Conditional Encompassing Test for Expected Shortfall Forecasts: A GMM Approach

Xiaochun Liu*†

November 27, 2020

Abstract

This paper proposes a conditional encompassing test for comparing expected shortfall forecasts in an out-of-sample framework. Particularly, the conditional encompassing test is developed with a tick loss function for Value at Risk (VaR) forecasts and a quadratic loss function conditional on VaR forecasts for expected shortfall forecasts. Then, a joint evaluation of the forecasts can be performed through a recursive GMM. In the case where encompassing is rejected, the proposed test provides a basis for the combination of the expected shortfall forecasts that outperforms individual forecasts. A simulation study obtains good size and power of the test in finite samples. Potential of the proposed test is empirically illustrated in comparing forecasts obtained from the recently developed risk models for the S&P 500 index returns.

Keywords: Conditional forecast efficiency, Out of sample forecast evaluation, Generalized method of moments, Optimal forecast combination weights, Asymmetric Laplace distribution

JEL: C12, C52, C58

*Department of Economics, Finance and Legal Studies, Culverhouse College of Business, University of Alabama, Tuscaloosa Alabama 35487 USA. Email: xliu121@ua.edu

†I acknowledge the International Institute of Forecasters and SAS Award grant to support this research on forecasting.

1 Introduction

The importance of expected shortfall has recently become more institutional since Basel Committee on Banking Supervision (BCBS) revised in 2016 the market risk framework to enhance a shift from Value at Risk (VaR) to an expected shortfall (ES) measure of risk under stress.[1] As a coherent measure of tail risks, expected shortfall is defined as the expected return in the part of the return distribution that is more extreme than a given quantile (Artzner et al., 1999; Tasche, 2002; Gerlach and Chen, 2016). Hence, the use of expected shortfall helps ensure a more prudent capture of tail risks and capital adequacy of commercial banks during periods of significant financial market stress. The current literature in financial econometrics and risk management has developed a variety of approaches to estimate expected shortfall.[2]

However, the evaluation and comparison of expected shortfall forecasts are still ongoing in its infant stage. For instance, in the direction of backtesting expected shortfall forecasts, studies have focused on the *absolute* evaluation, that is, on testing whether a forecasting model is correctly specified or whether a sequence of forecasts satisfies certain optimal properties, see, e.g., Kerkhof and Melenberg (2004), Wong (2008), Bayer and Dimitriadis (2020), among others.

A practical problem with the *absolute* evaluation, nonetheless, is that if different models are rejected as being misspecified or if more than one model are accepted, then the tests provide no guidance as to which one to choose. In this article I thus focus on the *relative* evaluation, which involves comparing the performance of competing, possibly misspecified models or sequences of forecasts for a variable and choosing the one that performs the best.

The goal of this paper is to develop a relative evaluation by testing conditional efficiency among expected shortfall forecasts. A forecast is said to be conditionally efficient if the expected loss of a combination of that forecast and a rival forecast is not significantly less than the expected loss of the original forecast alone.[3] In this regard, one concludes that the original forecast conditionally encompasses the rival, as it is able to explain the predictive ability of the

[1] The standards of minimum capital requirements for market risks published in January 2016 is available at https://www.bis.org/bcbs/publ/d352.htm.

[2] See, e.g., Zhu and Galbraith (2011), Chen et al. (2012), Gerlach and Chen (2016), Taylor (2019), Gerlach and Chen (2017), among others. Nadarajah (2014) provides a comprehensive review for the estimation methods of expected shortfall.

[3] Early studies have applied this encompassing principle for conditional mean forecasts. See, e.g., Lu and Mizon (1996), Clark and McCracken (2001), Fang (2003), Clements and Harvey (2010), among others,

rival.

Based on this encompassing principle, I therefore propose a conditional encompassing test, which involves a tick loss function for VaR forecasts and a quadratic loss function conditional on the VaR forecasts for ES forecasts. In this regard, the performance of ES forecasts can be evaluated conditional on the VaR forecasts. In particular, the test allows a standard recursive GMM to estimate optimal combination weights for the VaR and ES forecasts, and then the corresponding asymptotic properties of the GMM estimates are used to construct a Wald type test for the null hypothesis that a forecast conditionally encompasses a rival forecast.

An important feature of the conditional encompassing test is that it gives a theoretical basis for combinations of expected shortfall forecasts in cases when neither forecast encompasses its competitor.[4] Yet, expanding the information set through combination is particularly useful for evaluating expected shortfall conditional on a quantile usually at a small probability level, i.e., 1-5% in the Basel Accords regulations. VaR and ES at extreme probabilities are very sensitive to the few observations in the tails of a sample distribution, and hence, combining forecasts of different information sets could be an effective way to make the forecast performance more robust to the effects of sample-specific factors, such as fewer observations in extreme tails, outliers, and so on.

This paper conducts Monte Carlo simulations to examine asymptotic properties of the proposed test. Using daily S&P 500 index returns, I empirically illustrate the usefulness of the conditional expected shortfall forecast encompassing (CESFE) test in evaluating and comparing ES forecasts obtained from the recently developed risk models, including the parametric models of Chen et al. (2012) and the semiparametric models of Taylor (2019).

The work closely related to this paper is the study of Dimitriadis and Schnaitmann (2020). Nonetheless, this paper differs from their study in that they propose an unconditional encompassing test using the 2-elicitable loss functions, developed in Fissler and Ziegel (2016), for jointly evaluating VaR and ES forecasts. In addition, they implement the test through an M-estimation of optimal combination weights. The simulation study in Section 5.3 shows that the

[4]From a theoretical viewpoint, forecast combination can be seen as a way to pool the information contained in the individual forecasts, and its benefits have been widely advocated by a large amount of studies. See, e.g., e.g., Stock and Watson (1999&2003), Fang (2003), Eklund and Karlsson (2007), among others.

conditional and unconditional encompassing tests have their respective strengths in finite samples. In particular, the conditional test obtains better test size, while the unconditional test has stronger test power when the degree of model misspecification is relatively low. As discussed in Sections 4.2&5.3, the performance difference is partly due to the quadratic loss function chosen for ES forecasts conditional on VaR forecasts, which captures the interdependence between ES and VaR by a non-zero off-diagonal element in the Jacobian derivative matrix in the Wald test.

The remainder of the paper is organized as follows. Section 2 defines conditional expected shortfall and introduces forecast environment. Section 3 defines encompassing of conditional expected shortfall forecasts with the chosen loss functions for jointly evaluating VaR and ES forecasts. Section 4 constructs the conditional encompassing test through a recursive GMM and obtains its asymptotic properties to compute the test statistics. Section 5 examines the size and power properties of the test through a simulation study. Section 6 empirically illustrates the test in evaluating and comparing alternative ES forecasts for the S&P 500 index returns. Section 7 concludes the paper. The appendix presents proofs.

2 Conditional Expected Shortfall and Forecast Environment

Consider a stochastic process $\boldsymbol{\mathcal{V}} \equiv \left\{\boldsymbol{\mathcal{V}}_t : \Omega \to \mathbb{R}^{k+1},\ k \in \mathbb{N},\ t = 1, ..., T\right\}$ defined on a complete probability space $(\Omega, \mathcal{F}, P)$, where $\mathcal{F} \equiv \{\mathcal{F}_t, t = 1, ..., T\}$ and $\mathcal{F}_t \equiv \sigma\left\{\boldsymbol{\mathcal{V}}_s,\ s \leq t\right\}$ is a chosen σ-field. The observed vector $\boldsymbol{\mathcal{V}}_t$ is partitioned as $\boldsymbol{\mathcal{V}}_t \equiv \left(Y_t, \boldsymbol{X}_t^{'}\right)^{'}$, where $Y_t : \Omega \to \mathbb{R}$ is a continuous random variable of interest, and $\boldsymbol{X}_t : \Omega \to \mathbb{R}^k$ is a $k \times 1$ vector of explanatory variables. This paper is interested in the expected shortfall forecast of the distribution of Y_{t+1} at a given probability level, $\tau \in (0, 1)$, conditional on the information set $\mathcal{F}_t$, defined as

$$ES_{t+1}(\tau|\mathcal{F}_t) = \frac{1}{\tau}\int_0^{\tau} VaR_{t+1}\left(\iota|\mathcal{F}_t\right) d\iota \tag{2.1}$$

where $VaR_{t+1}\left(\tau|\mathcal{F}_t\right)$ is the $\tau \times 100\%$th VaR forecast of the distribution of Y_{t+1} conditional on $\mathcal{F}_t$, defined as

$$Pr\left(Y_{t+1} < VaR_{t+1}|\mathcal{F}_t\right) = \tau \tag{2.2}$$

or

$$VaR_{t+1}(\tau|\mathcal{F}_t) \equiv F^{-1}_{Y_{t+1}}(\tau|\mathcal{F}_t) \tag{2.3}$$

where $F^{-1}_{Y_{t+1}}(\cdot|\mathcal{F}_t)$ is the inverse of the conditional cumulative distribution function ($F_{Y_{t+1}}(\cdot|\mathcal{F}_t)$) of Y_{t+1}, which is assumed continuous. Conditional expected shortfall in (2.1) can alternatively be expressed as

$$ES_{t+1}(\tau|\mathcal{F}_t) = E_t[Y_{t+1}|Y_{t+1} < VaR_{t+1}(\tau|\mathcal{F}_t)] \tag{2.4}$$

Both (2.1) and (2.4) show that $ES_{t+1}(\tau|\mathcal{F}_t)$ is defined based on $VaR_{t+1}(\tau|\mathcal{F}_t)$.[5] For example, (2.4) represents the expected value of Y_{t+1} conditional on Y_{t+1} being more extreme than its $\tau \times 100\%$th quantile at time $t+1$.

The goal of this paper is to propose a test for comparing alternative sequences of one-step-ahead forecasts of $ES_{t+1}(\tau|\mathcal{F}_t)$. I perform the evaluation in an out-of-sample fashion. This involves dividing the sample of size T into an in-sample part of size m and an out-of-sample part of size n, so that $T = m + n$. The in-sample portion is used to produce the first set of forecasts, and the evaluation is performed over the remaining out-of-sample portion. In particular, the forecasts may be based on parametric models or be generated by semiparametric or nonparametric techniques. The forecasts can be produced using either a fixed forecasting scheme or a rolling window forecasting scheme.[6] For example, for a parametric model, a fixed forecasting scheme involves estimating the parameters only once on the first m observations and using these estimates to produce all of the forecasts for the out-of-sample period $t = m+1, ..., T$. In contrast, a rolling window forecasting scheme reestimates parameters at each out-of-sample point $t = m+1, ..., T$ using an estimation sample containing the m most recent observations, that is, the observation from $t-m+1$ to t.

To further simplify the notation, I hereafter drop the reference to the index τ and the conditioning information $\mathcal{F}_t$ to simply denote the $\tau \times 100\%$th ES and VaR at time $t+1$

[5] (2.1) is often referred to as the integrated conditional quantile function (ICQF), see, e.g., Peracchi and Tanase (2008), Leorato et al. (2012), among others

[6] The test requires the $\mathcal{F}_t$-measurable functions of VaR_{t+1} and ES_{t+1} constant over time. This implies that the use of an expanding estimation window (recursive forecasting scheme) is not allowed, whereas either a fixed or a rolling window of constant length satisfies the requirement. See also Giacomini and Komunjer (2005).

conditional on $\mathcal{F}_t$ as ES_{t+1} and VaR_{t+1}. As a general rule, a lower-case letter is used to denote observations of the corresponding random variable (i.e., $\boldsymbol{v}_t$ and $\boldsymbol{\mathcal{V}}_t$). The in-sample size m is a finite constant, chosen by the user a priori. As a consequence, all of the results in this paper should be interpreted as being conditional on the given choice of m, but for ease of notation I choose not to make this dependence explicit.

3 Encompassing Principles for Conditional Expected Shortfall Forecasts

The approach to comparing conditional expected shortfall forecasts is based on the principle of encompassing, see, e.g., Lu and Mizon (1996), Harvey et al. (1998), Clark and McCracken (2001) and West (2001), among others. Encompassing arises when one of two competing forecasts is able to explain the predictive ability of its rival. In this sense, a test for forecast encompassing is a test of the conditional efficiency of a forecast, where a forecast is said to be conditionally efficient if the expected loss of a combination of that forecast and a rival forecast is not significantly less than the expected loss of the original forecast alone (Clements and Hendry, 1998; Giacomini and Komunjer, 2005).

The two key ingredients of a forecast encompassing test are, therefore, (1) the loss function that is involved in the computation of the expected loss and (2) the weights of the forecast combination. The choice of the loss function is closely related to which characteristic of the unknown future distribution of the variable one wants to forecast. Let $\hat{f}_{t+1}$ be a forecast of some characteristic of interest of a random variable Y_{t+1}, conditional on the information set at time t. The forecast $\hat{f}_{t+1}$ is said to be optimal at time $t+1$ if it minimizes $E_t\left[\mathcal{L}\left(Y_{t+1}-\hat{f}_{t+1}\right)\right]$, where $\mathcal{L}$ is some loss function such that $\mathcal{L}:\mathbb{R}\to\mathbb{R}^+$.

3.1 The Loss Functions

The encompassing test requires two proper loss functions, one for VaR forecasts and the other for expected shortfall forecasts conditional on VaR forecasts. Specifically, I consider the con-

ventional quantile "tick" or "check" loss function for VaR_{t+1}, given by

$$\mathcal{T}_\tau\left(Y_{t+1}-\widehat{VaR}_{t+1}\right)\equiv\left[\tau-\mathbb{I}\left(Y_{t+1}<\widehat{VaR}_{t+1}\right)\right]\left(Y_{t+1}-\widehat{VaR}_{t+1}\right) \tag{3.1}$$

which is the asymmetric linear loss function of order τ. This tick function $\mathcal{T}$ is the implicit loss function whenever the object of interest is a forecast of a particular quantile of the conditional distribution of Y_{t+1}. Giacomini and Komunjer (2005) show that the focus on conditional (rather than unconditional) expected loss, such as (3.1), is a central feature of the treatment of both evaluation and combination of forecasts and distinguishes their approach from related literature, e.g., Granger (1989), Taylor and Bunn (1998), Elliott and Timmermann (2004), among others. This paper carries on this central feature to the evaluation and combination of conditional expected shortfall forecasts, as discussed in details later.

In particular, the loss function considered for the object of interest, ES_{t+1}, takes the form

$$\mathcal{L}_\tau\left(Y_{t+1}-\widehat{ES}_{t+1};\widehat{VaR}_{t+1}\right)=\left(Y_{t+1}-\widehat{ES}_{t+1}\right)^2\mathbb{I}\left(Y_{t+1}<\widehat{VaR}_{t+1}\right) \tag{3.2}$$

which depends on the forecast of conditional value-at-risk for time $t+1$. Specifically, the following lemma provides the basis for (3.2).

Lemma 1. *(Conditionally consistent criterion). Under the definitions (2.1)-(2.4), if* $\widehat{VaR}_{t+1}\xrightarrow{p}VaR_{t+1}$ *and* $\widehat{ES}_{t+1}\xrightarrow{p}ES_{t+1}$ *are consistent estimators as* $n\to\infty$*, then the residual sequence,* $\left\{e_{t+1}:\equiv Y_{t+1}-\widehat{ES}_{t+1}\right\}_{t=m}^{T-1}$ *should be i.i.d. and that, conditional on* $\mathbb{I}\left(Y_{t+1}<\widehat{VaR}_{t+1}\right)$*, it has expected value zero, such that*

$$E_t\left[\left(Y_{t+1}-\widehat{ES}_{t+1}\right)\mathbb{I}\left(Y_{t+1}\leq\widehat{VaR}_{t+1}\right)\right]=0,\qquad a.s.-P. \tag{3.3}$$

The proof of Lemma 1 is straightforward. Provided that $\widehat{VaR}_{t+1}$ and $\widehat{ES}_{t+1}$ are assumed consistent estimators of VaR_{t+1} and ES_{t+1} as $n\to\infty$, (3.3) can directly be obtained from the definition of expected shortfall, (2.4), as the first-order moment condition of the expected loss function of (3.2) given $\widehat{VaR}_{t+1}$.

If (3.3) does not hold, then $\widehat{ES}_{t+1}$ is an inconsistent estimator of ES_{t+1} conditional on

$\widehat{VaR}_{t+1}$. A negative value, $e_{t+1} < 0$, therefore represents underestimation of this measure of risk for $\tau < 0.5$. The relevant literature has applied (3.3) to backtesting expected shortfall in absolute evaluations, see, e.g., McNeil and Frey (2000), Ergun and Jun (2010), and Zhu and Galbraith (2011), among others.

Of course, the validity of *Lemma 1* also depends on the assumption of the consistent estimator, $\widehat{VaR}_{t+1}$. Specifically, the following lemma expresses the first-order moment condition of quantile regression.

Lemma 2. *The loss function, (3.1), provides the first-order moment condition of quantile regression as*

$$E_t\left[\tau - \mathbb{I}\left(Y_{t+1} - \widehat{VaR}_{t+1} < 0\right)\right] = 0 \qquad a.s. - P. \tag{3.4}$$

The proof for (3.4) can be found in Koenker (2005, §4). The literature has also used this first-order condition as the basis in a variety of value-at-risk backtests. See Nieto and Ruiz (2016) for a review.

3.2 The Definition of Encompassing

Consider two competing methods, $\mathcal{M}_1$ and $\mathcal{M}_2$, which produce forecasts of conditional value-at-risk and expected shortfall. Let $\widehat{\boldsymbol{ES}}_{t+1} = \left(\widehat{ES}_{1,t+1}, \widehat{ES}_{2,t+1}\right)'$ and $\widehat{\boldsymbol{VaR}}_{t+1} = \left(\widehat{VaR}_{1,t+1}, \widehat{VaR}_{2,t+1}\right)'$ denote the forecasts of expected shortfall and value-at-risk from $\mathcal{M}_1$ and $\mathcal{M}_2$. This paper is interested in testing whether $\widehat{ES}_{1,t+1}$ from $\mathcal{M}_1$ conditionally encompasses $\widehat{ES}_{2,t+1}$ from $\mathcal{M}_2$ over the entire out-of-sample period for $t = m, ..., T-1$.[7]

Further, let $\boldsymbol{\theta} = (\theta_1, \theta_2)'$ and $\boldsymbol{w} = (w_1, w_2)'$, which lie in some compact subsets of $\mathbb{R}^2$, denote the choices of weights in the combinations of conditional value-at-risk and expected shortfall forecasts, respectively. The common practice is to obtain a weighted average of forecasts with the weights adding up to unity, i.e., $\theta_1 + \theta_2 = 1$ and $w_1 + w_2 = 1$. However, Granger and Ramanathan (1984) find that the best method is to add a constant term and not to constrain the weights to add up to unity. Therefore, in this paper the unity restriction is not imposed on $\boldsymbol{\theta}$ and $\boldsymbol{w}$. See also Giacomini and Komunjer (2005).

[7] For simplicity, I restrict attention to pairwise comparisons, but all of the techniques can readily be extended to the general case of multiple forecasts.

Based on the general principles of forecast encompassing,[8] it is said that the forecast of $ES_{1,t+1}$ obtained from $\mathcal{M}_1$ conditionally encompasses the forecast of $ES_{2,t+1}$ from $\mathcal{M}_2$ for time $t+1$ if and only if

$$E_t\left[\mathcal{L}_\tau\left(Y_{t+1}-\widehat{ES}_{1,t+1};\widehat{VaR}_{1,t+1}\right)\right] \leq E_t\left[\mathcal{L}_\tau\left(Y_{t+1}-\boldsymbol{w}'\widehat{\boldsymbol{ES}}_{t+1};\boldsymbol{\theta}'\widehat{\boldsymbol{VaR}}_{t+1}\right)\right] \tag{3.5}$$
$$a.s.-P., \forall\,(w_1,w_2)\in\Theta\subset\mathbb{R}^2$$
$$\forall\,(\theta_1,\theta_2)\in\Theta\subset\mathbb{R}^2$$

where $\mathcal{L}(\cdot)$ is the loss function defined in (3.2) for ES forecasts. In practice, testing the inequality (3.5) is not feasible, because it involves computing the expected loss for all $(w_1,w_2)\in\Theta$ and $(\theta_1,\theta_2)\in\Theta$. Instead, let $\boldsymbol{w}^*=(w_1^*,w_2^*)'$ and $\boldsymbol{\theta}^*=(\theta_1^*,\theta_2^*)'$ denote the optimal sets of combination weights for ES and VaR forecasts, respectively, which are the solutions to jointly minimize (3.1) and (3.2). Therefore, I have the following definition of encompassing for conditional expected shortfall forecasts.

Definition 1. (Conditional expected shortfall forecast encompassing, CESFE). Let $\widehat{ES}_{1,t+1}$ and $\widehat{ES}_{2,t+1}$ be alternative forecasts for ES_{t+1} and $\widehat{VaR}_{1,t+1}$ and $\widehat{VaR}_{2,t+1}$ be alternative forecasts for VaR_{t+1} from the competing methods, $\mathcal{M}_1$ and $\mathcal{M}_2$, respectively. Then, $\widehat{ES}_{1,t+1}$ is said to encompass $\widehat{ES}_{2,t+1}$ at time $t+1$ if and only if (3.5) is binding for $\left(\boldsymbol{\theta}^{*'}\ \boldsymbol{w}^{*'}\right)'$

$$E_t\left[\mathcal{L}_\tau\left(Y_{t+1}-\widehat{ES}_{1,t+1};\widehat{VaR}_{1,t+1}\right)\right] = E_t\left[\mathcal{L}_\tau\left(Y_{t+1}-\boldsymbol{w}^{*'}\widehat{\boldsymbol{ES}}_{t+1};\boldsymbol{\theta}^{*'}\widehat{\boldsymbol{VaR}}_{t+1}\right)\right]$$
$$a.s.-P. \tag{3.6}$$

that is, if and only if

$$(\theta_1^*,\theta_2^*,w_1^*,w_2^*)=(1,0,1,0) \tag{3.7}$$

[8] See e.g., Clements and Hendry (1998), Harvey et al. (1998), McCracken (2000), Clark and McCracken (2001), among others.

where $\boldsymbol{\theta}^* = (\theta_1^*, \theta_2^*)'$ and $\boldsymbol{w}^* = (w_1^*, w_2^*)'$ jointly minimize (3.1) and (3.2), as

$$(\theta_1^*, \theta_2^*) \equiv arg \underset{(\theta_1,\theta_2)\in\Theta\subset\mathbb{R}^2}{min} E_t\left[\mathcal{T}_\tau\left(Y_{t+1} - \left(\theta_1\widehat{VaR}_{1,t+1} + \theta_2\widehat{VaR}_{2,t+1}\right)\right)\right] \tag{3.8}$$

$$(w_1^*, w_2^*) \equiv arg \underset{(w_1,w_2)\in\Theta\subset\mathbb{R}^2}{min} E_t\left[\mathcal{L}_\tau\left(Y_{t+1} - \left(w_1\widehat{ES}_{1,t+1} + w_2\widehat{ES}_{2,t+1}\right); \boldsymbol{\theta}^{*'}\widehat{\boldsymbol{VaR}}_{t+1}\right)\right] \tag{3.9}$$

In this paper, I restrict attention to linear combinations. And, the equivalence between (3.6) and (3.7) follows from the fact that the right side of (3.6) is the minimum of the conditionally expected loss over Θ and Θ.

Consequently, the optimal combination of VaR forecasts from (3.8) satisfies the first-order condition, (3.4), as

$$E_t\left[\tau - \mathbb{I}\left(Y_{t+1} - \boldsymbol{\theta}^{*'}\widehat{\boldsymbol{VaR}}_{t+1} < 0\right)\right] = 0 \qquad a.s. - P. \tag{3.10}$$

See *Appendix A* for the proof of (3.10). Similarly, the vector of optimal weights $\boldsymbol{w}^*$ obtained from the joint estimation satisfies the first-order condition, (3.3), such that

$$E_t\left[\left(Y_{t+1} - \boldsymbol{w}^{*'}\widehat{\boldsymbol{ES}}_{t+1}\right)\mathbb{I}\left(Y_{t+1} < \boldsymbol{\theta}^{*'}\widehat{\boldsymbol{VaR}}_{t+1}\right)\right] = 0 \qquad a.s. - P. \tag{3.11}$$

See *Appendix B* for the proof of (3.11).

Acerbi and Tasche (2002) show evidence that expected shortfall can be estimated effectively even in cases where the usual estimators for VaR fail. On the other hand, Chen (2008) shows that a better VaR estimation does not guarantee a corresponding better ES estimation. In this regard, an encompassing test based on *Definition 1* must be flexible to accommodate a variety of possible encompassing scenarios. For example, $\boldsymbol{\theta}^* = (1, 0)$ and $\boldsymbol{w}^* = (0, 1)$ imply an extreme scenario that $\widehat{ES}_{2,t+1}$ encompasses $\widehat{ES}_{1,t+1}$, whereas $\widehat{VaR}_{1,t+1}$ encompasses $\widehat{VaR}_{2,t+1}$. In the case where $\boldsymbol{\theta}^* = (1, 0)$ and $\boldsymbol{w}^* = (w_1^*, w_2^*)$, it suggests that $\widehat{ES}_{1,t+1}$ and $\widehat{ES}_{2,t+1}$ should be combined via the optimal weights, $\boldsymbol{w}^*$, although $\widehat{VaR}_{1,t+1}$ encompasses $\widehat{VaR}_{2,t+1}$. In the next section I discuss implementation of the CESFE test. While the CESFE test in this paper is illustrated for *Definition 1* with $H_0 : (\theta_1^*, \theta_2^*, w_1^*, w_2^*) = (1, 0, 1, 0)$, it can easily be implemented to test other null hypotheses of encompassing.

4 Conditional Expected Shortfall Forecast Encompassing Test

To test the encompassing hypothesis in *Definition 1* for whether $\widehat{ES}_{1,t+1}$ encompasses $\widehat{ES}_{2,t+1}$ over the entire out-of-sample period, I jointly solve the optimal weights, $\boldsymbol{\theta}^*$ and $\boldsymbol{w}^*$ by implementing a standard GMM with the optimization procedure appropriately modified to accommodate the nondifferentiable loss functions.[9] Next I describe the estimation procedure for the optimal combination weights.

4.1 Generalized Method-of-Moments Estimation for Optimal Combination Weights

According to *Definition 1*, $\widehat{ES}_{1,t+1}$ conditionally encompasses $\widehat{ES}_{2,t+1}$ for all t, $m \leq t \leq T-1$ if and only if $\left(\boldsymbol{\theta}_{m+1}^{*'}, \boldsymbol{w}_{m+1}^{*'}\right)' = \cdots = \left(\boldsymbol{\theta}_{T}^{*'}, \boldsymbol{w}_{T}^{*'}\right)' = (1,0,1,0)'$. In other words, the optimal combination weights are constant in time and equal to $(1,0,1,0)'$. By (3.10) and (3.11), it should therefore be the case that for $\boldsymbol{e}_1 = (1,0)'$, $E\left\{\left[\tau - \mathbb{I}\left(Y_{t+1} - \boldsymbol{e}_1'\widehat{\boldsymbol{VaR}}_{t+1} < 0\right)\right]\boldsymbol{Z}_{1t}\right\} = 0$ and $E\left\{\left[\left(Y_{t+1} - \boldsymbol{e}_1'\widehat{\boldsymbol{ES}}_{t+1}\right)\mathbb{I}\left(Y_{t+1} < \boldsymbol{e}_1'\widehat{\boldsymbol{VaR}}_{t+1}\right)\right]\boldsymbol{Z}_{2t}\right\} = 0$ for all $\mathcal{F}_t$-measurable information functions, $\{\boldsymbol{Z}_{1t}, \boldsymbol{Z}_{2t}\}$, and for all t, $m \leq t \leq T-1$. In particular, $\boldsymbol{Z}_{1t}$ and $\boldsymbol{Z}_{2t}$ are $k_1 \times 1$ and $k_2 \times 1$ vectors of instrumental variables, respectively, which are observed at time t. $\{\boldsymbol{Z}_{1t}, \boldsymbol{Z}_{2t}\}$ is assumed strictly stationary and mixing series, and can include previous forecasts (or measures of past forecast performance), provided that they are produced by either a fixed or a rolling window forecasting scheme. The reason for this is that in these two cases the forecasts are constant measurable functions of a finite window of data and thus inherit the properties of stationarity and mixing from the underlying series (Giacomini and Komunjer, 2005).

Further, define $\boldsymbol{g}_1$ as a k_1-vector-valued function $\boldsymbol{g}_1 : \Theta \times \mathbb{R} \times \mathbb{R}^{k_1} \to \mathbb{R}^{k_1}$ and $\boldsymbol{g}_2$ as a

[9] One may consider a two-step encompassing test approach by (i) solving the optimal weight vector $\boldsymbol{\theta}^*$ for conditional value-at-risk forecasts in the first step, and (ii) then estimating the optimal weight vector $\boldsymbol{w}^*$ given $\widehat{\boldsymbol{\theta}}^*$ for conditional expected shortfall forecasts in the second step. This two-step approach is based on the similar weak exogenous reasoning of Engle (2002) as the first-step estimation does not involve $\boldsymbol{w}$. However, this two-step approach that generally involves some loss of estimation efficiency is a special case of the general approach discussed in this section.

k_2-vector-valued function $\boldsymbol{g}_2 : \Theta \times \mathbb{R} \times \mathbb{R}^{k_2} \to \mathbb{R}^{k_2}$, such that

$$\boldsymbol{g}_1(\boldsymbol{\theta}; y_{t+1}, \boldsymbol{z}_{1t}) \equiv \left[\tau - \mathbb{I}\left(y_{t+1} < \boldsymbol{\theta}' \widehat{\boldsymbol{VaR}}_{t+1}\right)\right] \boldsymbol{z}_{1t} \tag{4.1}$$

$$\boldsymbol{g}_2(\boldsymbol{w}; y_{t+1}, \boldsymbol{z}_{2t}, \boldsymbol{\theta}) \equiv \left[\left(y_{t+1} - \boldsymbol{w}' \widehat{\boldsymbol{ES}}_{t+1}\right) \mathbb{I}\left(y_{t+1} < \boldsymbol{\theta}' \widehat{\boldsymbol{VaR}}_{t+1}\right)\right] \boldsymbol{z}_{2t} \tag{4.2}$$

The key element in the implementation of the encompassing test is that under the null of encompassing, it has, based on (3.10) and (3.11), the following moment conditions

$$\boldsymbol{g}_1^o(\boldsymbol{\theta}^*; Y_{t+1}, \boldsymbol{Z}_{1t}) \equiv E\left[\boldsymbol{g}_1(\boldsymbol{\theta}^*; Y_{t+1}, \boldsymbol{Z}_{1t})\right] = 0 \tag{4.3}$$

$$\boldsymbol{g}_2^o(\boldsymbol{w}^*; Y_{t+1}, \boldsymbol{Z}_{2t}, \boldsymbol{\theta}^*) \equiv E\left[\boldsymbol{g}_2(\boldsymbol{w}^*; Y_{t+1}, \boldsymbol{Z}_{2t}, \boldsymbol{\theta}^*)\right] = 0 \tag{4.4}$$

jointly to be true, or equivalently

$$\boldsymbol{g}^o(\boldsymbol{\theta}^*, \boldsymbol{w}^*;\ Y_{t+1}, \boldsymbol{Z}_{1t}, \boldsymbol{Z}_{2t}) = E\left[\boldsymbol{g}(\boldsymbol{\theta}^*, \boldsymbol{w}^*;\ Y_{t+1}, \boldsymbol{Z}_{1t}, \boldsymbol{Z}_{2t})\right] = 0 \tag{4.5}$$

where $\boldsymbol{g}(\boldsymbol{\theta}, \boldsymbol{w};\ Y_{t+1}, \boldsymbol{Z}_{1t}, \boldsymbol{Z}_{2t}) = \left(\boldsymbol{g}_1(\boldsymbol{\theta}; Y_{t+1}, \boldsymbol{Z}_{1t})',\ \boldsymbol{g}_2(\boldsymbol{w}; Y_{t+1}, \boldsymbol{Z}_{2t}, \boldsymbol{\theta})'\right)'$ is a $k \times 1$ vector of the moment conditions with $k = k_1 + k_2$.

Given the out-of-sample portion of size $n = T - m$, consisting of the sequence of observations $(\boldsymbol{z}_{1m}, \boldsymbol{z}_{2m}, y_{m+1}, ..., \boldsymbol{z}_{1,T-1}, \boldsymbol{z}_{2,T-1}, y_T)'$, I then use Hansen's (1982) GMM approach to estimate $\boldsymbol{\theta}^*$ and $\boldsymbol{w}^*$ as a solution to the minimization problem of (4.5), denoted by $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{w}}$, as

$$\min_{\boldsymbol{\theta} \in \Theta, \boldsymbol{w} \in \Theta} \left[\boldsymbol{g}_n(\boldsymbol{\theta}, \boldsymbol{w})\right]' \hat{\boldsymbol{S}}_n^{-1} \left[\boldsymbol{g}_n(\boldsymbol{\theta}, \boldsymbol{w})\right] \tag{4.6}$$

where $\boldsymbol{g}_n(\cdot)$ is the sample moment function, $\boldsymbol{g}_n(\boldsymbol{\theta}, \boldsymbol{w}) \equiv n^{-1} \sum_{t=m}^{T-1} \boldsymbol{g}(\boldsymbol{\theta}, \boldsymbol{w};\ y_{t+1}, \boldsymbol{z}_{1t}, \boldsymbol{z}_{2t})$, and $\hat{\boldsymbol{S}}_n$ is a consistent estimator of the asymptotic variance matrix $\boldsymbol{S}$,

$$\boldsymbol{S} \equiv E\left[\boldsymbol{g}(\boldsymbol{\theta}^*, \boldsymbol{w}^*;\ Y_{t+1}, \boldsymbol{Z}_{1t}, \boldsymbol{Z}_{2t})\, \boldsymbol{g}(\boldsymbol{\theta}^*, \boldsymbol{w}^*;\ Y_{t+1}, \boldsymbol{Z}_{1t}, \boldsymbol{Z}_{2t})'\right] \tag{4.7}$$

which is a $k \times k$ positive semi-definite matrix. Using the fact that the first-order conditions (3.10) and (3.11) imply that $\{g(\boldsymbol{\theta}^*, \boldsymbol{w}^*; Y_{t+1}, \boldsymbol{Z}_{1t}, \boldsymbol{Z}_{2t}), \mathcal{F}_t\}$ is a martingale difference sequence,

I obtain a consistent estimator of $\boldsymbol{S}$ as

$$\begin{aligned}\hat{\boldsymbol{S}}_n\left(\tilde{\boldsymbol{\theta}}_n,\tilde{\boldsymbol{w}}_n\right) &\equiv \frac{1}{n}\sum_{t=m}^{T-1}\boldsymbol{g}\left(\tilde{\boldsymbol{\theta}}_n,\tilde{\boldsymbol{w}}_n;\, y_{t+1},\boldsymbol{z}_{1t},\boldsymbol{z}_{2t}\right)\boldsymbol{g}\left(\tilde{\boldsymbol{\theta}}_n,\tilde{\boldsymbol{w}}_n;\, y_{t+1},\boldsymbol{z}_{1t},\boldsymbol{z}_{2t}\right)^{'} \\ &= \frac{1}{n}\sum_{t=m}^{T-1}\begin{pmatrix} \boldsymbol{g}_1\left(\tilde{\boldsymbol{\theta}}_n;y_{t+1},\boldsymbol{z}_{1t}\right)\boldsymbol{g}_1\left(\tilde{\boldsymbol{\theta}}_n;y_{t+1},\boldsymbol{z}_{1t}\right)^{'} & \boldsymbol{g}_1\left(\tilde{\boldsymbol{\theta}}_n;y_{t+1},\boldsymbol{z}_{1t}\right)\boldsymbol{g}_2\left(\tilde{\boldsymbol{w}}_n;y_{t+1},\boldsymbol{z}_{2t},\tilde{\boldsymbol{\theta}}_n\right)^{'} \\ \boldsymbol{g}_2\left(\tilde{\boldsymbol{w}}_n;y_{t+1},\boldsymbol{z}_{2t},\tilde{\boldsymbol{\theta}}_n\right)\boldsymbol{g}_1\left(\tilde{\boldsymbol{\theta}}_n;y_{t+1},\boldsymbol{z}_{1t}\right)^{'} & \boldsymbol{g}_2\left(\tilde{\boldsymbol{w}}_n;y_{t+1},\boldsymbol{z}_{2t},\tilde{\boldsymbol{\theta}}_n\right)\boldsymbol{g}_2\left(\tilde{\boldsymbol{w}}_n;y_{t+1},\boldsymbol{z}_{2t},\tilde{\boldsymbol{\theta}}_n\right)^{'} \end{pmatrix}\end{aligned} \tag{4.8}$$

with

$$\boldsymbol{g}_1\left(\tilde{\boldsymbol{\theta}}_n;y_{t+1},\boldsymbol{z}_{1t}\right)\boldsymbol{g}_1\left(\tilde{\boldsymbol{\theta}}_n;y_{t+1},\boldsymbol{z}_{1t}\right)^{'} = \left[\tau-\mathbb{I}\left(y_{t+1}<\tilde{\boldsymbol{\theta}}_n^{'}\widehat{\boldsymbol{VaR}}_{t+1}\right)\right]^2\boldsymbol{z}_{1t}\boldsymbol{z}_{1t}^{'}$$

$$\begin{aligned}\boldsymbol{g}_1\left(\tilde{\boldsymbol{\theta}}_n;y_{t+1},\boldsymbol{z}_{1t}\right)\boldsymbol{g}_2\left(\tilde{\boldsymbol{w}}_n;y_{t+1},\boldsymbol{z}_{2t},\tilde{\boldsymbol{\theta}}_n\right)^{'} &= \left[\tau-\mathbb{I}\left(y_{t+1}<\tilde{\boldsymbol{\theta}}_n^{'}\widehat{\boldsymbol{VaR}}_{t+1}\right)\right][(y_{t+1}- \\ &\quad \tilde{\boldsymbol{w}}_n^{'}\widehat{\boldsymbol{ES}}_{t+1})\,\mathbb{I}\left(y_{t+1}<\tilde{\boldsymbol{\theta}}_n^{'}\widehat{\boldsymbol{VaR}}_{t+1}\right)\Big]\boldsymbol{z}_{1t}\boldsymbol{z}_{2t}^{'}\end{aligned}$$

$$\begin{aligned}\boldsymbol{g}_2\left(\tilde{\boldsymbol{w}}_n;y_{t+1},\boldsymbol{z}_{2t},\tilde{\boldsymbol{\theta}}_n\right)\boldsymbol{g}_1\left(\tilde{\boldsymbol{\theta}}_n;y_{t+1},\boldsymbol{z}_{1t}\right)^{'} &= \left[\tau-\mathbb{I}\left(y_{t+1}<\tilde{\boldsymbol{\theta}}_n^{'}\widehat{\boldsymbol{VaR}}_{t+1}\right)\right][(y_{t+1}- \\ &\quad \tilde{\boldsymbol{w}}_n^{'}\widehat{\boldsymbol{ES}}_{t+1})\,\mathbb{I}\left(y_{t+1}<\tilde{\boldsymbol{\theta}}_n^{'}\widehat{\boldsymbol{VaR}}_{t+1}\right)\Big]\boldsymbol{z}_{2t}\boldsymbol{z}_{1t}^{'}\end{aligned}$$

$$\boldsymbol{g}_2\left(\tilde{\boldsymbol{w}}_n;y_{t+1},\boldsymbol{z}_{2t},\tilde{\boldsymbol{\theta}}_n\right)\boldsymbol{g}_2\left(\tilde{\boldsymbol{w}}_n;y_{t+1},\boldsymbol{z}_{2t},\tilde{\boldsymbol{\theta}}_n\right)^{'} = \left[\left(y_{t+1}-\tilde{\boldsymbol{w}}_n^{'}\widehat{\boldsymbol{ES}}_{t+1}\right)^2\mathbb{I}\left(y_{t+1}<\tilde{\boldsymbol{\theta}}_n^{'}\widehat{\boldsymbol{VaR}}_{t+1}\right)\right]\boldsymbol{z}_{2t}\boldsymbol{z}_{2t}^{'}$$

where $\tilde{\boldsymbol{\theta}}_n$ and $\tilde{\boldsymbol{w}}_n$ are some initial consistent estimates of $\boldsymbol{\theta}^*$ and $\boldsymbol{w}^*$, respectively.

Let $\boldsymbol{\beta}=\left(\boldsymbol{\theta}^{'},\boldsymbol{w}^{'}\right)^{'}$ denote the vector of weighting parameters to be estimated by GMM. The computation of $\hat{\boldsymbol{\beta}}_n$ and $\hat{\boldsymbol{S}}_n$ is typically done recursively. I first choose a conformable identity-weighting matrix in (4.6), and then estimate the corresponding $\hat{\boldsymbol{\beta}}_n^{(1)}$. The resulting new weighting matrix, $\hat{\boldsymbol{S}}_n\left(\hat{\boldsymbol{\beta}}_n^{(1)}\right)$, is more efficient than the previous one, and solving (4.6) leads to a new estimator $\hat{\boldsymbol{\beta}}_n^{(2)}$. These steps are repeated until the sequence of $\hat{\boldsymbol{\beta}}_n$ converges.

In practice, the choice of $\boldsymbol{Z}_{1t}$ and $\boldsymbol{Z}_{2t}$ depends on the nature of the application considered, which is discussed in more details in Section 6. $\boldsymbol{Z}_{1t}$ and $\boldsymbol{Z}_{2t}$ may or may not be the same in the identification of the weighting parameters. In cases where the information vectors fail to incorporate all of the relevant information, condition $\boldsymbol{g}^o\left(\boldsymbol{\theta}^*,\boldsymbol{w}^*\right)=0$ is no longer equivalent to the first-order condition (4.5), and $\left\{\boldsymbol{g}\left(\boldsymbol{\theta}^*,\boldsymbol{w}^*;Y_{t+1},\boldsymbol{Z}_{1t},\boldsymbol{Z}_{2t}\right),\mathcal{F}_t\right\}$ is no longer a martingale difference sequence. However, $\boldsymbol{S}$ can still be consistently estimated using some heteroscedasticity- and autocorrelation-robust estimator, like Newey and West's (1987) estimator. I next focus on the asymptotic properties of the GMM estimator, $\hat{\boldsymbol{\beta}}_n=\left(\hat{\boldsymbol{\theta}}_n^{'},\hat{\boldsymbol{w}}_n^{'}\right)^{'}$.

4.2 Asymptotic Properties of the GMM Estimator

The following assumptions are considered for the asymptotic properties of the GMM estimator, $\hat{\boldsymbol{\beta}}_n = \left(\hat{\boldsymbol{\theta}}_n^{'}, \hat{\boldsymbol{w}}_n^{'}\right)^{'}$.

Assumption 1. *(Consistency). Assume that Proposition 1 of Giacomini and Komunjer (2005) holds with the extension to conditional expected shortfall.*[10] *That is, for every* t, $m \leq t \leq T-1$, *(a) the conditional density of* Y_{t+1}, $f_t(\cdot)$ *is continuous, strictly positive and bounded, and the conditional cumulative distribution function of* Y_{t+1}, $F_t(\cdot)$ *is continuous and lies in* $[0,1]$*; (b) for* $i = 1,2$, $\widehat{VaR}_{i,t+1} \neq 0$, *a.s.-P, and corr* $\left(\widehat{VaR}_{1,t+1}, \widehat{VaR}_{2,t+1}\right) \neq \pm 1$. *Similarly, given* $\widehat{\boldsymbol{VaR}}_{t+1}$, $\widehat{ES}_{i,t+1} \neq 0$, *a.s.-P, and corr* $\left(\widehat{ES}_{1,t+1}, \widehat{ES}_{2,t+1}\right) \neq \pm 1$*; (c)* $\left\{\left(\mathbf{Z}_{1t}^{'}, \mathbf{Z}_{2t}^{'}, \boldsymbol{\mathcal{V}}_t^{'}\right)^{'}\right\}$ *is strictly stationary and* α*-mixing with* α *of size* $-r/(r-2)$ *and* $r > 2$*; (d)* $E\left[\mathbf{Z}_{1t}\mathbf{Z}_{1t}^{'}\right]$ *and* $E\left[\mathbf{Z}_{2t}\mathbf{Z}_{2t}^{'}\right]$ *are nonsingular; and (e) there exist some* $\delta > 0$ *such that* $E\left\|\mathbf{Z}_{it}\right\|^{2r+\delta} < \infty$ *for* $i = 1,2$. *Then,* $\hat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}^*$ *and* $\hat{\boldsymbol{w}}_n \xrightarrow{p} \boldsymbol{w}^*$ *or equivalently* $\hat{\boldsymbol{\beta}}_n \xrightarrow{p} \boldsymbol{\beta}^*$, *as* $n \to \infty$. *(f)* $E\left\|\widehat{\boldsymbol{VaR}}_{t+1}\right\|^4 < \infty$ *and* $E\left\|\widehat{\boldsymbol{ES}}_{t+1}\right\|^4 < \infty$*; (g)* $\boldsymbol{\beta}^*$ *is an interior point of* $\Xi \equiv (\Theta, \Theta)$.

Assumption 2. *(Convergence). The data generating process is assumed to meet the conditions for a law of large numbers to apply, so that one may assume that the empirical moments converge in probability to their expectation.*

$$\boldsymbol{g}_n(\boldsymbol{\theta}, \boldsymbol{w}) \equiv \frac{1}{n}\sum_{t=m}^{T-1} \boldsymbol{g}(\boldsymbol{\theta}_n, \boldsymbol{w}_n;\, y_{t+1}, \boldsymbol{z}_{1t}, \boldsymbol{z}_{2t}) \xrightarrow{p} \boldsymbol{g}^o(\boldsymbol{\theta}, \boldsymbol{w};\, Y_{t+1}, \boldsymbol{Z}_{1t}, \boldsymbol{Z}_{2t}) = 0$$

Assumption 3. *(Asymptotic distribution of empirical moments). Assume that the empirical moments obey a central limit theorem. This assumes that the moments have a finite asymptotic covariance matrix,* $\boldsymbol{S}$, *in (4.7), so that*

$$\sqrt{n}\boldsymbol{g}_n(\boldsymbol{\theta}, \boldsymbol{w}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{S})$$

[10] The extension to expected shortfall is followed straightforward to the proof of *Proposition 1 in* Giacomini and Komunjer (2005).

Assumption 1(b) is a mild condition ruling out the possibility that the sequences of forecasts are perfectly correlated, which would happen if, for example, the models were proportional or differed only by a constant. One could in principle relax the assumption of strict stationarity in *Assumption 1(c)* and rely on existing results on the consistency and asymptotic normality of GMM estimators for mixing sequences. However, as discussed in Giacomini and Komunjer (2005), relaxing this strict assumption would cause the optimal weights to depend on the sample size, and thus result in a less intuitive formulation of the null hypothesis of encompassing. *Assumption 1(d)&(e)* are fairly standard and imply in particular that all of the components of the information vector are not linearly dependent. *Assumption 1(f)* implicitly places conditions on the existence of the finite-sample moments of the estimators on which $\widehat{\boldsymbol{VaR}}_{t+1}$ and $\widehat{\boldsymbol{ES}}_{t+1}$ are based.

The underlying requirements on the data for *Assumption 3* to hold will vary and will be complicated if the observations comprising the empirical moments are not independent. For samples of independent observations, assuming the conditions underlying the Lindeberg-Feller or Liapounov central limit theorem will suffice (Greene, 2012). For the more general case, it is necessary to make some assumptions about the data, including *Assumption 2*. If one can go a step further and assume that the function $\boldsymbol{g}(\boldsymbol{\theta}, \boldsymbol{w})$ is an ergodic, stationary martingale difference series, then it can invoke the central limit theorem for the martingale difference series (Greene (2012), Theorem 20.3, p.916). It is generally fairly complicated to verify this martingale assumption for nonlinear models, so it is usually assumed outright.

With the assumptions in place, I have the asymptotic distribution of $\hat{\boldsymbol{\beta}}_n$ in the following theorem.

Theorem 1. *(Asymptotic distribution of the GMM estimator). Let Assumptions 1-3 hold. Then,* $\hat{\boldsymbol{\beta}}_n$ *is asymptotically normal,*

$$\left(\boldsymbol{\gamma}'\boldsymbol{S}^{-1}\boldsymbol{\gamma}\right)^{-1/2}\sqrt{n}\left(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*\right) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{I}) \tag{4.9}$$

with

$$\begin{aligned}\boldsymbol{\gamma} &= E\left[\nabla_{\boldsymbol{\beta}}\boldsymbol{g}(\boldsymbol{\beta})\right] \\ &= -\begin{pmatrix} E\left[f_t\left(\boldsymbol{\theta}^{*'}\widehat{\boldsymbol{VaR}}_{t+1}\right)\boldsymbol{Z}_{1t}\widehat{\boldsymbol{VaR}}_{t+1}^{'}\right] & \boldsymbol{0} \\ E\left[f_t\left(\boldsymbol{\theta}^{*'}\widehat{\boldsymbol{VaR}}_{t+1}\right)\left(y_{t+1}-\boldsymbol{w}^{*'}\widehat{\boldsymbol{ES}}_{t+1}\right)\boldsymbol{Z}_{2t}\widehat{\boldsymbol{VaR}}_{t+1}^{'}\right] & \tau E\left[\boldsymbol{Z}_{2t}\widehat{\boldsymbol{ES}}_{t+1}^{'}\right] \end{pmatrix}\end{aligned} \tag{4.10}$$

where $\nabla_{\boldsymbol{\beta}}\boldsymbol{g}(\boldsymbol{\beta})$ *is a Jacobian derivative matrix of* $\boldsymbol{g}(\boldsymbol{\beta})$ *with respect to* $\boldsymbol{\beta}$*, and* $\boldsymbol{S}$ *is defined in (4.7).*

Proof. See Appendix C. □

It should be noticed that the off-diagonal element of γ in (4.10) is non-zero, different from the (approximately) zero off-diagonal element of Λ in Dimitriadis and Schnaitmann (2020). This non-zero off-diagonal element is mainly due to the quadratic loss function chosen for ES forecasts conditional on VaR forecasts, which captures the interdependence between ES and VaR.

Theorem 1 requires that $\boldsymbol{g}_n(\boldsymbol{\beta})$ be once differentiable, which is not the case here due to the indicator function. Nonetheless, Newey and McFadden (1994) can be used to obtain the asymptotic normality for nonsmooth moment functions, which is also applied to *Theorem 1*. The basic insight of their approach is that a smoothness condition on $\boldsymbol{g}_n(\boldsymbol{\beta})$ can be replaced by the smoothness of its limit $\boldsymbol{g}^o(\boldsymbol{\beta})$, with the requirement that certain remainder terms are small.

Specifically in (4.10), the expression for $\boldsymbol{\gamma}$ depends on the value of the conditional density function f_t evaluated at the optimal combination of VaRs. If data distribution is assumed, the density value can easily be evaluated. Otherwise, I adopt the idea in Giacomini and Komunjer (2005) to use a smooth approximation to the indicator function (see e.g., Bracewell (2000), p. 63-65) to estimate the conditional density f_t for $\boldsymbol{\gamma}$ in (4.10) as

$$f_t\left(\hat{\boldsymbol{\theta}}_n^{'}\widehat{\boldsymbol{VaR}}_{t+1}\right) = \frac{1}{\varsigma}exp\left(\frac{y_{t+1}-\hat{\boldsymbol{\theta}}_n^{'}\widehat{\boldsymbol{VaR}}_{t+1}}{\varsigma}\right)\mathbb{I}\left(y_{t+1}<\hat{\boldsymbol{\theta}}_n^{'}\widehat{\boldsymbol{VaR}}_{t+1}\right) \tag{4.11}$$

where $\varsigma>0$. Convergences of $\hat{\boldsymbol{\gamma}}_{11,n}$ and $\hat{\boldsymbol{\gamma}}_{21,n}$ in $\boldsymbol{\gamma}$ to their expected values are uniform in ς in a neighborhood of 0, which ensures that $lim_{\varsigma\to 0}\hat{\boldsymbol{\gamma}}_{11,n}\xrightarrow{p}\boldsymbol{\gamma}_{11}$ and $lim_{\varsigma\to 0}\hat{\boldsymbol{\gamma}}_{21,n}\xrightarrow{p}\boldsymbol{\gamma}_{21}$.

In principle, ς is the choice of a researcher over an arbitrage range of small values, such as a range from 0.2×10^{-2} to 10^{-2} considered in Giacomini and Komunjer (2005). Alternatively, I propose the following result to determine the value of ς,

$$\hat{\varsigma} = \frac{1}{n} \sum_{t=m}^{T-1} \left(y_{t+1} - \hat{\boldsymbol{\theta}}_n^{'} \widehat{\boldsymbol{VaR}}_{t+1} \right) \left[\tau - \mathbb{I} \left(y_{t+1} \leq \hat{\boldsymbol{\theta}}_n^{'} \widehat{\boldsymbol{VaR}}_{t+1} \right) \right] \tag{4.12}$$

Note that (4.12) is the minimized "tick" loss function of (3.1) from (4.6) evaluated at the GMM estimates of optimal weights, so that under *Assumption 2* $\hat{\varsigma} \to 0$ ensures convergences of $\hat{\boldsymbol{\gamma}}_{11,n}$ and $\hat{\boldsymbol{\gamma}}_{21,n}$ in $\boldsymbol{\gamma}$. Taylor (2019) shows that the recent literature has used an asymmetric Laplace distribution as a quasi-maximum likelihood function to regression quantiles, where the maximum likelihood estimator for its scale parameter is obtained the same as $\hat{\varsigma}$ in (4.12).[11] An advantage to use (4.12) is to avoid the grid search for ς, where no criterion is available for determining the best value of ς among the search. The performance of (4.12) in evaluating (4.11) will be examined in Section 5 by the simulation study and then applied to the empirical illustration in Section 6.

4.3 CESFE Test Statistics

This subsection considers the tests for two null hypotheses: $H_{10} : (\theta_1^*, \theta_2^*, w_1^*, w_2^*) = (1, 0, 1, 0)$ against $H_{1a} : (\theta_1^*, \theta_2^*, w_1^*, w_2^*) \neq (1, 0, 1, 0)$, and $H_{20} : (\theta_1^*, \theta_2^*, w_1^*, w_2^*) = (0, 1, 0, 1)$ against $H_{2a} : (\theta_1^*, \theta_2^*, w_1^*, w_2^*) \neq (0, 1, 0, 1)$, which correspond to testing whether forecasts $\widehat{ES}_{1t+1}$and $\widehat{VaR}_{1t+1}$ from $\mathcal{M}_1$ encompass $\widehat{ES}_{2t+1}$ and $\widehat{VaR}_{2t+1}$ from $\mathcal{M}_2$ or whether $\widehat{ES}_{2t+1}$ and $\widehat{VaR}_{2t+1}$ from $\mathcal{M}_2$ encompass $\widehat{ES}_{1t+1}$and $\widehat{VaR}_{1t+1}$ from $\mathcal{M}_1$. This section provides the test statistics and the limiting distributions.

Let $\boldsymbol{r}_1 = (1, 0, 1, 0)^{'}$ and $\boldsymbol{r}_2 = (0, 1, 0, 1)^{'}$. Using the GMM estimators, I then propose a

[11] For an asymmetric Laplace distribution (ALD), its density function takes the form,

$$f(y; \mu, \tau, \varsigma) = \frac{\tau(1-\tau)}{\varsigma} exp\left\{ -\frac{(y-\mu)}{\varsigma} \left[\tau - \mathbb{I}(y < \mu) \right] \right\}$$

where $\mu \in \mathbb{R}$, $\tau \in (0, 1)$ and $\varsigma > 0$ are the location, asymmetric and scale parameters, respectively. Having $\mu = \hat{\boldsymbol{\theta}}_n^{'} \widehat{\boldsymbol{VaR}}_{t+1}$ and the chosen τ, *ALD* implies $Pr\left(Y_{t+1} < \hat{\boldsymbol{\theta}}_n^{'} \widehat{\boldsymbol{VaR}}_{t+1} \right) = \tau$. Taylor (2019) shows that (4.12) is the quasi-maximum likelihood estimate of ς from the ALD distribution, which eventually is the average of the tick loss function and can be interpreted as an estimator of the expectation of the tick loss function.

Wald test of the hypotheses H_{10} and H_{20} in the following theorem.

Theorem 2. *(CESFE test). Apply Theorem 1 to construct the test statistics*

$$CESFE1_n = n\left(\hat{\boldsymbol{\beta}}_n - \boldsymbol{r}_1\right)' \hat{\boldsymbol{\Omega}}_n^{-1} \left(\hat{\boldsymbol{\beta}}_n - \boldsymbol{r}_1\right) \tag{4.13}$$

and

$$CESFE2_n = n\left(\hat{\boldsymbol{\beta}}_n - \boldsymbol{r}_2\right)' \hat{\boldsymbol{\Omega}}_n^{-1} \left(\hat{\boldsymbol{\beta}}_n - \boldsymbol{r}_2\right) \tag{4.14}$$

where $\hat{\boldsymbol{\beta}}_n = \left(\hat{\boldsymbol{\theta}}_n', \hat{\boldsymbol{w}}_n'\right)'$ *solves (4.6) and* $\hat{\boldsymbol{\Omega}}_n$ *is some consistent estimate of* $\boldsymbol{\Omega} \equiv \left(\boldsymbol{\gamma}' \boldsymbol{S}^{-1} \boldsymbol{\gamma}\right)^{-1}$. *Then, for* $i = 1, 2$, *(a) under* H_{i0} : $CESFEi_n \xrightarrow{d} \chi_4^2$ *as* $n \to \infty$, *and (b) under* H_{ia} : $CESFEi_n \to +\infty$, *as* $n \to \infty$.

Proof. See Appendix D. □

The $CESFE$ test can then be implemented as follows. For a desired level of confidence, one first chooses the corresponding critical value c from the χ_4^2 distribution. Then, H_{10} is rejected if $CESFE1_n > c$ and H_{20} is rejected if $CESFE2_n > c$. In the context of real-time forecast selection, that is, for selecting at time T a best forecast method for time $T+1$, I propose the following decision rule. Perform the two tests of H_{10} and H_{20} on data up to time T. Hence, there are four possible scenarios. (1) If neither H_{10} nor H_{20} are rejected, then the test is not helpful for forecast selection (one could decide either to use the more parsimonious model or to conservatively set equal weights to the forecasts, i.e., $\hat{w}_{1n} = \hat{w}_{2n} = 0.5$). (2) If H_{10} is rejected while H_{20} is not rejected, then one would choose $\widehat{ES}_{2,T+1}$ as the best forecast so that $\hat{w}_{1n} = 0$ and $\hat{w}_{2n} = 1$. (3) If H_{20} is rejected while H_{10} is not rejected, then one would choose $\widehat{ES}_{1,T+1}$ as the best forecast so that $\hat{w}_{1n} = 1$ and $\hat{w}_{2n} = 0$. (4) If both H_{10} and H_{20} are rejected, then one would choose the combination of $\widehat{ES}^*_{T+1} = \hat{w}_{1n}\widehat{ES}_{2,T+1} + \hat{w}_{2n}\widehat{ES}_{2,T+1}$ as the best forecast, where $\hat{w}_{1n}$ and $\hat{w}_{2n}$ are out-of-sample estimates of the combination weights from (4.6). In this paper I illustrate *Theorem 2* for encompassing of conditional expected shortfall forecasts, which can be easily generalized to compare more than two alternative forecasts.

5 Simulation Study

I evaluate the performance of the proposed CESFE test in finite samples along three dimensions: the size of the test, its power, and the choice of ς for evaluating f_t. The simulation experiment is designed to match the problem of ES evaluation and combination in the empirical application.

Specifically, I consider the following data generating process (DGP)

$$y_{t+1} = \sigma_{t+1}\varepsilon_{t+1} \tag{5.1}$$

where $\varepsilon_{t+1} \sim \mathcal{D}(0,1)$. σ_{t+1} follows a standard deviation version of either the GARCH(1,1) model (Zakoian, 1994)

$$\sigma_{t+1} = \beta_0 + \beta_1\sigma_t + \beta_2 |y_t| \tag{5.2}$$

or the GJR-GARCH(1,1) model (Glosten et al., 1993)

$$\sigma_{t+1} = \beta_0 + \beta_1\sigma_t + \beta_2^+ |y_t| \mathbb{I}(y_t > 0) + \beta_2^- |y_t| \mathbb{I}(y_t < 0) \tag{5.3}$$

(5.3) allows conditional variance to respond differently to past positive and negative innovations. This asymmetry is sometimes referred to in the literature as a "leverage effect."

Particularly, Xiao and Koenker (2009) and Gerlach et al. (2011) show that quantile dynamics implied by (5.2) and (5.3) are the special cases of symmetric absolute value (SAV) and asymmetric slope (AS) CAViaR models, respectively, proposed by Engle and Manganelli (2004), as

$$VaR_{t+1} = \beta_0(\tau) + \beta_1(\tau)VaR_t + \beta_2(\tau)|y_t| \tag{5.4}$$

$$VaR_{t+1} = \beta_0(\tau) + \beta_1(\tau)VaR_t + \beta_2^+(\tau)|y_t|\mathbb{I}(y_t > 0) + \beta_2^-(\tau)|y_t|\mathbb{I}(y_t < 0) \tag{5.5}$$

where $\beta_0(\tau) = \beta_0\mathcal{D}_\varepsilon^{-1}(\tau)$, $\beta_1(\tau) = \beta_1$, $\beta_2(\tau) = \beta_2\mathcal{D}_\varepsilon^{-1}(\tau)$, $\beta_2^+(\tau) = \beta_2^+\mathcal{D}_\varepsilon^{-1}(\tau)$, and $\beta_2^-(\tau) = \beta_2^-\mathcal{D}_\varepsilon^{-1}(\tau)$. $\mathcal{D}_\varepsilon^{-1}(\tau)$ is the $\tau \times 100\%$th theoretical quantile of ε_{t+1} under a distribution assumption, $\mathcal{D}$. Note that (5.4) and (5.5) implied respectively by (5.2) and (5.3) are restricted in that

$\beta_1(\tau) = \beta_1$ is independent of τ, while Engle and Manganelli (2004) allow the quantile persistence coefficient $\beta_1(\tau)$ in the SAV- and AS-CAViaR models dependent of τ so as to capture asymmetric quantile persistence across τ.

However, for the purpose of data simulation, I consider the DGP by (5.1)-(5.3), as theoretical values of VaR and ES at time $t+1$ can explicitly be obtained as follows[12]

$$VaR_{t+1} = \sigma_{t+1}\mathcal{D}_{\varepsilon}^{-1}(\tau) \tag{5.6}$$

$$ES_{t+1} = \sigma_{t+1}E\left[\varepsilon_{t+1}|\varepsilon_{t+1} < \mathcal{D}_{\varepsilon}^{-1}(\tau)\right] \tag{5.7}$$

For example, in the case where $\varepsilon_{t+1} \sim \mathcal{N}(0,1)$ is assumed, then the theoretical VaR and ES values can be computed as

$$VaR_{t+1}^{N} = \sigma_{t+1}\Phi^{-1}(\tau) \tag{5.8}$$

$$ES_{t+1}^{N} = -\sigma_{t+1}\frac{\phi\left(\Phi^{-1}(\tau)\right)}{\tau} \tag{5.9}$$

where $\Phi(\cdot)$ and $\phi(\cdot)$ are the CDF and PDF of the standard normal distribution, respectively. See also, e.g., Bertsimas et al. (2004), Broda and Paolella (2011), Nadarajah et al. (2014), among others.

In addition, if a Student-t distribution, $\varepsilon_{t+1} \sim t_v$, is assumed with v degrees of freedom, then the theoretical VaR values can be obtained as

$$VaR_{t+1}^{T_v} = \sigma_{t+1}\frac{T_v^{-1}(\tau)}{\sqrt{v/(v-2)}} \tag{5.10}$$

where $T_v(\cdot)$ is the CDF of a Student-t distribution with v degrees of freedom. And, the theoretical ES values are given by

$$ES_{t+1}^{T_v} = -\sigma_{t+1}\frac{f_T(q_\tau, v)(v+q_\tau^2)}{\tau(v-1)} \tag{5.11}$$

where $q_\tau = T_v^{-1}(\tau)/\sqrt{v/(v-2)}$, and $f_T(x,v) = \left[v^{-1/2}/B\left(\frac{v}{2},\frac{1}{2}\right)\right](1+x^2/v)^{-(v+1)/2}$ is the

[12] See e.g., McNeil and Frey (2000), Righi and Ceretta (2015), Martins-Filho et al. (2018), among others.

standard (location-zero, scale-one) Student-t PDF with v degrees of freedom and the beta function, $B(\cdot,\cdot)$. See Broda and Paolella (2011), §2.2.2 and Nadarajah et al. (2014) for more details.

I consider the following parameter values: $(\beta_0, \beta_1, \beta_2) = (0.005, 0.85, 0.1)$ for (5.2) and $\left(\beta_0, \beta_1, \beta_2^+, \beta_2^-\right) = (0.005, 0.85, -0.02, 0.1)$ for (5.3), $v = 4$ for the Student-t distribution, and a range of values for the out-of-sample size, $n = (1000, 2500, 5000)$. In these particular cases, the in-sample size m is 0 and $T = n$. A range of values for the parameter ς in (4.11) are considered from 0.002 to 0.02 in increments of 0.002. The proposed approach to approximate ς by (4.12) is also experimented. For each sample size, I generate 5,000 Monte Carlo replications for each of the time series $\mathcal{M}_1 = \{y_{t+1}, VaR_{t+1}, ES_{t+1}; GJR, \mathcal{N}(0,1)\}_{t=0}^{n-1}$, $\mathcal{M}_2 = \{y_{t+1}, VaR_{t+1}, ES_{t+1}; GARCH, \mathcal{N}(0,1)\}_{t=0}^{n-1}$, $\mathcal{M}_3 = \{y_{t+1}, VaR_{t+1}, ES_{t+1}; GJR, t_{v=4}\}_{t=0}^{n-1}$, and $\mathcal{M}_4 = \{y_{t+1}, VaR_{t+1}, ES_{t+1}; GARCH, t_{v=4}\}_{t=0}^{n-1}$. The tail risk level, $\tau = 0.025$, is used for VaR and ES forecasts.[13]

5.1 Size of the Test

The combinations for an encompassing test between $\mathcal{M}_i$ and $\mathcal{M}_j$ are thus given by

$$
\begin{aligned}
VaR_{t+1}^{(i,j)} &= \theta_0 + \theta_{\mathcal{M}_i} VaR_{\mathcal{M}_i,t+1} + \theta_{\mathcal{M}_j} VaR_{\mathcal{M}_j,t+1} \\
ES_{t+1}^{(i,j)} &= \left(w_0 + w_{\mathcal{M}_i} ES_{\mathcal{M}_i,t+1} + w_{\mathcal{M}_j} ES_{\mathcal{M}_j,t+1}\right) \mathbb{I}\left(y_{\mathcal{M}_i,t+1} < VaR_{t+1}^{(i,j)}\right)
\end{aligned}
$$

for the null hypothesis, $H_{10}^{(i,j)}$: $\mathcal{M}_i$ encompasses $\mathcal{M}_j$ with $i, j = 1, 2, 3, 4$ and $i \neq j$. According to the procedure described in Section 4, the GMM estimators are constructed as $\hat{\boldsymbol{\beta}}_n^{(i,j)} = \left(\hat{\theta}_{0,ij}, \hat{\theta}_{\mathcal{M}_i}, \hat{\theta}_{\mathcal{M}_j}, \hat{w}_{0,ij}, \hat{w}_{\mathcal{M}_i}, \hat{w}_{\mathcal{M}_j}\right)$. Granger and Ramanathan (1984) find that the best method of combinations is to add a constant term and not to constrain the weights to add up to unity. See also Giacomini and Komunjer (2005). Therefore, I include constant terms in the forecast combinations and do not restrict combination weights to sum up to one.

In particular, this simulation study considers the null hypothesis that forecasts from the

[13] Basel Committee on Banking Supervision (BCBS) of Bank for International Settlements (BIS) published in January 2016 the regulation document, "Standards: Minimum capital requirements for market risk." Section C.3 on pp. 52 of the document requires that in calculating the expected shortfall, individual banks should use a one-tailed 2.5th percentile.

GJR model which implies AS-CAViaR quantile dynamics encompass forecasts from the GARCH model which implies SAV-CAViaR quantile dynamics. Therefore, the hypotheses are tested for $H_{10}^{(1,2)}$ and $H_{10}^{(3,4)}$ with normal and Student-t distributions, respectively. In these particular cases, forecasts from the GJR model will display correct empirical coverage by construction, whereas forecasts from the misspecified GARCH model will in general be biased. The information vectors are $\boldsymbol{Z}_{1t}^{(i,j)} = \left(1, y_{\mathcal{M}_i,t}, VaR_{\mathcal{M}_i,t}, VaR_{\mathcal{M}_j,t}\right)$ and $\boldsymbol{Z}_{2t}^{(i,j)} = \left(1, y_{\mathcal{M}_i,t}, ES_{\mathcal{M}_i,t}, ES_{\mathcal{M}_j,t}\right)$ for $H_{10}^{(i,j)}$. The test statistics that are given in *Theorem 2* are used to compute the proportion of rejections at the 5% nominal level for the null hypotheses $H_{10}^{(1,2)}$ and $H_{10}^{(3,4)}$.

The simulation results show that the nominal 5% test appears to be well sized. For $n = (1000, 2500, 5000)$, the rejection probabilities are, respectively, $(7.1\%, 6.5\%, 5.7\%)$ for the null hypothesis $H_{10}^{(1,2)}$ (normal distribution) and $(8.4\%, 7.2\%, 5.7\%)$ for the null hypothesis $H_{10}^{(3,4)}$ (t-distribution), when the true density, f_t, of y_{t+1} in (5.1) is used to evaluate $\boldsymbol{\gamma}$ in (4.10).

In a more plausible setup in which the true density f_t is unknown and where (4.11) and (4.12) are used to estimate f_t for computing $\boldsymbol{\gamma}$ in (4.10), the empirical rejection probabilities vary with the sample size n and the smoothing parameter ς, as shown in Figure (1), which plots size curves for the rejection probabilities against the range of ς values. The cycle dots, $\hat{\varsigma}$, in Figure 1 are estimates of ς from (4.12) to evaluate f_t in (4.11) for computing γ in (4.10).

A general pattern that emerges from Figure 1 is that the test appears generally well sized in that rejection probabilities approach the 5% nominal level as the sample size increases. Nonetheless, the size curves are relatively flat, indicating that the marginal effects of varying the value of ς on rejection probabilities are small. For instance, the ranges of rejection probabilities for $n = 2500$ are from 7.2% to 7.8% for the normal distribution and from 8.2% to 8.7% for the Student-t distribution across different values of ς. This result is mainly due to the component $\boldsymbol{\gamma}_{22}$ in (4.10) that is independent of the choice of ς. The simulation results also show that the test is well-sized for $\hat{\varsigma}$ estimated from (4.12).

5.2 Power of the Test

To generate data under the alternative hypothesis of no encompassing of GJR forecasts with respect to GARCH forecasts, I first replicate data simulations following the procedure described

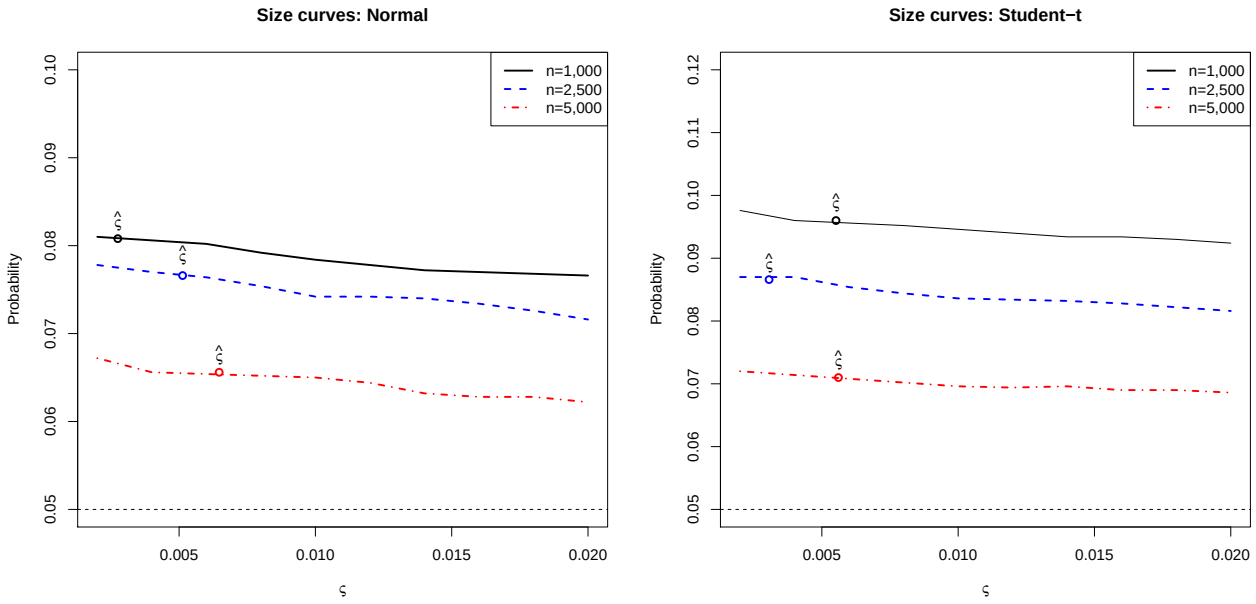

Figure 1: Size Curves of the CESFE test from the simulation experiment for the 5% nominal level. Rejection frequencies are computed over 5,000 Monte Carlo replications of the null hypothesis that forecasts from the GJR model encompass forecasts from the the GARCH model when the DGP is the GJR model. n is the sample size. ς is a user-defined constant required in evaluating f_t in (4.11) for computing γ in (4.10). A range of values for ς are evaluated for (4.11) from 0.002 to 0.02 in increments of 0.002. (4.11) is also evaluated by $\hat{\varsigma}$ estimated from (4.12).

in the previous section, and then let the DGP be

$$y_{t+1} = \delta y_{t+1}^{GARCH} + (1-\delta)\, y_{t+1}^{GJR} \tag{5.12}$$

where y_{t+1}^{GARCH} and y_{t+1}^{GJR} are simulated from (5.1) with (5.2) and (5.3), respectively, and $0 < \delta \leq 1$. (5.12) implies that

$$VaR_{t+1} = \delta VaR_{t+1}^{GARCH} + (1-\delta)\, VaR_{t+1}^{GJR}$$
$$ES_{t+1} = \left[\delta ES_{t+1}^{GARCH} + (1-\delta)\, ES_{t+1}^{GJR}\right] \mathbb{I}\left(y_{t+1} < VaR_{t+1}\right)$$

Note that the size study is obtained when the data are generated according to (5.12) with $\delta = 0$. Accordingly, increasing δ toward 1 allows to obtain the power curve for the CESFE test. I consider a number of different values for δ, ranging from 0.05 to 1 in increments of 0.05. For each parameterization, I generate 5,000 Monte Carlo replications of the time series from (5.12)

for normal and Student-t distributions, and proceed as previously by computing the proportion of rejections of the null hypothesis that forecasts from the GJR model encompass forecasts from the GARCH model at the 5% nominal level.

Figure 2 plots the power curves for $n = (1000, 2500, 5000)$ when using the true conditional density f_t in the expression (4.10) for $\boldsymbol{\gamma}$. As expected, the power increases with n. The loss of power induced by estimating $\boldsymbol{\gamma}$ with the estimator $\hat{\boldsymbol{\gamma}}_{n,\varsigma}$ in (4.11) is shown in Figure 3 for the case where $n = 2,500$ and for different values of the smoothing parameter ς and the estimated smoothing parameter, $\hat{\varsigma}$, from (4.12). This figure shows that the powers are very similar based on different values of ς and $\hat{\varsigma}$, while smaller values of ς have slightly higher powers. A possible explanation to the power insensitive to the values of ς is mainly due to the component $\boldsymbol{\gamma}_{22}$ in (4.10) that is independent of the choice of ς.[14]

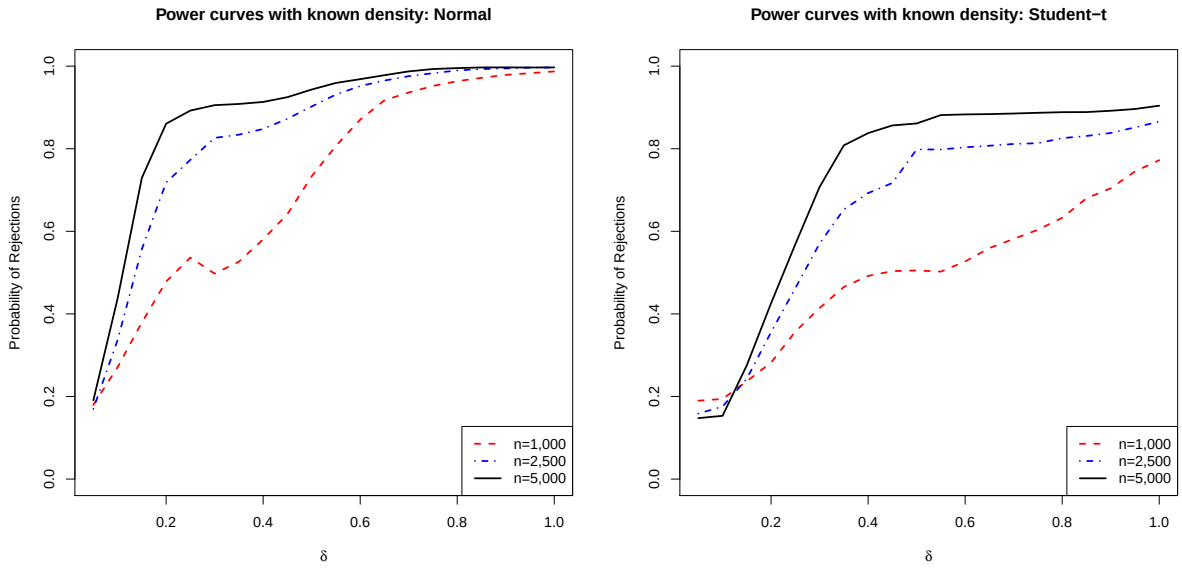

Figure 2: Power Curves of the CESFE test from the simulation experiment for known densities. Each curve represents the rejection frequency - computed by assuming f_t in (4.10) known - over 5,000 Monte Carlo replications. The null hypothesis being tested is that forecasts from the GJR model encompass forecasts from the GARCH model when the DGP is a convex combination of the two, with weights δ and $1 - \delta$.

[14] I verify this conjecture by testing VaR_{t+1} only for the null hypothesis that forecasts of VaR_{t+1} from the GJR model encompass those from the GARCH model, and obtain the conclusions for the test power similar to those drawn by Giacomini and Komunjer (2005). The test results are available upon request.

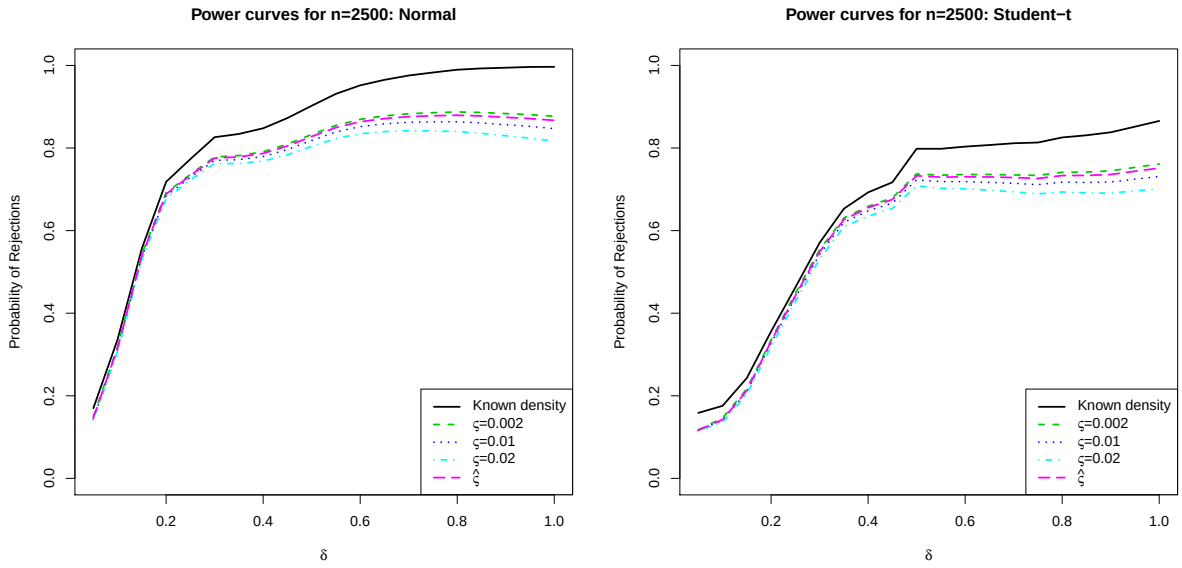

Figure 3: Power Curves of the CESFE test from the simulation experiment for $n = 2,500$. Each curve represents the rejection frequency over 5,000 Monte Carlo replications. The unknown f_t in (4.10) is evaluated by chosen ς and estimated $\hat{\varsigma}$ values. The null hypothesis being tested is that forecasts from the GJR model encompass forecasts from the GARCH model when the DGP is a convex combination of the two, with weights δ and $1 - \delta$.

5.3 Test Comparisons

This subsection conducts simulation to compare the conditional encompassing test proposed in this paper to the unconditional encompassing test of Dimitriadis and Schnaitmann (2020) in order to understand their relative strengths. Different from this study, Dimitriadis and Schnaitmann (2020) use the 2-elicitable loss functions, developed in Fissler and Ziegel (2016), for jointly evaluating the VaR and ES through an M-estimation. The comparison between the two encompassing tests is based on the data simulated in Sections 5.1&5.2 for the same null hypotheses.

Figure 4 plots the size and power comparisons of the tests. The figure shows that: (i) the conditional encompassing test obtains a consistently better test size than the unconditional test for each n considered. On average, the conditional encompassing test has the test size 8.7% closer to the 5% nominal level, relative to 11.3% obtained from the unconditional encompassing test; and (ii) the conditional encompassing test presents an increasing test power with the increased degree of misspecification, especially for $\delta > 0.4$ beyond which the conditional

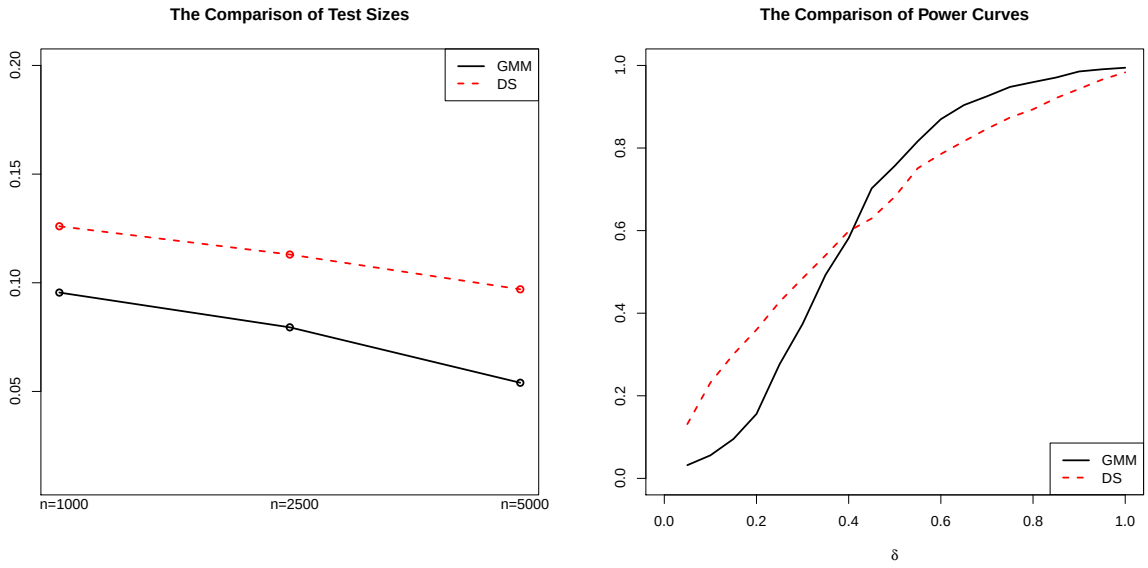

Figure 4: Plots of the size and power comparisons between the conditional encompassing test proposed in this paper based on a recursive GMM estimation and the unconditional test of Dimitriadis and Schnaitmann (2020) (DS) based on an M-estimation. The joint encompassing test of Dimitriadis and Schnaitmann (2020) is reported here. δ takes the values from 0.05 to 1 in increments of 0.05.

test power exceeds the unconditional one. Nonetheless, the unconditional test has stronger power when the degree of model misspecification is relatively low ($\delta < 0.4$). I claim that the performance difference is partly attributed to the non-zero off-diagonal element of γ in (4.10) of *Theorem 1*, which captures the interdependence between ES and VaR, different from the corresponding zero off-diagonal element of Λ in Dimitriadis and Schnaitmann (2020).

6 Empirical Illustration

This section illustrates the potential usefulness of the proposed CESFE test by applying it to evaluate and compare expected shortfall forecasts for daily S&P 500 index returns. Expected shortfall must be computed on a daily basis for the bank-wide internal model for regulatory capital purposes. In calculating expected shortfall, BCBS requires a one-tailed 2.5th percentile to be used, so that $\tau = 0.025$ is considered in this empirical illustration.

The daily S&P 500 price index was taken from Yahoo.Finance to compute returns from June 5, 1998 to April 18, 2018 ($T = 5000$ observations). The first 40% of the sample, corresponding to the period from June 5, 1998 to May 17, 2006 ($m = 2000$ observations), is used as the

in-sample period, while the remaining 60% ($n = 3000$ observations) are reserved to evaluate the out-of-sample forecasting performance. I adopt a fixed forecasting scheme, which means that all forecasts depend on the same set of parameters estimated over the first m observations, while the information set is daily updated for forecasts.

6.1 Risk Models

In addition to the four models in Section 5 used for the simulation study, in this empirical application I also consider the recently developed risk models by Chen et al. (2012) and Taylor (2019). In particular, Taylor (2019) uses a semiparametric method to forecast VaR and ES based on the asymmetric Laplace distribution with the probability density function of the form

$$f(y_t) = \frac{\tau - 1}{ES_t} exp\left(\frac{(y_t - VaR_t)(\tau - \mathbb{I}(y_t \leq VaR_t))}{\tau ES_t}\right) \tag{6.1}$$

where VaR_t follows either SAV- or AS-CAViaR process of Engle and Manganelli (2004) who generalize (5.4) and (5.5) by allowing $\beta_1(\tau)$ dependent of τ. The simple formulation for ES in (6.1) takes the form

$$ES_t = [1 + exp(\gamma_0)] VaR_t \tag{6.2}$$

where γ_0 is a constant parameter to be estimated. Since dynamics of VaR may not be the same as dynamics of ES, I suggest an alternative formulation for ES as[15]

$$ES_t = (1 + \delta_t) VaR_t \tag{6.3}$$

[15]Taylor (2019) suggests a different dynamic process for modeling expected shortfall as $ES_t = VaR_t - x_t$ where

$$x_t = \begin{cases} \gamma_0 + \gamma_1 (VaR_{t-1} - y_{t-1}) + \gamma_2 x_{t-1} & if\ y_{t-1} \leq VaR_{t-1} \\ x_{t-1} & otherwise \end{cases}$$

However, given that $\tau = 0.025$ is an extreme lower tail, 97.5% of probability mass of y_t thus have $x_t = x_{t-1}$, which behaves like a unit root. In the empirical application of this paper, this issue has caused nonstationary behavior in expected shortfall forecasts from a fixed forecast scheme.

with an autoregressive process for δ_t

$$log\left(\delta_t\right) = \gamma_0 + \gamma_1 log\left(\delta_{t-1}\right) + \gamma_2 \left|y_{t-1}\right| \tag{6.4}$$

Note that (6.2) and (6.3) ensure that the estimates of VaR and ES do not cross each other. I denote the model with (5.4) and (6.2) as $SAV - CAViaRES$, the model with (5.4) and (6.3) as $SAV - AR - CAViaRES$, the model with (5.5) and (6.2) as $AS - CAViaRES$, and the model with (5.5) and (6.3) as $AS - AR - CAViaRES$.

It should be noted that (6.1) is used as a quasi-maximum likelihood to infer the values of VaR_t and ES_t. In this context, the observations y_t are not assumed to follow the asymmetric Laplace distribution. To emphasize this, Gerlach et al. (2011), Liu (2016) and Liu and Luger (2018) clarify that the parameter τ is not estimated, but is a chosen fixed value, and that it is only a quantile that is estimated. The asymmetric Laplace quasi-likelihood simply provides a computationally convenient basis with which to enable their Bayesian approach to quantile regression.

By contrast, to capture potential skewness and heavy tails, Chen et al. (2012) assume that y_t follows an asymmetric Laplace distribution as

$$y_t = \left(\varepsilon_t - \mu_\varepsilon\right)\sigma_t, \quad \varepsilon_t \overset{i.i.d}{\sim} AL(0,1,p) \tag{6.5}$$

where $AL\left(0,1,p\right)$ represents the standard asymmetric Laplace distribution with mode 0 and variance 1, and the shape parameter p which is defined such that $p = Pr\left(\varepsilon_t < 0\right)$. The $AL\left(0,1,p\right)$ probability density takes the following form

$$f\left(\varepsilon;p\right) = b_p exp\left[-b_p\left|\varepsilon\right|\left(\frac{1}{p}\mathbb{I}\left(\varepsilon<0\right) + \frac{1}{1-p}\mathbb{I}\left(\varepsilon>0\right)\right)\right] \tag{6.6}$$

where $b_p = \sqrt{p^2 + \left(1+p\right)^2}$. The variance is 1 in (6.6), but the mean is $E(\varepsilon_t) = \frac{1-2p}{b_p}$ denoted as μ_ε. Thus, $e_t = \varepsilon_t - \mu_\varepsilon$ has an AL distribution with mean 0, variance 1, and the shape parameter p. Note that $p = 0.5$ implies an symmetric AL distribution. Specifically, if $p < 0.5$, the density is skewed to the right, while the opposite applies for $p > 0.5$. Different from τ in

(6.1), the shape parameter p in (6.6) will be estimated from data.

The time-varying variance in (6.5) follows either a standard GARCH(1,1) process as

$$\sigma_t^2 = \beta_0 + \beta_1 \sigma_{t-1}^2 + \beta_2 y_{t-1}^2 \tag{6.7}$$

or a GJR-GARCH(1,1) process as

$$\sigma_t^2 = \beta_0 + \beta_1 \sigma_{t-1}^2 + \beta_2^+ \mathbb{I}\left(y_{t-1} > 0\right) y_{t-1}^2 + \beta_2^- \mathbb{I}\left(y_{t-1} < 0\right) y_{t-1}^2 \tag{6.8}$$

The forecasts of VaR can then be obtained as

$$VaR_{t+1}\left(\tau|\mathcal{F}_t\right) = \begin{cases} \sigma_{t+1} \frac{p}{b_p} log\left(\frac{\tau}{p}\right) - \mu_\varepsilon \sigma_{t+1}, & for \quad 0 \leq \tau < p \\ -\sigma_{t+1} \frac{1-p}{b_p} log\left(\frac{1-\tau}{1-p}\right) - \mu_\varepsilon \sigma_{t+1}, & for \quad p \leq \tau < 1 \end{cases}$$

and the forecasts of expected shortfall conditional on y_{t+1} being below VaR_{t+1} is given by

$$ES_{t+1}\left(\tau|\mathcal{F}_t\right) = \left[1 - \frac{1}{log\left(\frac{\tau}{p}\right)}\right] VaR_{t+1}\left(\tau|\mathcal{F}_t\right); \quad 0 \leq \tau < p$$

where only the relevant case $\tau < p$ is shown. I denote the models of (6.7) and (6.8) with a constant shape parameter, p, as $ALGARCH - CP$ and $ALGJR - CP$, respectively.

To allow dynamics of ES different from dynamics of VaR, the following specification for the shape parameter p is specified to allow a time-varying shape

$$p_t = \frac{1}{1 + \sqrt{\frac{u_t}{v_t}}}$$

where

$$u_t = (1 - \lambda) \left|e_{t-1}\right| \mathbb{I}\left(e_{t-1} \geq 0\right) + \lambda u_{t-1}$$

$$v_t = (1 - \lambda) \left|e_{t-1}\right| \mathbb{I}\left(e_{t-1} < 0\right) + \lambda v_{t-1}$$

and $0 \leq \lambda \leq 1$ is an exponential smoothing parameter. The dynamic specification of the shape

parameter allows all higher moments to change over time, in a manner directly influenced by the standardized data sample $e_t = y_t/\sigma_t$. I denote the models of (6.7) and (6.8) with a time-varying shape parameter, p_t, as $ALGARCH-TVP$ and $ALGJR-TVP$, respectively.

For each of the twelve models, I first construct a vector of estimates of the unknown parameters by using the first $m = 2000$ observations. I then use this vector of parameter estimates to form out-of-sample VaR and ES forecasts according to a fixed forecasting scheme. In other words, at each data time period t, $m \leq t \leq T-1$, I compute one-step-ahead forecasts, $VaR_{i,t+1}$ and $ES_{i,t+1}$, for $i = 1, 2, ..., 12$, based on the twelve models by updating the information set $\mathcal{F}_{t-1}$ to $\mathcal{F}_t$.

For illustration, I report the parameter estimates of the twelve models in Table 1 where t-statistics are included in parentheses. The table shows that the parameter estimates are statistically significant at 5% level, except the insignificant estimates of $\beta_2^+(\tau)$ from the ALGJR-TVP model and a few $\beta_0(\tau)$'s. The estimates of λ that determines the time-varying shape parameter p_t are significant at all conventional levels, while estimates for the constant shape parameter p suggest that the returns are left-skewed, as $\hat{p} > 0.5$. The results also show that the quantile persistence, $\beta_1(\tau = 2.5\%)$, estimated from the SAV- and AS-CAViaRES models, appears to be higher than the volatility persistence, β_1, from the GARCH and GJR models.

As a quick check of the out-of-sample performance of individual risk models, I compute the empirical coverage ratio, $\hat{\tau}/\tau$, for VaR, where $\hat{\tau} = n^{-1}\sum_{t=0}^{n-1}\mathbb{I}\left(y_{t+1} < \widehat{VaR}_{t+1}\right)$, and the empirical loss ratio, $ELR = \sum_{t=0}^{n-1}\left(\widehat{ES}_{t+1}\mathbb{I}\left(y_{t+1} < \widehat{VaR}_{t+1}\right)\right) / \sum_{t=0}^{n-1}\left(y_{t+1}\ \mathbb{I}\left(y_{t+1} < \widehat{VaR}_{t+1}\right)\right)$ for ES. Note that these ratios are *absolute* evaluations. If a risk model under consideration performs well to satisfy (3.3) and (3.4), then both ratios are expected to equal one. Table 2 reports the empirical out-of-sample coverage and loss ratios. Models with the ratios closer to 1 are preferred. Average deviation is measured as the average of the absolute deviations of the empirical coverage and loss ratios from the value of 1, such as $\left(|1-\hat{\tau}/\tau| + |1-ELR|\right)/2$.

Table 2 shows that the best model for forecasting VaR is ALGARCH-TVP with empirical coverage ratio, 0.907, followed by ALGJR-TVP with the ratio of 0.813. Nonetheless, the table

Table 1: Parameter Estimates of the Risk Models

Volatility Models	$\beta_0 \times 10^4$	β_1	β_2	β_2^+	β_2^-	v	p	λ
GARCH-N	4.123	0.853	0.147					
	(2.192)	(15.31)	(7.143)					
GJR-N	1.582	0.955		-0.018	0.098			
	(2.264)	(62.79)		(-1.956)	(4.375)			
GARCH-t	3.365	0.879	0.120			11.79		
	(2.264)	(18.79)	(2.131)			(9.067)		
GJR-t	1.671	0.854		-0.011	0.101	16.32		
	(3.790)	(16.79)		(-1.913)	(10.29)	(8.298)		
ALGARCH-CP	0.022	0.895	0.104				0.541	
	(2.356)	(12.15)	(3.987)				(11.07)	
ALGJR-CP	0.022	0.906		0.001	0.185		0.541	
	(3.124)	(26.37)		(1.889)	(5.537)		(16.91)	
ALGARCH-TVP	0.031	0.894	0.105					0.961
	(1.317)	(18.94)	(6.013)					(102.4)
ALGJR-TVP	0.031	0.902		0.001	0.192			0.971
	(1.926)	(19.02)		(1.432)	(9.631)			(78.33)
CAViaRES Models	$\beta_0(\tau) \times 10^4$	$\beta_1(\tau)$	$\beta_2(\tau)$	$\beta_2^+(\tau)$	$\beta_2^-(\tau)$	$\gamma_0(\tau)$	$\gamma_1(\tau)$	$\gamma_2(\tau)$
SAV	-3.144	0.923	-0.152			-1.317		
	(-1.404)	(52.39)	(-2.213)			(-40.23)		
SAV-AR	-1.398	0.977	-0.048			-0.144	0.916	3.229
	(-1.445)	(32.59)	(-2.001)			(-2.346)	(7.511)	(20.19)
AS	-1.325	0.975		-0.113	0.030	-1.441		
	(-1.367)	(86.45)		(-6.051)	(3.161)	(-29.50)		
AS-AR	-1.337	0.974		-0.116	0.029	-0.811	0.469	8.396
	(-1.487)	(45.45)		(-3.025)	(2.788)	(-6.385)	(2.934)	(49.86)

This table reports parameter estimates of the twelve risk models for $\tau = 0.025$. t-statistics are reported in parentheses. The models are estimated for daily S&P 500 index returns sampled from June 5, 1998 to May 17, 2006 (2000 observations). SAV, SAV-AR, AS, and AS-AR represent the risk models proposed by Taylor (2019). ALGARCH-CP, ALGJR-CP, ALGARCH- TVP, and ALGJR-TVP are the risk models proposed by Chen et al. (2012). These models are specified in Section 6.1.

shows that the models with the assumed AL distribution have generally overestimated risks, as their empirical coverage ratios are less than 1. Therefore, as being conservative, these models are favored by regulators in requiring higher level of minimum capital from commercial banks. Among models that underestimate risks with coverage ratios greater than 1, the best model for forecasting VaR is the SAV model with empirical coverage ratio, 1.453, followed by the GARCH-t model which has the ratio of 1.48. Commercial banks would prefer these models with empirical coverage ratios greater than 1 in that lower levels of minimum capital are required.

Table 2: Out-of-Sample Empirical Coverage and Loss Ratios

Models	VaR Coverage Ratio	ES Loss Ratio	Average Deviation
GARCH-N	1.507	0.904	0.302
GJR-N	1.733	0.890	0.422
GARCH-t	1.480	0.952	0.264
GJR-t	1.680	0.922	0.379
SAV	1.453	0.960	0.247
SAV-AR	1.680	0.940	0.370
AS	1.747	0.890	0.429
AS-AR	1.773	0.911	0.431
ALGARCH-CP	0.747	1.063	0.158
ALGJR-CP	0.653	1.068	0.208
ALGARCH-TVP	0.907	1.070	0.082
ALGJR-TVP	0.813	1.079	0.133

$\hat{\tau} = n^{-1}\sum_{t=0}^{n-1}\mathbb{I}\left(y_{t+1} < \widehat{VaR}_{t+1}\right)$ as the empirical coverage is expected to equal the nominal coverage τ. $\hat{\tau}/\tau$ is referred to as empirical coverage ratio for VaR forecasts. The empirical loss ratio is computed for expected shortfall forecasts as $ELR = \sum_{t=0}^{n-1}\left(\widehat{ES}_{t+1}\mathbb{I}\left(y_{t+1} < \widehat{VaR}_{t+1}\right)\right) / \sum_{t=0}^{n-1}\left(y_{t+1}\mathbb{I}\left(y_{t+1} < \widehat{VaR}_{t+1}\right)\right)$. Models with these ratios closer to 1 are preferred. Average deviation is measured as the average of the absolute deviations of the empirical coverage and loss ratios from the value of 1, such as $\left(|1-\hat{\tau}/\tau| + |1-ELR|\right)/2$.

On the other hand, Table 2 shows that the best model for forecasting expected shortfall is the SAV model with the loss ratio, 0.960, followed by the GARCH-t model with the loss ratio of 0.952. Similar to the VaR forecasts, the SAV and GARCH-t models tend to underestimate tail risks due to their loss ratios smaller than 1, while the models with the assumed AL distribution have overestimated tail risks as their loss ratios are greater than 1. Among the twelve competing models, the ALGARCH-TVP model has the smallest average deviation (0.082) of the ratios from the value of 1, followed by the model of ALGJR-TVP with the average deviation of 0.133.

6.2 CESFE Test Results

To assess the relative performance of the models with the best empirical coverage and loss ratios as identified in Table 2, I perform the proposed CESFE test for the model set, $\mathcal{M} = (ALGARCH-TVP, ALGJR-TVP, SAV, GARCH-t)$. Specifically, I test the following null hypotheses: (1) ALGARCH-TVP encompasses SAV, (2) ALGARCH-TVP encompasses GARCH-t, (3) ALGARCH-TVP encompasses ALGJR-TVP, and (4) SAV encompasses GARCH-t. The optimal combination weights, $\boldsymbol{\beta}^*_{ij} = \left(\theta^*_{0,ij}, \theta^*_i, \theta^*_j, w^*_{0,ij}, w^*_i, w^*_j\right)$ are estimated for the

forecast combinations $\theta_{0,ij} + \theta_i VaR_{i,t+1} + \theta_j VaR_{j,t+1}$ and $w_{0,ij} + w_i ES_{i,t+1} + w_j ES_{j,t+1}$ using the GMM approach described in Section 4. For the purposes of this empirical application, I set $\boldsymbol{Z}_{1t} = (1, y_t, VaR_{i,t}, VaR_{j,t})$ and $\boldsymbol{Z}_{2t} = (1, y_t, ES_{i,t}, ES_{j,t})$ for $i, j \in \mathcal{M}$ and $i \neq j$.

Table 3 reports the estimated combination weights, $\hat{\theta}_{i,n}, \hat{\theta}_{j,n}, \hat{w}_{i,n}, \hat{w}_{j,n}$, together with t-statistics in parentheses. It is important to note that the computation of t-statistics is based on the estimator $\hat{\boldsymbol{\gamma}}_{n,\tau}$ of (4.10) to obtain the standard errors from *Theorem 1*. In particular, based on the simulation results from the previous section, I report the test results for several selected values of ς, including $0.002, 0.01, 0.02$ and the estimate of ς from (4.12). For these values of ς, the CESFE test has reasonable size and power properties, as shown in the simulation exercise. Table 3 also contains the corresponding test statistics $CESFE_{1n}$ and $CESFE_{2n}$ defined in *Theorem 2*, which are marked with * if they are statistically significant at the 5% level.

The CESFE test results in Table 3 reject the null hypotheses that the tail risk forecasts from the ALGARCH-TVP model encompass the forecasts from either SAV or GARCH-t model, since both $CESFE_{1n}$ and $CESFE_{2n}$ are statistically significant at 5% level using χ^2_4. These results imply that the forecast combinations via the estimated optimal weights will outperform the individual forecasts. While the optimal VaR weights, 0.029 and 0.002, respectively, for SAV and GARCH-t are small and statistically insignificant, their optimal weights, -0.993 and -0.647, of expected shortfall forecasts are statistically significant. The estimated negative weights tend to correct the overestimated tail risks by the ALGARCH-TVP model, consistent with the results from Table 2 .

In addition, the CESFE test results reject the null hypothesis that ALGARCH-TVP encompasses ALGJR-TVP. While the forecasts from the ALGJR-TVP model have received much higher positive and significant optimal combination weights, the negative optimal combination weight (-0.132) of ALGARCH-TVP, which is significant at 10% level for the expected shortfall forecast, appears to be important in correcting the overestimated risk by the ALGJR-TVP model. Note that in Table 2 the overestimated risk from the ALGJR-TVP model is implied by its empirical coverage ratio, 0.813, which is lower than the value of 0.907 from the ALGARCH-TVP model.

Table 3: Conditional Expected Shortfall Forecast Encompassing Test Results

Models	θ_{1n}	θ_{2n}	w_{1n}	w_{2n}	$CESFE_{1n}$	$CESFE_{2n}$
ALGARCH-TVP vs. SAV	0.982	0.029	1.629	-0.993		
$\varsigma = 0.002$	(2.564)	(0.041)	(5.730)	(-2.659)	64.99*	92.27*
$\varsigma = 0.01$	(2.195)	(0.031)	(5.694)	(-2.653)	66.02*	73.90*
$\varsigma = 0.02$	(1.960)	(0.024)	(5.645)	(-2.638)	66.34*	66.49*
$\hat{\varsigma} = 0.0007$	(2.717)	(0.045)	(5.749)	(-2.670)	63.99*	111.5*
ALGARCH-TVP vs. GARCH-t	0.982	0.002	1.319	-0.647		
$\varsigma = 0.002$	(2.226)	(0.003)	(2.907)	(-2.114)	86.94*	51.68*
$\varsigma = 0.01$	(1.971)	(0.001)	(2.899)	(-2.110)	83.13*	24.34*
$\varsigma = 0.02$	(1.732)	(0.001)	(2.887)	(-2.106)	81.90*	17.58*
$\hat{\varsigma} = 0.0007$	(3.316)	(0.004)	(2.893)	(-2.112)	91.10*	104.5*
ALGARCH-TVP vs. ALGJR-TVP	0.042	0.932	-0.132	0.960		
$\varsigma = 0.002$	(0.097)	(2.261)	(-1.753)	(1.913)	54.64*	105.9*
$\varsigma = 0.01$	(0.060)	(1.970)	(-1.760)	(1.960)	59.03*	96.91*
$\varsigma = 0.02$	(0.046)	(1.767)	(-1.761)	(1.969)	60.34*	94.48*
$\hat{\varsigma} = 0.0007$	(0.131)	(3.190)	(-1.746)	(1.857)	53.43*	120.9*
SAV vs. GARCH-t	1.116	-0.077	0.625	0.268		
$\varsigma = 0.002$	(1.773)	(-0.122)	(0.300)	(0.126)	8.709	9.370
$\varsigma = 0.01$	(1.322)	(-0.091)	(0.301)	(0.126)	9.166	9.079
$\varsigma = 0.02$	(1.034)	(-0.071)	(0.299)	(0.126)	9.342	8.373
$\hat{\varsigma} = 0.0008$	(1.535)	(-0.104)	(0.298)	(0.125)	7.935	8.978

This table reports out-of-sample CESFE test results for risk measures. The combination weights are estimated using the GMM approach described in Section 4. t-statistics are reported in parentheses and calculated based on Theorem 1 with $\varsigma = 0.002, 0.01, 0.02$ and $\hat{\varsigma}$ estimated from (4.12). The marked (*) values of $CESFE_{1n}$ and $CESFE_{2n}$ are statistically significant at the 5% level.

By contrast, Table 3 shows that both $CESFE_{1n}$ and $CESFE_{2n}$ are statistically insignificant in the case of SAV versus GARCH-t, so that the CESFE test results are inconclusive for forecast selection between these two competing models. Since the quantile function of the GARCH-t model is a special case of the SAV model, this inconclusive test result indicates that the SAV model might collapse to the GARCH-t model for daily S&P 500 index returns.

Table 4 summarizes the results of the CESFE test which is applied to 66 pairwise comparisons among the 12 competing models. Average optimal combination weights, $\bar{\theta}$ and $\bar{w}$, are reported for each competing model with average t-statistics in parentheses. The columns of Inconclusive, Encompassing, Encompassed and Combination contain the proportion of times among 11 comparisons that the row-heading model: (1) has inconclusive CESFE test results, (2) encompasses other competing models, (3) is encompassed by other competing models, and

(4) is combined with other competing models using the estimated optimal combination weights, respectively.

Table 4: Summary of Conditional Expected Shortfall Forecast Encompassing Test Results

Models	$\bar{\theta}$	$\bar{w}$	Inconclusive	Encompassing	Encompassed	Combination
GARCH-N	0.056	0.339	18.2%	0.0%	36.4%	45.4%
	(0.253)	(1.156)				
GJR-N	0.184	-0.693	9.1%	4.5%	22.7%	63.7%
	(0.680)	(-9.314)				
GARCH-t	0.228	0.110	6.8%	4.5%	31.8%	56.9%
	(1.427)	(0.095)				
GJR-t	0.368	0.868	6.8%	9.1%	11.4%	72.7%
	(1.308)	(2.329)				
SAV	0.409	0.010	18.2%	18.2%	13.6%	50.0%
	(1.695)	(0.231)				
SAV-AR	0.368	0.051	2.3%	0.0%	18.2%	79.5%
	(1.372)	(0.151)				
AS	0.478	-0.112	15.9%	6.8%	34.1%	43.2%
	(1.496)	(-5.325)				
AS-AR	0.615	0.388	9.1%	34.1%	18.2%	38.6%
	(2.046)	(1.383)				
ALGARCH-CP	0.737	0.696	0.0%	63.6%	0.0%	36.4%
	(1.917)	(1.771)				
ALGJR-CP	0.815	1.091	0.0%	27.3%	9.1%	63.6%
	(4.359)	(5.421)				
ALGARCH-TVP	0.876	1.023	0.0%	18.2%	0.0%	81.8%
	(2.531)	(1.863)				
ALGJR-TVP	0.962	1.166	0.0%	9.1%	0.0%	90.9%
	(4.798)	(3.325)				

This table summarizes the CESFE test results for 66 pairwise comparisons among the 12 competing models. Average optimal combination weights, $\bar{\theta}$ and $\bar{w}$, are reported for each model with average t-statistics in parentheses. The columns of Inconclusive, Encompassing, Encompassed and Combination contain the proportion of times that the row-heading model: (1) has inconclusive CESFE test results, (2) encompasses other competing models, (3) is encompassed by other competing models, and (4) is combined with other competing models using the estimated optimal combination weights, respectively.

Table 4 shows that among the competing models, the ALGJR-TVP model has obtained the highest and significant average optimal combination weights, 0.962 and 1.166, respectively for its VaR and ES forecasts. The CESFE test results show that in about 90.9% of the comparisons, the ALGJR-TVP model significantly contributes useful information to improve risk forecasting performance through combinations. In about 63.6% of comparisons, the forecasts from the

ALGARCH-CP model encompass those from competing models. The ALGARCH-CP model has significant combination weights, 0.737 and 0.696, on average for its VaR and ES forecasts, respectively. By contrast, about 36.4% of the CESFE test results show that the forecasts from the GARCH-N model are encompassed by the forecasts of other competing models.

Overall, Table 4 shows that the risk models with the assumption of an asymmetric Laplace distribution outperform other competing models considered in this empirical application for daily S&P 500 index returns, although these models tend to overestimate tail risks so that higher levels of minimum capital will be required for commercial banks. However, when encompassing is rejected, forecast combination via the estimated optimal weights can to some extent correct the overestimation of tail risks.

7 Conclusion

In this paper I propose a conditional encompassing test for comparing alternative VaR and ES forecasts in an out-of-sample framework. I base the evaluation on the concept of encompassing, which requires that a forecast be able to explain the predictive ability of a rival forecast. The test thus can be viewed as a test of superior predictive ability. The setup proposed in this paper also allows for discussing the benefit of forecast combination for VaR and ES forecasts, which becomes relevant in cases where neither forecast encompasses its competitor.

The test relies on a conditional, rather than unconditional, approach to out-of-sample evaluation, and the proposed test is derived in an environment with asymptotically nonvanishing estimation uncertainty. These features allow comparison of forecasts based on both nested and nonnested models and of forecasts produced by general estimation procedures. A fairly standard GMM estimation technique is implemented for the encompassing test, with the optimization procedure appropriately modified to accommodate the non-differentiable criterion functions. The proposed test displays good size and power properties for samples of sizes typically available in financial applications.

I apply the new encompassing test to evaluate and compare forecasts of conditional value-at-risk and expected shortfall for daily S&P 500 index returns. In addition to standard GARCH

models with the assumptions of normal and Student-t distributions, several recently developed risk models are considered to forecast tail risks, including the semiparametric approach of Taylor (2019) based on a quasi-maximum likelihood function and the parametric approach of Chen et al. (2012) that assumes that error terms follow an asymmetric Laplace distribution with a time-varying shape parameter to capture potential skewness and heavy tails.

The encompassing test results have revealed that the risk models of Chen et al. (2012) not only often encompass other competing models, but also contribute useful information to improve risk forecasting performance through forecast combinations. Nonetheless, the risk models of Chen et al. (2012) tend to overestimate tail risks so that higher levels of minimum capital are required for commercial banks. On the other hand, when neither forecast encompasses its competitor (for example, encompassing is rejected), a forecast combination through the estimated optimal weights outperforms individual forecasts.

References

[1] Acerbi, C., Tasche, D. (2002) On the coherence of expected shortfall. Journal of Banking & Finance 26: 1487-1503

[2] Artzner, P., Delbaen, F., Eber, J., Heath, D. (1999) Coherent measures of risk. Mathematical Finance 9(3): 203-228

[3] Bayer, S., Dimitriadis, T. (2020) Regression-Based Expected Shortfall Backtesting. Journal of Financial Econometrics, nbaa013, https://doi.org/10.1093/jjfinec/nbaa013

[4] Bertsimas, D., Lauprete, G., Samarov, A. (2004) Shortfall as a risk measure: properties, optimization and applications. Journal of Economic Dynamics & Control 28: 1353-1381

[5] Bracewell, R.N. (2000) The Fourier Transform and Its Applications. 3rd ed. New York: McGowan-Hill

[6] Broda, S.A., Paolella, M.S. (2011) Expected shortfall for distributions in finance. In Statistical Tools for Finance and Insurance, edited by P., Cizek, W.K. Haerdle and R. Weron. Springer, Berlin.

[7] Chen, S. (2008) Nonparametric Estimation of Expected Shortfall. Journal of Financial Econometrics 6(1): 87-107

[8] Chen, Q., Gerlach, R., Lu, Z. (2012) Bayesian Value-at-Risk and expected shortfall forecasting via the asymmetric Laplace distribution. Computational Statistics and Data Analysis 56: 3498-3516

[9] Clark, T., McCracken, M (2001) Tests of equal forecast accuracy and encompassing for nested models. Journal of Econometrics 105: 85-110

[10] Clements, M.P., Hendry, D.F. (1998) Forecasting economic time series. Cambridge: Cambridge U.K.: University Press

[11] Clements, M., Harvey, D. (2010) Forecast encompassing tests and probability forecasts. Journal of Applied Econometrics 25: 1028-1062

[12] Dimitriadis, T., Schnaitmann, J. (2020) Forecast encompassing tests for the expected shortfall. International Journal of Forecasting, forthcoming

[13] Eklund, J., Karlsson, S. (2007) Forecast Combination and Model Averaging Using Predictive Measures. Econometric Reviews 26(2-4): 329-363

[14] Elliott, G., Timmermann, A. (2004) Optimal forecast combinations under general loss functions and forecast error distributions. Journal of Econometrics 122: 47-80

[15] Engle, R. (2002) Dynamic Conditional Correlation: A Simple Class of Multivariate Generalized Autoregressive Conditional Heteroskedasticity Models. Journal of Business & Economic Statistics 20(3): 339-350

[16] Engle, R. and S. Manganelli (2004) CAViaR: Conditional Autoregressive Value at Risk by Regression Quantiles, Journal of Business & Economic Statistics, 22: 367-381

[17] Fang, Y. (2003) Forecasting combination and encompassing tests. International Journal of Forecasting 19: 87-94

[18] Fissler, T., Ziegel, J. A. (2016) Higher Order Elicitability and Osband's Principle. Annals of Statistics 44: 1680–1707

[19] Gerlach, R., Chen, C., Chan, N. (2011) Bayesian time-varying quantile forecasting for value-at-risk in financial markets. Journal of Business and Economic Statistics 29: 481-492

[20] Gerlach, R., Chen, C. (2016) Bayesian Expected Shortfall Forecasting Incorporating the Intraday Range. Journal of Financial Econometrics 14(1): 128-158

[21] Gerlach, R., Chen, C. (2017) Semi-parametric expected shortfall forecasting in financial markets. Journal of Statistical Computation and Simulation 87(6): 1084-1106

[22] Giacomini, R. and I. Komunjer (2005) Evaluation and Combination of Conditional Quantile Forecasts. Journal of Business & Economic Statistics 23(4): 416-431

[23] Glosten, L., Jagannathan, R., Runkle, D. (1993) On the relation between the expected value and the volatility of the nominal excess return on stock. Journal of Finance 48: 1179-1801

[24] Granger, C.W.J., Ramanathan, R. (1984) Improved methods of combining forecasts. Journal of Forecasting 3: 197-204

[25] Granger, C.W.J. (1989) Combining forecasts-twenty years later. Journal of Forecasting 8: 167-173

[26] Greene, W.H. (2012) Econometric Analysis. Prentice Hall, Upper Saddle River, New Jersey 07458. ISBN: 10: 0-13-139538-6

[27] Harvey, D.I., S.J. Leybourne and P. Newbold (1998) Tests for Forecast Encompassing. Journal of Business & Economic Statistics 16(2): 254-259

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

[29] Koenker, R. (2005) Quantile Regression. Cambridge University Press, New York

[30] Leorato, S., Peracchi, F., Tanase, A. (2012) Asymptotically efficient estimation of the conditional expected shortfall. Computational Statistics and Data Analysis 56: 768-784

[31] Liu, X. (2016) Markov switching quantile autoregression. Statistica Neerlandica 70(4): 356-395

[32] Liu, X., Luger, R. (2018) Markov-Switching Quantile Autoregression: A Gibbs Sampling Approach. Studies in Nonlinear Dynamics and Econometrics. 22(2): 1-33

[33] Lu, M., Mizon, G.E. (1996) The encompassing principle and hypothesis testing. Econometric Theory 12(5): 845-858

[34] Martins-Filho, C., Yao, F., Torero, M. (2018) Nonparametric estimation of conditional value-at-risk and expected shortfall based on extreme value theory. Econometric Theory 34: 23-67

[35] McCracken, M.W. (2000) Robust out of sample inference. Journal of Econometrics 99: 195-223

[36] McNeil, A. J. and R. Frey. 2000. Estimation of Tail-Related Risk Measures for Heteroscedastic Financial Time Series: An Extreme Value Approach. Journal of Empirical Finance7: 271-300

[37] Nadarajah, S., Zhang, B., Chan, S. (2014) Estimation methods for expected shortfall. Quantitative Finance 14(2): 271-291

[38] Newey, W.K., West, K.D. (1987) A simple positive semidefinite, heteroscedasticity and autocorrelation consistent covariance matrix. Econometrica 55: 703-708

[39] Newey, W.K., McFadden, D.L. (1994) Large-sample estimation and hypothesis testing. Handbook of Econometrics Vol. 4, eds, R.F. Engle and D.L. McFadden. New York: North-Holland, p. 2113-2247

[40] Nieto, M., Ruiz, E. (2016) Frontiers in VaR forecasting and backtesting. International Journal of Forecasting 32: 475-501

[41] Peracchi, F., Tanase, A.V. (2008) On estimating the conditional expected shortfall. Applied Stochastic Models in Business and Industry 24: 471–493

[42] Righi, M., Ceretta, P. (2015) A comparison of expected shortfall estimation models. Journal of Economics and Business 78: 14-47

[43] Stock, J., Watson, M. (1999) A comparison of linear and nonlinear univariate models for forecasting macroeconomic times series. In cointegration, causality and forecasting, eds. R.F. Engle and H. White, Oxford, U.K. Oxford University Press, p. 1-44.

[44] Stock, J., Watson, M. (2003) Forecasting output and inflation: the role of asset prices. Journal of Economic Literature 41: 788-823

[45] Tasche, D. (2002) Expected shortfall and beyond. Journal of Banking & Finance 26: 1519-1533

[46] Taylor, J., Bunn, D.W. (1998) Combining forecast quantiles using quantile quantile regression: investigating the derived weights, estimator bias and imposing constraints. Journal of Applied Statistics 25: 193-206

[47] Taylor, J. (2019) Forecasting Value at Risk and Expected Shortfall Using a Semiparametric Approach Based on the Asymmetric Laplace Distribution. Journal of Business & Economic Statistic 37(1): 121-133

[48] West, K.D. (2001) Tests for Forecast Encompassing When Forecasts Depend on Estimated Regression Parameters. Journal of Business & Economic Statistics 19(1): 29-33

[49] Wong, W.K. (2008) Backtesting trading risk of commercial banks using expected shortfall. Journal of Banking & Finance 32: 1404-1415

[50] Xiao, Z., Koenker, R. (2009) Conditional quantile estimation for generalized autoregressive conditional heteroscedasticity models. Journal of the American Statistical Association 104(488): 1696-1712

[51] Zakoian, J. (1994) Threshold heteroskedastic model. Journal of Economic Dynamics and Control 18: 931-955

[52] Zhu, D., Galbraith, J.W. (2011) Modeling and forecasting expected shortfall with the generalized asymmetric Student-t and asymmetric exponential power distributions. Journal of Empirical Finance 18: 765-778

A Proof for (3.10)

Proof. Let

$$\begin{aligned}\xi_t(\boldsymbol{\theta}) &\equiv E_t\left[\tau - \mathbb{I}\left(Y_{t+1} - \boldsymbol{\theta}'\widehat{\boldsymbol{VaR}}_{t+1} < 0\right)\right]\left(Y_{t+1} - \boldsymbol{\theta}'\widehat{\boldsymbol{VaR}}_{t+1}\right) \\ &= \int_{\mathbb{R}} \tau\left(y_{t+1} - \boldsymbol{\theta}'\widehat{\boldsymbol{VaR}}_{t+1}\right) dF_t\left(y_{t+1}\right) \\ &- \int_{\mathbb{R}} \mathbb{I}\left(Y_{t+1} - \boldsymbol{\theta}'\widehat{\boldsymbol{VaR}}_{t+1} < 0\right)\left(Y_{t+1} - \boldsymbol{\theta}'\widehat{\boldsymbol{VaR}}_{t+1}\right) dF_t\left(y_{t+1}\right) \\ &= \int_{-\infty}^{+\infty} \tau\left(y_{t+1} - \boldsymbol{\theta}'\widehat{\boldsymbol{VaR}}_{t+1}\right) dF_t\left(y_{t+1}\right) - \int_{-\infty}^{0} \varepsilon_{t+1} dF_t\left(\varepsilon_{t+1} + \boldsymbol{\theta}'\widehat{\boldsymbol{VaR}}_{t+1}\right)\end{aligned}$$

where $\varepsilon_{t+1} = Y_{t+1} - \boldsymbol{\theta}'\widehat{\boldsymbol{VaR}}_{t+1}$. Thus, $\nabla_\theta \xi_t(\theta) = -\tau\widehat{\boldsymbol{VaR}}_{t+1} - \int_{-\infty}^{0} \widehat{\boldsymbol{VaR}}_{t+1}\varepsilon_{t+1} f_t\left(\varepsilon_{t+1} + \boldsymbol{\theta}'\widehat{\boldsymbol{VaR}}_{t+1}\right) d\varepsilon_{t+1}$, because I assume that the random variable Y_{t+1} has a continuously differentiable density f_t, that is, $dF_t\left(y_{t+1}\right) = f_t(y_{t+1})dy_{t+1}$ and f_t continuous. By arranging the previous equality, I obtain

$$\begin{aligned}\nabla_{\boldsymbol{\theta}}\xi_t(\boldsymbol{\theta}) = -\tau\widehat{\boldsymbol{VaR}}_{t+1} &- \left[\widehat{\boldsymbol{VaR}}_{t+1}\varepsilon_{t+1} f\left(\varepsilon_{t+1} + \boldsymbol{\theta}'\widehat{\boldsymbol{VaR}}_{t+1}\right)\right]_{-\infty}^{0} \\ &+ \int_{-\infty}^{0} \widehat{\boldsymbol{VaR}}_{t+1} f_t\left(\varepsilon_{t+1} + \boldsymbol{\theta}'\widehat{\boldsymbol{VaR}}_{t+1}\right) d\varepsilon_{t+1}\end{aligned}$$

so that $\nabla_{\boldsymbol{\theta}}\xi_t(\boldsymbol{\theta}) = -\tau\widehat{\boldsymbol{VaR}}_{t+1} + \widehat{\boldsymbol{VaR}}_{t+1}\int_{-\infty}^{\boldsymbol{\theta}'\widehat{VaR}_{t+1}} f_t\left(y_{t+1}\right) dy_{t+1}$. I can then write $\nabla_{\boldsymbol{\theta}}\xi_t(\boldsymbol{\theta}) = -E\left\{\left[\tau - \mathbb{I}\left(Y_{t+1} - \boldsymbol{\theta}'\widehat{\boldsymbol{VaR}}_{t+1} < 0\right)\right]\widehat{\boldsymbol{VaR}}_{t+1}\right\}$. If $\boldsymbol{\theta}^*$ is a solution to the minimization problem of quantile regression, then

$$\nabla_{\boldsymbol{\theta}}\xi_t(\boldsymbol{\theta})|_{\boldsymbol{\theta}^*} = 0 \quad a.s. - P$$

that is,

$$E_t\left\{\left[\tau - \mathbb{I}\left(Y_{t+1} - \boldsymbol{\theta}^{*'}\widehat{\boldsymbol{VaR}}_{t+1} < 0\right)\right]\widehat{\boldsymbol{VaR}}_{t+1}\right\} = 0, \quad a.s. - P$$

Because $\widehat{VaR}_{t+1}$ is $\mathcal{F}_t-$measurable, I can rewrite the previous equation as

$$E_t\left[\tau - \mathbb{I}\left(Y_{t+1} - \boldsymbol{\theta}^{*'}\widehat{\boldsymbol{VaR}}_{t+1} < 0\right)\right] = 0, \quad a.s. - P$$

□

B Proof for (3.11)

Proof. Consider the first-order condition of optimization,

$$\nabla_w \mathcal{L}_\tau\left(Y_{t+1} - \boldsymbol{w}'\widehat{\boldsymbol{ES}}_{t+1}; \boldsymbol{\theta}^{*'}\widehat{\boldsymbol{VaR}}_{t+1}\right) = -2\widehat{\boldsymbol{ES}}_{t+1}\left(Y_{t+1} - \boldsymbol{w}'\widehat{\boldsymbol{ES}}_{t+1}\right)\mathbb{I}\left(Y_{t+1} < \boldsymbol{\theta}^{*'}\widehat{VaR}_{t+1}\right) = 0$$

so that the solution, $\boldsymbol{w}^*$, satisfies

$$E_t\left[\widehat{\boldsymbol{ES}}_{t+1}\left(Y_{t+1}-\boldsymbol{w}^{*'}\widehat{\boldsymbol{ES}}_{t+1}\right)\mathbb{I}\left(Y_{t+1}<\boldsymbol{\theta}^{*'}\widehat{VaR}_{t+1}\right)\right]=0,\quad a.s.-P$$

Because $\widehat{\boldsymbol{ES}}_{t+1}$ is $\mathcal{F}_t-$measurable, I can rewrite the previous equation as

$$E_t\left[\left(Y_{t+1}-\boldsymbol{w}^{*'}\widehat{\boldsymbol{ES}}_{t+1}\right)\mathbb{I}\left(Y_{t+1}<\boldsymbol{\theta}^{*'}\widehat{VaR}_{t+1}\right)\right]=0,\quad a.s.-P$$

□

C Proof of Theorem 1

Proof. The derivative of $\boldsymbol{g}\left(\boldsymbol{\beta};\, Y_{t+1}, \boldsymbol{Z}_{1t}, \boldsymbol{Z}_{2t}\right)$ with respect to $\boldsymbol{\beta}$ is given by

$$\nabla_{\boldsymbol{\beta}}\boldsymbol{g}(\boldsymbol{\beta})=\begin{pmatrix}\frac{\partial \boldsymbol{g}_1(\boldsymbol{\theta};Y_{t+1},\boldsymbol{Z}_{1t})}{\partial \boldsymbol{\beta}'}\\ \frac{\partial \boldsymbol{g}_2(\boldsymbol{w};Y_{t+1},\boldsymbol{Z}_{2t},\boldsymbol{\theta})}{\partial \boldsymbol{\beta}'}\end{pmatrix} \tag{C.1}$$

which is a $k\times 4$ derivative matrix. (C.1) requires that $\boldsymbol{g}(\boldsymbol{\beta})$ be once differentiable, which is not the case here. Applying Newey and McFadden (1994), Giacomini and Komunjer (2005) have shown that $\partial\left[\tau-\mathbb{I}\left(Y_{t+1}<\boldsymbol{\theta}'\widehat{\boldsymbol{VaR}}_{t+1}\right)\right]\boldsymbol{Z}_{1t}/\partial\boldsymbol{\theta}'=-\delta\left(\boldsymbol{\theta}'\widehat{\boldsymbol{VaR}}_{t+1}-Y_{t+1}\right)$ $\boldsymbol{Z}_{1t}\widehat{\boldsymbol{VaR}}_{t+1}'$ is a $k_1\times 2$ derivative matrix, where $\delta(\cdot)$ represents the Dirac function, that is, $\delta(x)=0$ if $x\neq 0$ and $\int_{\mathbb{R}}\delta(x)dx=1$. Using this result, I obtain the following matrix of two partitions

$$\nabla_{\boldsymbol{\beta}}\boldsymbol{g}_1(\boldsymbol{\theta})=\left(-\delta\left(\boldsymbol{\theta}'\widehat{\boldsymbol{VaR}}_{t+1}-Y_{t+1}\right)\boldsymbol{Z}_{1t}\widehat{\boldsymbol{VaR}}_{t+1}'\vdots\,\mathbf{0}_{k_1\times 2}\right)$$

which is a $k_1\times 4$ derivative matrix. Furthermore,

$$\nabla_{\boldsymbol{\beta}}\boldsymbol{g}_2(\boldsymbol{\beta})=\left(-\delta\left(\boldsymbol{\theta}'\widehat{\boldsymbol{VaR}}_{t+1}-Y_{t+1}\right)\boldsymbol{Z}_{2t}\widehat{\boldsymbol{VaR}}_{t+1}'\left(y_{t+1}-\boldsymbol{w}'\widehat{\boldsymbol{ES}}_{t+1}\right)\vdots\right.$$
$$\left.-\boldsymbol{Z}_{2t}\widehat{\boldsymbol{ES}}_{t+1}'\mathbb{I}\left(y_{t+1}<\boldsymbol{\theta}'\widehat{\boldsymbol{VaR}}_{t+1}\right)\right)$$

is a $k_2\times 4$ derivative matrix of two partitions.

The proof of *Proposition 2* in Giacomini and Komunjer (2005) shows that $E_t\left[\delta\left(\boldsymbol{\theta}^{*'}\right.\right.$ $\left.\left.\widehat{\boldsymbol{VaR}}_{t+1}-Y_{t+1}\right)\right]=f_t\left(\boldsymbol{\theta}^{*'}\widehat{\boldsymbol{VaR}}_{t+1}\right)$ where $f_t(\cdot)$ is the density of Y_{t+1} conditional on the information set $\mathcal{F}_t$. From (3.10), it is easy to show that $E_t\left[\mathbb{I}\left(y_{t+1}<\boldsymbol{\theta}^{*'}\widehat{\boldsymbol{VaR}}_{t+1}\right)\right]=$ $F_t\left(\boldsymbol{\theta}^{*'}\widehat{\boldsymbol{VaR}}_{t+1}\right)\xrightarrow{p}\tau$ where $F_t(\cdot)$ is the continuous cumulative distribution function of Y_{t+1} conditional on $\mathcal{F}_t$. Then, I have

$$\boldsymbol{\gamma}_1=E\left[\nabla_{\boldsymbol{\beta}}\boldsymbol{g}_1(\boldsymbol{\theta})\right]=\left(-E\left[f_t\left(\boldsymbol{\theta}^{*'}\widehat{\boldsymbol{VaR}}_{t+1}\right)\boldsymbol{Z}_{1t}\widehat{\boldsymbol{VaR}}_{t+1}'\right]\vdots\,\mathbf{0}_{k_1\times 2}\right)$$

and

$$\begin{aligned}\boldsymbol{\gamma}_2 &= E\left[\nabla_{\boldsymbol{\beta}} \boldsymbol{g}_2(\boldsymbol{\beta})\right] \\ &= \left(-E\left[f_t\left(\boldsymbol{\theta}^{*'}\widehat{\boldsymbol{VaR}}_{t+1}\right)\left(y_{t+1} - \boldsymbol{w}^{*'}\widehat{\boldsymbol{ES}}_{t+1}\right)\boldsymbol{Z}_{2t}\widehat{\boldsymbol{VaR}}_{t+1}^{'}\right] \vdots -\tau E\left[\boldsymbol{Z}_{2t}\widehat{\boldsymbol{ES}}_{t+1}^{'}\right]\right)\end{aligned}$$

such that

$$\begin{aligned}\boldsymbol{\gamma} &= E\left[\nabla_{\boldsymbol{\beta}} \boldsymbol{g}(\boldsymbol{\beta})\right] = \begin{pmatrix} \boldsymbol{\gamma}_{11} & \boldsymbol{\gamma}_{12} \\ \boldsymbol{\gamma}_{21} & \boldsymbol{\gamma}_{22} \end{pmatrix} \\ &= -\begin{pmatrix} E\left[f_t\left(\boldsymbol{\theta}^{*'}\widehat{\boldsymbol{VaR}}_{t+1}\right)\boldsymbol{Z}_{1t}\widehat{\boldsymbol{VaR}}_{t+1}^{'}\right] & \boldsymbol{0}_{k_1\times 2} \\ E\left[f_t\left(\boldsymbol{\theta}^{*'}\widehat{\boldsymbol{VaR}}_{t+1}\right)\left(y_{t+1} - \boldsymbol{w}^{*'}\widehat{\boldsymbol{ES}}_{t+1}\right)\boldsymbol{Z}_{2t}\widehat{\boldsymbol{VaR}}_{t+1}^{'}\right] & \tau E\left[\boldsymbol{Z}_{2t}\widehat{\boldsymbol{ES}}_{t+1}^{'}\right] \end{pmatrix}\end{aligned}$$

From *Assumptions 2&3*, I can now apply the *Delta* method to show that $\sqrt{n}\left(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*\right) \xrightarrow{d} \mathcal{N}\left(\boldsymbol{0}, \left(\boldsymbol{\gamma}'\boldsymbol{S}^{-1}\boldsymbol{\gamma}\right)^{-1}\right)$.

□

D Proof of Theorem 2

Proof. From *Theorem 1*, it follows that $\sqrt{n}\left(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*\right) \xrightarrow{d} \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{\Omega}\right)$, with nonsingular $\Omega = \left(\boldsymbol{\gamma}'\boldsymbol{S}^{-1}\boldsymbol{\gamma}\right)^{-1}$. A consistent estimate $\hat{\boldsymbol{\Omega}}$ of $\boldsymbol{\Omega}$ provides that

$$n\left(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*\right)' \hat{\boldsymbol{\Omega}}_n^{-1}\left(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*\right) \xrightarrow{d} \chi_4^2 \tag{D.1}$$

see, e.g., Theorem 4.30 of White (2001), Greene (2012), §13. From (D.1) *Theorem 2(a)&(b)* follow.

□