Backtesting Expected Shortfall via Multi-Quantile Regression

Ophélie Couperier, Jérémy Leymarie

Backtesting Expected Shortfall via Multi-Quantile Regression

Ophélie Couperier * Jérémy Leymarie †

November 28, 2019

Abstract

In this article, we propose a new approach to backtest Expected Shortfall (ES) exploiting the definition of ES as a function of Value-at-Risk (VaR). Our methodology examines jointly the validity of the VaR forecasts along the tail distribution of the risk model, and encompasses the Basel Committee recommendation of verifying quantiles at risk levels 97.5%, and 99%. We introduce four easy-to-use backtests in which we regress the ex-post losses on the VaR forecasts in a multi-quantile regression model, and test the resulting parameter estimates. Monte-Carlo simulations show that our tests are powerful to detect various model misspecifications. We apply our backtests on S&P500 returns over the period 2007-2012. Our tests clearly identify misleading ES forecasts in this period of financial turmoil. Empirical results also show that the detection abilities are higher when the evaluation procedure involves more than two quantiles, which should accordingly be taken into account in the current regulatory guidelines.

Keywords: Banking regulation; Financial risk management; Forecast evaluation; Hypothesis testing; Tail risk.

JEL classification: C12, C52, G18, G28, G32

*Ensae (CREST, UMR CNRS 9194), 5 avenue Henry Le Chatelier, 91120 Palaiseau, France. Email: ophelie.couperier@ensae.fr

†University of Orléans (LEO, FRE CNRS 2014), 11 rue de Blois, 45067 Orléans, France. Email: jeremy.leymarie@univ-orleans.fr. Corresponding author.

1 Introduction

In response to the market failures revealed by the global 2007-2008 financial crisis, the Basel Committee on Banking Supervision (BCBS) has adopted the Basel III accords to improve the banking sector's ability to absorb shocks arising from financial and economic stress (BCBS, 2010). Among the number of fundamental reforms that must be implemented until January 1st, 2022 (BCBS, 2019), the BCBS has substituted Value-at-Risk (VaR) by Expected Shortfall (ES) for the calculation of market risk capital requirements. Expected Shortfall, also referred to as Conditional VaR (CVaR) or Tail VaR (TVaR), measures the expected loss incurred on an asset portfolio given that the loss exceeds VaR. That is, if L_t is the (integrable) ex-post loss on a portfolio at time t, Ω_{t-1} is the information at time $t-1$, and $Q_{L_t}(.)$ is the quantile function of L_t, the τ-level ES and VaR are given by

$$ES_t(\tau) = \mathbb{E}\left[L_t \mid L_t \geq VaR_t(\tau); \Omega_{t-1}\right],$$

$$VaR_t(\tau) = Q_{L_t}(\tau; \Omega_{t-1}).$$

As an alternative tail risk measure, ES offers a number of appealing properties that overcomes the deficiencies of the more-familiar VaR. In particular, ES is *coherent* meaning that it satisfies the properties of monotonicity, sub-additivity, homogeneity, and translational invariance (see Artzner et al., 1999; Acerbi and Tasche, 2002). Furthermore, ES provides information about the expected size of the potential loss given that a loss bigger than VaR is experienced, while VaR only captures the likelihood of an incurred loss, and tells us nothing about tail sensitivity. In its revised standards for market risk, the BCBS emphasizes the important role of ES in place of VaR "*to ensure a more prudent capture of "tail risk" and capital adequacy during periods of significant financial market stress*" (BCBS, 2016, page 1).

Although ES is now considered as the new standard for risk management and regulatory requirements, there are still outstanding questions about the modeling of ES (see e.g. Taylor,

2019; Patton et al., 2019), and the validation of the ES forecasts, or backtesting. Jorion (2006) defines backtesting as a formal statistical framework that consists in verifying if actual losses are in line with projected losses. Because ES is unobservable, its evaluation cannot be performed conventionally as a direct comparison of the observed value with its forecast, and thus generally relies on the elicitability property. A risk measure is said to be *elicitable* if there exists a loss function such that the solution of minimizing the expected loss is the risk measure itself. However, it has been established that, in contrast to VaR, ES does not meet the general property of elicitability (Gneiting, 2011), but satisfies narrower properties such as conditional elicitability (Emmer et al., 2015), or joint elicitability with VaR (Acerbi and Szekely, 2014; Fissler and Ziegel, 2016), making its evaluation trickier than VaR in practice. Several contributions are tied to these properties, and provide backtests by making explicit reference of the ES forecasts in the testing procedure (McNeil and Frey, 2000; Acerbi and Szekely, 2014; Nolde and Ziegel, 2017; Bayer and Dimitriadis, 2019).

To circumvent the lack of elicitability of ES, several alternative testing strategies have been proposed in the literature. Following the recent classification of Kratz et al. (2018), these backtests enter the category of *implicit* backtests, as they focus on the tail distribution characteristics of the model rather than directly on ES. They generally exploit the fact that ES can be expressed as a function of VaR, which itself is elicitable. Assume the law of L_t is continuous. Definition of a conditional probability and a change of variable yield a useful representation of ES in terms of VaR

$$ES_t(\tau) = \frac{1}{1-\tau}\int_{\tau}^{1} VaR_t(u)\,du. \tag{1}$$

Based on this analogy, Costanzino and Curran (2015) derive a coverage backtest for spectral risk measures such as ES in the spirit of the traditional VaR coverage backtests. Du and Escanciano (2017) define a cumulative violation process for ES that generalizes the violation

process for VaR and propose two backtests of ES. Starting with the same process, Löser et al. (2019) develop a backtest of ES that is theoretically valid in finite out-of-sample size and that can be easily extended to a multivariate setting. Costanzino and Curran (2018) provide a Trafic Light backtest for ES which extends the so-called Traffic Light backtest for VaR. More largely, several additional techniques have been proposed to assess the whole return distribution encompassing ES as a special case (Berkowitz, 2001; Kerkhof and Melenberg, 2004; Wong, 2008). See the survey of Argyropoulos and Panopoulou (2016) for more details.

In this article, we also propose to exploit the relationship that prevails between ES and VaR, but contrary to the existing literature, our procedure aims at focusing on a finite number of VaRs. Definition of a Riemann sum gives a handy approximation of ES,

$$ES_t(\tau) \approx \frac{1}{p}\sum_{j=1}^{p} VaR_t(u_j),$$

where the risk level u_j is defined by $u_j = \tau + (j-1)\frac{1-\tau}{p}$ for $j = 1, 2, \ldots, p$. This representation suggests that p quantiles with appropriate risk levels would be convenient to assess the performance of an ES model. In other words, an estimate/forecast of $ES_t(\tau)$ issued from a given model could be considered valid if the sequence of $VaR_t(u_j)$ estimates/forecasts issued from the same model is itself valid. This testing strategy is fully consistent with the general recommendation of financial supervisors, indicating that "*Backtesting requirements [for ES] are based on comparing each desk's 1-day static value-at-risk measure [...] at both the 97.5th percentile and the 99th percentile*" (BCBS, 2016, page 57).

The main contribution of this article is to propose an original backtesting methodology to ES based on the theory of multi-quantile regression. We develop a multivariate framework, focusing on multi-quantile regression, to jointly assess VaR at multiple levels in the tail distribution of the risk model. The method extends the seminal idea of Gaglianone et al. (2011) to evaluate the validity of a single VaR relying on a single quantile regression.

Our backtesting procedure has many advantages. First, our approach encompasses the regulatory standards that consist of verifying the validity of two given quantiles. Second, our validation strategy offers flexibility since the risk manager or the supervisor may select both the number of risk levels and their magnitude depending on the objective in mind (regulatory guidelines, ES statistical approximation, etc.). Third, our testing strategy enters the category of regression-based backtests and complements the existing literature on regression-based risk forecast evaluation (see Engle and Manganelli, 2004; Christoffersen, 2011; Bayer and Dimitriadis, 2019, among others). Finally, our approach represents an alternative to the multiple VaR exceptions backtests (see Colletaz et al., 2013; Kratz et al., 2018).

Formally, we show that the parameters of the multi-quantile regression model have specific properties under the hypothesis of valid ES forecasts. We propose four backtests which correspond to various linear restrictions on these parameters. These restrictions are implications of a Mincer-Zarnowitz representation (Mincer and Zarnowitz, 1969). Then, we test the resulting parameter restrictions using Wald-type inference. Finally, we introduce a procedure deduced from our regression framework to adjust the invalid risk forecasts.

Several approaches to estimation and statistical inference in multi-quantile regression are suggested. Our baseline procedure is to apply the QML estimation method (White et al., 2008, 2015), and then, to implement a pairs bootstrap algorithm (Freedman, 1981) in order to correct the finite sample size distortions of our backtests. A second approach is to consider the estimation method of Jun and Pinkse (2009) which is designed to improve estimation efficiency in presence of correlated generalized errors and to apply the pairs bootstrap. An ultimate approach, although only available for single quantile models, consists of applying the procedure of Chernozhukov and Fernández-Val (2011) based on the extreme value theory.

Several Monte Carlo experiments are provided and an empirical application with the S&P500 series is conducted. Our backtests deliver good performances to detect misleading

ES forecasts. We also find that the use of asymptotic critical values is prone to substantial size distortions, while the implementation of bootstrap critical values provides satisfactory size performances regardless of the sample size. The latter should hence be preferred when asymptotic theory does not apply conveniently.

Our empirical results suggest an update of the regulatory guidelines. First, we show that the BCBS recommendation of assessing quantiles at risk levels 97.5% and 99% is not always sufficient to identify misspecified ES models. The use of additional quantiles is recommended to improve the soundness of the decision. Second, our results suggest to limit the number p of quantiles in small samples (with typically $p \leq 6$) and to consider higher values if the historical sample covers longer periods. Finally, we show numerically that our approximation of ES as a combination of several VaRs is close to its theoretical counterpart, which strongly supports its implementation in a risk management viewpoint.

The rest of the paper is organized as follows. In Section 2, we introduce the multi-quantile regression framework. Section 3 describes the null hypotheses of our tests, the test statistics, their asymptotic properties, and the procedure to implement the bootstrap critical values. Section 4 examines the finite sample performance of the proposed backtests through a set of Monte Carlo experiments. In Section 5, we apply our backtesting methodology on the S&P500 index and introduce the procedure to adjust the imperfect ES forecasts. This section also contains a number of robustness checks with alternative estimation and statistical inference approaches. Finally, we conclude the paper in Section 6.

2 Multi-quantile regression framework

This section describes our proposed multi-quantile regression approach. In the first part, we discuss the usefulness of approximating ES via a finite sum of VaRs. In a second part, we describe the multi-quantile regression model that we employ in our testing strategy. The

last part is devoted to the description of the estimation method and the asymptotic theory.

2.1 ES as an approximation of VaRs

Our backtesting procedure exploits the relationship between VaR and ES. We suppose that ES can be approximated as an average of VaRs. This assertion stems from the representation of ES as the limit of a Riemann sum when the partition becomes infinitely fine.

Definition 1 (ES approximation)**.** *Let $\tau \in]0,1[$ denote the coverage level. The τ-level ES approximation is defined as a finite Riemann sum involving p VaRs such as*

$$ES_t(\tau) \approx \frac{1}{p}\sum_{j=1}^{p} VaR_t(u_j), \tag{2}$$

where risk levels u_j, $j = 1, 2, \ldots, p$, satisfy $u_j = \tau + (j-1)\frac{1-\tau}{p}$, and p denotes the number of subdivisions taken in the definite integral.

Our approximation of ES averages VaRs in the upper tail distribution of the risk model. The number of quantiles involved in the sum is given by p and characterizes the approximation accuracy. In particular, $p = 1$ involves a single VaR at coverage level τ, while increasing p to infinity leads Equation (2) to converge to the theoretical ES. As we rely on a Riemann sum, the approximation assigns equal weights $1/p$ to each element in the sum, and the risk levels u_j, $j = 1, 2, \ldots, p$, are determined so that the interval is equally partitioned between the two boundaries τ and 1. Several alternatives for the approximation of a definite integral are available. Here, we rely on a Riemann sum for its simplicity and ease of implementation. We show how to derive the above formula in Appendix A.

In practice, p may be chosen small as the interval of the definite integral is restricted to the extreme upper tail distribution. For instance, Gouriéroux and Liu (2012) identify for a large class of distributions a common linear conversion pattern between VaR and ES, so that a few VaRs are generally enough to get a good approximation of ES. Daníelsson and Zhou

(2016) empirically show that VaR and ES are in most cases related by a small constant and are hence almost equally informative. Kratz et al. (2018) provide multinomial backtests of VaRs, and show that backtesting exceptions jointly at four to eight risk levels yields a very effective test in terms of balancing simplicity and reasonable size and power properties.

Our approximation is useful for at least two reasons. First, this simple formula is appealing in a regulatory and risk management viewpoint since the estimation of VaR is well-established and its computation is easier compared to ES. Secondly, and it is the purpose of this paper, the above relationship greatly simplifies the assessment of ES, by focusing on the validity of several VaRs, and is more intelligible in the context of banking regulation. This approach is fully consistent with the BCBS guidelines on ES assessment stating that "*Backtesting requirements [for ES] are based on comparing each desk's 1-day static value-at-risk measure [...] at both the 97.5th percentile and the 99th percentile*" (BCBS, 2016, page 11).

2.2 Multi-quantile regression model

In the sequel, we consider an asset or a portfolio, and denote by L_t the corresponding loss observed at time t, for $t = 1, 2, \ldots, T$. In addition, we denote by Ω_{t-1} the information set available at time $t-1$, with $(L_{t-1}, L_{t-2}, \ldots) \subseteq \Omega_{t-1}$. Formally, the Ω_{t-1} conditional VaR at level u_j of the L_t distribution is the quantity $VaR_t(u_j)$ such that

$$\Pr\left(L_t \geq VaR_t(u_j) \,|\Omega_{t-1}\right) = u_j. \tag{3}$$

A VaR model is said to be correctly specified (at coverage level u_j) as soon as Equation (3) holds for all t. In practice, VaR forecasts are assessed through the evaluation of this simple equality. Given the ES approximation introduced in Definition 1, this equality may arguably be adapted for the assessment of ES models. The chief insight is to evaluate Equation (3) for a number p of risk levels as set out in Definition 1. Accordingly, one should conclude to the appropriateness of a given ES model as soon as the sequence $VaR_t(u_j)$, $t = 1, 2, \ldots, T$,

issued by the ES model satisfies Equation (3) jointly for $j = 1, 2, \ldots, p$.

We refer to the original idea of Gaglianone et al. (2011) who derive a backtest of VaR at a single coverage level, introducing VaR as a regressor of a quantile regression model. We generalize their approach for the assessment of multiple VaRs. To do so, we regress the ex-post losses $\{L_t, t = 1, 2, \ldots, T\}$ on the p VaR forecasts $\{VaR_t(u_j), t = 1, 2, \ldots, T\}_{j=1,2,\ldots,p}$ in a multi-quantile regression model.

$$L_t = \beta_0(u_j) + \beta_1(u_j) VaR_t(u_j) + \epsilon_{j,t} \quad \forall\ j = 1, 2, \ldots, p, \tag{4}$$

where $\beta_0(u_j)$, and $\beta_1(u_j)$, respectively, denote the intercept and the slope parameters at level u_j, and where $\epsilon_{j,t}$ is the error term at risk level u_j and time t, such that the u_j-th conditional quantile of $\epsilon_{j,t}$ satisfies $Q_{\epsilon_{j,t}}(u_j; \Omega_{t-1}) = 0$. This specification could be interpreted as a multi-quantile regression version of Koenker and Xiao (2002). More specifically, the representation is tightly related to the multi-quantile CaViAR model (MQ-CaViAR) of White et al. (2008, 2015) which allows a joint modeling of multiple conditional VaRs. Given the multi-quantile regression model of Equation (4), the u_j-th conditional quantile of L_t is defined as

$$Q_{L_t}(u_j; \Omega_{t-1}) = \beta_0(u_j) + \beta_1(u_j) VaR_t(u_j) \quad \forall\ j = 1, 2, \ldots, p. \tag{5}$$

This equation is central for our backtesting methodology as it establishes a direct link between the VaR forecasts (issued from the external ES model), with the true unknown conditional quantile (issued from the ex-post observed losses). Our procedure consists in verifying if there exists a perfect match between $VaR_t(u_j)$ and $Q_{L_t}(u_j; \Omega_{t-1})$. Consistently with Gaglianone et al. (2011), we rely on the regression parameters, and test if the intercept parameter $\beta_0(u_j)$, and the slope parameter $\beta_1(u_j)$, are respectively equal to zero, and one, for $j = 1, 2, \ldots, p$. For these values, and given Definition 1, the risk model is accepted as a valid proxy of the true unknown data generating process to deliver the ES forecasts.

2.3 Parameter estimation and asymptotic properties

Our backtesting procedure requires to consistently estimate the parameters $\beta_0(u_j)$, and $\beta_1(u_j)$, for $j = 1, 2, \ldots, p$. Under the hypothesis that a sequence of VaR is valid, coefficients satisfy $\beta_0(u_j) = 0$, and $\beta_1(u_j) = 1$, for $j = 1, 2, \ldots, p$. In what follows, we denote by $\beta(u_j) = (\beta_0(u_j),\ \beta_1(u_j))'$ the vector of parameters for the u_j-th quantile index, and we write $\beta = \left(\beta(u_1)',\ \beta(u_2)',\ \ldots, \beta(u_p)'\right)'$ the stacked vector of $2p$ coefficients. We assume that the sequence $\{u_j, j = 1, 2, \ldots, p\}$ is ordered in the sense that $u_1 < u_2 < \ldots < u_p < 1$.

In order to estimate β, we consider the QML estimator proposed by White et al. (2008, 2015) dedicated to multi-quantile regression, given by

$$\widehat{\beta} = arg \min_{\beta \in \mathbb{R}^{2p}} T^{-1} \sum_{t=1}^{T} \left(\sum_{j=1}^{p} \rho_{u_j} \left(L_t - \beta_0(u_j) - \beta_1(u_j) VaR_t(u_j) \right) \right),$$

where $\rho_{u_j}(x) = x\psi_{u_j}(x)$ is the standard "check function", and $\psi_{u_j}(x) = u_j - \mathbb{1}(x \leq 0)$ is the usual quantile step function. Under suitable regularity conditions, White et al. (2008, 2015) show that this estimator is consistent and asymptotically normally distributed. The conditions are described in Appendix B and a discussion is provided on how these assumptions are fulfilled in our context. However, in case of correlated generalized errors $\psi(\epsilon_{j,t})$ between different quantiles, the QML estimator is not necessarily efficient. The procedure of Jun and Pinkse (2009) is designed to improve efficiency in the presence of dependent cross-equation errors. An application to this procedure is provided in Section 5.2 to gauge potential interest.

Under Assumptions A0-A2 in Appendix B, the asymptotic distribution of the QML estimator is given by

$$\sqrt{T}\left(\widehat{\beta} - \beta\right) \xrightarrow{d} \mathcal{N}(0, \Sigma),$$

where Σ denotes the asymptotic covariance matrix which takes the form of a Huber (1967) sandwich. Its expression is given by $\Sigma = A^{-1}VA^{-1}$, with $V = \mathbb{E}[\eta_t \eta_t']$, $\eta_t = \sum_{j=1}^{p} \nabla Q_{L_t}(u_j; \Omega_{t-1})\, \psi_{u_j}(\epsilon_{j,t})$, $A = \sum_{j=1}^{p} \mathbb{E}[f_{j,t}(0)\, \nabla Q_{L_t}(u_j; \Omega_{t-1})\, \nabla' Q_{L_t}(u_j; \Omega_{t-1})]$, where

$\nabla Q_{L_t}(u_j;\Omega_{t-1})$ denotes the $2p$ gradient vector differentiated with respect to β, $\epsilon_{j,t} = L_t - Q_{L_t}(u_j;\Omega_{t-1})$, and $f_{j,t}(0)$ denotes the pdf of $\epsilon_{j,t}$ evaluated at zero. In Appendix C, we provide a consistent estimator $\widehat{\Sigma}$ of Σ that will be used to compute our test statistics.

Finally, Appendix D provides a discussion on the rate of convergence and interplay of p and T when T tends to infinity. Under this asymptotic framework, we show that p is increasing with T. Then, we consider a simple illustration assuming p takes the form of a power function. Under this assumption, T needs to diverge faster than p to preserve the asymptotic theory of White et al. (2008, 2015). This condition is not importantly restrictive but suggests the existence of an (asymptotic) upper limit for p which depends on the sample size. Section 4 provides several Monte-Carlo experiments with various values for p and T to give guidelines on how to choose these parameters jointly in finite samples.

3 Backtesting ES

In this section, we present our backtests for ES. Our procedures assess whether the parameters $\beta_0(u_j)$ and $\beta_1(u_j)$ coincide with their expected values for risk levels u_j, $j = 1, 2, \ldots, p$. To this end, we propose four backtests that analyze various settings on the regression coefficients. In the sequel, we introduce the null hypotheses, the test statistics, and establish their asymptotic properties. Finally, we discuss the use of finite sample critical values and provide a bootstrap algorithm when the asymptotic theory does not apply conveniently.

3.1 The backtests

Formally, our goal is to test $\beta_0(u_j) = 0$, and $\beta_1(u_j) = 1$, for $j = 1, 2, \ldots, p$. As highlighted by Gaglianone et al. (2011) for a unique quantile regression, the aforementioned set of restrictions retains a Mincer and Zarnowitz (1969) interpretation for each quantile regression in (4). Here, we propose to test various implications of these coefficient restrictions by taking

into consideration four distinct null hypotheses based on a reduced number of constraints. Many backtests test implications of a more general hypothesis. In this context, Du and Escanciano (2017) assess two implications for the martingale difference sequence of their cumulative violation process. McNeil and Frey (2000) and Nolde and Ziegel (2017) propose to test the zero mean hypothesis of their residuals which more largely behave as white noise.

Definition 2 (Null hypotheses)**.** *Denote by J_1, J_2, I, and S, the four backtests. The corresponding null hypotheses H_{0,J_1}, H_{0,J_2}, $H_{0,I}$, $H_{0,S}$, are defined as follows:*

$$H_{0,J_1} : \sum_{j=1}^{p} \left(\beta_0\left(u_j\right) + \beta_1\left(u_j\right)\right) = p, \tag{6}$$

$$H_{0,J_2} : \sum_{j=1}^{p} \beta_0\left(u_j\right) = 0, \text{ and, } \sum_{j=1}^{p} \beta_1\left(u_j\right) = p, \tag{7}$$

$$H_{0,I} : \sum_{j=1}^{p} \beta_0\left(u_j\right) = 0, \tag{8}$$

$$H_{0,S} : \sum_{j=1}^{p} \beta_1\left(u_j\right) = p, \tag{9}$$

where notations J_1 and J_2 indicate the "joint" backtests, and where I and S refer to the "intercept" backtest and to the "slope" backtest, respectively.

Equations (6)-(9) of Definition 2 gives the null hypotheses H_{0,J_1}, H_{0,J_2}, $H_{0,I}$, $H_{0,S}$. They are devised to assess various implications that the regression coefficients should satisfy when the ES forecasts are valid. The coefficients are summed across risk levels u_j, $j = 1, 2, \ldots, p$. This aggregation substantially reduces the number of constraints. H_{0,J_2} is hence characterized by two constraints, and H_{0,J_1}, $H_{0,I}$, $H_{0,S}$ involve a single constraint. Furthermore, aggregating coefficients along the quantile curve allows addressing the quantile crossing problem that appears when multiple quantiles are jointly estimated. As stressed by Chernozhukov et al. (2010), quantile crossing is a problem when the ultimate goal of a researcher is modeling the quantile curve. Alternately our testing procedure focuses on the parameter estimates of the quantile models. As the procedure of Chernozhukov et al. (2010) does not require any

re-estimation of $\beta_0(u_j)$ and $\beta_1(u_j)$, the statistics J_1, J_2, I, S, remain unchanged.

Our null hypotheses analyze various settings on the regression coefficients. The null of the joint backtests, H_{0,J_1} and H_{0,J_2}, look at the expected value of both the intercept and slope parameters $\beta_0(u_j)$ and $\beta_1(u_j)$ for $j = 1, 2, \ldots, p$. H_{0,J_1} sums the two types of coefficient together, while H_{0,J_2} sums the coefficients separately depending on whether they are slope parameters or intercept parameters. Finally, the null hypotheses of the intercept backtest and the slope backtest, $H_{0,I}$ and $H_{0,S}$, focus solely on one of the two parameter components. $H_{0,I}$ is built to examine the intercept parameters $\beta_0(u_j)$, $j = 1, 2, \ldots, p$, and $H_{0,S}$ is devoted to the analysis of the slope parameters $\beta_1(u_j)$, $j = 1, 2, \ldots, p$. These additional null hypotheses complement the joint backtests to identify the nature of the misspecification. If the joint hypotheses are rejected, separate tests for these two types of measurement error should be considered. They are inspired by the prediction-realization framework of Mincer and Zarnowitz (1969). When $H_{0,I}$ is rejected, the intercept parameters, $\beta_0(u_j), j = 1, 2, \ldots, p$, do not sum to 0, and hence, the average of VaR forecasts either underestimate or overestimate the true quantiles, if the sign of the sum is positive or negative, respectively. The rejection of $H_{0,S}$ indicates that the sum of the slope parameters $\beta_1(u_j), j = 1, 2, \ldots, p$, does not equal p, which highlights correlation between the forecasting errors and the quantile series.

Definition 3 (Wald-test statistics)**.** *Let us denote by $W \in \{J_1, J_2, I, S\}$ the generic notation for the test statistic, and consider the classical formulation of a Wald-type test such as $H_{0,W}$: $R_W\beta = q_W$. The general expression of the test statistics is given by*

$$W = T\left(R_W\widehat{\beta} - q_W\right)'\left(R_W\widehat{\Sigma}R_W'\right)^{-1}\left(R_W\widehat{\beta} - q_W\right), \tag{10}$$

where T is the out-of-sample size, and $\widehat{\Sigma}$ denotes a consistent estimator of the asymptotic covariance matrix.

To assess our null hypotheses we consider Wald-type inference. Equation (10) of Defini-

tion 3 gives the general expression of the test statistics. According to our notations, substituting W by J_1, J_2, I, and S, yields the four test statistics. For ease of presentation, the null hypotheses are now presented in a classical formulation, such that $H_{0,W} : R_W\beta = q_W$. Given the null hypotheses of Definition 2, the quantities R_W and q_W are as follows: $R_{J_1} = \iota_p \otimes \begin{pmatrix} 1 & 1 \end{pmatrix}$, $q_{J_1} = p$, $R_{J_2} = \iota_p \otimes I_2$, $q_{J_2} = \begin{pmatrix} 0 & p \end{pmatrix}'$, $R_I = \iota_p \otimes \begin{pmatrix} 1 & 0 \end{pmatrix}$, $q_I = 0$, $R_S = \iota_p \otimes \begin{pmatrix} 0 & 1 \end{pmatrix}$, $q_S = p$, where ι_p is a p-row unit vector, and I_2 denotes the identity matrix of size 2.

Proposition 1 (Chi-squared distribution)**.** *Consider the multi-quantile regression model in Equation (4), Assumptions A0-A3 in Appendix B, and the null hypotheses of Definition 2, the test statistics J_1, I, and S, converge to a chi-squared distribution with 1 degree of freedom, and the test statistic J_2 converges to a chi-squared distribution with 2 degrees of freedom.*

Proposition 1 gives the asymptotic distribution of the Wald statistics J_1, J_2, I, S under their respective null hypotheses H_{0,J_1}, H_{0,J_2}, $H_{0,I}$, $H_{0,S}$. As a result of coefficients' aggregation, the asymptotic distributions are based on a small and fixed number of degrees of freedom no matter how p is chosen. Thus, the four backtests have unchanged critical values whatever the number of quantiles considered in the ES approximation. Finally, we provide in Appendix E the proof for consistency of the tests under fixed untrue hypothesis.

3.2 Finite sample inference

Our four backtests are asymptotically chi-squared distributed and we can employ them if the asymptotic conditions are fulfill for realistic sample sizes. However, in the case of ES assessment, the focus is on the extreme tail distribution, that is for risk levels above the regulatory coverage level, i.e. $\tau = 0.975$. This may induce scarce information and affect the inference when the sample size is not large enough. Furthermore, the asymptotic framework of White et al. (2008, 2015) implicitly assumes that $(1 - u_p)T$ diverges to infinity, where u_p denotes the highest level of the multi-quantile regression. Chernozhukov (2005) and

Chernozhukov and Fernández-Val (2011) provide a refinement of this assumption based on the extreme value theory allowing $(1-u_p)T \to k < \infty$. However, to date this literature has only considered single quantile models and it is not obvious how the results for the single quantile models extend to multi-quantile models. To overcome these typical deficiencies, we implement a bootstrap procedure to adjust the critical values of our test statistics in finite samples.

In the following, we propose a pairs bootstrap algorithm (Freedman, 1981) in order to correct the finite sample size distortions of our backtests. This is a fully non-parametric procedure that can be applied to a very wide range of models, including quantile regression model (Koenker et al., 2018). This approach consists in resampling the data, keeping the dependent and independent variables together in pairs. The procedure is valid for any sample sizes T, and large levels u_j, $j = 1, 2, \ldots, p$, and ideally applies in our case when the constraints of the null hypothesis are linear in the parameters. The algorithm is as follows:

1. Estimate β and Σ on the original data $\{L_t, VaR_t(u_j)\}_{j=1,2,\ldots,p}$, $t = 1, 2, \ldots, T$, to obtain $\widehat{\beta}$ and $\widehat{\Sigma}$, and compute the unconstrained test statistic W given by

$$W = T\left(R_W\widehat{\beta} - q_W\right)'\left(R_W\widehat{\Sigma}R_W'\right)^{-1}\left(R_W\widehat{\beta} - q_W\right).$$

2. Build a bootstrap sample by drawing with replacement T pairs of observations from the original data $\{L_t, VaR_t(u_j)\}_{j=1,2,\ldots,p}$, $t = 1, 2, \ldots, T$.

3. Estimate the model on the bootstrap sample, to obtain $\widehat{\beta}^b$ and $\widehat{\Sigma}^b$, and compute the bootstrapped test statistic W^b under the null hypothesis as follows:

$$W^b = T\left(R_W\widehat{\beta}^b - R_W\widehat{\beta}\right)'\left(R_W\widehat{\Sigma}^bR_W'\right)^{-1}\left(R_W\widehat{\beta}^b - R_W\widehat{\beta}\right).$$

4. Repeat $B-1$ times steps 2 and 3, to obtain the bootstrap statistics W^b, $b = 1, 2, \ldots, B$.

Two remarks should be made about the algorithm. First, when we use the pairs boot-

strap we cannot impose the null hypothesis on the bootstrap data generating process since imposing restrictions on β is unfeasible. To overcome this issue, we calculate the bootstrap statistics by considering the difference $R_W\beta - R_W\widehat{\beta}$ rather than $R_W\beta - q$. Since the estimate of β from the bootstrap samples should, on average, be equal to $\widehat{\beta}$, at least asymptotically, the null hypothesis tested by W^b becomes "true" for the pairs bootstrap data generating process. Second, the critical value c_α is obtained as the α-quantile of the bootstrap statistics W^b, $b = 1, 2, \ldots, B$. The decision rule is as follows. If the original test statistic W is greater than the α-level bootstrapped critical value c_α, we conclude to the rejection of the null hypothesis. In addition, we compute the p-value of the test as $P = B^{-1}\sum_{b=1}^{B}\mathbb{1}\left(W^b > W\right)$.

4 Simulation study

In this section, we provide Monte Carlo simulations to illustrate the finite sample properties (empirical size and power) of our four backtests. The simulation study is performed on 5000 replications, and we consider sample sizes $T = 250, 500, 1000, 2500$. The results associated with the bootstrap critical values are based on $B = 1000$ bootstrap samples. Finally, the backtests are computed with $\tau = 0.975$ that is the current banking regulation coverage level.

Beyond the traditional size and power analysis, a second important objective of this section is to characterize the influence of the number p of quantiles used to assess the ES forecasts. We aim at examining whether an ES backtest based on a large number of quantiles may provide better performances than a backtest based on a small number of quantiles, as it is recommended by the current BCBS guidelines. For that, we consider different choices for the number of risk levels, namely $p = 1, 2, 4, 6, 8, 10, 12$. The p risk levels $u_1, u_2, \ldots, u_p$ are computed in accordance with Definition 1. Notice that $p = 1$ coincides with the VaR backtest at level τ of Gaglianone et al. (2011). With $p = 2$ risk levels, our backtests are in accordance with the number of quantiles of the regulatory guidances. Finally, the case $p = 4$

corresponds to the framework considered by Emmer et al. (2015).

The correct data generating process is given by the AR(1)-GARCH(1,1) specification with Student innovations. This model has been widely used for assessing tail risk measures (see e.g. McNeil and Frey, 2000; Du and Escanciano, 2017; Löser et al., 2019, among others). The ex-post portfolio loss L_t, $t = 1, 2, \ldots T$, is given by

$$\begin{aligned} L_t &= \delta_0 + \delta_1 L_{t-1} + \epsilon_t, \\ \epsilon_t &= \sigma_t \eta_t, \ \eta_t \sim t_v, \\ \sigma_t^2 &= \gamma_0 + \gamma_1 \epsilon_{t-1}^2 + \gamma_2 \sigma_{t-1}^2, \end{aligned} \tag{11}$$

where t_v denotes the Student's t distribution with v degrees of freedom. Given the model in Equation (11), the true ES and VaR at coverage level τ are given by

$$ES_t(\tau) = \delta_0 + \delta_1 L_{t-1} + \sigma_t m(\tau), \tag{12}$$

$$VaR_t(\tau) = \delta_0 + \delta_1 L_{t-1} + \sigma_t F_v^{-1}(\tau), \tag{13}$$

with $m(\tau) = \mathbb{E}\left[\eta_t | \eta_t \geq F_v^{-1}(\tau)\right]$, and where $F_v^{-1}(\tau)$ denotes the τ-quantile of the Student distribution with v degrees of freedom. As a robustness check of the above model, Appendix F provides simulation results for the simple case of a GARCH(1,1) model that excludes the conditional mean component with $L_t = \epsilon_t$ where ϵ_t is as in Equation (11). Both models are calibrated using the opposite of the daily log-returns of the S&P500 index over the period from January 2, 2013 to December 29, 2017, with $\left(\widehat{\delta}_0, \widehat{\delta}_1, \widehat{\gamma}_0, \widehat{\gamma}_1, \widehat{\gamma}_2, \widehat{v}\right) = (-0.085, -0.093, 0.034, 0.214, 0.748, 5)$ and $(\widehat{\gamma}_0, \widehat{\gamma}_1, \widehat{\gamma}_2, \widehat{v}) = (0.034, 0.197, 0.763, 5)$, respectively for the AR(1)-GARCH(1,1) model and the GARCH(1,1) model. Finally to investigate the power, we consider several misspecified alternatives for L_t:

A_1 : AR(1)-GARCH(1,1) model with underestimated conditional variances: L_t is as Equation (11), with $\sigma_t^2 = \left(\gamma_0 + \gamma_1 \epsilon_{t-1}^2 + \gamma_2 \sigma_{t-1}^2\right) \times (1 - \kappa)$, where $\kappa = 0.25, 0.50, 0.75,$ respectively.

A_2 : GARCH in mean model: $L_t = \kappa \times \sigma_t^2 + \epsilon_t,\ \epsilon_t = \sigma_t \eta_t,\ \sigma_t^2 = \gamma_0 + \gamma_1 \epsilon_{t-1}^2 + \gamma_2 \sigma_{t-1}^2,\ \eta_t \sim t_v$, where $\kappa = +2.5, -2.5$, respectively.

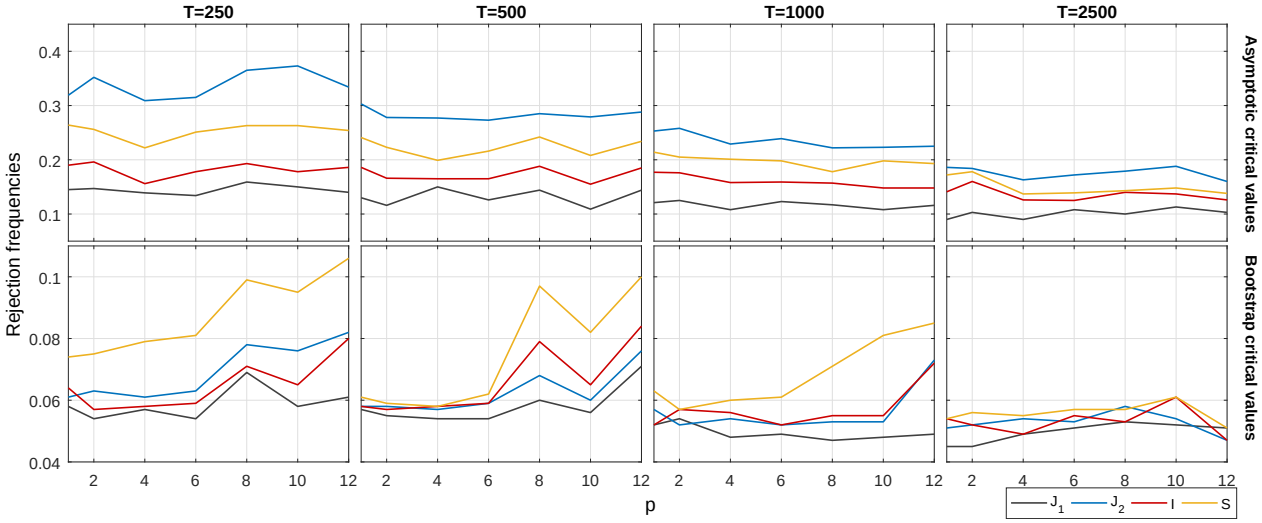
A_3 : AR(1)-GARCH(1,1) model with mixed normal innovations: L_t satifies Equation (11), with $\eta_t \sim \left(0.5X^+ + 0.5X^-\right)/\sqrt{10}$, where $X^+ \sim \mathcal{N}(3,1)$ and $X^- \sim \mathcal{N}(-3,1)$.

A_4 : 12-month historical simulation model : VaR and ES are given by their empirical counterparts from the 250 previous trading days such that $VaR_t(\tau) = \text{percentile}\left(\{L_{t-i}\}_{i=1}^{250}, 100\tau\right)$, and $ES_t(\tau) = \frac{1}{\sum_{i=1}^{250} \mathbb{1}_{(L_{t-i} \geq VaR_{t-i}(\tau))}} \sum_{i=1}^{250} L_{t-i} \times \mathbb{1}_{(L_{t-i} \geq VaR_{t-i}(\tau))}$.

In A_1, the conditional variance of the series σ_t is alternately underestimated of 25%, 50%, and 75% to examine whether our tests are able to detect an underestimation of ES stemming from a misleading appreciation of volatility. In A_2, the misspecification occurs in the conditional mean by assuming a GARCH in mean model. In A_3, the distribution of the innovations η_t is incorrect and should imply misleading ES predictions compared to the t-distribution. Finally in scenario A_4, the time-varying dynamics is incorrectly captured by the historical simulation method. It should be noticed that our alternatives are in line with the existing literature on tail risk assessment. Bayer and Dimitriadis (2019) look at an alternative close to A_1 by varying the coefficients related to the GARCH component. A_2 and A_3 were applied by Du and Escanciano (2017) to illustrate the performance of their unconditional and conditional ES backtests. Finally, scenario A_4 was extensively studied by Kratz et al. (2018), Bayer and Dimitriadis (2019), Gaglianone et al. (2011), among others.

Figure 1 displays graphically empirical sizes of the tests at 5% significance level. The first row reports the results of the asymptotic tests and the second row embeds those of the bootstrap based tests. Each column is for a given sample size T, and the results are shown as a function of p for comparison. As previously discussed, the use of asymptotic critical values (based on a χ^2 distribution) induces important size distortions. For instance, with

Figure 1: Empirical size of the tests at 5% significance level (AR(1)-GARCH(1,1) model)



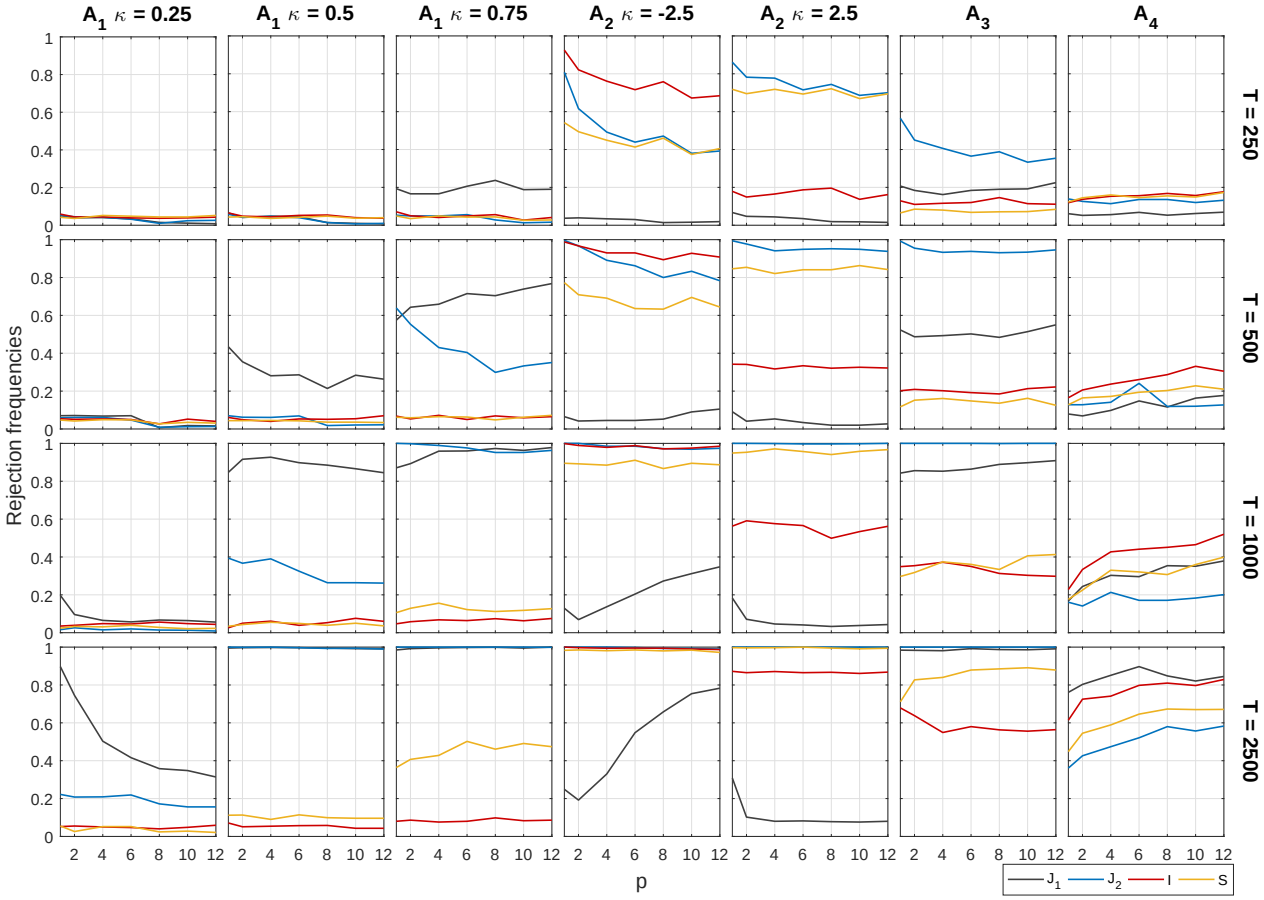
Note: Size of the four backtests are displayed as a function of p. The first row reports the results computed with the asymptotic critical values, and the second row those computed with the bootstrap critical values. The columns correspond to different sample sizes T.

sample size $T = 500$, and $p = 6$, the four test statistics J_1, J_2, I, and S, display empirical sizes equal to 0.126, 0.273, 0.165, 0.216, respectively. These distortions are caused by poor inference made on regression parameters in the extreme upper tail when the sample size is not sufficiently large. On the contrary, the backtests based on bootstrap critical values display empirical sizes that are close to the nominal size of 5% for all reported sample sizes and risk levels. For large coverage levels and moderate samples, we thus recommend to use bootstrap critical values rather than asymptotic ones.

In reply to questions concerning the link between p and T, we note that the size of the four bootstrap-based backtests slightly deteriorates for $p > 6$ with $T = 250$ and $T = 500$ revealing that the tests are sensitive to the choice of p in small samples. In details, the slope backtest is the most affected by these distortions, while the J_1 backtest is well-sized most of the time. On the contrary, for larger sample sizes, typically $T = 1000$ and $T = 2500$, these distortions are negligible. Our recommendation is hence to restrict the number p of quantiles when applying the tests in small samples, with typically $p \leq 6$, and to consider higher values if the historical sample covers longer periods.

To provide robustness check of these results, Figure 7 in Appendix F reports empirical sizes when the data generating process is given by a GARCH(1,1) model. We observe the same findings as those provided with the AR(1)-GARCH(1,1) model. The asymptotic tests are largely oversized, while the bootstrap tests are close to the nominal size of 5% for all reported sample sizes and risk levels. Finally, there is also an asymptotic refinement of the empirical sizes as T goes to infinity for both asymptotic and bootstrap tests.

Figure 2: Empirical power of the tests at 5% significance level (AR(1)-GARCH(1,1) model, bootstrap critical values)



Note: Power of the four backtests are displayed as a function of p. The rows correspond to different sample sizes T, and the columns to the different misspecified alternatives A_1-A_4. Reported powers are size corrected.

Figure 2 reports the empirical powers (size-corrected) associated with our seven alternatives. Here, we only present the simulation results associated with the bootstrap critical values. The simulation results obtained with the asymptotic critical values are overall the

same (see Figure 6 in Appendix F). Overall, the tests correctly detect the misspecified alternatives A_1, A_2, A_3, A_4, and we verify that there is a general improvement of powers as the sample size T increases (from row 1 to row 4), suggesting that these tests are consistent for these alternatives. For instance, with $T = 500$, and $p = 4$, the test statistic J_1 identifies the misleading scenario A_3 in 49.3% of times, while it reaches 98.1% of times with $T = 2500$.

Second, the joint test statistics, J_1 and J_2, generally deliver higher power performances compared to the intercept and slope test statistics I and S. This finding comes from the definition of the joint null hypotheses that focus on both intercept and slope coefficients and are thus more conservative than the null of the intercept and slope backtests. In details for the two joint tests, we find that J_1 performs generally better to detect A_1 and A_4, while J_2 more often identifies A_2 and A_3, which suggests complementarity between the two joint backtests. Although the intercept and slope backtests exhibit lower power performances, they provide useful informations on the type of misspecification. In details, the slope backtest performs better in alternatives A_1 and A_3, while the intercept backtest is superior for alternative A_4. Thus, A_1 and A_3 mainly affect the expected value of the slope parameters meaning that the errors are correlated and proportional to the true quantiles. In contrast, alternative A_4 induces distortions in the expected value of the intercept coefficients suggesting that the origin of errors is more global as they are not related to the true quantiles.

Third, we observe that the selection of the number p of risk levels is difficult to link with the rejection frequencies in alternatives A_1, A_2, A_3, since reported powers are slightly affected by p in general. This finding may be explained by the nature of these alternatives for which the misspecification is relatively uniform along the tail, and does not require many levels. On the contrary, in alternative A_4, we conclude that an increase of p is beneficial for detecting the misleading one-year historical simulation method as power is unequivocally increasing with p, especially when T is large. This is due to the fact that, for this alternative,

the error made along the tail is more irregular and requires the use of additional levels. Thus, it is helpful to consider, $p = 1, 2, \ldots, p_{\max}$, successively, with typically $p_{\max} = 12$ as provided above. This may come in handy for improving the statistical decision.

Finally, we provide a robustness check of the powers with the GARCH(1,1) model (see Figures 8 and 9 in Appendix F). The rejection frequencies are very close to those associated with the AR(1)-GARCH(1,1) model. Consequently, the decision whether to introduce or not a conditional mean in the risk model does not affect the power performances.

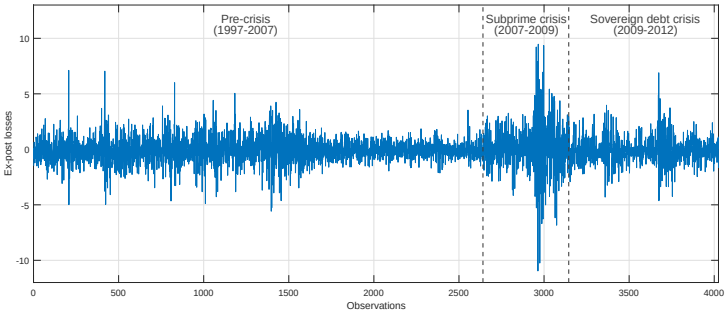
5 Empirical application

In this section, we apply our backtests to the daily returns of the S&P500 index. In addition, we provide a method for the adjustment of imperfect forecasts relying on our backtesting framework. In the sequel, we set $\tau = 0.975$ to coincide with the regulatory ES coverage level. The probability levels u_j, $j = 1, 2, \ldots, p$, are calculated accordingly with Definition 1. In addition, we consider the risk levels suggested by the BCBS, i.e. $u_1 = 0.975$, and $u_2 = 0.990$, respectively. Finally, for comparison purposes and to provide useful backtesting recommendations, we consider several values $p = 1, 2, 4, 6, 8, 10, 12$.

5.1 Data

We consider the daily adjusted closing prices of the S&P500 index over the period January 1, 1997 - December 31, 2012. The in-sample period spans from January 1, 1997 to June 30, 2007, and we use two out-of-sample periods (1) from July 1, 2007 to June 30, 2009, corresponding to the subprime mortgage crisis, and (2) from July 1, 2007 to December 31, 2012, which pools the subprime mortgage crisis and the European sovereign debt crisis, two major episodes of financial instability. We compute the daily log-returns and denote by L_t the opposite returns. In line with our notations, a positive value indicates a loss.

Figure 3: S&P500 daily losses (%), and descriptive statistics



	1997-2007	2007-2009	2007-2012
Nb of obs	2641	504	1384
Mean	-0.028	0.105	0.005
Median	-0.059	-0.032	-0.067
Variance	1.280	4.932	2.643
Skewness	0.090	0.059	0.242
Kurt. exc.	3.171	4.028	6.943
Minimum	-5.574	-10.957	-10.957
Maximum	7.113	9.470	9.470

Note: The sample covers the period from January 1, 1997 to December 31, 2012. Source: *finance.yahoo.com* website.

The S&P500 series is depicted in Figure 3 with the three aforementioned sub-periods. The in-sample period (1997-2007) is weakly volatile, while the out-of-sample crisis periods (2007-2009 and 2007-2012) display more severe levels of volatility, with several extreme events. Figure 3 also provides some descriptive statistics. The variance and the average ex-post losses are higher in the out-of-sample periods than in the in-sample period, especially for the period 2007-2009. In addition, the series is right-skewed and has a kurtosis excess.

To predict the ES risk measure, we fit an AR(1)-GARCH(1,1) model with Student innovations, as defined in (11), using the S&P500 daily losses of the in-sample period. The ES and VaR forecasts are defined as in Equations (12) and (13), respectively. The set of unknown parameters is estimated by maximum likelihood. We obtain the following coefficient estimates $\left\{\widehat{\delta}_0, \widehat{\delta}_1, \widehat{\gamma}_0, \widehat{\gamma}_1, \widehat{\gamma}_2, \widehat{v}\right\} = \{-0.057, -0.032, 0.007, 0.060, 0.936, 9\}$. As a robustness check, we also fit a GARCH(1,1) model on the same period as defined in the simulation study and for which we obtain the following estimates $\{\widehat{\gamma}_0, \widehat{\gamma}_1, \widehat{\gamma}_2, \widehat{v}\} = \{0.007, 0.059, 0.937, 9\}$.

5.2 Empirical results

We start by evaluating the relevancy of the ES approximation of Definition 1, consisting in averaging several quantiles in the tail of the risk model. To do so, we compare the approximation considering $p = 1, 2, 4, 6, 8, 10, 12$ quantiles, with what we refer to as "exact ES". The latter corresponds to an ES which is computed via an exact method of calculation.

The technique relies on simulations and is described in Appendix G.

Figure 4: In-sample ES estimates issued from the approximation and the exact calculation method (AR(1)-GARCH(1,1) model)

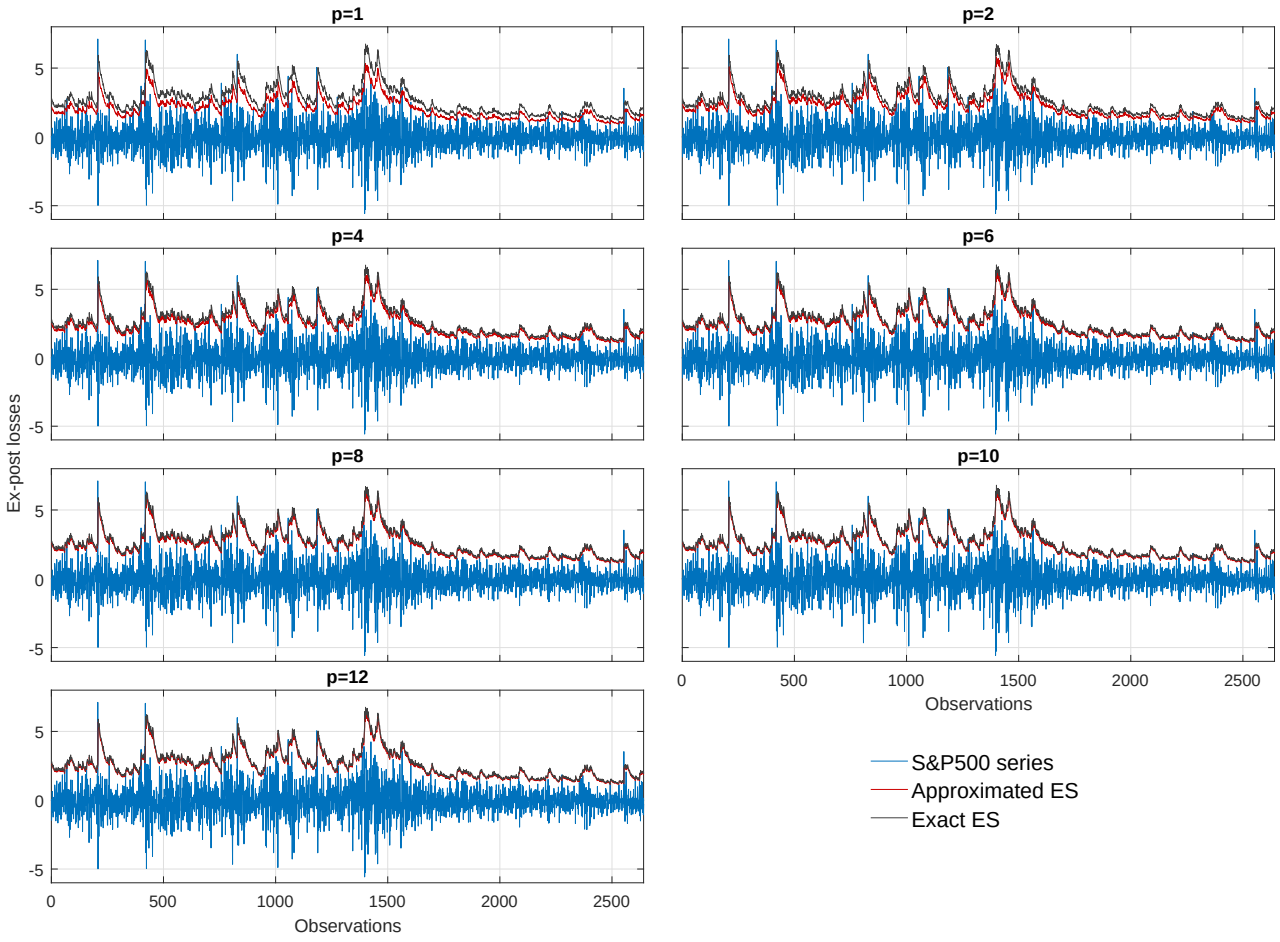


Figure 4 reports the in-sample ES estimates obtained with the approximation and the exact calculation method. Two remarks should be made here. First, the ES forecasts issued from the approximation and the exact method strongly correlate regardless of the value p. The approximation performs very well to capture the ex-post losses information. Second, we observe that the approximation is substantially improved when p slightly increases and coincides almost completely with the exact ES using six (or more) quantiles.

Because the approximation is obtained by combining VaRs, our finding is in accordance with several papers. Gouriéroux and Liu (2012) study the relationship between VaR and ES and show that they are related through their risk levels by some link function. Daníelsson

and Zhou (2016) argue that the two measures of risk are related by a small constant and are conceptually equally informative. This similarity also comes from the structure of the model used to compute the risk measure. For instance, VaR and ES issued by an AR(1)-GARCH(1,1) model have common conditional mean and variance across risk levels implying that these risk measures are closely related (see Equations (12) and (13)). Finally, Figure 10 in Appendix H displays the same results using a GARCH(1,1) model. Removing the conditional mean component does not affect the approximation accuracy as the two computation methods match almost perfectly for $p \geq 6$. For its ease of implementation and accuracy, the approximation is appealing to compute and evaluate the performance of ES forecasts.

Table 1: p-values of the backtesting tests (AR(1)-GARCH(1,1) model)

p	$J_1^{(b)}$	$J_2^{(b)}$	$I^{(b)}$	$S^{(b)}$
		Panel A. 2007-2009		
1	0.035	0.051	0.125	0.949
2	0.014	0.041	0.038	0.200
4	0.009	0.040	0.023	0.103
6	0.009	0.038	0.021	0.123
8	0.099	0.049	0.154	0.564
10	0.029	0.061	0.053	0.432
12	0.023	0.052	0.038	0.223
2 *(regulatory levels)*	0.024	0.047	0.053	0.351
		Panel B. 2007-2012		
1	0.056	0.040	0.176	0.554
2	0.004	0.013	0.014	0.215
4	0.002	0.004	0.003	0.096
6	0.004	0.005	0.009	0.196
8	0.008	0.008	0.041	0.538
10	0.007	0.010	0.021	0.410
12	0.004	0.006	0.008	0.245
2 *(regulatory levels)*	0.006	0.012	0.032	0.448

Note: p-values of the four backtests computed with $p = 1, 2, 4, 6, 8, 10, 12$ risk levels successively, and the two regulatory levels $u_1 = 0.975$, $u_2 = 0.990$. Reported p-values are obtained using bootstrap critical values. Panel A gives the results for the period 2007-2009 and Panel B provides results for the period 2007-2012.

Table 1 reports the p-values of the backtests. For a sake of clarity, we only report the p-values obtained with the bootstrap critical values and the results are discussed at 5% significance level. Panel A provides the results over the sample 2007-2009. The test statistic J_1 leads to reject the validity of the ES predictions regardless of the number p of quantiles (except for $p = 8$ where the rejection occurs at a 10% significance level). Interestingly, we

observe that the larger p, the smaller the p-value until $p = 6$, indicating that the rejections are more severe when the number of risk levels increases until an optimal number p. This supports the existence of an upper limit for p which depends on the sample size since T is relatively small ($T = 504$), and thus, p should not be chosen too large. The test statistic J_2 displays higher p-values in general. The backtest based on a single VaR no longer rejects the validity of the ES predictions, and the p-value based on the regulatory levels of the BCBS is close to 5%, making the decision rule more unclear for those number of risk levels. Finally, given the p-values of the test statistic I for $p = 2, 4, 6, 12$, we tend to reject the expected value on the intercept coefficients, and as a result, there is a global bias in the quantile estimates issued by the ES model. On the contrary, the test statistic S leads to the conclusion that the slope parameters are as expected under the null hypothesis, and thus, the magnitude of errors is not related to the true quantiles. Panel B contains the p-values for the period 2007-2012. Overall, we obtain similar results, but the rejections are found more severe in this enlarged sample. Interestingly, the rejections of J_1 are now experienced at a 1% significance level and even for $p > 6$, as opposed to panel A. This highlights the underlying link between p and T as panel B uses $T = 1384$ observations enabling a larger number p of quantiles to be used. Table 3 of Appendix H displays the p-values of the backtests when applying a GARCH(1,1) model. The results are similar. Note however, for $p = 1$, that the p-values are generally higher with the GARCH(1,1) model than for the AR(1)-GARCH(1,1) model. For instance, the p-value of the statistic J_1 in panel B equals 0.056 with the AR(1)-GARCH(1,1) model, while it reaches 0.199 with the GARCH(1,1) model. For that model, additional quantiles are indicated to increase the rejection abilities of the tests.

In sum, we should be cautious in using a single quantile to assess the tail distribution of the risk model. Such procedures may lead market practitioners to select a model that generates mistaken ex-post forecasts. Furthermore, the results issued from the regulatory

guidelines are contrasted. Two risk levels are not always enough to provide a sound conclusion about the correctness of the ES forecasts. We recommend the use of additional risk levels beyond the regulatory coverage level $\tau = 0.975$ to improve the reliability of the decision.

Table 2: QML coefficient estimates ($p = 6$, AR(1)-GARCH(1,1) model)

	u_1	u_2	u_3	u_4	u_5	u_6
			Panel A. 2007-2009			
β_0	0.661	0.696	$0.808\ {}^{\star\star}_{\circ\circ\circ}$	$0.846\ {}^{\star\star}_{\circ\circ\circ}$	$0.965\ {}^{\star}_{\circ\circ\circ}$	$1.076\ {}^{\star}_{\circ\circ\circ}$
	(0.295)	(0.296)	(0.227)	(0.240)	(0.429)	(0.265)
β_1	1.005	0.953	$0.911\ {}^{\star}$	$0.847\ {}^{\star\star}_{\circ\circ\circ}$	0.804	$0.689\ {}^{\star\star}_{\circ\circ\circ}$
	(0.093)	(0.088)	(0.056)	(0.053)	(0.142)	(0.042)
joint	*	*	**	**		**
			Panel B. 2007-2012			
β_0	0.376	$0.510\ {}^{\star}$	$0.692\ {}^{\star\star\star}_{\circ\circ\circ}$	$0.808\ {}^{\star\star\star}_{\circ\circ\circ}$	$0.777\ {}^{\star\star}_{\circ\circ\circ}$	0.784
	(0.200)	(0.182)	(0.195)	(0.186)	(0.284)	(0.611)
β_1	1.031	0.974	0.902	$0.851\ {}^{\star\star}$	0.826	0.787
	(0.073)	(0.067)	(0.065)	(0.050)	(0.107)	(0.232)
joint	**	**	**	**	**	

Note: Standard errors are reported in parentheses. *, **, and *** indicate statistical significance at the 10%, 5% and 1% level, respectively, and are obtained with the pairs bootstrap algorithm. $^{\circ}$, $^{\circ\circ}$, and $^{\circ\circ\circ}$, indicate statistical significance at the same levels and are obtained with the procedure of Chernozhukov and Fernández-Val (2011). Panel A gives estimation results for the period 2007-2009 and Panel B provides estimation results for the period 2007-2012.

Table 2 displays the coefficient estimates of the multi-quantile regression of Equation (4) for $p = 6$ risk levels, to help understand the reasons that explain the rejections of the ES forecasts. Panel A and B provide the results for periods 2007-2009 and 2007-2012, respectively. It must be recalled that, if the risk model is correctly specified, the intercept coefficient β_0 and the slope coefficient β_1 take values zero and one, respectively. We observe in both panels that the coefficients β_0 are overestimated for all the risk levels $u_1, u_2, \ldots, u_6$, while the coefficient β_1 is overestimated for the first level u_1, and it becomes underestimated for all the remaining risk levels $u_2, u_3, \ldots, u_6$. The average errors of β_0 and β_1 are respectively equal to 0.84 and -0.13 in panel A, and 0.66 and -0.10 in panel B, indicating that the magnitude of errors is more important in panel A than in panel B, and that the intercept coefficients are more affected than the slope coefficients. Finally, we observe that the distortion of the regression coefficients with respect to their expected values is more pronounced for the

highest risk levels suggesting that the errors are more severe far in the tail.

Furthermore, we provide in Table 2 one by one inference on the regression parameters with the pairs bootstrap algorithm. The results are depicted with the symbol "$*$" and are discussed at a 5% significance level. We observe that the intercept parameters are statistically not equal to zero for the intermediary levels u_3 and u_4 in panel A, and the additional u_5 risk level is also significantly different from zero in panel B. For the slope coefficients, the u_4 and u_6 order quantiles are statistically different from one in panel A, and only the level u_4 is misspecified in panel B. In addition, we report joint inference, i.e. looking at both the intercept and slope coefficients. The results are provided in the row labeled as "joint" (bottom of the panels). Similarly to the previous findings, we find that the intermediary, and highest order quantiles u_3, u_4 and u_6 are misleading in panel A, whereas in panel B, all the quantiles are misspecified (except for the highest, presumably because the coefficients have large standard errors), meaning that the entire tail distribution is incorrectly estimated.

Next, we propose a variety of robustness checks to our baseline estimation method. Several alternatives to the QML estimator (White et al., 2008, 2015) and pairs bootstrap (Freedman, 1981) are available and should be regarded as well. In the sequel, we suggest a number of avenues to be explored. First, we apply the procedure of Chernozhukov and Fernández-Val (2011) based on the extreme value theory (EVT) that allows testing individual restrictions. The results are depicted in Table 2 with the symbol "$\circ$". Overall, we find similar results between EVT and pairs bootstrap. Rejection of the null is mostly experienced at the same levels in panel A and panel B. However, we observe that the procedure of Chernozhukov and Fernández-Val (2011) is generally more powerful than pairs bootstrap at the highest risk levels. For instance, the expected value of the intercept parameters $\beta_0(u_j)$ in panel A is rejected at level 1% for $j = 3, 4, 5, 6$ with the EVT procedure, while the pairs bootstrap rejects the null at larger levels (5% or 10%). This illustrates the superiority of EVT in

multi-quantile regression models. A robustness check of these results is provided with the GARCH(1,1) model where we overall get the same results (see Table 4 in Appendix H).

Second, we apply the seemingly unrelated regressions (SUR) estimation method for quantile models of Jun and Pinkse (2009). The procedure is designed to improve estimation efficiency in presence of correlated generalized errors. In our framework, the risk levels u_j, $j = 1, 2, \ldots, p$, are closed to each other, and the correlation may be important between different quantiles. Consequently, the QML estimation method may loose a lot in accuracy. In the sequel, we compute the backtesting tests based on the quantile regression parameter estimates of Jun and Pinkse (2009). Table 5 of Appendix H reports the corresponding bootstrap p-values. The test statistics J_1, J_2, I, lead to reject the validity of the ES estimates while the statistic S does not. These findings are similar with those of the QML estimates (see Table 1). Our conclusions are not affected by the choice of the estimation method (SUR vs. QML estimation). Table 6 of Appendix H displays the coefficient estimates computed with the SUR-estimation procedure. We observe that the SUR estimates $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are close to the QML ones, explaining why our conclusions of our backtests are robust. Second, we note that the asymptotic standard errors of the SUR-estimator are not always close to those of the QML estimation method. The largest differences occur at risk levels u_5 for the slope parameter (in panels A and B) and u_6 for both parameters (in panel B), where standard errors are typically more than twice lower than those of the QML estimator. This confirms the efficiency improvement achieved by the SUR-estimator. Our conclusions with the GARCH(1,1) model are the same (see Tables 7 and 8 in Appendix H).

5.3 Adjusted ES forecasts

In what follows, we exploit our testing strategy to provide adjusted ES forecasts. Our routine is designed to take into account both misspecification and estimation uncertainty,

without having to change the misspecified risk model. Furthermore, the procedure may serve to identify whether the model overestimates, or underestimates the true unknown ES, by comparing the initial forecast with its adjusted counterpart, which appears useful in a risk management and regulatory viewpoint.

The correction of imperfect risk forecasts is not a novel concept in the financial literature. Gouriéroux and Zakoïan (2013) propose to adjust the VaR forecasts affected by estimation uncertainty. Similarly, Boucher et al. (2014) adjust imperfect VaR forecasts based on backtesting frameworks, and recently Lazar and Zhang (2019) apply the same strategy to adjust imperfect ES forecasts. The method typically consists in modifying the coverage level τ of the risk measure so as to meet the null hypothesis of valid risk forecasts. The originality of our technique stems from the fact that we employ a regression-based framework to correct the ex-ante forecasts, while available techniques are generally based on the concept of violation. This allows us to directly adjust the risk forecasts by application of a regression model, without having to rescale the coverage level τ.

For ease of notation, we assume the parameters of the multi-quantile regression to be known. Formally, the adjusted VaR forecast at level u_j, and time t, is defined as the ex-ante prediction of the multi-quantile regression model, namely $Q_{L_t}\left(u_j; \Omega_{t-1}\right)$. In view of Equation (5), the initial imperfect VaR forecast is subsequently weighted by the regression parameters $\beta_0\left(u_j\right)$ and $\beta_1\left(u_j\right)$, which provides an adjustment corresponding to the global bias caused by misspecification and estimation uncertainty. The adjusted ES forecast at coverage level τ and time t is derived from the ES approximation as follows:

$$ES_t^*\left(\tau\right) = \frac{1}{p}\sum_{j=1}^{p} Q_{L_t}\left(u_j; \Omega_{t-1}\right).$$

The adjusted ES forecasts are robust to model risk, as they meet the desirable properties on the regression coefficients. Indeed, if we compute the backtesting procedure with the

sequence $\{Q_{L_t}(u_j;\Omega_{t-1})\}_{j=1}^{p}$ instead of the initial misleading $\{VaR_t(u_j)\}_{j=1}^{p}$, the parameters would exactly coincide with the expected values under the null hypothesis, i.e. $\beta_0(u_j)=0$, and $\beta_1(u_j)=1$, for the risk levels $u_1,u_2,\ldots,u_p$.

Figure 5: ES forecasts and adjusted ES forecasts over the period 2007-2009 (AR(1)-GARCH(1,1) model)

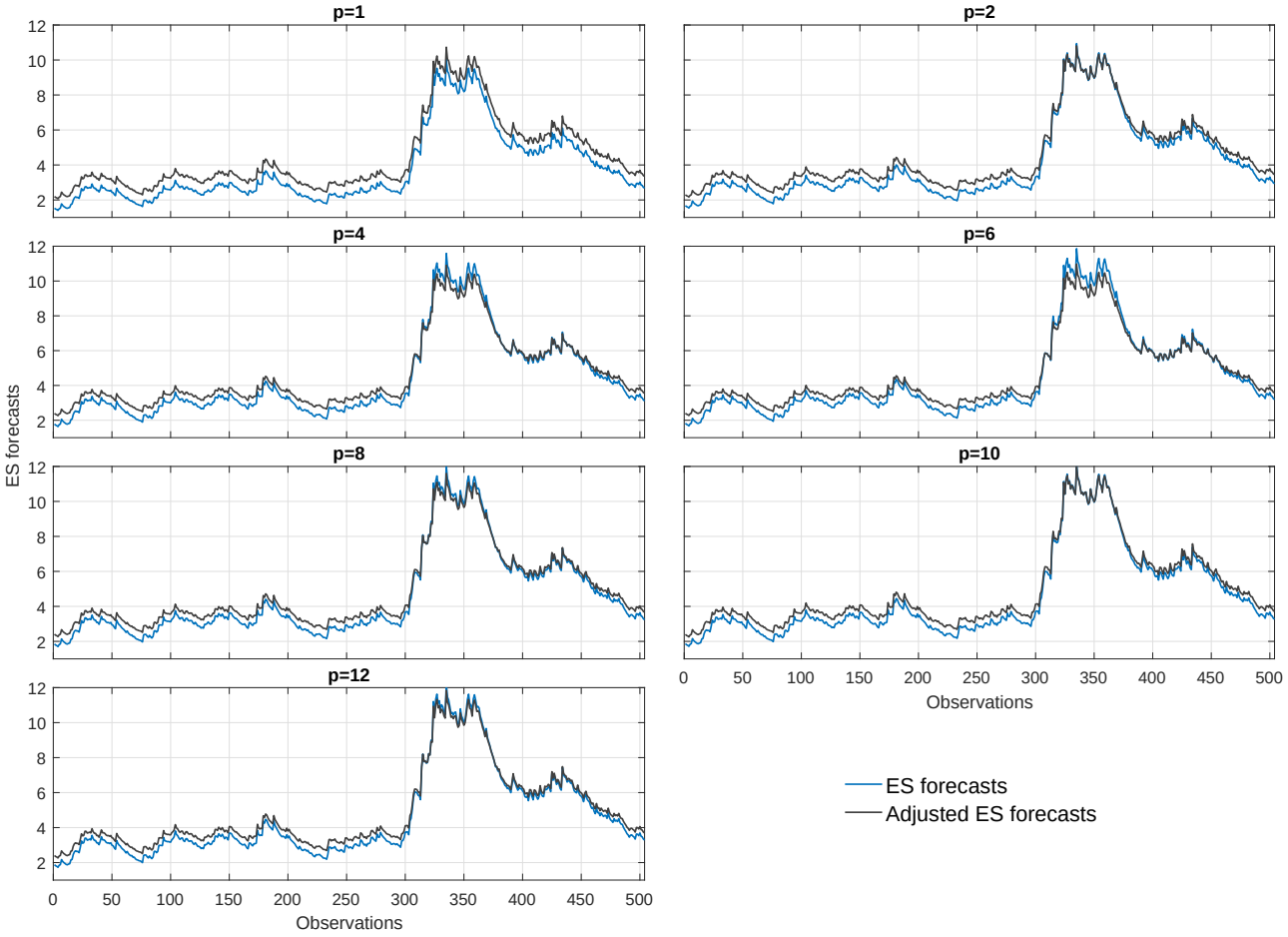


Figure 5 reports the ES predictions and adjusted ES predictions for the period 2007-2009. The forecasts are built using the approximation with $p=1,2,4,6,8,10,12$. We observe that the AR(1)-GARCH(1,1) model generally provides underestimated forecasts compared to the adjusted predictions. The underestimation is more pronounced for the smallest predictions, the error being more severe when the risk forecasts are originally small. Thus, our procedure serves at identifying whether the model generates overestimates or underestimates, the latter case being more worrisome in a financial stability perspective. Finally, the ES forecasts are

slightly overestimated when the variance of the series is larger, suggesting that the risk model may overestimate the true volatility in turbulent financial times. This is due to the volatility persistence in the GARCH component. Our findings are robust to (1) the use of a simple GARCH(1,1) model, (2) the use of the two BCBS regulatory levels, and (3) the extended period 2007-2012 (see Figures 11, 12, 13, and 14 in Appendix H).

6 Conclusion

The financial crisis of 2007-2008 and its aftermath has led to a reassessment of risk-management practices and financial market regulation through the Basel III accords (BCBS, 2010). Among the number of fundamental reforms for the market risk, the BCBS has adopted ES in place of VaR as the new standard for risk management. One of the major obstacle to its implementation was the deficit of simple tools for the evaluation of its forecasts. This article introduces four easy-to-use regression-based backtests of ES. Our econometric approach consists in regressing the ex-post losses on the VaRs forecasts in a multi-quantile regression model, and then, testing the resulting parameter estimates using Wald-type inference.

Several simulation studies are provided. We find that the use of asymptotic critical values may lead to important size distortions if the sample size is not large enough. We propose a pairs bootstrap algorithm to correct these small-sample biases (Freedman, 1981) and show that our regression-based tests are reasonably sized within this bootstrap framework. We consider several misleading alternatives in line with the existing literature on risk assessment (Gaglianone et al., 2011; Du and Escanciano, 2017; Bayer and Dimitriadis, 2019; Kratz et al., 2018, etc.). Our methodology detect misspecifications in all considered simulation experiments. In particular, they identify the most frequent inaccuracies in risk modeling, namely mean, variance, tail, and dynamic misspecifications.

We apply our tests on the S&P500 index over the period 2007-2012. During this period

of financial turmoil, our backtests clearly reject the validity of the ES forecasts based on a AR(1)-GARCH(1,1) and a GARCH(1,1) model. We also highlight the importance of choosing a sufficient number of quantiles to assess ES. The use of one or two quantiles is inadvisable as they are not always enough to identify improper risk forecasts. On the contrary, four or more quantiles (until an optimal number) deliver much more sound decisions, suggesting an update of the regulatory guidelines to the evaluation of more than two quantiles.

7 Acknowledgments

We would like to thank for their valuable comments Denisa Banulescu-Radu, Ansgar Belke, Sylvain Benoit, Massimiliano Caporin, Christian Francq, Sullivan Hué, Christophe Hurlin, Sébastien Laurent, Yang Lu, Richard Luger, Štefan Lyócsa, Mathias Pohl, Daniel Platte, Luca Riccetti, Olivier Scaillet, Sessi Tokpavi, Jean-Michel Zakoïan, and two anonymous referees. We also thank the participants of the 17th annual conference "Développements Récents de l'Econométrie Appliquée à la Finance" (University of Nanterre, Paris), 12th international conference on Computational and Financial Econometrics (University of Pisa, Italy), 12th Financial Risks International Forum (Paris, France), workshop in Financial Econometrics (University of Nantes, France), 8th PhD student conference in International Macroeconomics and Financial Econometrics (University of Nanterre, Paris), 7th spring conference of the Multinational Finance Society (Chania, Greece), 5th international conference on Applied Theory, Macro and Empirical Finance (Thessaloniki, Greece), workshop ANR MultiRisk (Florence, Italy), 9th international conference of the Financial Engineering and Banking Society (Prague, Czech Republic), 4th international workshop on "Financial Markets and Nonlinear Dynamics" (Paris, France), 2nd Quantitative Finance and Financial Econometrics international conference (Aix-Marseille School of Economics, France), 2019 INFINITI conference on International Finance (Adam Smith Business School, Scotland), 12th

annual pre-conference of the Society for Financial Econometrics (Shanghai, China), 36th international conference of the French Finance Association (University of Laval, Canada), 6th annual conference of the International Association for Applied Econometrics (Nicosia, Cyprus), 72nd European Meeting of the Econometric Society (Manchester, UK), and 26th Annual Meeting of the German Finance Association (Essen, Germany). We thank the ANR MultiRisk (ANR-16-CE26-0015-01) for supporting our research.

Appendix

A - Application of a finite Riemann sum to ES

In the sequel, we show how to derive the approximation of ES suggested in Definition 1. Consider the following improper Riemann integral,

$$\int_a^b f(t)dt, \tag{14}$$

where $f(.)$ is given by the increasing function $\frac{1}{1-\tau}VaR_t(.)$ and where a and b are respectively τ and 1 so that the above expression is identical to the ES defined in Equation (1). Definition of a Riemann sum yields a useful approximation of Equation (14),

$$S_p(f) = \frac{b-a}{p}\sum_{j=1}^{p} f\left(a + (j-1)\frac{b-a}{p}\right),$$

where p is the number of subdivisions or quantiles taken in the definite integral to approximate ES. Replacing a, b, and $f(.)$, by their corresponding quantities leads,

$$\frac{1}{1-\tau}\int_\tau^1 VaR_t(u)du \approx \frac{1}{1-\tau}S_p(VaR_t) = \frac{1}{p}\sum_{j=1}^{p} VaR_t\left(\tau + (j-1)\frac{1-\tau}{p}\right).$$

This verifies the ES formula of Definition 1 where risk levels u_j are given by $\tau + (j-1)\frac{1-\tau}{p}$.

B - Assumptions

This section introduces the assumptions needed to establish the asymptotic normality and the consistency of the QML estimator and to ensure the validity of Proposition 1.

Assumption A0: $\{L_t, VaR_t(u_j)\}_{j=1}^{p}$ is a stationary and ergodic process and measurable with respect to Ω_{t-1}.

Assumption A1: L_t has conditional (on Ω_{t-1}) distribution function F_t, with continuous and positive density f_t at conditional quantile $Q_{L_t}(u;\Omega_{t-1}) = F_t^{-1}(u|\Omega_{t-1})$ for all $u \in (0,1)$.

Assumption A2: We have $\mathbb{E}\left[|L_t|\right] < \infty$. Furthermore, consider the quantity $D_{0,t} = \max_{t=1,\ldots,T} \max_{j=1,\ldots,p} \sup|Q_{L_t}\left(u_j;\Omega_{t-1}\right)|$, then we have $\mathbb{E}\left[D_{0,t}\right] < \infty$.

Assumption A3: The matrices $A = \sum_{j=1}^{p} \mathbb{E}\left[f_{j,t}\left(0\right)\nabla Q_{L_t}\left(u_j;\Omega_{t-1}\right)\nabla' Q_{L_t}\left(u_j;\Omega_{t-1}\right)\right]$ and $V = \mathbb{E}\left[\eta_t\eta_t'\right]$ are positive definite.

Assumption A0 is standard in modeling financial times series. It is broadly accepted that asset prices are integrated at order one, so that financial returns are stationary. This data assumption is hence satisfied. Assumption A1 allows for nonidentical distributions as we enable L_t to be conditional on an unknown information set Ω_{t-1}. Assumption A2 imposes moment conditions, and in particular ensures finite expectation for L_t. This is satisfied by the vast majority of financial time series models, including stationary and invertible ARMA processes, GARCH processes, etc. Assumption A3 is standard in Wald-type inference to ensure that the variance-covariance matrix Σ is positive definite. Furthermore, Assumptions A0 through A2 are standard in QML estimation (e.g., White, 1994), and are also widely used in the literature on quantile regression models (e.g., Koenker and Machado, 1999; Koenker and Xiao, 2002). They are of great importance to establish consistency and to apply the central limit theorem of White (2001, theorem 5.24) based on the method proposed by Huber (1967).

C - Consistent variance-covariance matrix estimation

In what follows, we provide a consistent estimator of the variance-covariance matrix Σ. The methodology is derived from White et al. (2008, 2015). A consistent estimate of Σ can be obtained from the decomposition of the Huber (1967) sandwich form and is thus given by $\widehat{\Sigma} = \widehat{A}^{-1}\widehat{V}\widehat{A}^{-1}$. In the sequel, we provide consistent estimators $\widehat{A}$ and $\widehat{V}$. To obtain $\widehat{V}$, we

apply a simple plug-in estimator as follows:

$$\widehat{V} = T^{-1} \sum_{t=1}^{T} \widehat{\eta}_t \widehat{\eta}_t',$$

where $\widehat{\eta}_t$ is given by its estimated counterpart $\widehat{\eta}_t = \sum_{j=1}^{p} \nabla \widehat{Q}_{L_t}(u_j, \Omega_{t-1}) \psi_{u_j}(\widehat{\epsilon}_{j,t})$, with $\widehat{Q}_{L_t}(u_j, \Omega_{t-1}) = \widehat{\beta}_0(u_j) + \widehat{\beta}_1(u_j) VaR_t(u_j)$, and $\widehat{\epsilon}_{j,t} = L_t - \widehat{Q}_{L_t}(u_j, \Omega_{t-1})$.

The estimation of A is trickier because it requires to consistently estimate $f_{j,t}(0)$, namely the density of the error term $\epsilon_{j,t}$ given Ω_{t-1} evaluated at zero. Because the function is unknown, we follow Powell (1984) and use a non parametric estimator. The method was also implemented by Engle and Manganelli (2004) to estimate the variance-covariance matrix of a set of coefficients issued from the so-called CaViaR model. Then, $\widehat{A}$ is given by

$$\widehat{A} = (2\widehat{c}_T T)^{-1} \sum_{t=1}^{T} \sum_{j=1}^{p} \mathbb{1}(|\widehat{\epsilon}_{j,t}| \leq \widehat{c}_T) \nabla \widehat{Q}_{L_t}(u_j, \Omega_{t-1}) \nabla' \widehat{Q}_{L_t}(u_j, \Omega_{t-1}),$$

where $\widehat{c}_T$ is a bandwidth parameter that must verify $\widehat{c}_T / c_T \xrightarrow{p} 1$, with c_T a nonstochastic positive sequence satisfying $c_T = o(1)$, and $c_T^{-1} = o(T^{1/2})$. Throughout the paper, we select a bandwidth parameter $\widehat{c}_T = T^{-1/7}$ which verifies the above properties.

D - On the rate of convergence and interplay of p and T

Let us consider the highest risk level u_p issued from the sequence $u_j, j = 1, 2, \ldots, p$. Given Definition 1, we have

$$u_p = 1 - (1 - \tau)/p,$$

where p is the number of VaRs used to approximate ES and τ is a constant number representing the coverage level of ES. For a number p of subdivisions large enough, the approximation of Definition 1 is close to the theoretical ES and u_p is close to one. In what follows, we study

this limiting case. Let us define u_p as a function of the sample size T such as,

$$u_p = 1 - \epsilon_T, \tag{15}$$

where the nonstochastic positive sequence $\epsilon_T = (1 - \tau)/p$ satisfies $\epsilon_T \to 0$ when $T \to \infty$. Equation (15) is a common representation to extremal quantile regression (Chernozhukov, 2005; Chernozhukov and Fernández-Val, 2011). Given the definition of ϵ_T and since τ is a constant parameter, it follows that p is increasing with T. To illustrate this point, assume p takes the form of a power function,

$$p = T^{\gamma}, \tag{16}$$

with $\gamma > 0$. Next, consider the estimation procedure of White et al. (2008, 2015) which implicitly assumes that $T(1-u_p) \to \infty$ as T goes to infinity. Chernozhukov (2005) and Chernozhukov and Fernández-Val (2011) relate this condition to an intermediate order quantile regression. Our goal is to identify a suitable rate of convergence of p and T which ensures the above condition. We have

$$T\epsilon_T \to \infty. \tag{17}$$

Then, combining Equations (16) and (17), we get

$$(1 - \tau)T^{1-\gamma} \to \infty. \tag{18}$$

Equation (18) is only satisfied when $\gamma < 1$. Looking at Equation (16), this condition implies that T needs to diverge faster than p to guarantee the asymptotic theory of White et al. (2008, 2015).

E - Proof of consistency under fixed untrue hypothesis

Proof. In line with our previous notations, we term W the generic notation of the test statistic such that $W \in \{J_1, J_2, I, S\}$. The test statistic is given by

$$W = T(R_W\widehat{\beta} - q_W)'(R_W\widehat{\Sigma}R_W')^{-1}(R_W\widehat{\beta} - q_W).$$

The null hypothesis of the proposed Wald-type test can be written as $H_{0,W} : R_W\beta - q_W = 0$, against the two-sided alternative $H_{1,W} : R_W\beta - q_W \neq 0$. The continuous mapping theorem implies under $H_{1,W}$ that

$$R_W\widehat{\beta} - q_W \xrightarrow{p} R_W\beta - q_W \neq 0.$$

Rearranging the term T in the test statistic and using the continuous mapping theorem leads

$$WT^{-1} \xrightarrow{p} (R_W\beta - q_W)'(R_W\Sigma R_W')^{-1}(R_W\beta - q_W).$$

Because $(R_W\Sigma R_W')^{-1}$ is positive definite, we get under $H_{1,W}$: $(R_W\beta - q_W)'(R_W\Sigma R_W')^{-1}(R_W\beta - q_W) > 0$. Multiplying $(R_W\beta - q_W)'(R_W\Sigma R_W')^{-1}(R_W\beta - q_W)$ by T under $H_{1,W}$ hence gives

$$\lim_{T\to+\infty} W = +\infty,$$

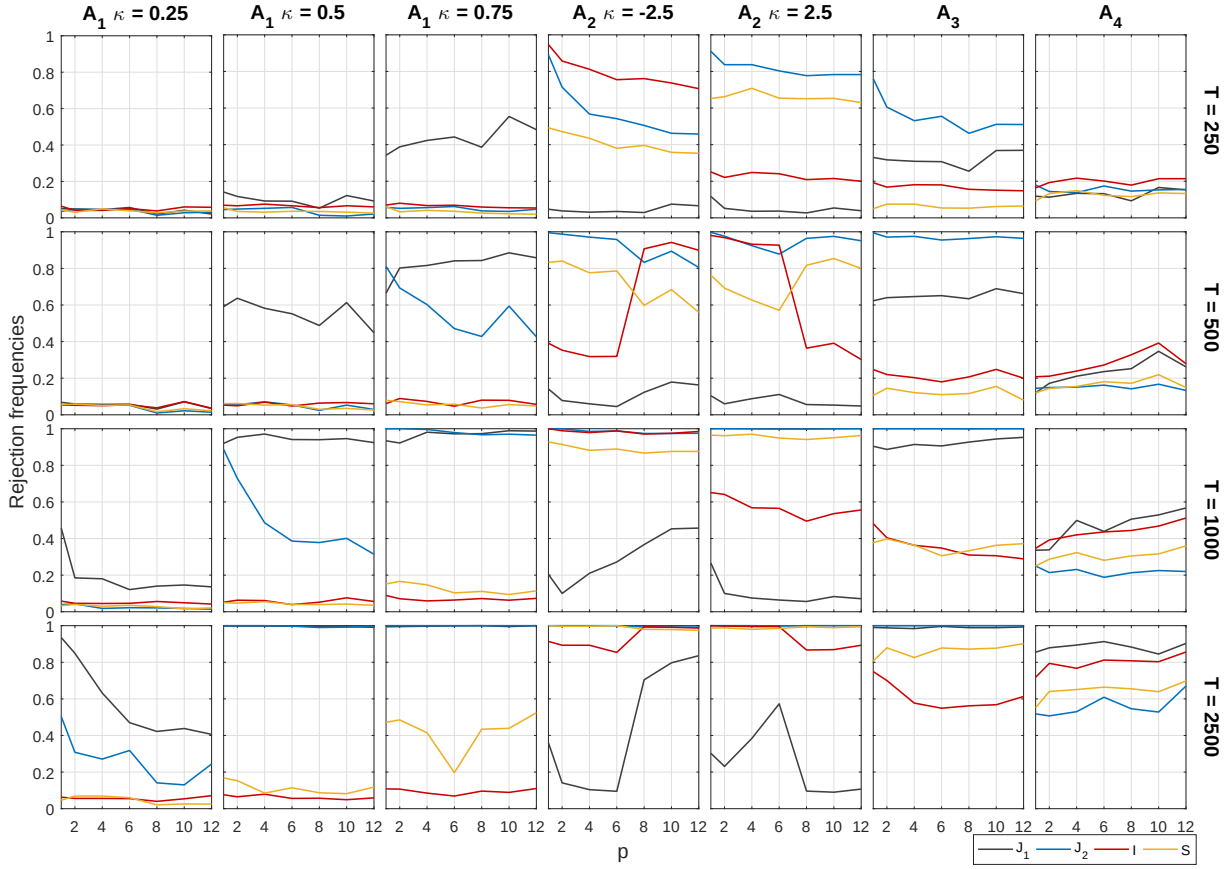
and therefore we get

$$\lim_{T\to+\infty} \mathbb{P}(W > \chi^2_{1-\alpha}(d_W) \,|H_{1,W}) = 1,$$

where $\chi^2_{1-\alpha}(d_W)$ is the fractile of order $1-\alpha$ of the chi-square distribution with d_W degrees of freedom, and where α is the significance level of the test. ■

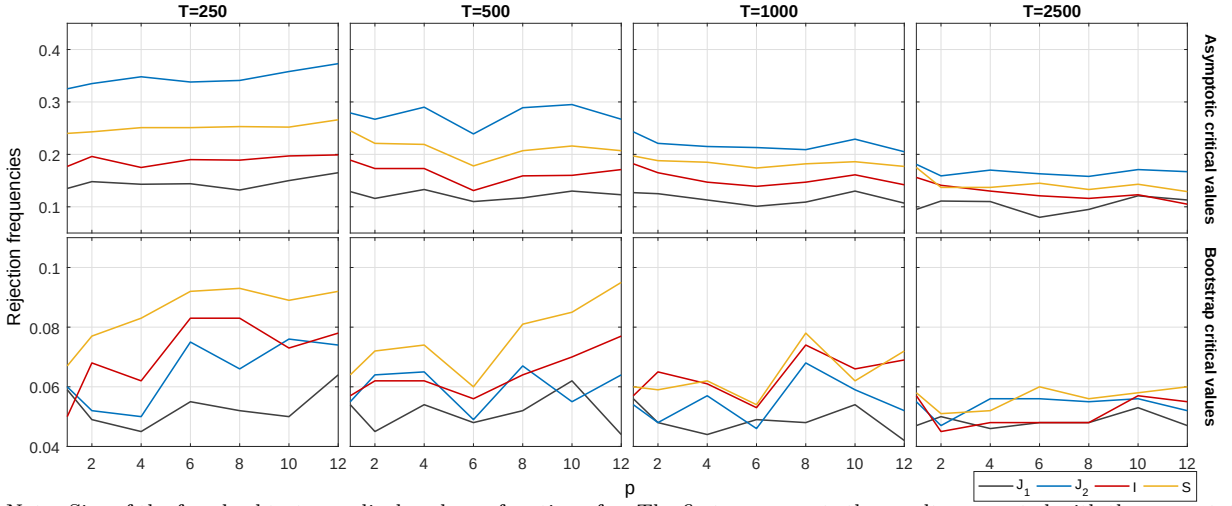
F - Robustness checks of the simulation study

Figure 6: Empirical power of the tests at 5% significance level (AR(1)-GARCH(1,1) model, asymptotic critical values)



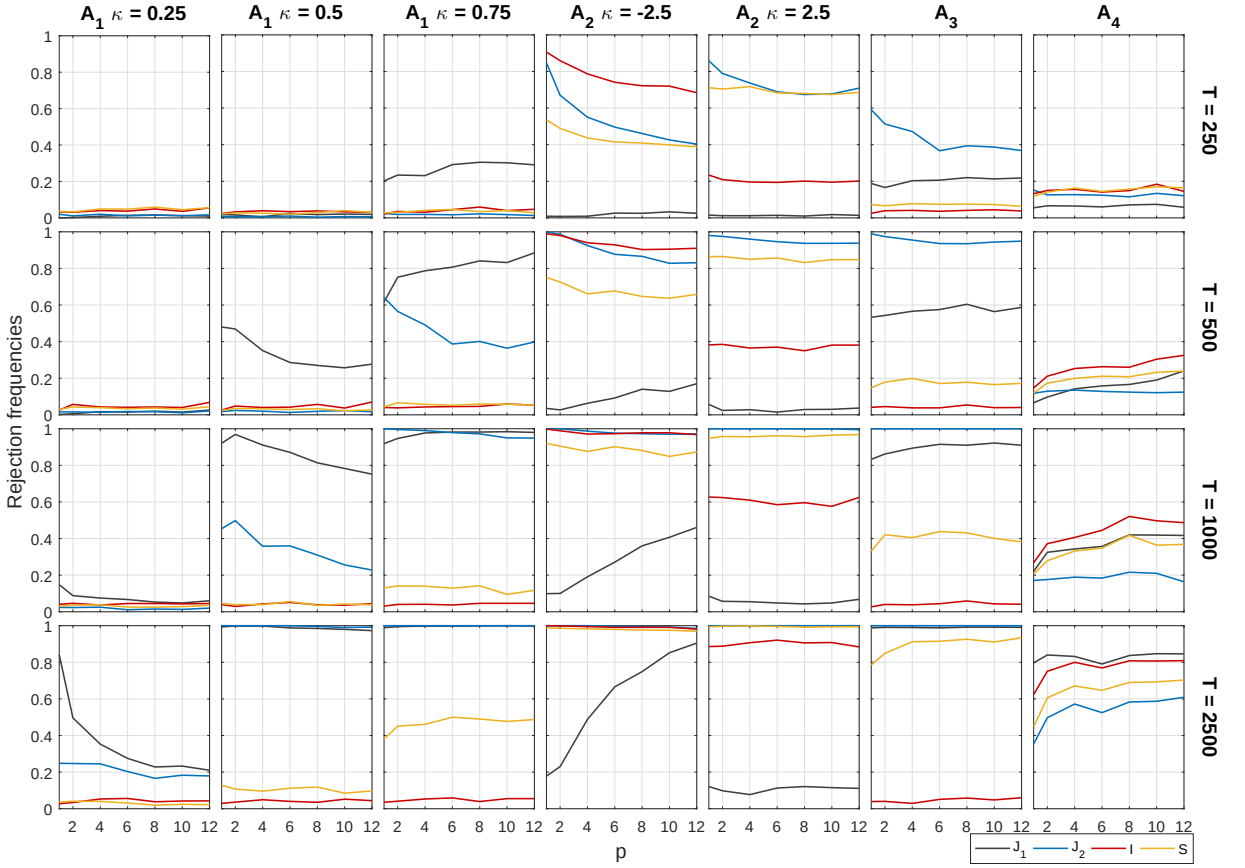
Note: Power of the four backtests are displayed as a function of p. The rows correspond to different sample sizes T, and the columns to the different misspecified alternatives A_1-A_4. Reported powers are size corrected.

Figure 7: Empirical size of the tests at 5% significance level (GARCH(1,1) model)



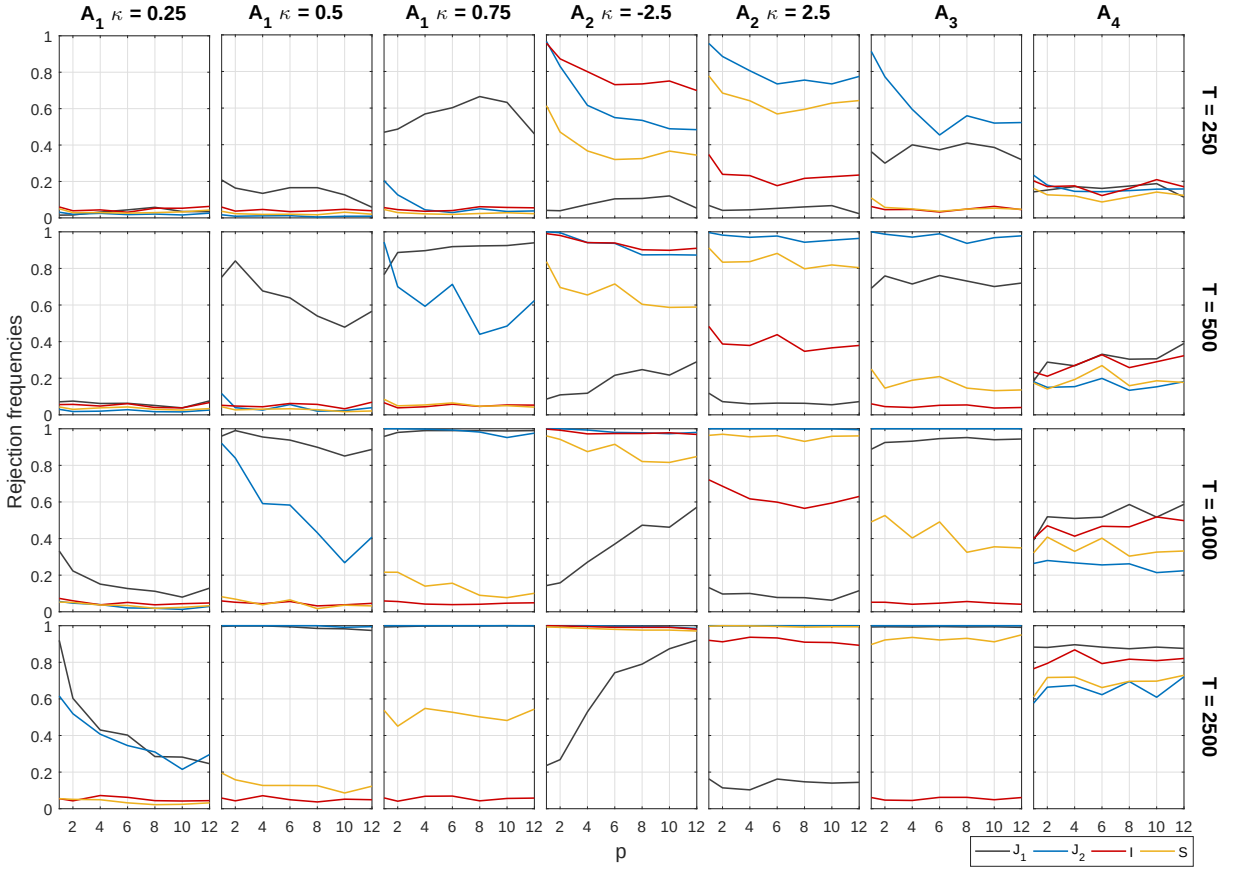
Note: Size of the four backtests are displayed as a function of p. The first row reports the results computed with the asymptotic critical values, and the second row those computed with the bootstrap critical values. The columns correspond to different sample sizes T.

Figure 8: Empirical power of the tests at 5% significance level (GARCH(1,1) model, bootstrap critical values)



Note: Power of the four backtests are displayed as a function of p. The rows correspond to different sample sizes T, and the columns to the different misspecified alternatives A_1-A_4. Reported powers are size corrected.

Figure 9: Empirical power of the tests at 5% significance level (GARCH(1,1) model, asymptotic critical values)



Note: Power of the four backtests are displayed as a function of p. The rows correspond to different sample sizes T, and the columns to the different misspecified alternatives A_1-A_4. Reported powers are size corrected.

G - Exact calculation method of ES

This section describes the methodology for the exact computation of ES forecasts at coverage level τ. Several techniques are available in practice. As the distribution of the innovations is parametric, we rely on Monte Carlo simulations. For ease of notation, we assume parameters to be known while in practice we use estimated parameters. The algorithm is as follows:

1. Randomly draw S pseudo standardized innovations $\{\eta_t^s\}_{s=1}^S$ from the Student distribution, with degrees of freedom v. We set the number $S = 100000$ in the empirical application.

2. Compute the ES at time t of the standardized innovation η_t as the Monte Carlo

average of the simulated innovations such that $m(\tau) = \frac{1}{\sum_{s=1}^{S} \mathbb{1}\left(\eta_t^s \geq F_v^{-1}(\tau)\right)} \sum_{s=1}^{S} \eta_t^s \times \mathbb{1}\left(\eta_t^s \geq F_v^{-1}(\tau)\right)$, where $F_v^{-1}(\tau)$ is the τ-quantile of the innovation distribution and is obtained as $F_v^{-1}(\tau) = \text{percentile}\left(\{\eta_t^s\}_{s=1}^{S}, 100\tau\right)$.

3. Compute the ES at time t as $ES_t(\tau) = \delta_0 + \delta_1 L_{t-1} + \sigma_t m(\tau)$.

H - Robustness checks of the empirical application

Figure 10: In-sample ES estimates issued from the approximation and the exact calculation method (GARCH(1,1) model)

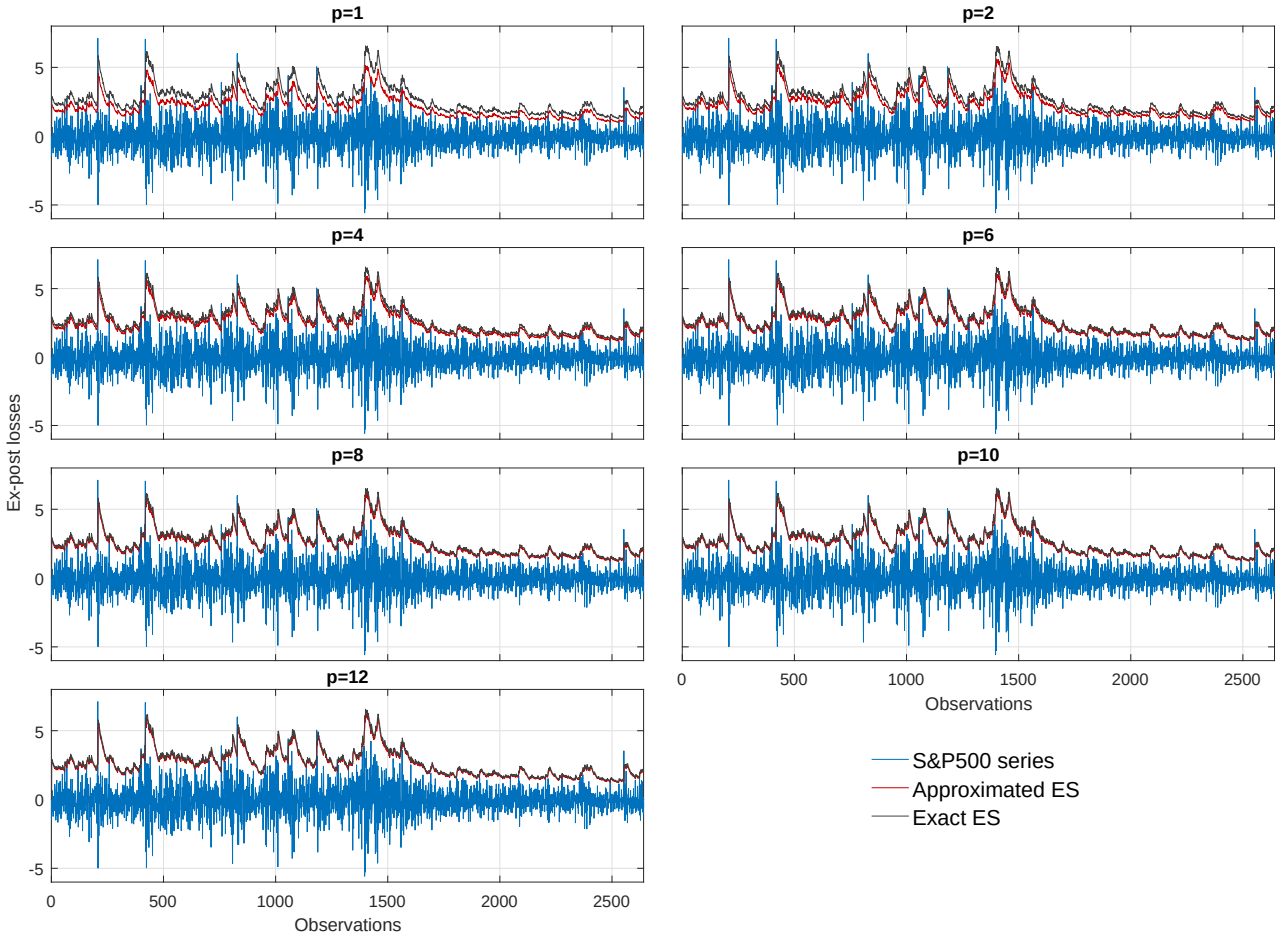


Table 3: p-values of the backtesting tests (GARCH(1,1) model)

p	$J_1^{(b)}$	$J_2^{(b)}$	$I^{(b)}$	$S^{(b)}$
		Panel A. 2007-2009		
1	0.060	0.082	0.141	0.853
2	0.027	0.054	0.045	0.233
4	0.014	0.048	0.027	0.106
6	0.016	0.046	0.026	0.135
8	0.153	0.053	0.229	0.673
10	0.036	0.067	0.052	0.448
12	0.026	0.058	0.042	0.223
2 (regulatory levels)	0.036	0.064	0.085	0.406
		Panel B. 2007-2012		
1	0.199	0.076	0.401	0.839
2	0.011	0.029	0.050	0.458
4	0.005	0.012	0.015	0.196
6	0.008	0.011	0.036	0.317
8	0.029	0.015	0.102	0.563
10	0.010	0.020	0.021	0.360
12	0.012	0.020	0.043	0.415
2 (regulatory levels)	0.005	0.015	0.014	0.230

Note: p-values of the four backtests computed with $p = 1, 2, 4, 6, 8, 10, 12$ risk levels successively, and the two regulatory levels $u_1 = 0.975$, $u_2 = 0.990$. Reported p-values are obtained using bootstrap critical values. Panel A gives the results for the period 2007-2009 and Panel B provides results for the period 2007-2012.

Table 4: QML coefficient estimates ($p = 6$, GARCH(1,1) model)

	u_1	u_2	u_3	u_4	u_5	u_6
			Panel A. 2007-2009			
β_0	0.600	0.683 *	0.769 **	0.811 $^{\star\star}_{\circ\circ\circ}$	0.972 $^{\star}_{\circ\circ\circ}$	1.065 $_{\circ\circ\circ}$
	(0.307)	(0.298)	(0.264)	(0.257)	(0.446)	(0.266)
β_1	1.011	0.955	0.917	0.853 **	0.804	0.692 $^{\star\star}_{\circ\circ\circ}$
	(0.093)	(0.089)	(0.059)	(0.055)	(0.143)	(0.043)
joint	*	*	*	**		**
			Panel B. 2007-2012			
β_0	0.338	0.388 $_{\circ\circ\circ}$	0.601 $^{\star\star}_{\circ\circ\circ}$	0.743 $^{\star\star\star}_{\circ\circ\circ}$	0.753 $^{\star\star}_{\circ\circ\circ}$	0.603
	(0.303)	(0.198)	(0.197)	(0.189)	(0.293)	(0.799)
β_1	1.025	0.987	0.911	0.860 **	0.829	0.832
	(0.122)	(0.069)	(0.066)	(0.056)	(0.109)	(0.308)
joint	*	*	*	**	**	

Note: Standard errors are reported in parentheses. *, **, and *** indicate statistical significance at the 10%, 5% and 1% level, respectively, and are obtained with the pairs bootstrap algorithm. $^{\circ}$, $^{\circ\circ}$, and $^{\circ\circ\circ}$, indicate statistical significance at the same levels and are obtained with the procedure of Chernozhukov and Fernández-Val (2011). Panel A gives estimation results for the period 2007-2009 and Panel B provides estimation results for the period 2007-2012.

Table 5: p-values of the backtesting tests using Jun and Pinkse (2009) estimates (AR(1)-GARCH(1,1) model)

p	$J_1^{(b)}$	$J_2^{(b)}$	$I^{(b)}$	$S^{(b)}$
		Panel A. 2007-2009		
1	0.024	0.057	0.161	0.928
2	0.013	0.048	0.024	0.471
4	0.005	0.024	0.023	0.285
6	0.006	0.029	0.031	0.222
8	0.018	0.053	0.147	0.676
10	0.031	0.069	0.060	0.585
12	0.030	0.063	0.047	0.354
2 (regulatory levels)	0.020	0.051	0.055	0.578
		Panel B. 2007-2012		
1	0.052	0.040	0.260	0.594
2	0.007	0.017	0.013	0.286
4	0.004	0.005	0.004	0.154
6	0.004	0.004	0.010	0.233
8	0.007	0.015	0.039	0.607
10	0.010	0.017	0.031	0.499
12	0.005	0.008	0.010	0.315
2 (regulatory levels)	0.009	0.019	0.044	0.309

Note: p-values of the four backtests computed with $p = 1, 2, 4, 6, 8, 10, 12$ risk levels successively, and the two regulatory levels $u_1 = 0.975$, $u_2 = 0.990$. Reported p-values are obtained using bootstrap critical values. Panel A gives the results for the period 2007-2009 and Panel B provides results for the period 2007-2012.

Table 6: Coefficient estimates issued from Jun and Pinkse (2009) estimation procedure ($p = 6$, AR(1)-GARCH(1,1) model)

	u_1	u_2	u_3	u_4	u_5	u_6
			Panel A. 2007-2009			
β_0	0.664	0.696	0.810 *	0.846 *	0.975 **	1.077 **
	(0.400)	(0.378)	(0.409)	(0.366)	(0.375)	(0.251)
β_1	0.972	0.954	0.890 **	0.847 ***	0.764 **	0.689 **
	(0.095)	(0.084)	(0.089)	(0.071)	(0.071)	(0.045)
joint		*	**	**	**	**
			Panel B. 2007-2012			
β_0	0.369	0.621 **	0.669 **	0.757 **	0.783 **	1.072 **
	(0.197)	(0.217)	(0.219)	(0.224)	(0.221)	(0.210)
β_1	1.032	0.934	0.904 *	0.856 **	0.823 *	0.690 *
	(0.063)	(0.063)	(0.063)	(0.063)	(0.053)	(0.045)
joint	*	**	**	**	**	*

Note: Standard errors are reported in parentheses. *, **, and *** indicate statistical significance at the 10%, 5% and 1% level, respectively, and are obtained with the pairs bootstrap algorithm. Panel A gives estimation results for the period 2007-2009 and Panel B provides estimation results for the period 2007-2012.

Table 7: p-values of the backtesting tests using Jun and Pinkse (2009) estimates (GARCH(1,1) model)

p	$J_1^{(b)}$	$J_2^{(b)}$	$I^{(b)}$	$S^{(b)}$
		Panel A. 2007-2009		
1	0.062	0.096	0.246	0.822
2	0.024	0.065	0.044	0.305
4	0.013	0.052	0.022	0.127
6	0.018	0.056	0.021	0.147
8	0.069	0.059	0.154	0.664
10	0.046	0.069	0.053	0.559
12	0.052	0.063	0.046	0.314
2 (regulatory levels)	0.044	0.076	0.099	0.518
		Panel B. 2007-2012		
1	0.135	0.117	0.318	0.740
2	0.010	0.017	0.060	0.241
4	0.001	0.009	0.010	0.095
6	0.006	0.009	0.019	0.269
8	0.018	0.011	0.029	0.495
10	0.022	0.027	0.040	0.330
12	0.018	0.023	0.049	0.401
2 (regulatory levels)	0.007	0.010	0.018	0.201

Note: p-values of the four backtests computed with $p = 1, 2, 4, 6, 8, 10, 12$ risk levels successively, and the two regulatory levels $u_1 = 0.975$, $u_2 = 0.990$. Reported p-values are obtained using bootstrap critical values. Panel A gives the results for the period 2007-2009 and Panel B provides results for the period 2007-2012.

Table 8: Coefficient estimates issued from Jun and Pinkse (2009) estimation procedure ($p = 6$, GARCH(1,1) model)

	u_1	u_2	u_3	u_4	u_5	u_6
			Panel A. 2007-2009			
β_0	0.852	0.683	0.768	0.811 *	0.886 *	1.065
	(0.424)	(0.415)	(0.385)	(0.346)	(0.371)	(0.336)
β_1	0.907	0.955	0.894	0.853 **	0.774 *	0.692 **
	(0.100)	(0.089)	(0.084)	(0.071)	(0.063)	(0.055)
joint		*	*	**	*	**
			Panel B. 2007-2012			
β_0	0.367 *	0.388	0.676 **	0.743 **	0.804 **	1.024 *
	(0.212)	(0.210)	(0.232)	(0.235)	(0.212)	(0.224)
β_1	1.012	0.987	0.904	0.860 *	0.811 *	0.696
	(0.060)	(0.061)	(0.061)	(0.060)	(0.054)	(0.044)
joint	*	*	*	**	**	*

Note: Standard errors are reported in parentheses. *, **, and *** indicate statistical significance at the 10%, 5% and 1% level, respectively, and are obtained with the pairs bootstrap algorithm. Panel A gives estimation results for the period 2007-2009 and Panel B provides estimation results for the period 2007-2012.

Figure 11: ES forecasts and adjusted ES forecasts over the period 2007-2012 (AR(1)-GARCH(1,1) model)

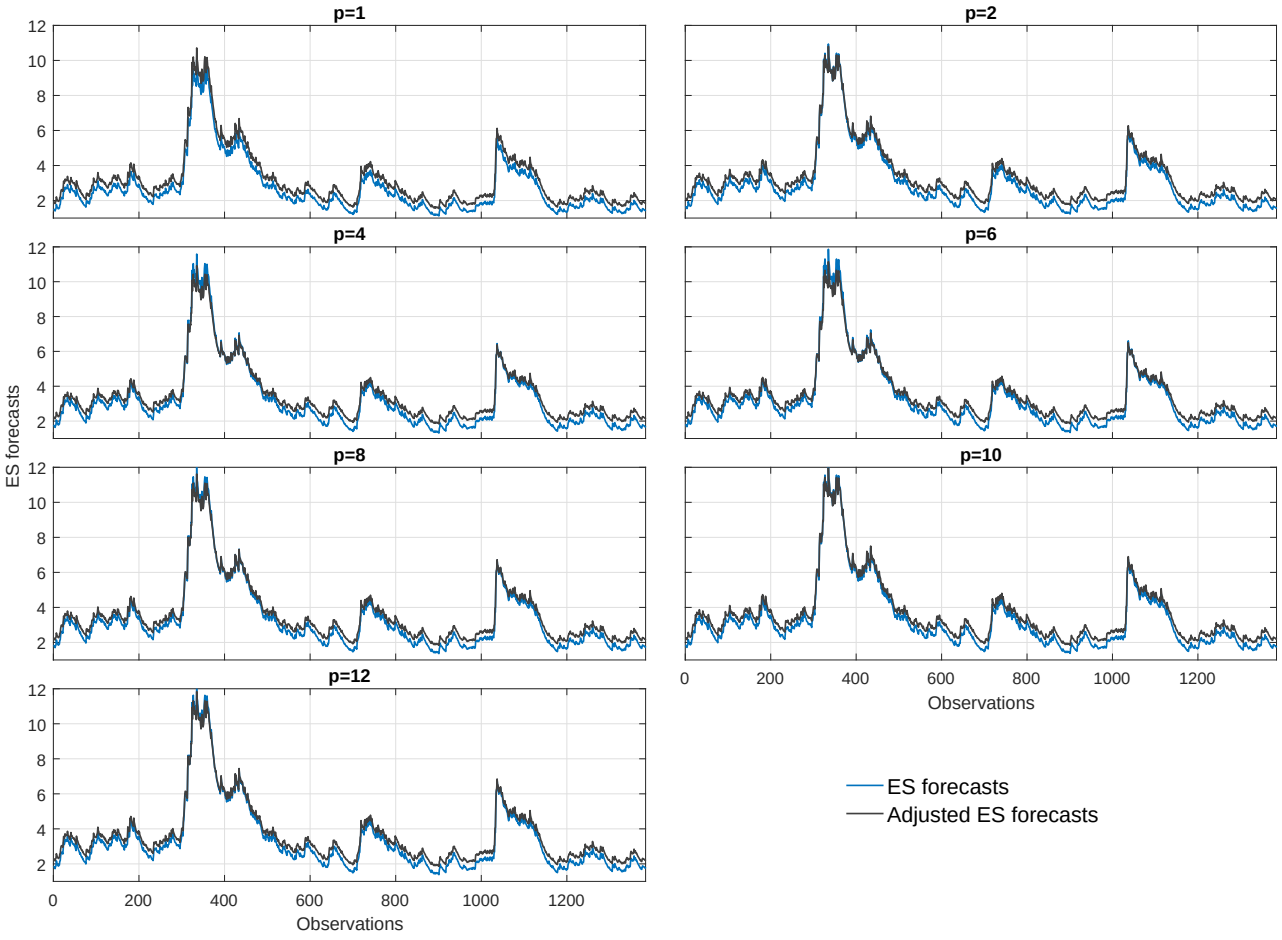


Figure 12: ES forecasts and adjusted ES forecasts over the period 2007-2009 (GARCH(1,1) model)

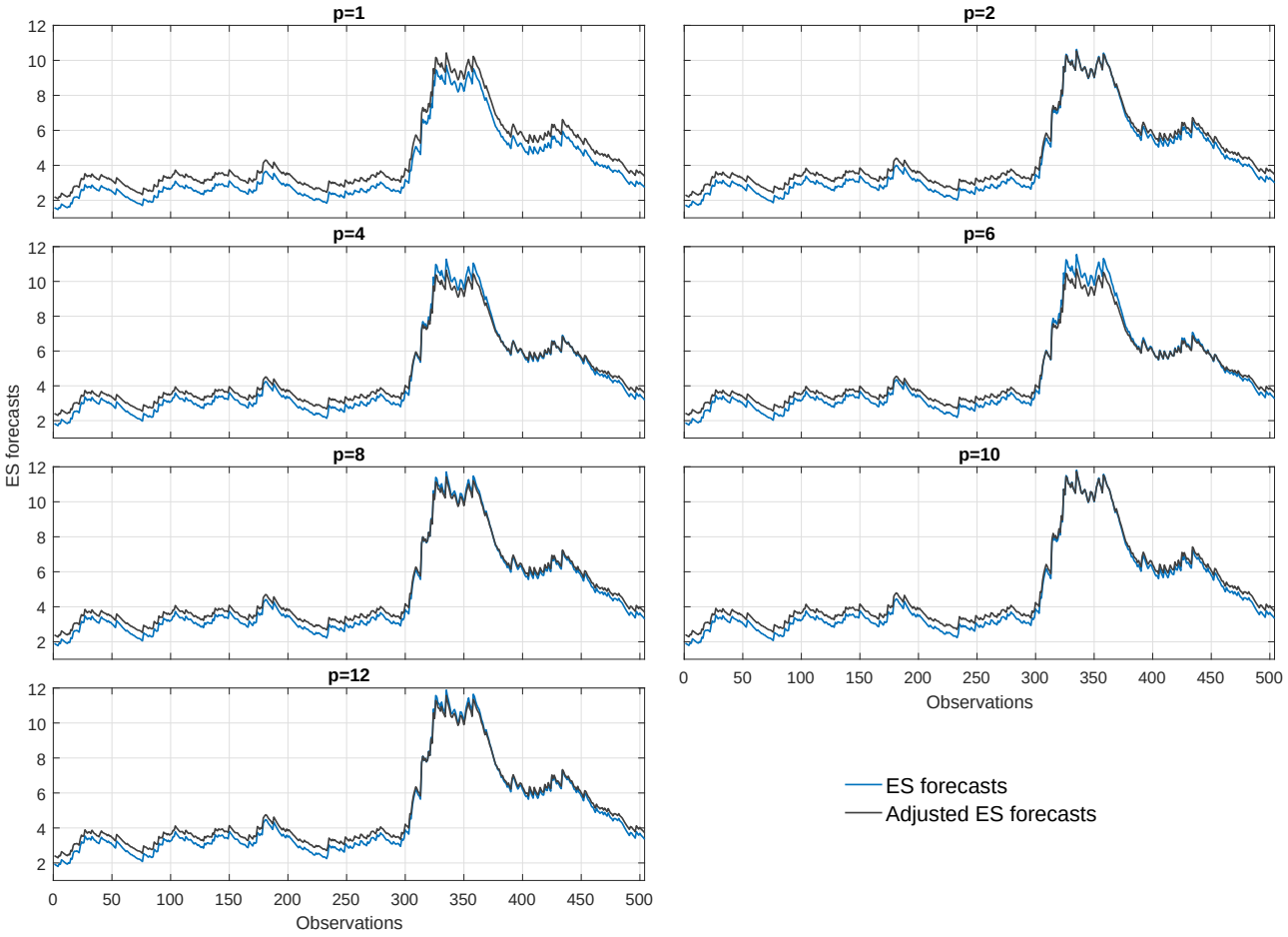


Figure 13: ES forecasts and adjusted ES forecasts over the period 2007-2012 (GARCH(1,1) model)

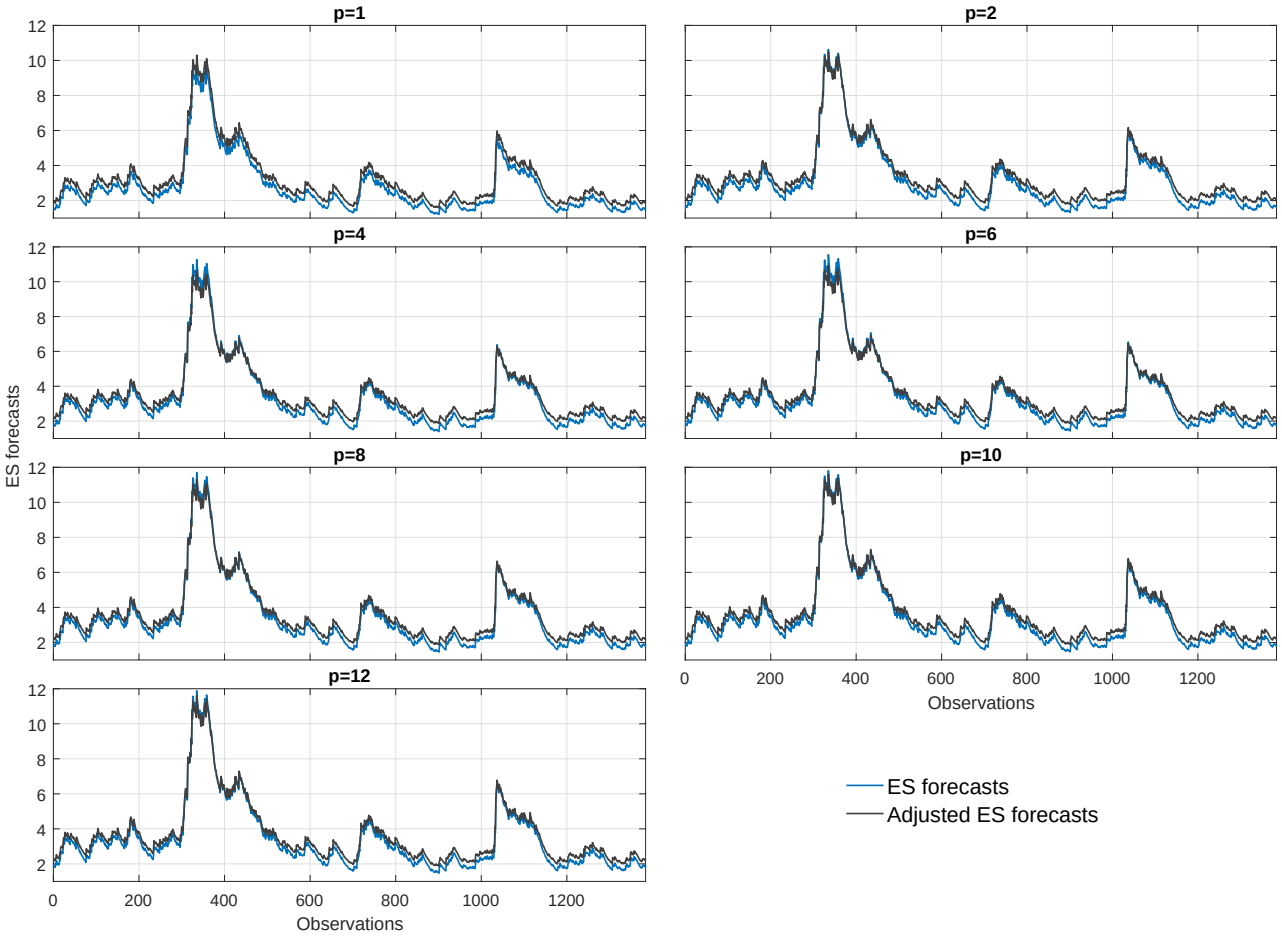
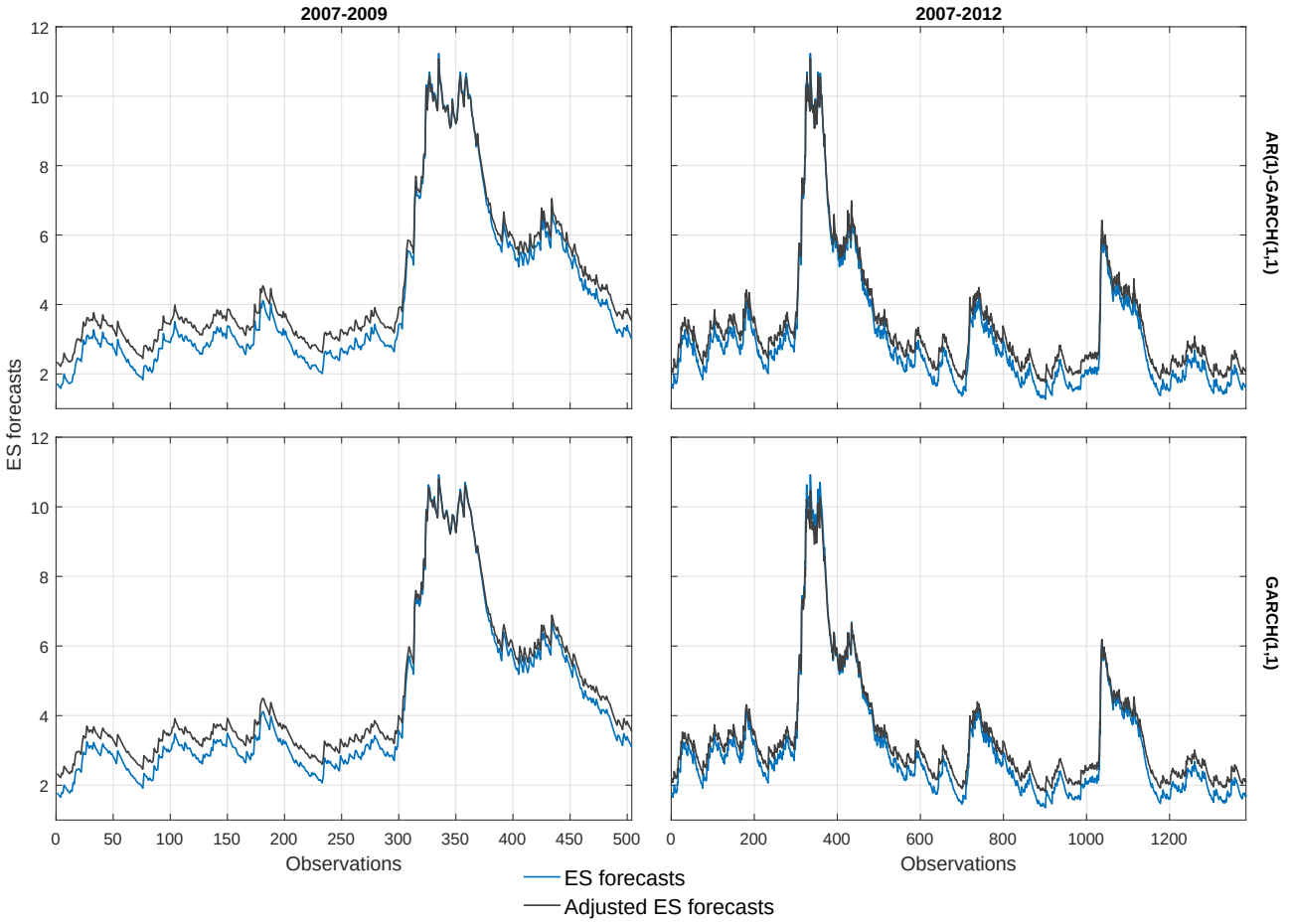


Figure 14: ES forecasts and adjusted ES forecasts over the periods 2007-2009 (on the left) and 2007-2012 (on the right) with the two BCBS regulatory risk levels (AR(1)-GARCH(1,1) model, GARCH(1,1) model)



References

Acerbi, C. and Szekely, B. (2014). Backtesting expected shortfall. *Risk*, 27(11):76–81.

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

Argyropoulos, C. and Panopoulou, E. (2016). A survey on risk forecast evaluation. Working paper.

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

Basel Committee on Banking Supervision (2010). Basel iii: A global regulatory framework for more resilient banks and banking systems. Consultation paper, December.

Basel Committee on Banking Supervision (2016). Minimum capital requirements for market risk. Consultation paper, January.

Basel Committee on Banking Supervision (2019). Explanatory note on the minimum capital requirements for market risk. Consultation paper, January.

Bayer, S. and Dimitriadis, T. (2019). Regression based expected shortfall backtesting. Working paper.

Berkowitz, J. (2001). Testing density forecasts, with applications to risk management. *Journal of Business & Economic Statistics*, 19(4):465–474.

Boucher, C., Daníelsson, J., Kouontchou, P., and Maillet, B. (2014). Risk models-at-risk. *Journal of Banking & Finance*, 44:72–92.

Chernozhukov, V. (2005). Extremal quantile regression. *Annals of Statistics*, 33:806–839.

Chernozhukov, V. and Fernández-Val, I. (2011). Inference for extremal conditional quantile models, with an application to market and birthweight risks. *Review of Economic Studies*, 78:559–589.

Chernozhukov, V., Fernández-Val, I., and Galichon, A. (2010). Quantile and probability curves without crossing. *Econometrica*, 78:1093–1125.

Christoffersen, P. (2011). Elements of financial risk management. Academic Press, Second edition.

Colletaz, G., Hurlin, C., and Pérignon, C. (2013). The risk map: A new tool for validating risk models. *Journal of Banking & Finance*, 37(10):3843–3854.

Costanzino, N. and Curran, M. (2015). Backtesting general spectral risk measures with application to expected shortfall. *Journal of Risk Model Validation*, 9(1):21–31.

Costanzino, N. and Curran, M. (2018). A simple traffic light approach to backtesting expected shortfall. *Risks*, 6(1):1–7.

Daníelsson, J. and Zhou, C. (2016). Why risk is so hard to measure. DNB working paper.

Du, Z. and Escanciano, J. C. (2017). Backtesting expected shortfall: Accounting for tail risk. *Management Science*, 63(4):901–1269.

Emmer, S., Kratz, M., and Tasche, D. (2015). What is the best risk measure in practice? a 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.

Fissler, T. and Ziegel, J. (2016). Higher order elicitability and osband's principle. *Annals of Statistics*, 44(4):1680–1707.

Freedman, D. (1981). Bootstrapping regression models. *Annals of Statistics*, 9(6):1218–1228.

Gaglianone, W. P., Lima, L. R., Linton, O., and Smith, D. R. (2011). Evaluating value-at-risk models via quantile regression. *Journal of Business & Economic Statistics*, 29(1):150–160.

Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494):746–762.

Gouriéroux, C. and Liu, W. (2012). Converting tail-var to var: An econometric study. *Journal of Financial Econometrics*, 10(2):233–264.

Gouriéroux, C. and Zakoïan, J.-M. (2013). Estimation-adjusted var. *Econometric Theory*, 29(4):735–770.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1:221–233, University of California Press, Berkeley.

Jorion, P. (2006). Value at risk: the new benchmark for managing financial risk. McGraw-Hill, Third edition.

Jun, S. and Pinkse, J. (2009). Efficient semiparametric seemingly unrelated quantile regression estimation. *Econometric Theory*, 25:1392–1414.

Kerkhof, J. and Melenberg, B. (2004). Backtesting for risk-based regulatory capital. *Journal of Banking & Finance*, 28(4):1845–1865.

Koenker, R., Chernozhukov, V., He, X., and Peng, L. (2018). Handbook of quantile regression. Chapman and Hall/CRC Handbooks of Modern Statistical Methods.

Koenker, R. and Machado, J. (1999). Goodness of fit and related inference processes for quantile regression. *Journal of the American Statistiscal Association*, 94(448):1296–1310.

Koenker, R. and Xiao, Z. (2002). Inference on the quantile regression process. *Econometrica*, 70(4):1583–1612.

Kratz, M., Lok, Y. H., and McNeil, A. J. (2018). Multinomial var backtests: A simple implicit approach to backtesting expected shortfall. *Journal of Banking & Finance*, 88:393–407.

Lazar, E. and Zhang, N. (2019). Model risk of expected shortfall. *Journal of Banking & Finance*, 105:74–93.

Löser, R., Wied, D., and Ziggel, D. (2019). New backtests for unconditional coverage of expected shortfall. *Journal of Risk*, 21(4):39–59.

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.

Mincer, J. A. and Zarnowitz, V. (1969). The evaluation of economic forecasts. In *Economic forecasts and expectations: Analysis of forecasting behavior and performance*, pages 3–46. NBER.

Nolde, N. and Ziegel, J. F. (2017). Elicitability and backtesting: Perspectives for banking regulation. *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). *Forthcoming in Journal of Econometrics*.

Powell, J. L. (1984). Least absolute deviations estimation for the censored regression model. *Journal of Econometrics*, 25(3):303–325.

Taylor, J. W. (2019). Forecasting value at risk and expected shortfall using a semiparametric approach based on the asymmetric laplace distribution. *Journal of Business & Economic Statistics*, 37(1):121–133.

White, H. (1994). Estimation, inference and specification analysis. Cambridge University Press, New York.

White, H. (2001). Asymptotic theory for econometricians. Academic Press, San Diego.

White, H., Kim, T. H., and Manganelli, S. (2008). Modeling autoregressive conditional skewness and kurtosis with multi-quantile caviar. *Volatility and Time Series Econometrics: Essays in Honor of Robert F. Engle*, In Bollerslev T., Russell, J., and Watson, M. editors.

White, H., Kim, T.-H., and Manganelli, S. (2015). Var for var: Measuring tail dependence using multivariate regression quantiles. *Journal of Econometrics*, 187(1):169–188.

Wong, W. K. (2008). Backtesting trading risk of commercial banks using expected shortfall. *Journal of Banking & Finance*, 32(7):1404–1415.