100 index returns

	Full sample	Crisis sample
No. obs	2386	507
Mean	0.02	-0.08
Median	0.04	-0.07
Std. dev.	0.93	1.96
Skewness	-0.22	0.08
Kurtosis	5.33	6.75
Max.	4.51	9.48
10%	-1.07	-2.29
5%	-1.51	-3.08
1%	-2.74	-5.71
Min.	-4.95	-8.93

This table provide descriptive statistics of the daily returns (in %) on the FTSE 100 index. The long sample runs from November 8, 2009 to April 17, 2019. The crisis sample runs from June 30, 2007 to June 30, 2009.

Table 4: Summary statistics of parameter estimates

	AR-GARCH			AR-GJR-GARCH			AR-HEAVY		
	Mean	Median	St.dev.	Mean	Median	St.dev.	Mean	Median	St.dev.
a_0	0.004	0.003	0.022	0.002	0.000	0.023	0.006	0.003	0.023
ω_0	0.029	0.030	0.012	0.053	0.055	0.010	0.043	0.048	0.019
α_0	0.109	0.096	0.029	0.014	0.014	0.014	-	-	-
δ_0	-	-	-	-	-	-	0.414	0.410	0.130
γ_0	-	-	-	0.143	0.149	0.047	-	-	-
β_0	0.863	0.882	0.044	0.844	0.849	0.037	0.585	0.553	0.125
ν	9.432	10.000	2.885	13.048	13.171	4.310	13.668	13.402	6.310

Note: This table provides summary statistics for the parameter estimates of the AR-GARCH, AR-GJR-GARCH, and AR-HEAVY models, estimated using rolling windows of $1,000$ observations. The first estimation window runs until November 7, 2009, the last until 16 April 2019 (2,386 windows).

Table 5: Sample fraction of VaR violations and mean ES errors for a coverage level of 2.5%

Panel A: Sample fraction of VaR violations		
R	Crisis sample	Long sample
AR-GARCH		
500	0.047	0.033
1000	0.059	0.031
2500	-	0.034
AR-GJR-GARCH		
500	0.069	0.038
1000	0.081	0.032
2500	-	0.035
AR-HEAVY		
500	0.045	0.038
1000	0.055	0.036
2500	-	0.036
Panel B: Sample mean of ES error (in %) given a VaR violation		
R	Crisis sample	Long sample
AR-GARCH		
500	-0.330	-0.012
1000	-0.234	0.026
2500	-	-0.002
AR-GJR-GARCH		
500	-0.179	-0.006
1000	-0.170	-0.019
2500	-	-0.044
AR-HEAVY		
500	0.003	-0.026
1000	-0.038	0.023
2500	-	-0.034

Note: The AR-GARCH AR-GJR-GARCH, and AR-HEAVY models are estimated over a rolling window of R observations. Panel A reports the sample fraction of VaR violations defined in Eq. (22). Panel B reports the mean ES error given a VaR violation, which is defined in Eq. (23). We consider the coverage level $1 - \alpha = 97.5\%$. The crisis and long samples contain 507 and 2386 out-of-sample observations.

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

Panel A: Crisis sample								
R	VaR tests		Joint (VaR,ES) tests			ES tests		
	Unc.	Cond.	Unc.	Cond.		Unc.	Cond.	
	$EO_P^{(1)}$	$EO_P^{(2)}$	$T_P^{(1)}$	$T_P^{(3)}$	$T_P^{(4)}$	$DE_P^{(1)}$	$DE_P^{(2)}$	$DE_P^{(3)}$
Standard tests								
500	0.02	0.39	0.06	0.00	0.05	0.01	0.54	0.01
1000	0.00	0.88	0.00	0.00	0.13	0.00	0.54	0.01
Robust tests								
500	0.12	0.39	0.19	0.00	0.04	0.04	0.54	0.01
1000	0.02	0.90	0.05	0.01	0.13	0.01	0.54	0.02
Panel B: Long Sample								
Standard tests								
500	0.04	0.80	0.12	0.02	0.93	0.03	0.10	0.02
1000	0.08	0.23	0.16	0.56	0.59	0.08	0.09	0.01
2500	0.02	0.03	0.06	0.31	0.90	0.02	0.02	0.01
Robust tests								
500	0.31	0.81	0.60	0.04	0.94	0.17	0.12	0.07
1000	0.16	0.37	0.26	0.60	0.59	0.17	0.10	0.02
2500	0.06	0.15	0.15	0.43	0.90	0.06	0.05	0.01

Note: This table presents p-values for the standard and robust versions of the tests introduced in Section 3. The VaR and ES forecasts are generate by the AR-GARCH model estimated over rolling windows of R observations. We consider coverage level $1 - \alpha = 97.5\%$. In panel A we report results for the crisis sample, whose out-of-sample period runs from June 30, 2007 to June 30, 2009 and the long sample whose ouf-of-sample period runs from November 8, 2009 to April 17, 2019.

Table 7: p-values of VaR and ES tests applied to forecasts generated by the AR-GJR-GARCH model

R	VaR tests		Joint (VaR,ES) tests			ES tests		
	Unc.	Cond.	Unc.	Cond.		Unc.	Cond.	
	$EO_P^{(1)}$	$EO_P^{(2)}$	$T_P^{(1)}$	$T_P^{(3)}$	$T_P^{(4)}$	$DE_P^{(1)}$	$DE_P^{(2)}$	$DE_P^{(3)}$
Panel A: Crisis sample								
Standard tests								
500	0.00	0.77	0.00	0.00	0.10	0.00	0.76	0.07
1000	0.00	0.87	0.00	0.00	0.09	0.00	0.94	0.14
Robust tests								
500	0.03	0.79	0.02	0.00	0.14	0.64	0.77	0.11
1000	0.00	0.90	0.00	0.01	0.09	0.45	0.97	0.25
Panel B: Long sample								
Standard tests								
500	0.00	0.57	0.01	0.87	0.90	0.00	0.67	0.09
1000	0.05	0.67	0.13	0.70	0.62	0.02	0.42	0.04
2500	0.01	0.83	0.03	0.25	0.36	0.00	0.16	0.06
Robust tests								
500	0.06	0.77	0.16	0.96	0.91	0.06	0.83	0.10
1000	0.11	0.67	0.28	0.73	0.64	0.61	0.46	0.05
2500	0.03	0.83	0.08	0.29	0.43	0.78	0.21	0.07

Note: This table presents p-values for the standard and robust versions of the tests introduced in Section 3. The VaR and ES forecasts are generate by the AR-GJR-GARCH model estimated over rolling windows of R observations. We consider coverage level $1 - \alpha = 97.5\%$. In panel A we report results for the crisis sample, whose out-of-sample period runs from June 30, 2007 to June 30, 2009 and the long sample whose ouf-of-sample period runs from November 8, 2009 to April 17, 2019.

Table 8: p-values of VaR and ES tests applied to forecasts generated by the AR-HEAVY model

Crisis sample								
In-sample length R	VaR tests		Joint (VaR,ES) tests			ES tests		
	Unc.	Cond.	Unc.	Cond.		Unc.	Cond.	
	$EO_P^{(1)}$	$EO_P^{(2)}$	$T_P^{(1)}$	$T_P^{(3)}$	$T_P^{(4)}$	$DE_P^{(1)}$	$DE_P^{(2)}$	$DE_P^{(3)}$
Standard tests								
500	0.03	0.39	0.07	0.00	0.84	0.05	0.52	0.00
1000	0.00	0.94	0.01	0.01	0.57	0.01	0.51	0.01
Robust tests								
500	0.07	0.40	0.13	0.00	0.84	0.16	0.52	0.00
1000	0.01	0.94	0.03	0.01	0.57	0.04	0.51	0.01
Panel B: Long sample								
Standard tests								
500	0.00	0.26	0.01	0.59	0.47	0.00	0.20	0.15
1000	0.00	0.19	0.01	0.55	0.82	0.02	0.30	0.15
2500	0.01	0.18	0.02	0.52	0.28	0.00	0.07	0.13
Robust tests								
500	0.01	0.28	0.04	0.61	0.47	0.02	0.22	0.17
1000	0.04	0.29	0.05	0.63	0.83	0.10	0.43	0.19
2500	0.02	0.19	0.07	0.53	0.29	0.02	0.08	0.14

Note: This table presents p-values for the standard and robust versions of the tests introduced in Section 3. The VaR and ES forecasts are generate by the AR-HEAVY model estimated over rolling windows of R observations. We consider coverage level $1 - \alpha = 97.5\%$. In panel A we report results for the crisis sample, whose out-of-sample period runs from June 30, 2007 to June 30, 2009 and the long sample whose ouf-of-sample period runs from November 8, 2009 to April 17, 2019.