



**HAL**  
open science

# Backtesting Expected Shortfall via Multi-Quantile Regression

Ophélie Couperier, Jérémy Leymarie

► **To cite this version:**

Ophélie Couperier, Jérémy Leymarie. Backtesting Expected Shortfall via Multi-Quantile Regression. 2018. halshs-01909375v1

**HAL Id: halshs-01909375**

**<https://shs.hal.science/halshs-01909375v1>**

Preprint submitted on 31 Oct 2018 (v1), last revised 1 Oct 2020 (v5)

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Backtesting Expected Shortfall via Multi-Quantile Regression

Ophélie Couperier <sup>\*</sup>      Jérémy Leymarie <sup>†</sup>

October 31, 2018

## 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 jointly assesses the quality of VaRs 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 which 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

---

<sup>\*</sup>Ensaie (CREST, UMRS CNRS 9194).

<sup>†</sup>University of Orléans (LEO, FRE CNRS 2014). 11 rue de Blois. 45000 FRANCE. Corresponding author: jeremy.leymarie@univ-orleans.fr

# 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, 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, measures the expected loss incurred on an asset portfolio given that the loss exceeds VaR. That is, if  $L_t$  is the ex-post loss on a portfolio at time  $t$ ,  $\Omega_{t-1}$  is the risk manager's information at time  $t - 1$ , and  $Q_{L_t}(\cdot)$  is the quantile function of  $L_t$ , the  $\tau$ -level ES and VaR are given by

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

$$VaR_t(\tau) = Q_{L_t}(\tau; \Omega_{t-1}). \quad (2)$$

As an alternative tail risk measure, ES offers a number of appealing properties that overcomes the theoretical deficiencies of the more-familiar VaR. In particular, ES is *coherent* meaning that this risk measure satisfies the properties of monotonicity, sub-additivity, homogeneity, and translational invariance (see Artzner et al., 1999, and 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 the recent contributions of Taylor, 2017, and Patton, Ziegel and Chen, 2018), 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 *elicibility* property. However, it has been established that, in contrast to VaR, ES does not meet the general property of elicibility (Gneiting, 2011), but satisfies narrower properties such as conditional elicibility (Emmer, Kratz, and Tasche, 2015), or joint elicibility with VaR (Acerbi and Szekely, 2014, and 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 (see for instance McNeil and Frey, 2000, Acerbi and Szekely, 2014, Nolde and Ziegel, 2017, and Bayer and Dimitriadis, 2018).

To circumvent the lack of elicibility of ES, several alternative testing strategies have been proposed in the literature. Following the recent classification of Kratz, Lok, and McNeil (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. Indeed, definition of a conditional probability and a change of variable yield a useful representation of ES in terms of VaR

$$ES_t(\tau) = \frac{1}{1-\tau} \int_{\tau}^1 VaR_t(u) du. \quad (3)$$

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

for ES, generalizing the violation process for VaR. Costanzino and Curran (2018) provide a Traffic 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 (see Berkowitz, 2001, Kerkhof and Melenberg, 2004, and Wong, 2008). For more details and references about risk forecast evaluation, see the survey of Argyropoulos and Panopoulou (2016).

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 as a finite sum of VaRs.

$$ES_t(\tau) \approx \frac{1}{p} \sum_{j=1}^p VaR_t(u_j), \quad (4)$$

where the risk level  $u_j$  is defined by  $u_j = \tau + (j - 1) \frac{1-\tau}{p}$  for  $j = 1, 2, \dots, p$ . This representation suggests that  $p$  quantiles with appropriate risk levels would be convenient to assess the quality of an ES model. As a result, an estimate of  $ES_t(\tau)$  issued from a given model could be considered valid if the sequence of  $VaR_t(u_j)$  estimates 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 paper is to propose an original backtesting methodology for ES based on the theory of multi-quantile regression. Our validation procedure investigates the quality of quantiles at different places in the tail of the ES model within a regression-based framework. This approach has many advantages. First, our procedure is flexible since the user may choose the number and values of quantiles to be investigated, and thus can easily focus on various aspects of the tail distribution. Second, our testing strategy

encompasses the regulatory standards which consist in verifying the validity of two given quantiles. Finally, our new tool enters the category of regression-based backtests, which complements the existing literature on regression-based forecast evaluation as proposed by Engle and Manganelli (2004), Christoffersen (2011), Bayer and Dimitriadis (2018), among others. In addition, this original approach represents an alternative to the multiple VaR exceptions backtests developed by Kratz, Lok, and McNeil (2018).

Our procedure extends the seminal idea of Gaglianone et al. (2011) to evaluate the quality of VaR applying quantile regression. We develop a more general theory and focus on multi-quantile regression to jointly assess VaR at multiple levels in the tail distribution of the risk model. We show that the parameter estimates issued from the multi-quantile regression model satisfy specific properties under the hypothesis of correct ES forecasts. We propose four backtests to assess various settings on these regression parameters. Finally, to address the issue of invalid risk forecasts, we develop a procedure to correct the imperfect predictions relying on our multi-quantile regression framework.

We provide several Monte Carlo experiments and an empirical application using the S&P500 series in support of the new technique introduced. 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, and address these deficiencies via the implementation of bootstrap critical values. The latter provides satisfactory size performances regardless of the sample size, and hence should be preferred when asymptotic theory does not apply conveniently. We also show that the BCBS recommendation of verifying quantiles at coverage levels 97.5% and 99% is not always sufficient to reject the validity of the ES forecasts. We hence recommend the use of additional risk levels to improve the soundness of the decision. Finally, we confirm 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 our 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 provides the simulation study. In Section 5, we apply our backtesting methodology on the S&P500 index, and develop a procedure to adjust the imperfect ES forecasts. Finally, we conclude the paper in Section 6.

## 2 Multi-quantile regression framework

This section is devoted to the description of 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. Finally, the last part is devoted to the description of the estimation method, and asymptotic theory of our auxiliary regressions.

### 2.1 ES as an approximation of VaRs

Our backtesting procedure exploits the relationship between VaR and ES. We suppose that ES can be appropriately approximated by several 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  $\tau$ -level ES approximation is defined as a finite Riemann sum involving a number  $p$  of VaRs, and is given by*

$$ES_t(\tau) \approx \frac{1}{p} \sum_{j=1}^p VaR_t(u_j), \quad (5)$$

where risk levels  $u_j$ ,  $j = 1, 2, \dots, p$ , are defined by  $u_j = \tau + (j - 1) \frac{1-\tau}{p}$ , 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 applied in the approximation is given by  $p$ , and characterizes the degree of accuracy of the approximation. In particular,  $p = 1$  involves a single VaR at coverage level  $\tau$ , while increasing  $p$  to infinity leads the VaRs to be evaluated continuously along the tail, so that the Riemann sum converges to the theoretical ES. In practice,  $p$  may be chosen small as the interval of the definite integral is restricted to the extreme upper tail distribution, and therefore a few quantiles are generally enough to get good approximations. Finally, the risk levels  $u_j$ ,  $j = 1, 2, \dots, p$ , are determined so that the interval is equally partitioned between the two boundaries  $\tau$ , and 1.

Our approximation is useful for several reasons. First, this simple formula is appealing in a regulatory and risk management viewpoint since VaR is well-established compared to ES and pretty easier to compute. Secondly, and it is the purpose of this paper, the above relationship greatly simplifies the assessment of ES in practice, by focusing on the validity of several VaRs, and is more intelligible in the context of banking regulation. This approach is indeed fully consistent with the BCBS guidelines on ES assessment 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 11). Finally, our validation strategy offers a certain flexibility since the risk manager's or the supervisor may select both the number of probability levels and their magnitude depending on the objective in mind (regulatory guidelines, ES statistical approximation, etc.).

## 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  $t = 1, 2, \dots, T$ . In addition, we denote by  $\Omega_{t-1}$  the information set available at time  $t - 1$ , with  $(L_{t-1}, L_{t-2}, \dots) \subseteq \Omega_{t-1}$ . Formally, the  $u_j$ -level VaR of the  $L_t$  distribution is the quantity  $VaR_t(u_j)$  such that

$$Pr(L_t \geq VaR_t(u_j) | \Omega_{t-1}) = u_j. \quad (6)$$

According to Definition 1, this equation serves as an implicit backtest of ES. One should conclude to the appropriateness of a given ES model as soon as the sequence  $VaR_t(u_j)$ ,  $t = 1, 2, \dots, T$ , issued from the ES model satisfies the above joint equality for all  $j$ .

We refer to the original idea of Gaglianone et al. (2011), 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  $\{L_t, t = 1, 2, \dots, T\}$  on the  $p$  VaR forecasts  $\{VaR_t(u_j), t = 1, 2, \dots, T\}_{j=1,2,\dots,p}$  in a multi-quantile regression model.

$$L_t = \beta_0(u_j) + \beta_1(u_j) VaR_t(u_j) + \epsilon_{j,t} \quad \forall j = 1, 2, \dots, p, \quad (7)$$

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) proposed by White, Kim, and Manganeli (2008, 2015) which allows a joint modeling of multiple conditional VaRs. Given the multi-quantile regression model of Equation (7), the  $u_j$ -th conditional quantile of  $L_t$  is defined as

$$Q_{L_t}(u_j; \Omega_{t-1}) = \beta_0(u_j) + \beta_1(u_j) VaR_t(u_j) \quad \forall j = 1, 2, \dots, p. \quad (8)$$

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  $j = 1, 2, \dots, p$ . For these parameter values, and according to Definition 1, the tail distribution of the ES model and the corresponding ES forecasts are said to be valid.

### 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  $j = 1, 2, \dots, p$ . Under the hypothesis that a sequence of VaR is valid, coefficients satisfy  $\beta_0(u_j) = 0$ , and  $\beta_1(u_j) = 1$ , for  $j = 1, 2, \dots, p$ . In what follows, we denote by  $\beta(u_j) = (\beta_0(u_j), \beta_1(u_j))'$  the vector of unknown parameters for the  $u_j$ -th quantile index, and we write  $\beta = (\beta(u_1)', \beta(u_2)', \dots, \beta(u_p)')$  the stacked vector of  $2p$  coefficients. In addition, we assume that the sequence  $\{u_j, j = 1, 2, \dots, p\}$  is ordered in the sense that  $u_1 < u_2 < \dots < u_p < 1$ .

In order to estimate  $\beta$ , we consider the multi-quantile regression approach recently proposed by White, Kim and Manganelli (2008, 2015). A consistent QMLE estimator is given by

$$\hat{\beta} = \arg \min_{\beta \in \mathbb{R}^{2p}} T^{-1} \sum_{t=1}^T \left( \sum_{j=1}^p \rho_{u_j}(L_t - \beta_0(u_j) - \beta_1(u_j) VaR_t(u_j)) \right), \quad (9)$$

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, Kim, and Manganelli (2008, 2015) show that this estimator is asymptotically normally distributed,

with covariance matrix that takes the form of a Huber (1967) sandwich:

$$\sqrt{T}(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(0, \Sigma), \quad (10)$$

where  $\Sigma$  denotes the asymptotic covariance matrix. Its expression is given by  $\Sigma = A^{-1}VA^{-1}$ , with:

$$V = \mathbb{E}[\eta_t \eta_t'], \quad (11)$$

$$\eta_t = \sum_{j=1}^p \nabla Q_{L_t}(u_j; \Omega_{t-1}) \psi_{u_j}(\epsilon_{j,t}), \quad (12)$$

$$A = \sum_{j=1}^p \mathbb{E}[f_{j,t}(0) \nabla Q_{L_t}(u_j; \Omega_{t-1}) \nabla' Q_{L_t}(u_j; \Omega_{t-1})], \quad (13)$$

where  $\nabla Q_{L_t}(u_j; \Omega_{t-1})$  denotes the  $2p$  gradient vector differentiated with respect to  $\beta$ , and  $\epsilon_{j,t} = L_t - Q_{L_t}(u_j; \Omega_{t-1})$ , and where  $f_{j,t}(0)$  denotes the pdf of  $\epsilon_{j,t}$  evaluated at zero. In Appendix A, we provide a consistent estimator  $\hat{\Sigma}$  of the true unknown quantity  $\Sigma$  that will be used for the computation of our test statistics.

### 3 Backtesting ES

In this section, we present our backtests for ES. Our procedures assess whether the parameters  $\beta_0(u_j)$  and  $\beta_1(u_j)$  meet their expected values for the risk levels  $u_j$ ,  $j = 1, 2, \dots, p$ . 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  $j = 1, 2, \dots, p$ . We propose to assess various implications of these coefficient restrictions by taking into consideration four

distinct null hypotheses based on a reduced number of constraints. Many backtests assess implications of a more general hypothesis. In this context, 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, while Du and Escanciano (2017) assess two implications for the martingale difference sequence of their cumulative violation process.

**Definition 2** (The 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:*

$$H_{0,J_1} : \sum_{j=1}^p (\beta_0(u_j) + \beta_1(u_j)) = p, \quad (14)$$

$$H_{0,J_2} : \sum_{j=1}^p \beta_0(u_j) = 0, \quad \text{and,} \quad \sum_{j=1}^p \beta_1(u_j) = p, \quad (15)$$

$$H_{0,I} : \sum_{j=1}^p \beta_0(u_j) = 0, \quad (16)$$

$$H_{0,S} : \sum_{j=1}^p \beta_1(u_j) = p, \quad (17)$$

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.

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. In accordance with the definition of ES as a finite Riemann sum of VaRs, we sum the coefficients across risk levels  $j$ ,  $j = 1, 2, \dots, p$ . Apart from being consistent with Definition 1, this coefficients' aggregation substantially reduces the number of constraints to be tested. The structure of  $H_{0,J_2}$  is hence characterized by two constraints, and those of  $H_{0,J_1}$ ,  $H_{0,I}$ ,  $H_{0,S}$  involve a single constraint.

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  $j = 1, 2, \dots, p$ .  $H_{0,J_1}$  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 of the slope backtest,  $H_{0,I}$  and  $H_{0,S}$ , focus solely on one of the two parameter component.  $H_{0,I}$  examines the intercept parameters  $\beta_0(u_j)$ ,  $j = 1, 2, \dots, p$ , and  $H_{0,S}$  investigates the slope parameters  $\beta_1(u_j)$ ,  $j = 1, 2, \dots, p$ . These additional null hypotheses complement the joint backtests to identify the nature of the misspecification. When  $H_{0,I}$  is rejected, this indicates that the forecasting errors are constant across time. In contrast, the rejection of  $H_{0,S}$  suggests that the errors are time-varying since they change with respect to the VaR predictions.

**Definition 3** (Wald-test statistics). *Let us indicate by  $W$ ,  $W \in \{J_1, J_2, I, S\}$ , 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*

$$W = T \left( R_W \hat{\beta} - q_W \right)' \left( R_W \hat{\Sigma} R_W' \right)^{-1} \left( R_W \hat{\beta} - q_W \right), \quad (18)$$

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

To assess our null hypotheses we consider Wald-type inference. Definition 3 provides the general expression of the test statistics. In accordance with our notations, substituting  $W$  by  $J_1$ ,  $J_2$ ,  $I$ , and  $S$ , yields the four test statistics. For simplicity, 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  $q_W$ ,  $W \in \{J_1, J_2, I, S\}$ , are as follows:  $R_{J_1} = \iota_p \otimes (1 \ 1)$ ,  $q_{J_1} = p$ ,  $R_{J_2} = \iota_p \otimes I_2$ ,  $q_{J_2} = (0 \ p)'$ ,  $R_I = \iota_p \otimes (1 \ 0)$ ,  $q_I = 0$ ,  $R_S = \iota_p \otimes (0 \ 1)$ ,  $q_S = p$ , where  $\iota_p$  is a  $p$ -row unit vector, and  $I_2$  denotes the identity matrix of size 2.

**Proposition 1** (Chi-squared distribution). *Suppose the covariance matrix  $\Sigma$  is non singular. Under the Normality condition of Equation (10), 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  $H_{0,J_1}$ ,  $H_{0,J_2}$ ,  $H_{0,I}$ ,  $H_{0,S}$ . The test statistics are asymptotically chi-squared distributed, with one degree of freedom for  $J_1$ ,  $I$ ,  $S$ , and two degrees of freedom for  $J_2$ . 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. In consequence, the four backtests have unchanged critical values whatever the degree of accuracy used in the ES approximation.

### 3.2 Finite sample inference

Our four test statistics are asymptotically chi-squared distributed, and it is thus possible to employ them if the asymptotic conditions are fulfilled for realistic sample sizes. However, in the specific context of ES assessment, focus is on extreme tail distribution, i.e. for risk levels above the regulatory coverage level  $\tau = 0.975$ . In practice, this may induce scarce information, and affect the inference when the sample size is not sufficiently large. To overcome these typical deficiencies, we implement a finite sample procedure via bootstrap to adjust the critical values in small samples.

The backtests of tail risk measures such as VaR and ES are typically affected by these distortions. For instance, using their regression-based procedure, Gaglianone et al. (2011) obtain size distortions for moderate sample sizes. In the same vein, the regression procedure to assess ES proposed by Bayer and Dimitriadis (2018) induces size distortions even for large sample sizes, e.g.  $T = 1000$ . The authors also show that the conditional calibration test of

Nolde and Ziegel (2017), and the exceedance residual test of McNeil and Frey (2000) display poor results for realistic samples. They propose a bootstrap algorithm to correct these biases. Also intertwined with this literature, Hurlin et al. (2017) introduce a bootstrap procedure dedicated to the inference of the risk measures themselves.

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

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

2. Build a bootstrap sample by drawing with replacement  $T$  pairs of observations from the original data  $\{L_t, VaR_t(u_j)\}_{j=1,2,\dots,p}$ ,  $t = 1, 2, \dots, T$ .
3. Estimate the model on the bootstrap sample, to obtain  $\hat{\beta}^b$  and  $\hat{\Sigma}^b$ , and compute the bootstrapped test statistic  $W^b$  under the null hypothesis as follows:

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

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

Two remarks should be made about the algorithm. First, when we use the pairs bootstrap we cannot impose the null hypothesis on the bootstrap data generating process since

imposing restrictions on  $\beta$  is unfeasible. To overcome this issue, we calculate the bootstrap statistics by considering the difference  $R_W\beta - R_W\hat{\beta}$  rather than  $R_W\beta - q$ . Since the estimate of  $\beta$  from the bootstrap samples should, on average, be equal to  $\hat{\beta}$ , at least asymptotically, the null hypothesis tested by  $W^b$  becomes "true" for the pairs bootstrap data generating process. Second, the critical value  $c_\alpha$  is obtained as the  $\alpha$ -quantile of the bootstrap statistics  $W^b$ ,  $b = 1, 2, \dots, B$ . The decision rule is as follows. If the original test statistic  $W$  is greater than the  $\alpha$ -level bootstrapped critical value  $c_\alpha$ , we conclude to the rejection of the null hypothesis. In addition, we compute the p-value of the test as  $P = B^{-1} \sum_{b=1}^B \mathbb{1}(W^b > W)$ .

## 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 5,000 replications, and we consider two sample sizes  $T = 500, 2500$ . The results are reported using both the asymptotic critical values (based on a  $\chi^2$  distribution), and the bootstrap critical values. We calculate the latter with  $B = 1000$  bootstrap samples. Finally, the backtests are computed with  $\tau = 0.975$ , which is the coverage level applied in the context of the current banking regulation.

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 for assessing 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. To answer this question, we consider different choices for the number of risk levels, namely  $p = 1, 2, 4, 6$ . The  $p$  risk levels  $u_1, u_2, \dots, u_p$  are calculated in accordance with Definition 1. Notice that  $p = 1$  coincides with the VaR backtest at level  $\tau$  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, it must be emphasized that  $p = 4$  meets the number of risk levels discussed by Emmer, Kratz, and Tasche (2015).

The correct data generating process is given by the popular AR(1)-GARCH(1,1) structure with Student innovations. Accordingly, we define the ex-post portfolio loss  $L_t$ ,  $t = 1, 2, \dots, T$  as

$$\begin{aligned} L_t &= \delta_0 + \delta_1 L_{t-1} + \epsilon_t, \\ \epsilon_t &= \sigma_t \eta_t, \quad \eta_t \sim t_v, \\ \sigma_t^2 &= \gamma_0 + \gamma_1 \epsilon_{t-1}^2 + \gamma_2 \sigma_{t-1}^2, \end{aligned} \tag{19}$$

where  $t_v$  denotes the Student's  $t$  distribution with  $v$  degrees of freedom. Given the model in Equation (19), the true ES and VaR at coverage level  $\tau$  are given by

$$ES_t(\tau) = \delta_0 + \delta_1 L_{t-1} + \sigma_t m(\tau), \tag{20}$$

$$VaR_t(\tau) = \delta_0 + \delta_1 L_{t-1} + \sigma_t F_v^{-1}(\tau), \tag{21}$$

with  $m(\tau) = \mathbb{E}[\eta_t | \eta_t \geq F_v^{-1}(\tau)]$ , and where  $F_v^{-1}(\tau)$  denotes the  $\tau$ -quantile of the Student distribution with  $v$  degrees of freedom. Finally, we calibrate the parameters  $(\delta_0, \delta_1, \gamma_0, \gamma_1, \gamma_2, v)$  using the S&P500 series over the period 2013-2017, which leads us to consider the following numeric values in the simulation study  $(-0.085, -0.093, 0.034, 0.214, 0.748, 5)$ . Finally to investigate the power, we consider several misspecified alternatives for  $L_t$ :

$A_1$  : AR(1)-GARCH(1,1) model with undervalued conditional variances:  $L_t$  is as Equation (19),

with  $\sigma_t^2 = (\gamma_0 + \gamma_1\epsilon_{t-1}^2 + \gamma_2\sigma_{t-1}^2) \times (1 - \kappa)$ , 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:  $L_t$  satisfies Equation (19), with  $\eta_t \sim (0.5X^+ + 0.5X^-) / \sqrt{10}$ , 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  $VaR_t(\tau) = \text{percentile}(\{L_{t-i}\}_{i=1}^{250}, 100\tau)$ , 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  $\sigma_t$  is alternately undervalued of 25%, 50%, and 75% to examine whether our tests are able to detect an underestimation of ES stemming from a poor 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  $\eta_t$  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 risk assessment. Bayer and Dimitriadis (2018) 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, Lok, McNeil (2018), Bayer and Dimitriadis (2018), Gaglianone et al. (2011), among many others.

Table 1, and 2, report the rejection frequencies of the tests at 5% significance level for sample sizes  $T = 500$ , and  $T = 2500$ , respectively. The first four columns report the results

Table 1: Empirical rejection rates of the backtests at 5% significance level,  $T = 500$

		$J_1$	$J_2$	$I$	$S$	$J_1^{(b)}$	$J_2^{(b)}$	$I^{(b)}$	$S^{(b)}$
$p = 1$									
$H_0$		0.130	0.303	0.186	0.241	0.057	0.058	0.058	0.061
$A_1$	$\kappa = 0.25$	0.068	0.053	0.054	0.055	0.070	0.059	0.054	0.047
	$\kappa = 0.50$	0.591	0.055	0.053	0.061	0.434	0.070	0.061	0.044
	$\kappa = 0.75$	0.665	0.811	0.061	0.079	0.575	0.639	0.068	0.062
$A_2$	$\kappa = +2.5$	0.141	0.995	0.391	0.833	0.091	0.994	0.342	0.846
	$\kappa = -2.5$	0.104	0.997	0.989	0.763	0.065	0.996	0.988	0.773
$A_3$		0.623	0.994	0.247	0.106	0.523	0.992	0.202	0.117
$A_4$		0.120	0.145	0.208	0.120	0.079	0.128	0.165	0.128
$p = 2$									
$H_0$		0.116	0.278	0.166	0.223	0.055	0.058	0.057	0.059
$A_1$	$\kappa = 0.25$	0.059	0.054	0.052	0.058	0.071	0.061	0.051	0.042
	$\kappa = 0.50$	0.637	0.055	0.049	0.062	0.355	0.062	0.049	0.044
	$\kappa = 0.75$	0.802	0.693	0.089	0.072	0.643	0.553	0.053	0.059
$A_2$	$\kappa = +2.5$	0.078	0.987	0.353	0.840	0.041	0.977	0.341	0.854
	$\kappa = -2.5$	0.060	0.976	0.968	0.692	0.042	0.967	0.967	0.709
$A_3$		0.640	0.971	0.220	0.145	0.487	0.955	0.209	0.152
$A_4$		0.172	0.149	0.211	0.144	0.069	0.128	0.206	0.164
$p = 4$									
$H_0$		0.150	0.277	0.165	0.199	0.054	0.057	0.058	0.058
$A_1$	$k = 0.25$	0.057	0.054	0.050	0.051	0.068	0.061	0.053	0.049
	$\kappa = 0.50$	0.582	0.071	0.069	0.053	0.281	0.061	0.040	0.045
	$\kappa = 0.75$	0.816	0.602	0.073	0.054	0.659	0.430	0.072	0.065
$A_2$	$\kappa = +2.5$	0.060	0.971	0.318	0.776	0.053	0.941	0.317	0.821
	$\kappa = -2.5$	0.088	0.924	0.932	0.627	0.045	0.891	0.930	0.691
$A_3$		0.646	0.975	0.203	0.121	0.493	0.933	0.202	0.161
$A_4$		0.211	0.151	0.239	0.155	0.098	0.141	0.237	0.172
$p = 6$									
$H_0$		0.126	0.273	0.165	0.216	0.054	0.059	0.059	0.062
$A_1$	$\kappa = 0.25$	0.058	0.052	0.054	0.054	0.070	0.047	0.050	0.048
	$\kappa = 0.50$	0.552	0.054	0.048	0.055	0.286	0.069	0.053	0.043
	$\kappa = 0.75$	0.841	0.471	0.047	0.058	0.715	0.404	0.050	0.063
$A_2$	$\kappa = +2.5$	0.045	0.958	0.319	0.786	0.034	0.949	0.334	0.841
	$\kappa = -2.5$	0.111	0.878	0.927	0.571	0.045	0.862	0.930	0.636
$A_3$		0.651	0.955	0.180	0.110	0.502	0.938	0.192	0.148
$A_4$		0.236	0.162	0.272	0.181	0.148	0.241	0.261	0.194

Note: The results based on asymptotic critical values are reported in the first four columns. The results using bootstrap critical values are displayed in the last four columns, and indicated by (b) in the table. Reported powers are size-corrected.

of the asymptotic backtests, and the last four columns embed the bootstrap based tests.

As previously discussed, the use of asymptotic critical values (based on a  $\chi^2$  distribution) induces important size distortions. For instance, with sample size  $T = 500$ , and  $p = 6$ , the four test statistics  $J_1$ ,  $J_2$ ,  $I$ , and  $S$ , display empirical sizes 0.126, 0.273, 0.165, 0.216, respectively. These distortions are caused by poor inference made on regression parameters

Table 2: Empirical rejection rates of the backtests at 5% significance level,  $T = 2500$

		$J_1$	$J_2$	$I$	$S$	$J_1^{(b)}$	$J_2^{(b)}$	$I^{(b)}$	$S^{(b)}$
$p = 1$									
$H_0$		0.090	0.186	0.141	0.172	0.045	0.051	0.054	0.054
$A_1$	$\kappa = 0.25$	0.935	0.503	0.063	0.046	0.897	0.222	0.052	0.055
	$\kappa = 0.50$	0.998	1.000	0.074	0.168	0.995	1.000	0.071	0.112
	$\kappa = 0.75$	0.994	1.000	0.108	0.472	0.985	1.000	0.080	0.364
$A_2$	$\kappa = +2.5$	0.360	1.000	0.914	1.000	0.308	1.000	0.872	0.997
	$\kappa = -2.5$	0.303	1.000	1.000	0.987	0.249	1.000	1.000	0.983
$A_3$		0.990	1.000	0.749	0.807	0.984	1.000	0.679	0.711
$A_4$		0.855	0.518	0.718	0.552	0.761	0.361	0.614	0.448
$p = 2$									
$H_0$		0.103	0.184	0.160	0.178	0.045	0.052	0.052	0.056
$A_1$	$\kappa = 0.25$	0.849	0.308	0.058	0.069	0.745	0.208	0.055	0.026
	$\kappa = 0.50$	0.998	1.000	0.065	0.153	0.997	0.999	0.051	0.113
	$\kappa = 0.75$	0.995	1.000	0.107	0.485	0.992	1.000	0.086	0.407
$A_2$	$\kappa = +2.5$	0.141	1.000	0.893	0.997	0.102	1.000	0.865	0.996
	$\kappa = -2.5$	0.231	1.000	0.998	0.988	0.192	1.000	0.997	0.985
$A_3$		0.988	1.000	0.700	0.879	0.983	1.000	0.638	0.827
$A_4$		0.879	0.507	0.794	0.640	0.803	0.426	0.725	0.545
$p = 4$									
$H_0$		0.090	0.163	0.126	0.137	0.049	0.054	0.049	0.055
$A_1$	$\kappa = 0.25$	0.632	0.271	0.058	0.069	0.503	0.209	0.050	0.052
	$\kappa = 0.50$	0.998	1.000	0.079	0.085	0.998	1.000	0.054	0.090
	$\kappa = 0.75$	0.997	1.000	0.085	0.415	0.996	1.000	0.076	0.428
$A_2$	$\kappa = +2.5$	0.104	1.000	0.893	0.997	0.080	1.000	0.871	0.997
	$\kappa = -2.5$	0.384	1.000	0.995	0.980	0.330	1.000	0.994	0.981
$A_3$		0.984	1.000	0.577	0.826	0.981	1.000	0.549	0.840
$A_4$		0.894	0.530	0.767	0.651	0.851	0.474	0.741	0.589
$p = 6$									
$H_0$		0.108	0.172	0.125	0.139	0.051	0.053	0.055	0.057
$A_1$	$\kappa = 0.25$	0.470	0.318	0.055	0.060	0.416	0.219	0.047	0.052
	$\kappa = 0.50$	0.997	0.997	0.058	0.114	0.997	0.997	0.057	0.114
	$\kappa = 0.75$	0.999	1.000	0.069	0.497	0.998	1.000	0.080	0.502
$A_2$	$\kappa = +2.5$	0.095	1.000	0.854	1.000	0.082	1.000	0.865	1.000
	$\kappa = -2.5$	0.573	0.998	0.995	0.985	0.548	0.998	0.995	0.985
$A_3$		0.996	1.000	0.549	0.878	0.992	1.000	0.580	0.879
$A_4$		0.913	0.609	0.812	0.664	0.897	0.521	0.798	0.646

Note: The results based on asymptotic critical values are reported in the first four columns. The results using bootstrap critical values are displayed in the last four columns, and indicated by (b) in the table. Reported powers are size-corrected.

in the extreme upper tail and arise from the fact that large sample theory for quantile regression does not apply sufficiently far in the tails. In contrast, the backtests based on bootstrap critical values give satisfactory size performances. We find the empirical sizes close to the nominal size of 5% for all reported sample sizes and risk levels. For moderate sample sizes, we thus recommend the use of bootstrap critical values rather than asymptotic ones.

Our backtests display good power performances. The results are discussed in details hereinafter. First, we find that the tests generally detect well the misspecified alternatives  $A_1$ ,  $A_2$ ,  $A_3$ , and  $A_4$ , and we note a general improvement of powers as the sample size  $T$  increases, 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.

Third, although the intercept and slope backtests exhibit lower power performances, they provide useful informations about the type of misspecification. We observe that the slope backtest performs better in alternatives  $A_1$  and  $A_3$ , while the intercept backtest is superior for alternative  $A_4$ . In line with these rejections,  $A_1$  and  $A_3$  mainly affect the expected value of the slope parameters indicating that the error is essentially multiplicative with respect to the true quantiles. On the contrary, alternative  $A_4$  induces distortions of the intercept coefficients, suggesting that the VaRs issued from the ES model are affected additively. In the latter case, the error is hence more global and not directly related to the true quantiles.

Finally, we show that the selection of the number  $p$  of risk levels is not crucial for detecting alternatives  $A_1$ ,  $A_2$ , and  $A_3$  since reported powers are weakly affected by  $p$ . 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. In contrast, for the last alternative  $A_4$ , we conclude that an increase of  $p$  is useful for detecting the misleading one-year historical

simulation method since we observe a general improvement of powers for larger values of  $p$ . 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.

## 5 Empirical application

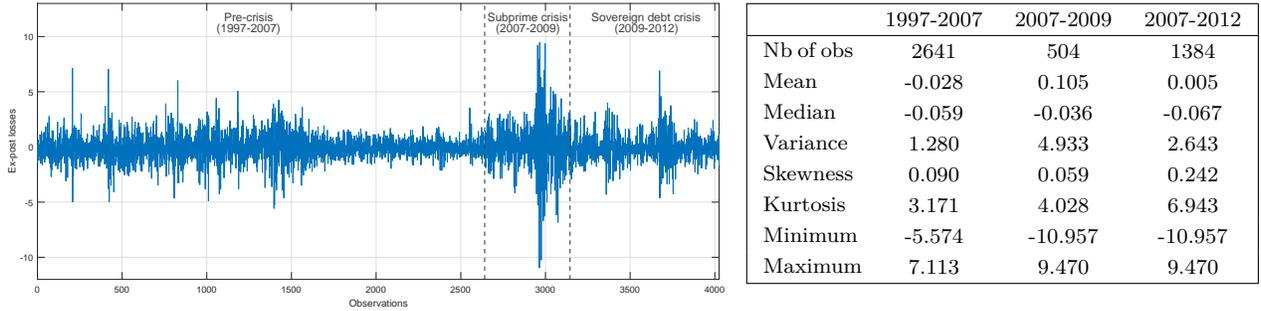
In this section, we apply our backtests to the daily returns of the S&P500 index, and illustrate their capabilities to identify a misspecified ES model. In addition, we provide a method for the adjustment of imperfect forecasts relying on our backtesting framework. In the sequel, we set  $\tau = 0.975$  to coincide with the regulatory ES coverage level. The probability levels  $u_j$ ,  $j = 1, 2, \dots, p$ , 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  $p = 1, 2, 4, 6$ .

### 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 until June 30, 2007, and we consider 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 together the subprime mortgage crisis and the European sovereign debt crisis, two major episodes of economic and 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.

The S&P500 series is depicted in Figure 1 with its three sub-periods. The in-sample period (1997-2007) is weakly volatile, while the out-of-sample periods (2007-2009, and 2007-

Figure 1: S&P500 daily losses (%), and descriptive statistics



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

2012) are characterized by more severe levels of volatility, with several extreme events. Figure 1 also provides some descriptive statistics. The variance and the average ex-post losses are more pronounced in the out-of-sample periods, especially for the period 2007-2009. In addition, we observe that the series is right-skewed with an excess kurtosis.

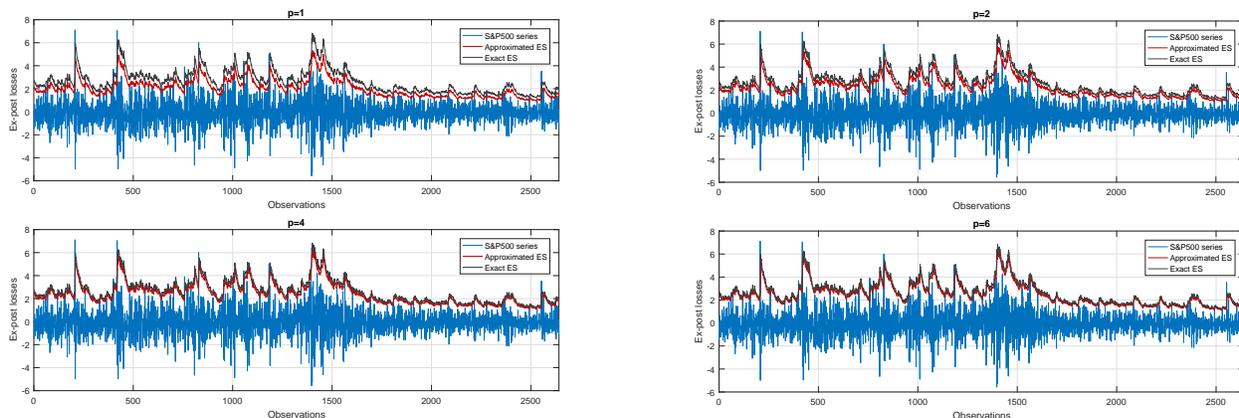
To predict the ES risk measure, we use the AR(1)-GARCH(1,1) model with Student innovations, as defined in (19). ES, and VaR, are respectively defined in Equations (20) and (21). The set of unknown parameters are estimated by maximum likelihood over the in-sample period. We obtain the following coefficient estimates  $\{\widehat{\delta}_0, \widehat{\delta}_1, \widehat{\gamma}_0, \widehat{\gamma}_1, \widehat{\gamma}_2, \widehat{v}\} = \{-0.0568, -0.0321, 0.0067, 0.0603, 0.9356, 9\}$ .

## 5.2 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  $p = 1, 2, 4, 6$  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 B.

Figure 2 reports the in-sample ES estimates obtained using the approximation and the exact calculation method. Two remarks should be made here. First, the approximation and

Figure 2: In-sample ES estimates issued from the approximation, and the exact calculation method



exact ES forecasts strongly correlate regardless of the value  $p$ , indicating that the approximation captures the ex-post losses information as properly as the exact calculation method. Because the approximation is obtained by combining VaRs, our finding is in accordance with Danielsson and Zhou (2016) who show that VaR and ES are equally informative. Second, we observe that the quality of the approximation is substantially improved when  $p$  slightly increases. As an illustration, we find that the ES estimates issued from the two competing approaches coincide almost completely using six quantiles. For its ease of implementation and its accuracy, the approximation appears hence to be highly appealing to compute the ES forecasts and to assess them.

Table 3 reports the p-values of the backtests. For a sake of clarity, we only report the p-values obtained with the bootstrap critical values. 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 of quantiles  $p$ . We also observe that the larger  $p$ , the smaller the p-value, indicating that the rejections are more severe when the number of risk levels increases. The test statistic  $J_2$  displays higher p-values. At 5% significance level, the backtest based on a single VaR no longer rejects the validity of the ES predictions, and the p-value based

Table 3: p-values of the backtests for several number  $p$  of quantiles

$p$	$J_1^{(b)}$	$J_2^{(b)}$	$I^{(b)}$	$S^{(b)}$
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
2 ( <i>regulatory levels</i> )	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
2 ( <i>regulatory levels</i> )	0.006	0.012	0.032	0.448

Note: p-values of the four backtests computed with  $p = 1, 2, 4, 6$  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.

on the regulatory levels of the BCBS is close to 5%, making the decision rule more unclear for those number of risk levels. Finally, with  $p = 2, 4, 6$ , the test statistic  $I$  invalidates the expected intercept coefficients, and hence the quality of the ES forecasts in an additive viewpoint. On the contrary, the test statistic  $S$  concludes that the slope parameters are as expected under the null hypothesis. 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 presumably due to the consistency of our backtesting methodology when applying it to larger sample sizes.

Given these results, it emerges that we should be very cautious in using a single quantile to assess the tail distribution of the risk model. Such procedures may lead practitioners to validate a model that generates mistaken forecasts. In addition, the results issued from the regulatory guidelines are contrasted. Two risk levels are not always enough to provide a sound conclusion about the quality 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. Typically, the proposed backtesting methodology gives satisfactory and informative results with four to six risk levels.

Table 4: Coefficient estimates issued from the multi-quantile regression,  $p = 6$

	$u_1$	$u_2$	$u_3$	$u_4$	$u_5$	$u_6$
Panel A. 2007-2009						
$\beta_0$	0.662 (0.295)	0.696 (0.296)	0.808** (0.227)	0.846** (0.240)	0.965* (0.429)	1.077* (0.265)
$\beta_1$	1.005 (0.093)	0.954 (0.089)	0.911* (0.056)	0.847** (0.053)	0.804 (0.142)	0.689** (0.042)
<i>Joint</i>	*	*	**	**		**
Panel B. 2007-2012						
$\beta_0$	0.376 (0.200)	0.510* (0.182)	0.692*** (0.195)	0.808*** (0.186)	0.777** (0.284)	0.784 (0.611)
$\beta_1$	1.031 (0.074)	0.974 (0.067)	0.902 (0.065)	0.851** (0.050)	0.826 (0.107)	0.787 (0.232)
<i>Joint</i>	**	**	**	**	**	**

Note: Standard errors are reported in parentheses. \*, \*\*, and \*\*\* indicate statistical significance at the 10%, 5% and 1% level, respectively, and they are obtained using bootstrap critical values.

Table 4 displays the coefficient estimates of the multi-quantile regression for  $p = 6$  risk levels, to help at understanding the reasons that explain the rejections of the ES forecasts. Panel A, and B, respectively provide the results for periods 2007-2009, and 2007-2012. In both panels, the intercept coefficients  $\beta_0$  are overestimated for all the risk levels  $u_1, u_2, \dots, u_6$ , while the slope coefficient  $\beta_1$  is overestimated for the first level  $u_1$ , and it becomes underestimated for all the remaining risk levels  $u_2, u_3, \dots, u_6$ . The average errors of  $\beta_0$ , and  $\beta_1$ , take respectively values 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 is more pronounced for the highest risk levels suggesting that the errors are more severe far in the tail.

Table 4 also provides inference on the regression parameters looking at each risk level separately. Hereinafter, we discuss the results at 5% significance level. We observe that the

intercept parameters are statistically not equal to zero for the intermediary risk levels  $u_3$  and  $u_4$  in panel A, and the additional  $u_5$  risk level is also significant 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. Finally, we also report joint inference, i.e. looking at both the intercept and slope coefficients. The results are provided in the row referred to as "joint". 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.

### 5.3 Adjusted ES forecasts

In what follows, we propose to 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 misleading ES model. Furthermore, in the case of invalid risk forecasts, the procedure allows to visually inspect whether the model overestimates, or underestimates the true unknown quantity, 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. More recently, Boucher et al. (2014) adjust imperfect VaR forecasts relying on backtesting frameworks, and Lazar and Zhang (2018) apply the same strategy to adjust imperfect ES forecasts. The method typically consists in modifying the coverage level  $\tau$  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  $\tau$ .

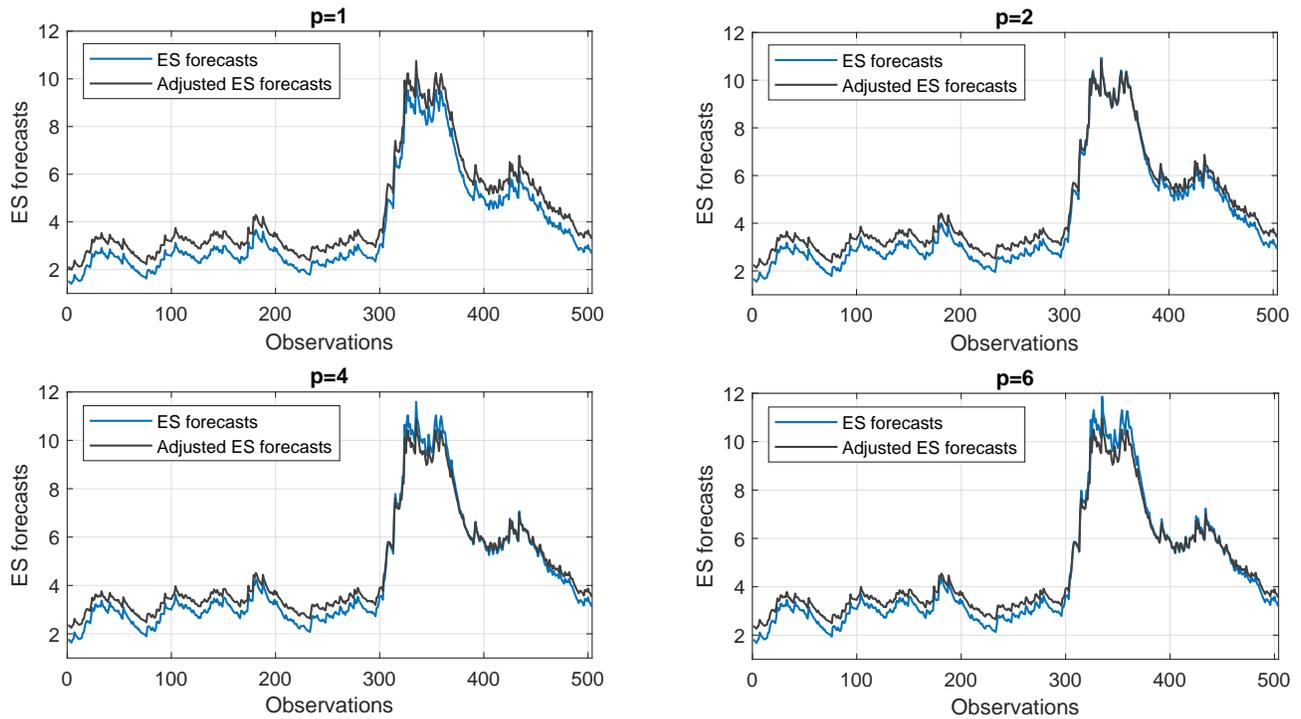
For ease of notation, we assume the parameters of the multi-quantile regression to be known while in practice we use estimated parameters. Formally, the adjusted VaR forecast at level  $u_j$ , and time  $t$ , is defined as the ex-ante prediction of the multi-quantile regression model, namely  $Q_{L_t}(u_j; \Omega_{t-1})$ . In view of representation (8), the initial imperfect VaR forecast is subsequently weighted by the regression parameters  $\beta_0(u_j)$  and  $\beta_1(u_j)$ , and thus incorporates the missing out-of-sample informations. The adjusted ES forecast at coverage level  $\tau$ , and time  $t$ , is derived from the approximation of ES in terms of several adjusted VaRs as follows:

$$ES_t^*(\tau) = \frac{1}{p} \sum_{j=1}^p Q_{L_t}(u_j; \Omega_{t-1}). \quad (22)$$

The newly adjusted ES forecasts are robust to model risk, as they meet the desirable properties on the regression coefficients. Indeed, if we compute our backtests 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  $u_1, u_2, \dots, u_p$ .

Figure 3 reports the ES predictions, and adjusted ES predictions for the period 2007-2009. The risk forecasts are built using the approximation with  $p = 1, 2, 4, 6$ . We observe that the AR(1)-GARCH(1,1) structure generally provides underestimated risk forecasts compared to the adjusted predictions. We note that the underestimation is more pronounced for the smallest predictions, the error being more severe when the risk forecasts are originally small. Our procedure thus also serve at identifying whether the model generates overestimates, or underestimates, the latter case being more worrisome in a financial stability perspective. Finally, with an increasing number  $p$  of quantiles, we find that the ES forecasts are slightly overestimated when the variance of the series is larger, suggesting that the risk model overshoots the true volatility during these periods. This is presumably caused by the volatility

Figure 3: ES forecasts and adjusted ES forecasts over the period 2007-2009



persistence in the GARCH component. We observe similar results for the period 2007-2012, as well as when applying the risk levels of the BCBS. The corresponding Figures 4, and 5 are reported in Appendix C.

## 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, the BCBS has adopted ES in place of VaR as the new standard for risk management and regulatory requirements. One of the major obstacle to its implementation was undoubtedly the deficit of simple tools for the evaluation of its forecasts. In this article, we have introduced four easy-to-use regression-based back-tests of ES. Our methodology jointly assesses the quality of VaRs along the tail distribution

of the risk model. This approach has the advantage of being consistent with the regulatory guidances to verify if the underlying ES model delivers correct quantiles at levels 0.975 and 0.990 (BCBS, 2016). To this end, we regress the ex-post losses on the VaRs forecasts in a multi-quantile regression model, and test 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, 2018, Kratz, Lok, and McNeil, 2018, etc.). Our proposed backtests 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.

Using the popular AR(1)-GARCH(1,1) model with Student innovations, we apply our tests on the S&P500 index over the period 2007-2012. Our backtests clearly reject the quality of the ES forecasts in this period of financial turmoil. Beyond this result, we highlight the importance of choosing a sufficient number of quantiles to assess ES. The use of one or two risk levels is especially discouraged as they are not always enough to identify improper risk forecasts. On the contrary, four to six risk levels deliver much more sound decisions, suggesting an update of the current regulatory guidelines in favor of the evaluation of more than two quantiles.

Within the current debate on risk assessment, the proposed regression-based backtests appear to be valuable diagnostic tools in line with the expertise and skills of financial institutions and regulators. They have the advantages to be easy to implement, and to complete

the toolbox commonly used by risk managers. They are therefore more likely to be embraced by financial institutions and regulators as new standards for financial risk management.

## 7 Acknowledgements

We would like to thank Denisa Banulescu-Radu, Sylvain Benoit, Christophe Hurlin, Olivier Scaillet, and Sessi Tokpavi for fruitful discussions. We thank the ANR MultiRisk (ANR-16-CE26-0015-01) for supporting our research.

## References

- [1] Acerbi, C., and Szekely, B. (2014). Backtesting Expected Shortfall. *Risk Magazine*, 27, 76–81.
- [2] Acerbi, C., and Tasche, D. (2002). On the Coherence of Expected Shortfall. *Journal of Banking and Finance*, 26(7), 1487-1503.
- [3] Argyropoulos, C., and Panopoulou, E. (2016). A Survey on Risk Forecast Evaluation. Working paper.
- [4] Artzner, P., Delbaen, F., Eber, J.M., and Heath, D. (1999). Coherent Measures of Risk. *Mathematical Finance*, 9(3), 203-228.
- [5] Basel Committee on Banking Supervision (2010). Basel III: A Global Regulatory Framework for More Resilient Banks and Banking Systems. Consultation paper, December.
- [6] Basel Committee on Banking Supervision (2016). Minimum Capital Requirements for Market Risk. Consultation paper, January.
- [7] Bayer, S., and Dimitriadis, T. (2018). Regression Based Expected Shortfall Backtesting. Working paper.
- [8] Berkowitz, J. (2001). Testing Density Forecasts, With Applications to Risk Management. *Journal of Business and Economic Statistics*, 19(4), 465–474.
- [9] Boucher, C., Danielsson, J., Kouontchou, P., and Maillet, B. (2014). Risk Models-at-Risk. *Journal of Banking and Finance*, 44, 72-92.

- [10] Christoffersen, P. (2011). Elements of Financial Risk Management. *Academic Press*, 2nd edition.
- [11] 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.
- [12] Costanzino, N., and Curran, M. (2018). A Simple Traffic Light Approach to Backtesting Expected Shortfall. *Risks*, 6(1), 1-7.
- [13] Danielsson, J., and Zhou, C. (2016). Why Risk is so Hard to Measure. Working paper.
- [14] Du, Z., and Escanciano, J.C. (2017). Backtesting Expected Shortfall: Accounting for Tail Risk. *Management Science*, 63(4), 901-1269.
- [15] 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.
- [16] Engle, R., and Manganelli, S. (2004). CAViaR: Conditional Autoregressive Value at Risk by Regression Quantiles. *Journal of Business and Economic Statistics*, 22(4), 367-381.
- [17] Fissler, T., and Ziegel, J. (2016). Higher Order Elicitability and Osband’s Principle. *Annals of Statistics*, 44(4), 1680-1707.
- [18] Freedman, D. (1981). Bootstrapping Regression Models. *Annals of Statistics*, 9(6), 1218-1228.
- [19] Gaglianone, W.P., Lima, L.R., Linton, O., and Smith, D.R. (2011). Evaluating Value-at-Risk Models via Quantile Regression. *Journal of Business and Economic Statistics*, 29(1), 150-160.
- [20] Gneiting, T. (2011). Making and Evaluating Point Forecasts. *Journal of the American Statistical Association*, 106(494), 746-762.
- [21] Gouriéroux, C., and Zakoïan, J.M. (2013). Estimation-Adjusted VaR. *Econometric Theory*, 29(4), 735-770.
- [22] 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.

- [23] Hurlin, C., Laurent, S., Quaadvlieg, R., and Smeekes, S. (2017). Risk Measure Inference. *Journal of Business and Economic Statistics*, 35(4), 499-512.
- [24] Jorion, P. (2006). Value at Risk: The New Benchmark for Managing Financial Risk. *McGraw-Hill*, third edition.
- [25] Kerkhof, J., and Melenberg, B. (2004). Backtesting for Risk-Based Regulatory Capital. *Journal of Banking and Finance*, 28(4), 1845–1865.
- [26] Koenker, R., Chernozhukov, V., He, X. and Peng, L. (2018). Handbook of Quantile Regression. *Chapman and Hall/CRC Handbooks of Modern Statistical Methods*.
- [27] Koenker, R., and Xiao, Z. (2002). Inference on the Quantile Regression Process. *Econometrica*, 70(4), 1583–1612.
- [28] Kratz, M., Lok, Y.H., and McNeil, A.J. (2018). Multinomial VaR backtests: A simple Implicit Approach to Backtesting Expected Shortfall. *Journal of Banking and Finance*, 88, 393-407.
- [29] Lazar, E., and Zhang, N. (2018). Model Risk of Expected Shortfall. Working paper.
- [30] 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.
- [31] Nolde, N., and Ziegel, J. (2017). Elicitability and Backtesting: Perspectives for Banking Regulation. *Annals of Applied Statistics*, 11(4), 1833-1874.
- [32] Patton, A.J., Ziegel, J.F., and Chen, R. (2018). Dynamic Semiparametric Models for Expected Shortfall (and Value-at-Risk). Forthcoming in *Journal of Econometrics*.
- [33] Powell, J.L. (1984). Least Absolute Deviations Estimation for the Censored Regression Model. *Journal of Econometrics*, 25(3), 303-325.
- [34] Taylor, J.W. (2017). Forecasting Value at Risk and Expected Shortfall using a Semiparametric Approach Based on the Asymmetric Laplace Distribution. Forthcoming in *Journal of Business and Economic Statistics*.
- [35] White, H.L., Kim, T.H., and Manganelli, S. (2008). Modeling Autoregressive Conditional Skewness and Kurtosis with Multi-Quantile CAViaR. In Bollerslev T., Russell, J.,

and Watson, M. editors, *Volatility and Time Series Econometrics: Essays in Honor of Robert F. Engle*.

- [36] White, H.L., Kim, T.H., and Manganelli, S. (2015). VAR for VaR: Measuring Tail Dependence using Multivariate Regression Quantiles. *Journal of Econometrics*, 187(1), 169-188.
- [37] Wong, W.K. (2008). Backtesting Trading Risk of Commercial Banks using Expected Shortfall. *Journal of Banking and Finance*, 32(7), 1404–1415.

# Appendix

## A - Consistent variance-covariance matrix estimation

In what follows, we provide a consistent estimator of the variance-covariance matrix  $\Sigma$ . The methodology is derived from White, Kim, and Manganelli (2008, 2015) who provide asymptotic theory in a multi-quantile regression framework. A consistent estimate of  $\Sigma$  can be obtained from the decomposition of the Huber (1976) sandwich form and is thus given by  $\widehat{\Sigma} = \widehat{A}^{-1}\widehat{V}\widehat{A}^{-1}$ . In what follows, we provide consistent estimators  $\widehat{A}$ , and  $\widehat{V}$ . To obtain  $\widehat{V}$ , we apply a simple plug-in estimator as follows:

$$\widehat{V} = T^{-1} \sum_{t=1}^T \widehat{\eta}_t \widehat{\eta}_t', \quad (23)$$

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  $\widehat{Q}_{L_t}(u_j, \Omega_{t-1}) = \widehat{\beta}_0(u_j) + \widehat{\beta}_1(u_j) \text{VaR}_t(u_j)$ , 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  $\Omega_{t-1}$  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.  $\widehat{A}$  is then given by

$$\widehat{A} = (2\widehat{c}_T T)^{-1} \sum_{t=1}^T \sum_{j=1}^p \mathbb{1}(|\widehat{\epsilon}_{j,t}| \leq \widehat{c}_T) \nabla \widehat{Q}_{L_t}(u_j, \Omega_{t-1}) \nabla' \widehat{Q}_{L_t}(u_j, \Omega_{t-1}), \quad (24)$$

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.

## B - Exact calculation method of ES

In this section, we describe the methodology used for the computation of exact ES forecasts at coverage level  $\tau$ . Several techniques are available in practice. In the following, because the distribution properties of the innovations are known, 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  $S = 500000$  in the empirical application.
2. Compute the ES at time  $t$  of the standardized innovation  $\eta_t$  as the Monte Carlo average of the simulated innovations:

$$m(\tau) = \frac{1}{\sum_{s=1}^S \mathbb{1}(\eta_t^s \geq F_v^{-1}(\tau))} \sum_{s=1}^S \eta_t^s \times \mathbb{1}(\eta_t^s \geq F_v^{-1}(\tau)),$$

where  $F_v^{-1}(\tau)$  is the  $\tau$ -quantile of the innovation distribution and is obtained as follows

$$F_v^{-1}(\tau) = \text{percentile}(\{\eta_t^s\}_{s=1}^S, 100\tau).$$

3. Compute the ES at time  $t$  as follows:

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

## C - Adjusted ES forecasts

Figure 4: ES forecasts and adjusted ES forecasts over the period 2007-2012

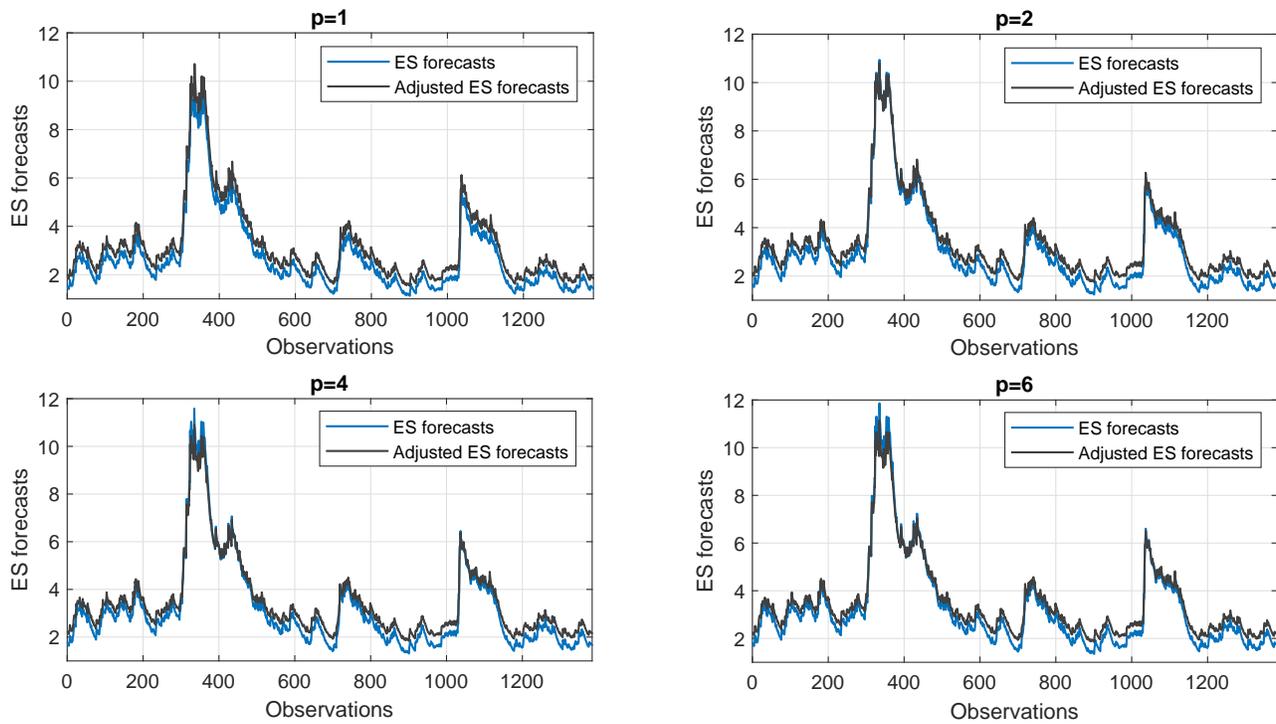


Figure 5: ES forecasts and adjusted ES forecasts over the periods 2007-2009 (on the left) and 2007-2012 (on the right) with the two regulatory risk levels

