Forecast encompassing tests for the expected shortfall

Timo Dimitriadis [a,*], Julie Schnaitmann [b]

[a] *Heidelberg Institute for Theoretical Studies, Institute of Economics, Universität Hohenheim, Germany*
[b] *Department of Economics, University of Konstanz, Germany*

A R T I C L E   I N F O

A B S T R A C T

We introduce new forecast encompassing tests for the risk measure Expected Shortfall (ES). The ES has received much attention since its introduction into the Basel III Accords, which stipulate its use as the primary market risk measure for international banking regulation. We utilize joint loss functions for the pair ES and Value at Risk to set up three ES encompassing test variants. The tests are built on an asymptotic theory that is robust to misspecifications. We investigate the finite sample properties of the tests in an extensive simulation study. Finally, we use the encompassing tests to illustrate the potential of forecast combination methods for different financial assets.

1. Introduction

The expected shortfall (ES) was recently introduced as the primary market risk measure for international banking regulation in the Basel III Accords (Basel Committee, 2016, 2017). Consequently, there has been great demand for reliable methods of evaluating and comparing the predictive ability of competing ES forecasts. The ES at probability level $\alpha \in (0, 1)$ is defined as the expectation of the returns smaller than the respective α-quantile (the value-at-risk, VaR), where α is usually set at 2.5%, as proposed by the Basel Accords. The ES is replacing the VaR in banking regulation, as it overcomes several shortcomings of the latter: the VaR lacks coherence and it cannot capture tail risks beyond the α-quantile (Artzner et al., 1999; Basel Committee, 2013; Danielsson et al., 2001). While the empirical properties favor the ES over the VaR as a measure of risk, the ES lacks elicitability, implying that no strictly consistent loss functions exist. The non-elicitability of the ES is overcome by considering the pair VaR and ES which are jointly elicitable, i.e. there

exist joint loss functions for the VaR and the ES (Fissler & Ziegel, 2016). This discovery triggered the development of several forecasting methods and forecast evaluation techniques for the ES, see Barendse (2020), Bayer and Dimitriadis (2020), Dimitriadis and Bayer (2019), Fissler et al. (2016), Nolde and Ziegel (2017), Patton et al. (2019), and Taylor (2019), among others.

A desirable tool for the comparison of ES forecasts are encompassing tests, which however build upon the existence of strictly consistent loss functions. Given two competing forecasts A and B, forecast encompassing involves testing the null hypothesis that forecast A does not perform worse than any (linear) combination of these forecasts. This is carried out by testing whether the optimal combination weight of forecast B deviates significantly from zero.[1] This null hypothesis allows for the convenient interpretation that forecast B does not add any information to forecast A, such that forecast A is superior to forecast B. Appropriate loss functions are needed for encompassing tests for two reasons. First, the *superior*

---

* Correspondence to: Schloss-Wolfsbrunnenweg 35, 69118 Heidelberg, Germany.
E-mail addresses: timo.dimitriadis@h-its.org (T. Dimitriadis), julie.schnaitmann@uni-konstanz.de (J. Schnaitmann).

---

[1] For the classical theory on forecast encompassing, see Clark and McCracken (2001), Clements and Harvey (2009), Diebold (1989), Ericsson (1993), Giacomini and Komunjer (2005), Harvey et al. (1998), Hendry and Richard (1982), Mizon and Richard (1986), and Newbold and Harvey (2007), among others.

performance of competing forecasts is defined in a statistical sense by using strictly consistent loss functions. Second, loss and identification functions are crucial for M- or GMM-estimation of the optimal forecast combination weights through an appropriate regression framework for the risk measure(s) under consideration.

In this paper, we introduce novel encompassing tests for the ES based on the joint loss functions for the VaR and ES developed in Fissler and Ziegel (2016). We introduce the following three test variants for the ES. First, we propose a joint VaR and ES forecast encompassing test. Second, we introduce an auxiliary ES encompassing test that estimates the optimal combination weights for the vector of the VaR and ES, yet only tests the parameters associated with the ES. Although it incorporates both VaR and ES forecasts, this variant only tests ES forecasts. The third variant is a strict ES encompassing test. This test overcomes the dependence of encompassing tests on VaR forecasts by testing competing ES forecasts exclusively. However, this comes at the cost of potential model misspecifications. This third variant is particularly relevant due to the current set of rules established by the Basel Committee on Banking Supervision, which only requires that financial institutions report ES forecasts (Basel Committee, 2016, 2017). Only this test variant can be applied in situations where the person evaluating the forecasts merely has forecasts for the ES at hand. However, in situations where both the VaR and ES forecasts are available (stemming from the same model or forecasting procedure), application of the joint or auxiliary tests is generally recommended.

We implement the encompassing tests through M-estimation of the optimal combination weights (Dimitriadis & Bayer, 2019; Patton et al., 2019) and in an environment with asymptotically non-vanishing estimation uncertainty of the forecasting procedures (Giacomini & Komunjer, 2005; Giacomini & White, 2006). Because the strict ES encompassing test is subject to potential model misspecification, we derive the asymptotic distribution of the test statistics in a general setting that allows for misspecified models. This generalizes the asymptotic theory of Bayer and Dimitriadis (2020), Dimitriadis and Bayer (2019) and Patton et al. (2019) to potentially misspecified (and nonlinear) models. We base the Wald test statistics of the encompassing tests on a covariance estimator that is robust to misspecifications. Our implementation further introduces a *link* or *combination function* that captures the different linear and nonlinear forecast combination methods in existing research on encompassing testing, see Clements and Harvey (2009, 2010), among others.

We analyze the finite sample behavior of our encompassing tests and the effect of the potential model misspecification in an extensive simulation study using models from various model classes associated with the ES. For this, we consider classical GARCH models, the generalized auto-regressive score (GAS) models with time-varying higher moments of Creal et al. (2013), the GAS models for the VaR and ES of Patton et al. (2019), and the ES-CAViaR models of Taylor (2019). Data stemming from the latter three model classes induce some model misspecification for the strict ES encompassing test, and this

allows us to evaluate the effect of misspecifications on our tests. We find that all tests exhibit approximately the correct size with good power properties for all considered simulations. This also holds for the strict ES encompassing test, demonstrating that this test is robust to the degree of model misspecification we usually encounter in financial applications.

Tests for forecast encompassing are commonly used to establish a theoretical basis for forecast combinations in cases where encompassing is rejected for both forecasts (Clements & Harvey, 2009; Giacomini & Komunjer, 2005; Newbold & Harvey, 2007). In such cases, neither of the forecasts alone performs as well as an optimal forecast combination, indicating that a forecast combination incorporates more information than the individual forecasts. Giacomini and Komunjer (2005), Halbleib and Pohlmeier (2012), Taylor (2020), and Timmermann (2006) advocate general forecast combination methods for multiple reasons, and particularly for risk measures with small probability levels, as are customary for the VaR and the ES.

We apply our encompassing tests to ES forecasts from classical GARCH and GAS models, and from the recently developed dynamic ES models of Patton et al. (2019) and Taylor (2019) for daily returns of the IBM stock, the S&P 500 index, and the DAX 30 index. For the IBM stock, forecast combination methods outperform the stand-alone forecasting models in many instances. This pattern is less pronounced for the S&P 500 and the DAX 30 indices, which are already well diversified through their versatile composition. Thus, classical diversification gains (Timmermann, 2006) of forecast combination methods might be less pronounced for stock indices. The two ES-based test variants show similar results, confirming that the strict ES test is robust to potential misspecifications in financial settings.

The classical idea of forecast encompassing goes back to Chong and Hendry (1986), Hendry and Richard (1982), and Mizon and Richard (1986) and was developed for mean forecasts under the squared loss function. Broad reviews of encompassing testing are provided by, e.g. Clements and Harvey (2009) and Newbold and Harvey (2007). Harvey and Newbold (2000) extended the encompassing technique, which classically focuses on two competing forecasts, to multiple forecasts. Giacomini and Komunjer (2005) developed (conditional) encompassing of quantile forecasts and focused on encompassing tests for *methods* instead of *models*. Clements and Harvey (2010) generalized encompassing tests to probabilistic forecasts by relying on strictly consistent scoring rules. Clements and Harvey (2010) and Giacomini and Komunjer (2005) investigated extensions of encompassing to more complicated functionals of the conditional distribution. Our work pursues this path by developing encompassing tests for the ES as a prominent example of higher-order elicitable functionals, when only joint loss functions for vector-valued functionals are available. Our testing approach can be adapted to further higher-order elicitable functionals, such as the Range-VaR and pair of mean and variance (Cont et al., 2010; Embrechts et al., 2018; Fissler & Ziegel, 2019).

The rest of the paper is organized as follows. In Section 2, we introduce encompassing tests for the ES and derive the asymptotic distribution of the associated test statistics under model misspecification. Section 3 presents an extensive simulation study analyzing the size and power properties of our tests. In Section 4, we apply the testing procedure to daily financial returns of the IBM stock and the S&P 500 and DAX 30 and Section 5 concludes. All proofs are deferred to Appendix A. Technical details of the proofs and additional results are provided in the Supplementary Material.

2. Theory

We consider a stochastic process $Z = \{Z_t : \Omega \to \mathbb{R}^{l+1}, l \in \mathbb{N}, t = 1, \ldots, T\}$, which is defined on some common and complete probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where $\mathcal{F} = \{\mathcal{F}_t, t = 1, \ldots, T\}$ and $\mathcal{F}_t = \sigma\{Z_s, s \leq t\}$. We partition the stochastic process as $Z_t = (Y_t, X_t)$, where $Y_t : \Omega \to \mathbb{R}$ is an absolutely continuous random variable of interest and $X_t : \Omega \to \mathbb{R}^l$ is a vector of explanatory variables. We denote the conditional distribution of Y_{t+1}, given the information set $\mathcal{F}_t$, by F_t. Accordingly, $\mathbb{E}_t$, Var_t, and h_t denote the expectation, variance, and density, respectively, corresponding to F_t. Following Giacomini and Komunjer (2005), we consider ($\mathcal{F}_t$-measurable) one-step-ahead forecasts, henceforth denoted by $\hat{f}_t$, $\hat{q}_t$ and $\hat{e}_t$, which are generated by a function $f(\gamma_{t,m}, Z_t, Z_{t-1}, \ldots)$ that is fixed over time. For this, $\gamma_{t,m}$ denotes the (estimated) model parameters at time t or alternatively the semi- or non-parametric estimator used in the construction of the forecasts. This construction allows for fixed forecasting schemes, where the model parameters $\gamma_{t,m}$ are only estimated once, and rolling window forecasting schemes, where the parameters $\gamma_{t,m}$ are re-estimated in each step. We denote general competing forecasts by $\hat{f}_t = (\hat{f}_{1,t}, \hat{f}_{2,t})$, specific VaR (quantile) forecasts by $\hat{q}_t = (\hat{q}_{1,t}, \hat{q}_{2,t})$, and ES forecasts by $\hat{e}_t = (\hat{e}_{1,t}, \hat{e}_{2,t})$.

In the context of evaluating point forecasts, an important property of risk measures (or more general statistical functionals) is *elicitability* (Gneiting, 2011). Elicitability means that there exist strictly consistent loss functions, i.e. loss functions $\rho(Y, f)$ that depend on the random variable $Y \sim F$ and the issued forecast f, whose expectation $\mathbb{E}[\rho(Y, \cdot)]$ is uniquely minimized by the true risk measure $\Gamma(F)$. Using such a loss function, one can assess the quality of issued forecasts by comparing their average losses induced by the realizations of the predicted variable. Evaluating forecasts through strictly consistent loss functions has the desired impact that it incentivizes financial institutions to report their correct forecasts truthfully (Fissler et al., 2016; Gneiting, 2011). As a direct consequence, the literature on tests for forecast comparison and forecast rationality revolves around the associated loss functions, see Clements and Harvey (2010), Diebold and Mariano (1995), Elliott et al. (2005), Giacomini and Komunjer (2005), Giacomini and White (2006), Gneiting (2011), Mizon and Richard (1986), Patton and Timmermann (2007), and Patton (2011), among many others.

Many important statistical functionals, such as the variance, the ES, the minimum, the maximum, and the

mode, are not elicitable, i.e. no strictly consistent loss functions exist (Fissler & Ziegel, 2016; Gneiting, 2011; Heinrich, 2014). This deficiency calls for generalized approaches in many academic disciplines. We built our test procedure for the ES on such an approach, which considers multiple functionals stacked as vectors and considers joint elicitability. Fissler and Ziegel (2016) showed that the ES is jointly elicitable with the VaR by constructing strictly consistent joint loss functions for this pair, which we utilize in our encompassing approach.

In the following section, we formally introduce the concept of forecast encompassing in the classical case of one-dimensional, real-valued, and elicitable functionals. Subsequently, we make use of the higher-order elicitability of the ES and generalize the encompassing approach to ES forecasts in Section 2.2.

2.1. The encompassing principle

Following Diebold (1989), Giacomini and Komunjer (2005), Hendry and Richard (1982), and Mizon and Richard (1986), we formally introduce the classical concept of linear forecast encompassing for one-dimensional, real-valued, and elicitable functionals. We assume that two competing forecasters predict the variable of interest Y_{t+1} and issue one-step-ahead point forecasts $\hat{\boldsymbol{f}}_t = (\hat{f}_{1,t}, \hat{f}_{2,t})$ for a given functional $\Gamma(F_t)$.[2] In order to conduct the forecast evaluation in an out-of-sample fashion, we divide the sample size T in an in-sample part of size m and an out-of-sample part of size n such that $T = m + n$. The in-sample period is used to generate the forecasts $\hat{f}_{1,t}$ and $\hat{f}_{2,t}$ as described at the beginning of Section 2, while the out-of-sample period is used for the evaluation of the forecasts. This procedure poses little restrictions on how to generate the forecasts, and it allows for parametric, semiparametric, or nonparametric techniques and for nested and non-nested forecasting procedures (Giacomini & Komunjer, 2005).

Let $\rho(Y_{t+1}, \hat{f}_t)$ be a strictly consistent loss function for $\Gamma(\cdot)$. Then, we say that forecast $\hat{f}_{1,t}$ encompasses $\hat{f}_{2,t}$ at time t, if

$$\mathbb{E}\left[\rho(Y_{t+1}, \hat{f}_{1,t})\right] \leq \mathbb{E}\left[\rho(Y_{t+1}, \theta_1 \hat{f}_{1,t} + \theta_2 \hat{f}_{2,t})\right], \quad (2.1)$$

for all $(\theta_1, \theta_2) \in \Theta \subseteq \mathbb{R}^2$. Eq. (2.1) implies that, in terms of the loss induced by ρ, the forecast $\hat{f}_{1,t}$ is at least as good as any (linear) combination of $\hat{f}_{1,t}$ and $\hat{f}_{2,t}$. Hence, forecast $\hat{f}_{2,t}$ does not add any information on Y_{t+1} that is not already incorporated in $\hat{f}_{1,t}$. We define (θ_1^*, θ_2^*) as the optimal combination parameters that minimize the expected loss:

$$(\theta_1^*, \theta_2^*) = \underset{(\theta_1, \theta_2) \in \Theta}{\arg\min} \ \mathbb{E}\left[\rho(Y_{t+1}, \theta_1 \hat{f}_{1,t} + \theta_2 \hat{f}_{2,t})\right]. \quad (2.2)$$

---

[2] While we focus our approach on one-step-ahead forecasts, extensions to multi-step-ahead forecasts are straightforward by employing a HAC-type estimator for the asymptotic covariance.

By definition, it holds that $\mathbb{E}\left[\rho\left(Y_{t+1}, \theta_1 \hat{f}_{1,t} + \theta_2 \hat{f}_{2,t}\right)\right] \geq \mathbb{E}\left[\rho\left(Y_{t+1}, \theta_1^* \hat{f}_{1,t} + \theta_2^* \hat{f}_{2,t}\right)\right]$ for all $\left(\theta_1, \theta_2\right) \in \Theta$. In particular, this implies that

$$\mathbb{E}\left[\rho\left(Y_{t+1}, \hat{f}_{1,t}\right)\right] \geq \mathbb{E}\left[\rho\left(Y_{t+1}, \theta_1^* \hat{f}_{1,t} + \theta_2^* \hat{f}_{2,t}\right)\right]. \qquad (2.3)$$

Combining (2.1) and (2.3) yields the following definition of forecast encompassing.

Definition 2.1 (*Linear Forecast Encompassing for Elicitable Functionals*). We say that forecast $\hat{f}_{1,t}$ encompasses $\hat{f}_{2,t}$ at time t with respect to the loss function ρ if and only if

$$\mathbb{E}\left[\rho\left(Y_{t+1}, \hat{f}_{1,t}\right)\right] = \mathbb{E}\left[\rho\left(Y_{t+1}, \theta_1^* \hat{f}_{1,t} + \theta_2^* \hat{f}_{2,t}\right)\right], \qquad (2.4)$$

which is equivalent to $\left(\theta_1^*, \theta_2^*\right) = \left(1, 0\right)$.

Tests for forecast encompassing are carried out through the following steps. First, we regress the realizations Y_{t+1} onto the forecasts $\hat{f}_{1,t}$ and $\hat{f}_{2,t}$ using an appropriate regression technique for the functional under consideration in order to obtain the estimated combination (or encompassing) parameters $\hat{\theta}_n$ and their asymptotic distribution. Then, we test whether these parameters equal one or zero.

As discussed in Clements and Harvey (2009, 2010), there are several different testing specifications available for the encompassing principle. These differ in terms of the admissible specifications of the linear (or nonlinear) forecast combination formula. We generalize and unify these approaches by introducing a general *link* or *combination function*,

$$g : \mathfrak{F} \times \Theta \to \mathbb{R}, \qquad (\hat{\boldsymbol{f}}_t, \theta) \mapsto g(\hat{\boldsymbol{f}}_t, \theta), \qquad (2.5)$$

which maps the forecasts and the respective parameters onto a linear or nonlinear forecast combination, and where $\mathfrak{F}$ denotes the random space of the issued forecasts. For this, the function g and the parameter space Θ have to be chosen such that there is a $\theta_0 \in \Theta$, such that $g(\hat{\boldsymbol{f}}_t, \theta_0) = \hat{f}_{1,t}$ almost surely. Thus, we can test whether $\hat{f}_{1,t}$ alone captures the full information provided by any forecast combination by testing the parametric restriction $\theta^* = \theta_0$.

Definition 2.2 (*General Forecast Encompassing for Elicitable Functionals*). We say that forecast $\hat{f}_{1,t}$ encompasses $\hat{f}_{2,t}$ at time t with respect to the loss function ρ and with respect to the link function g if and only if

$$\mathbb{E}\left[\rho\left(Y_{t+1}, \hat{f}_{1,t}\right)\right] = \mathbb{E}\left[\rho\left(Y_{t+1}, g(\hat{\boldsymbol{f}}_t, \theta^*)\right)\right], \qquad (2.6)$$

which is equivalent to $\theta^* = \theta_0$.

This general definition unifies the following existing specifications of forecast encompassing, but also allows for more general linear and nonlinear specifications, see e.g. Clements and Harvey (2009, 2010) and Ericsson (1993).

Example 2.3. Prominent examples of linear and nonlinear forecast encompassing are the following link functions and associated null hypotheses:

(1) $g(\hat{\boldsymbol{f}}_t, \theta) = \theta_1 + \theta_2 \hat{f}_{1,t} + \theta_3 \hat{f}_{2,t}$ and $\mathbb{H}_0 : (\theta_2^*, \theta_3^*) = (1, 0)$ or $\mathbb{H}_0 : (\theta_1^*, \theta_2^*, \theta_3^*) = (0, 1, 0)$,

(2) $g(\hat{\boldsymbol{f}}_t, \theta) = \theta_1 + \theta_2 \hat{f}_{1,t} + (1 - \theta_2) \hat{f}_{2,t}$ and $\mathbb{H}_0 : \theta_2^* = 1$ or $\mathbb{H}_0 : (\theta_1^*, \theta_2^*) = (0, 1)$,

(3) $g(\hat{\boldsymbol{f}}_t, \theta) = \theta_1 + \hat{f}_{1,t} + \theta_2 \hat{f}_{2,t}$ and $\mathbb{H}_0 : \theta_2^* = 0$ or $\mathbb{H}_0 : (\theta_1^*, \theta_2^*) = (0, 0)$,

(4) $g(\hat{\boldsymbol{f}}_t, \theta) = \theta_1 \hat{f}_{1,t} + \theta_2 \hat{f}_{2,t}$ and $\mathbb{H}_0 : (\theta_1^*, \theta_2^*) = (1, 0)$,

(5) $g(\hat{\boldsymbol{f}}_t, \theta) = \theta_1 \hat{f}_{1,t} + (1 - \theta_1) \hat{f}_{2,t}$ and $\mathbb{H}_0 : \theta_1^* = 1$,

(6) $g(\hat{\boldsymbol{f}}_t, \theta) = \hat{f}_{1,t} + \theta_1 \hat{f}_{2,t}$ and $\mathbb{H}_0 : \theta_1^* = 0$,

(7) $g(\hat{\boldsymbol{f}}_t, \theta) = \theta_1 \pm \exp\left(\theta_2 \log(\pm \hat{f}_{1,t}) + \theta_3 \log(\pm \hat{f}_{2,t})\right)$ and $\mathbb{H}_0 : (\theta_2^*, \theta_3^*) = (1, 0)$.

2.2. Forecast encompassing for the expected shortfall

In this section, we consider encompassing tests for the ES. For absolutely continuous distributions F_t, the ES is formally defined as

$$ES_{t,\alpha}(Y_{t+1}) = \mathbb{E}_t\left[Y_{t+1} | Y_{t+1} \leq Q_{t,\alpha}(Y_{t+1})\right], \qquad (2.7)$$

where $Q_{t,\alpha}(Y_{t+1})$ denotes the conditional α-quantile of Y_{t+1} given $\mathcal{F}_t$. As discussed in the previous section, the main ingredient of forecast encompassing tests is the specification of the underlying loss function, which has to be associated with the risk measures for which we consider forecasts. As such loss functions do not exist for the ES alone, we utilize a strictly consistent joint loss function for the pair consisting of the ES and the VaR, given by Fissler and Ziegel (2016):

$$\rho(Y, q_\alpha, e_\alpha) = -\frac{1}{e_\alpha}\left(e_\alpha - q_\alpha + \frac{(q_\alpha - Y)\mathbb{1}_{\{Y \leq q_\alpha\}}}{\alpha}\right) + \log(-e_\alpha), \qquad (2.8)$$

where the arguments Y, q_α, and e_α denote the return realization, the quantile, and the ES, respectively. As this loss function exhibits the desirable property of having loss differences that are homogeneous of order zero, it is often denoted as the FZ0-loss function, see e.g. Patton et al. (2019). While there are infinitely many strictly consistent loss functions for the pair VaR and ES, recent literature seems to agree upon this choice. Dimitriadis and Bayer (2019) found that it exhibits a stable numerical performance in M-estimation and empirically yields relatively efficient parameter estimates. Nolde and Ziegel (2017) discussed the desirable property of homogeneity of these loss functions, and Bayer and Dimitriadis (2020), Patton et al. (2019), and Taylor (2019) used this loss function to estimate dynamic ES models.

Following the specification of a link function in (2.5), we introduce the quantile- and ES-specific link functions:

$$g^q : \mathfrak{Q} \times \Theta^\beta \to \mathbb{R}, \qquad (\hat{\boldsymbol{q}}_t, \beta) \mapsto g^q(\hat{\boldsymbol{q}}_t, \beta), \qquad (2.9)$$

$$g^e : \mathfrak{E} \times \Theta^\eta \to \mathbb{R}, \qquad (\hat{\boldsymbol{e}}_t, \eta) \mapsto g^e(\hat{\boldsymbol{e}}_t, \eta), \qquad (2.10)$$

where $\mathfrak{Q}$ and $\mathfrak{E}$ denote the random spaces of the VaR and ES forecasts, respectively, $\Theta^\beta \subseteq \mathbb{R}^{k_\beta}$ and $\Theta^\eta \subseteq \mathbb{R}^{k_\eta}$, such that $\Theta = \Theta^\beta \times \Theta^\eta$ and $k_\beta + k_\eta = k \in \mathbb{N}$. We assume that the functions g^q and g^e and the parameter space Θ are chosen such that there are values $\beta_0 \in \Theta^\beta$ and $\eta_0 \in \Theta^\eta$, such that $g^q(\hat{\boldsymbol{q}}_t, \beta_0) = \hat{q}_{1,t}$ and $g^e(\hat{\boldsymbol{e}}_t, \eta_0) = \hat{e}_{1,t}$ almost surely.

In the following, we introduce the concept of joint forecast encompassing for the pair consisting of the VaR and the ES. Analogously to (2.2), we define the optimal combination parameters for the VaR and ES as

$$
\theta^* = (\beta^*, \eta^*) \\
= \underset{(\beta, \eta) \in \Theta}{\arg \min} \; \mathbb{E}\left[\rho\left(Y_{t+1}, g^q(\hat{\boldsymbol{q}}_t, \beta), g^e(\hat{\boldsymbol{e}}_t, \eta)\right)\right]. \quad (2.11)
$$

Definition 2.4 (*Joint VaR and ES Forecast Encompassing*). Let $(\hat{q}_{1,t}, \hat{e}_{1,t})$ and $(\hat{q}_{2,t}, \hat{e}_{2,t})$ denote pair-wise competing forecasts for the pair consisting of the conditional quantile and ES of F_t. We say that $(\hat{q}_{1,t}, \hat{e}_{1,t})$ encompasses $(\hat{q}_{2,t}, \hat{e}_{2,t})$ at time t with respect to the link functions g^q and g^e if and only if

$$
\mathbb{E}\left[\rho\left(Y_{t+1}, \hat{q}_{1,t}, \hat{e}_{1,t}\right)\right] \\
= \mathbb{E}\left[\rho\left(Y_{t+1}, g^q(\hat{\boldsymbol{q}}_t, \beta^*), g^e(\hat{\boldsymbol{e}}_t, \eta^*)\right)\right], \quad (2.12)
$$

where the loss function ρ is given in (2.8). This holds if and only if $(\beta^*, \eta^*) = (\beta_0, \eta_0)$.

We test whether the sequence of joint quantile and ES forecasts $(\hat{q}_{1,t}, \hat{e}_{1,t})$ encompasses the sequence $(\hat{q}_{2,t}, \hat{e}_{2,t})$ for all $t = m, \dots, T-1$ by estimating the parameters of the following semiparametric regression:

$$
Y_{t+1} = g^q(\hat{\boldsymbol{q}}_t, \beta) + u^q_{t+1}, \qquad \text{and} \\
Y_{t+1} = g^e(\hat{\boldsymbol{e}}_t, \eta) + u^e_{t+1}, \quad (2.13)
$$

where $Q_\alpha(u^q_{t+1}|\mathcal{F}_t) = 0$ and $ES_\alpha(u^e_{t+1}|\mathcal{F}_t) = 0$ almost surely for all $t = m, \dots, T-1$ by using the M-estimation technique introduced in Patton et al. (2019) and Dimitriadis and Bayer (2019). We then test for $(\beta^*, \eta^*) = (\beta_0, \eta_0)$ using a Wald-type test statistic.

Definition 2.4 develops a *joint* encompassing test for the VaR and ES that is reasonable given the joint elicitability property of the VaR and ES. However, a further objective of this paper is to construct encompassing tests for the ES alone, which we do in the following.

Definition 2.5 (*Auxiliary ES Forecast Encompassing*). Let $(\hat{q}_{1,t}, \hat{e}_{1,t})$ and $(\hat{q}_{2,t}, \hat{e}_{2,t})$ denote competing forecasts for the pair consisting of the conditional quantile and ES of F_t. We say that $\hat{e}_{1,t}$ *auxiliarily encompasses* $\hat{e}_{2,t}$ at time t with respect to the link functions g^q and g^e if and only if

$$
\mathbb{E}\left[\rho\left(Y_{t+1}, g^q(\hat{\boldsymbol{q}}_t, \beta^*), \hat{e}_{1,t}\right)\right] \\
= \mathbb{E}\left[\rho\left(Y_{t+1}, g^q(\hat{\boldsymbol{q}}_t, \beta^*), g^e(\hat{\boldsymbol{e}}_t, \eta^*)\right)\right], \quad (2.14)
$$

that is, if and only if $\eta^* = \eta_0$.[3]

This parameter restriction is tested using a Wald-type test statistic based on the estimates of the regression setup given in (2.13). As we do not test the quantile specific parameters β^*, we do not require that the underlying quantile forecast also encompasses its competitor under this null hypothesis. Hence, even though this test is based on the joint regression, it is only an encompassing

test of the ES forecasts. We call this test the *auxiliary ES encompassing test*, as it still depends on the auxiliary quantile forecasts that are used to estimate of the optimal combination of parameters.

Given that both the VaR and ES forecasts are available, application of either the joint or auxiliary test is the most plausible approach given their joint elicitability. However, even though the emphasis of the auxiliary encompassing test is on the ES, it still requires quantile forecasts for the implementation of the parameter estimation. This can be problematic for two reasons. First, quantile forecasts are still used in the estimation procedure and thus have an indirect effect on the parameter estimates of ES-specific parameters. For example, the previous tests are not applicable for ES forecasts that are based on the same VaR forecasts, as this implies perfect collinearity of the quantile regressors. Second, the auxiliary test is only applicable in the setup where the person applying the test has access to the quantile forecasts. In the current implementation of the regulatory framework of the Basel Committee (Basel Committee, 2016, 2017), banks are only obligated to report their ES forecasts (at probability level 2.5%), and not their corresponding VaR forecasts. Thus, the accompanying VaR forecasts, on which the ES forecasts are internally based, are not generally available to the regulator who has to decide on an adequate risk management of the financial institution at hand.

In order to account for these scenarios, we further introduce the *strict ES encompassing test*, which only requires ES forecasts. For this, we slightly modify the definition of (2.11) by replacing $g^q(\hat{\boldsymbol{q}}_t, \beta)$ through $g^q(\hat{\boldsymbol{e}}_t, \beta)$,[4]

$$
\theta^* = (\beta^*, \eta^*) = \underset{(\beta, \eta) \in \Theta}{\arg \min} \; \mathbb{E}\left[\rho\left(Y_{t+1}, g^q(\hat{\boldsymbol{e}}_t, \beta), g^e(\hat{\boldsymbol{e}}_t, \eta)\right)\right].
$$
$$(2.15)$$

Definition 2.6 (*Strict ES Forecast Encompassing*). Let $\hat{e}_{1,t}$ and $\hat{e}_{2,t}$ denote competing ES forecasts of the underlying predictive distribution F_t. We say that $\hat{e}_{1,t}$ strictly encompasses $\hat{e}_{2,t}$ at time t with respect to the link functions g^q and g^e if and only if

$$
\mathbb{E}\left[\rho\left(Y_{t+1}, g^q(\hat{\boldsymbol{e}}_t, \beta^*), \hat{e}_{1,t}\right)\right] \\
= \mathbb{E}\left[\rho\left(Y_{t+1}, g^q(\hat{\boldsymbol{e}}_t, \beta^*), g^e(\hat{\boldsymbol{e}}_t, \eta^*)\right)\right], \quad (2.16)
$$

that is, if and only if $\eta^* = \eta_0$.

We test whether $\hat{e}_{1,t}$ strictly encompasses $\hat{e}_{2,t}$ for all $t = m, \dots, T-1$ by setting up the following slightly transformed regression:

$$
Y_{t+1} = g^q(\hat{\boldsymbol{e}}_t, \beta) + u^q_{t+1}, \qquad \text{and} \\
Y_{t+1} = g^e(\hat{\boldsymbol{e}}_t, \eta) + u^e_{t+1}, \quad (2.17)
$$

where $Q_\alpha(u^q_{t+1}|\mathcal{F}_t) = 0$ and $ES_\alpha(u^e_{t+1}|\mathcal{F}_t) = 0$ almost surely for all $t = m, \dots, T-1$. The crucial difference

---

[3] It is important to notice that the optimal combination parameter β^* on the left-hand side of (2.14) is given by (2.11) and not in the sense of an optimal combination parameter of a *restricted* model.

[4] Note that the parameters denoted by θ^* in (2.11) and (2.15) can generally differ.

between this test and the joint and auxiliary encompassing tests is that, instead of using the quantile forecasts $\hat{\boldsymbol{q}}_t$ in the quantile link function g^q, we use the ES forecasts $\hat{\boldsymbol{e}}_t$ for both the quantile and ES link functions g^q and g^e. We argue that this can be seen as a best feasible solution, due to the lack of loss functions for the ES alone and the need to develop forecast evaluation methods given the current setup of the Basel III regulatory framework (Basel Committee, 2016, 2017).

The underlying idea of this test is mainly motivated by pure scale models, i.e. $Y_{t+1} = \sigma_t u_{t+1}$, $u_{t+1} \sim F(0, 1)$, which are still the most frequently used class of models for risk management with the GARCH and stochastic volatility models as prime examples. For this model class, the VaR and ES forecasts are perfectly colinear, $\hat{e}_t = \frac{\xi_\alpha}{z_\alpha}\hat{q}_t$, where z_α and ξ_α are the α-quantile and α-ES of the distribution $F(0, 1)$. Hence, the quantile model $g^q(\hat{\boldsymbol{e}}_t, \beta) = g^q(\hat{\boldsymbol{q}}_t \xi_\alpha/z_\alpha, \beta) = g^q(\hat{\boldsymbol{q}}_t, \tilde{\beta})$ is correctly specified, but with transformed quantile parameters $\tilde{\beta}$.[5] As we only test on the ES-specific parameters η described in Definition 2.6, our test is invariant to this (often linear) transformation of the parameter β, and thus, it is correctly specified for pure scale models.

In the general case, the quantile equation can be misspecified. Thus, we provide an asymptotic theory under general model misspecification for the M-estimator in the following section. Potential model misspecifications might bias the pseudo-true parameters and challenge the interpretability of the test decision, but we argue that this effect is negligible for this setup. First, misspecifications are only *slight* in the sense that daily financial return data are approximated well by pure scale processes. Second, misspecifications are *indirect* in the sense that while the quantile parameters are potentially misspecified, we only test the ES parameters, which are influenced by misspecifications only indirectly through the joint estimation. In the simulation study described in Section 3, we illustrate that the performance of our strict ES encompassing test is not negatively influenced by more general data generating processes. We consider the GAS models with time-varying higher moments of Creal et al. (2013) and the dynamic ES models of Patton et al. (2019) and Taylor (2019).

Tests for equal (superior) predictive ability in the sense of Clark and McCracken (2001), Diebold and Mariano (1995), Giacomini and White (2006), West (2006), and the model confidence set approach of Hansen et al. (2011) can be seen as general alternatives to encompassing tests. As these tests are directly based on the average loss difference, they can only test the predictive ability of the VaR and ES jointly. By contrast, encompassing tests are based on the regression coefficients of the semiparametric quantile and ES models, and hence only indirectly on the respective loss function. This fundamental difference allows for stand-alone encompassing tests for ES forecasts. This constitutes a great advantage for ES encompassing tests.

Strictly speaking, strict consistency of loss functions only implies that the *optimal* forecast exhibits the *smallest possible* loss in expectation. In reality, however, competing forecasts are often misspecified due to estimation errors or misspecified forecasting models. Patton (2019) showed that in these scenarios, the ranking induced by the loss functions can be sensitive to the choice of (strictly consistent) loss functions or even misleading. Holzmann and Eulert (2014) showed that, for competing forecasts that are based on nested information sets and that are correctly specified given their underlying (but usually incomplete) information set (auto-calibrated), applying any strictly consistent loss function results in a correct ranking of the forecasts. In our case of testing forecast encompassing, we indeed build on nested information sets, as it obviously holds that $\sigma\{\hat{f}_{1,t}, \hat{f}_{2,t}\} \supseteq \sigma\{\hat{f}_{1,t}\}$. Thus, by further assuming that the issued forecasts are auto-calibrated given the forecaster's information set, we can conclude that the ranking implied by (2.1) is indeed the correct one and invariant towards the choice of strictly consistent loss functions.

2.3. Asymptotic theory under model misspecification

In the following, we use the short notation $g_t^e(\eta) = g^e(\hat{\boldsymbol{e}}_t, \eta)$ and $g_t^q(\beta) = g^q(\hat{\boldsymbol{q}}_t, \beta)$ (or $g_t^q(\beta) = g^q(\hat{\boldsymbol{e}}_t, \beta)$ in the case of the strict test). We define the M-estimator as

$$\hat{\theta}_n = \underset{\theta \in \Theta}{\arg\min}\, Q_n(\theta),$$

where $\quad Q_n(\theta) = \frac{1}{n} \sum_{t=m}^{T-1} \rho\big(Y_{t+1}, g_t^q(\beta), g_t^e(\eta)\big),$ (2.18)

and the pseudo-true parameter as[6]

$$\theta_n^* = \underset{\theta \in \Theta}{\arg\min}\, Q_n^0(\theta),$$

where $\quad Q_n^0(\theta) = \frac{1}{n} \sum_{t=m}^{T-1} \mathbb{E}\big[\rho\big(Y_{t+1}, g_t^q(\beta), g_t^e(\eta)\big)\big].$

(2.19)

When the link (regression) functions $g^q(\cdot)$ and $g^e(\cdot)$ are correctly specified, we get that the pseudo-true parameter θ_n^* equals the classical true regression parameter and is independent of the sample size n. We further define the corresponding identification functions, which are almost surely the derivative of the loss function ρ with respect to θ:

$$\psi\big(Y_{t+1}, g_t^q(\beta), g_t^e(\eta)\big)$$
$$= \begin{pmatrix} -\frac{\nabla g_t^q(\beta)}{\alpha g_t^e(\eta)}\Big(\mathbb{1}_{\{Y_{t+1} \le g_t^q(\beta)\}} - \alpha\Big) \\ \frac{\nabla g_t^e(\eta)}{g_t^e(\eta)^2}\Big(g_t^e(\eta) - g_t^q(\beta) + \frac{1}{\alpha}(g_t^q(\beta) - Y_{t+1})\mathbb{1}_{\{Y_{t+1} \le g_t^q(\beta)\}}\Big) \end{pmatrix}.$$
(2.20)

We restrict our attention to processes that satisfy the following conditions:

Assumption 2.7. We assume that

---

[5] For the prominent case of linear encompassing link formulas $g^q(\cdot)$, it holds that $\tilde{\beta} = \beta z_\alpha/\xi_\alpha$.

[6] The pseudo-true parameter can generally depend on the issued loss function, i.e. in this case on the zero homogeneous choice in (2.8).

(a) the process Z_t is strongly mixing of size $-r/(r-2)$ for some $r > 2$,

(b) the parameter space $\Theta = \Theta^\beta \times \Theta^\eta \subseteq \mathbb{R}^k$ is compact and non-empty,

(c) the pseudo-true parameter θ_n^* defined in (2.19) is in the interior of Θ and is the unique minimizer of the objective function $Q_n^0(\theta)$, and the sequence $\mathbb{E}_t\big[\psi\big(Y_{t+1}, g_t^q(\beta), g_t^e(\eta)\big)\big]$, defined in (2.20), is uncorrelated,

(d) the distribution of Y_{t+1} given $\mathcal{F}_t$, denoted by F_t, is absolutely continuous with continuous and strictly positive density h_t, which is bounded from above almost surely on the whole support of F_t and Lipschitz continuous,

(e) for all θ in a neighborhood of θ_n^*, it holds that $\left|\frac{1}{g_t^e(\eta)}\right| \leq K < \infty$ for some constant $K > 0$,

(f) the link functions $g_t^q(\beta)$ and $g_t^e(\eta)$ are $\mathcal{F}_t$-measurable, twice continuously differentiable in $\theta = (\beta, \eta)$ on $\text{int}(\Theta)$ almost surely, and if $\mathbb{P}\big(g_t^q(\beta_1) = g_t^q(\beta_2) \cap g_t^e(\eta_1) = g_t^e(\eta_2)\big) = 1$, then $\theta_1 = \theta_2$,

(g) the matrices Λ_n and Σ_n, defined in Proposition 2.9, are positive definite with a determinant bounded away from zero for all sufficiently large n,

(h) it holds that $g_t^q(\beta) \leq Q$, $\nabla g_t^q(\beta) \leq Q_1$, $H_t^q(\beta) \leq Q_2$, $\nabla H_t^q(\beta) \leq Q_3$, and $g_t^e(\eta) \leq E$, $\nabla g_t^e(\eta) \leq E_1$, $H_t^e(\eta) \leq E_2$, $\nabla H_t^e(\eta) \leq E_3$, for all θ in a neighborhood of θ_n^*, where the random variables $Q, E, Q_1, E_1, Q_2, E_2, Q_3, E_3$ are all $\mathcal{F}_t$-measurable, and for some $r > 2$ (from condition (a)), the following moments are bounded: (i) $\mathbb{E}[Q_1^{r+1}]$, (ii) $\mathbb{E}[E_1^{r+1}]$, (iii) $\mathbb{E}[Q_2^{(r+1)/2}]$, (iv) $\mathbb{E}[E_2^{(r+1)/2}]$, (v) $\mathbb{E}[E_1 Q_2]$, (vi) $\mathbb{E}[Q_1 Q_2]$, (vii) $\mathbb{E}[Q_1 E_2]$, (viii) $\mathbb{E}[Q_1^2 E_1]$, (ix) $\mathbb{E}[EE_1^3]$, (x) $\mathbb{E}[EE_3]$, (xi) $\mathbb{E}[EE_1 E_2]$, (xii) $\mathbb{E}[QE_1 E_2]$, (xiii) $\mathbb{E}[QE_1^3]$, (xiv) $\mathbb{E}[Q_1 Q^r E_1^r]$, (xv) $\mathbb{E}[E_1^{-1} E_2 | Y_t|^r]$, (xvi) $\mathbb{E}[E_1^{r+1}|Y_t|^r]$, (xvii) $\mathbb{E}[Y_t^{2r}]$,

(i) for any n, the term $\sup_{\beta \in \Theta^\beta} \sum_{t=m}^{T-1} \mathbb{1}_{\{Y_{t+1}=g_t^q(\beta)\}}$ is almost surely bounded from above.

The following propositions show consistency and asymptotic normality of the M-estimator under potential model misspecifications.

Proposition 2.8. *Given the conditions in Assumption 2.7, it holds that $\hat{\theta}_n - \theta_n^* \overset{\mathbb{P}}{\longrightarrow} 0$.*

Proposition 2.9. *Given the conditions in Assumption 2.7, it holds that*

$$\Omega_n^{-1/2}(\theta_n^*)\sqrt{n}\big(\hat{\theta}_n - \theta_n^*\big) \overset{d}{\longrightarrow} \mathcal{N}(0, I_k), \quad (2.21)$$

with $\Omega_n(\theta_n^*) = \Lambda_n^{-1}(\theta_n^*)\Sigma_n(\theta_n^*)\Lambda_n^{-1}(\theta_n^*)$, where $\Lambda_n(\theta_n^*) = \begin{pmatrix} \Lambda_{n,qq}(\theta_n^*) & \Lambda_{n,qe}(\theta_n^*) \\ \Lambda_{n,eq}(\theta_n^*) & \Lambda_{n,ee}(\theta_n^*) \end{pmatrix}$, and $\Sigma_n(\theta_n^*) = \frac{1}{n}\sum_{t=m}^{T-1} \mathbb{E}\big[\psi\big(Y_{t+1}, g_t^q(\beta_n^*), g_t^e(\eta_n^*)\big) \cdot \psi\big(Y_{t+1}, g_t^q(\beta_n^*), g_t^e(\eta_n^*)\big)^\top\big]$. Furthermore, the components of $\Lambda_n(\theta_n^*)$ are given by

$$\Lambda_{n,qq}(\theta_n^*) = -\frac{1}{n}\sum_{t=m}^{T-1} \mathbb{E}\left[\frac{H_t^q(\beta_n^*)}{\alpha g_t^e(\eta_n^*)}\big(F_t(g_t^q(\beta_n^*)) - \alpha\big)\right.$$
$$\left. + \frac{\nabla g_t^q(\beta_n^*)\nabla g_t^q(\beta_n^*)^\top}{\alpha g_t^e(\eta_n^*)} h_t(g_t^q(\beta_n^*))\right], \quad (2.22)$$

$$\Lambda_{n,qe}(\theta_n^*) = \Lambda_{n,eq}(\theta_n^*)^\top$$
$$= \frac{1}{n}\sum_{t=m}^{T-1} \mathbb{E}\left[\frac{\nabla g_t^q(\beta_n^*)\nabla g_t^e(\eta_n^*)^\top}{\alpha g_t^e(\eta_n^*)^2}\big(F_t(g_t^q(\beta_n^*)) - \alpha\big)\right], \quad (2.23)$$

$$\Lambda_{n,ee}(\theta_n^*) = \frac{1}{n}\sum_{t=m}^{T-1} \mathbb{E}\left[\frac{\nabla g_t^e(\eta_n^*)\nabla g_t^e(\eta_n^*)^\top}{g_t^e(\eta_n^*)^2}\right.$$
$$+ \left(\frac{H_t^e(\eta_n^*)}{g_t^e(\eta_n^*)^2} - 2\frac{\nabla g_t^e(\eta_n^*)\nabla g_t^e(\eta_n^*)^\top}{g_t^e(\eta_n^*)^3}\right) \times \quad (2.24)$$
$$\left.\left(g_t^e(\eta_n^*) - g_t^q(\beta_n^*) + \frac{1}{\alpha}(g_t^q(\beta_n^*) - Y_{t+1})\mathbb{1}_{\{Y_{t+1}\leq g_t^q(\beta_n^*)\}}\right)\right], \quad (2.25)$$

where $H_t^q(\beta)$ and $H_t^e(\eta)$ are the Hessian matrices of $g_t^q(\beta)$ and $g_t^e(\eta)$, respectively.

The two preceding propositions extend the asymptotic theory of Patton et al. (2019) to the case of possibly misspecified models, and the misspecification theory for linear models of Bayer and Dimitriadis (2020) to nonlinear models. The proofs in Appendix A combine, extend, and proceed along the lines of the ideas of Engle and Manganelli (2004) and Patton et al. (2019). The conditions closely resemble the regularity conditions of Patton et al. (2019). Because we allow for model misspecifications, we impose a unique minimization condition (c) and slightly strengthen the moment conditions (h). In the baseline case of linear encompassing link functions g^q and g^e, the required moment conditions simplify to those given in Bayer and Dimitriadis (2020).

For the estimation of the asymptotic covariance matrix $\widehat{\Omega}_n$ under possible model misspecifications, we follow the approach of Bayer and Dimitriadis (2020) and Dimitriadis and Bayer (2019). We deal with the three nuisance quantities in $\widehat{\Omega}_n$ as follows. In order to estimate the density quantile function $h_t(g_t^q(\beta_n^*))$, we follow the *nid*-estimator of Hendricks and Koenker (1992). As the degree of misspecification in the investigated financial time series is small (Bayer & Dimitriadis, 2020), we approximate $F_t(g_t^q(\beta_n^*)) \approx \alpha$. For the conditional truncated variance, $\text{Var}_t\big(g_t^q(\beta_n^*) - Y_{t+1}\big|Y_{t+1} \leq g_t^q(\beta_n^*)\big)$, we employ the *scl-sp* estimator of Dimitriadis and Bayer (2019).

We now consider the asymptotic distributions of our three ES encompassing tests, proposed in Section 2.2, under the null hypotheses and for general link functions, where we test certain s-dimensional ($s \in \mathbb{N}, s \leq k$) subvectors of θ. For this, let $R \in \mathbb{R}^{k \times s}$ be a selection matrix whose columns consist of k-dimensional Cartesian unit (column) vectors $e_j \in \mathbb{R}^k$, which are zero apart from a one in dimension j. For example, when $g_t^q(\beta)$ and $g_t^e(\eta)$ equal the linear link functions with intercept, given in the first point of Example 2.3, $\theta = (\beta_1, \beta_2, \beta_3, \eta_1, \eta_2, \eta_3)$. Then, for the strict and auxiliary ES encompassing tests, $R = (e_5, e_6)$, and for the joint test $R = (e_2, e_3, e_5, e_6)$. These choices pick the respective parameters from θ. Then, we define the respective test statistics by

$$Z_{R,n} = n\big(\hat{\theta}_n R - \theta_n^* R\big) R^\top \widehat{\Omega}_n^{-1} R \big(\hat{\theta}_n R - \theta_n^* R\big)^\top. \quad (2.26)$$

Theorem 2.10 (*ES Encompassing Tests*). *Given the conditions of Assumption 2.7 and given that $\widehat{\Omega}_n - \Omega_n \overset{\mathbb{P}}{\to} 0$ under the respective null hypotheses given in Definition 2.4–2.6, it*

holds that

$$Z_{R,n} \xrightarrow{d} \chi_s^2. \tag{2.27}$$

For linear link functions, this theorem implies that the limiting χ^2 distribution of the joint test has four degrees of freedom, while that of the strict and auxiliary tests has two degrees of freedom.

An important application of these ES encompassing tests is in the context of selecting the best-performing forecast, i.e. selecting at time T a superior forecasting method for the future. This is particularly relevant because the ES was recently introduced into the Basel regulations without having proper forecast selection procedures at hand. Following Giacomini and Komunjer (2005), we propose the following decision rule. We test two encompassing hypotheses, $\mathbb{H}_0^{(1)}$: $\hat{e}_{1,t}$ encompasses $\hat{e}_{2,t}$; and $\mathbb{H}_0^{(2)}$: $\hat{e}_{2,t}$ encompasses $\hat{e}_{1,t}$; for $t = m, \ldots, T-1$. Then, there are four possible scenarios: (1) if neither $\mathbb{H}_0^{(1)}$ nor $\mathbb{H}_0^{(2)}$ are rejected, the test is not helpful for forecast selection. (2) If $\mathbb{H}_0^{(1)}$ is rejected and $\mathbb{H}_0^{(2)}$ is not rejected, we can conclude that forecast $\hat{e}_{2,t}$ does add information to forecast $\hat{e}_{1,t}$, but we cannot conclude the reverse. Thus, we use the forecasting method of $\hat{e}_{2,t}$. (3) If $\mathbb{H}_0^{(2)}$ is rejected and $\mathbb{H}_0^{(1)}$ is not rejected, the same logic applies inversely, and we use the forecasting method of $\hat{e}_{1,t}$. (4) If both, $\mathbb{H}_0^{(1)}$ and $\mathbb{H}_0^{(2)}$ are rejected, the test delivers statistical evidence that both forecasts contain exclusive information and that a forecast combination outperforms the stand-alone forecasts. Consequently, we use a combined forecast $\hat{e}_{c,t} = g^e(\hat{e}_t, \hat{\eta}_n)$, where the estimated combination parameters $\hat{\eta}_n$ are obtained from the M-estimator proposed here.

Forecast encompassing tests that are *conditional* on some information set $\tilde{\mathcal{G}}_t = \sigma\{W_t\}$ based on some $\mathcal{F}_t$-measurable vector of instruments W_t in the sense of Giacomini and Komunjer (2005) can be facilitated by estimating the regression parameters through (over-identified) GMM-estimation instead of M-estimation. However, for the strict ES test, this approach requires the asymptotic theory under model misspecification for the over-identified GMM estimator based on *non-smooth* objective functions. Whereas this is theoretically available for smooth moment conditions (see e.g. Hall and Inoue (2003) and Hansen and Lee (2019)), its generalization to non-smooth objective functions is not straightforward. Thus, we leave *conditional* ES encompassing tests based on misspecified GMM-estimation for future research. The moment conditions of our unconditional approach can be interpreted as *conditional* encompassing with respect to the instruments $\nabla g_t^q(\beta)$ and $\nabla g_t^e(\eta)$. In the classical baseline case of linear forecast encompassing, these instruments simplify to $\hat{q}_t$ and $\hat{e}_t$. Thus, our approach tests *conditional* encompassing with respect to the information set $\mathcal{G}_t = \sigma\{1, \hat{q}_t, \hat{e}_t\} \subseteq \mathcal{F}_t$, which in most cases already captures the most relevant information available.

3. Simulation study

In this section, we evaluate the size and power properties of our three proposed ES encompassing tests and compare them to the VaR encompassing test of Giacomini and Komunjer (2005). For this, we describe the simulation setup in Section 3.1 and we report and discuss the simulation results in Section 3.2. Section 3.3 considers three extensions of the simulation setup with respect to additional data generating processes (DGPs) and loss and link functions.

3.1. The simulation setup

We employ the three encompassing tests based on the linear link functions $g^q(\hat{f}_t, \beta) = \beta_1 + \beta_2 \hat{f}_{1,t} + \beta_3 \hat{f}_{2,t}$ and $g^e(\hat{e}_t, \eta) = \eta_1 + \eta_2 \hat{e}_{1,t} + \eta_3 \hat{e}_{2,t}$, where $\hat{f}_t = \hat{q}_t$ for the joint and auxiliary tests and $\hat{f}_t = \hat{e}_t$ for the strict test, together with the parameter space $\Theta = \{\theta = (\beta, \eta) \in \mathbb{R}^6 : \|\theta\| \leq K\}$.[7] For the respective encompassing tests, in each case we test the following two opposing hypotheses:

$$
\begin{aligned}
\text{Joint}: \quad & \mathbb{H}_0^{(1)}: (\beta_2^*, \beta_3^*, \eta_2^*, \eta_3^*) = (1, 0, 1, 0), \\
& \mathbb{H}_0^{(2)}: (\beta_2^*, \beta_3^*, \eta_2^*, \eta_3^*) = (0, 1, 0, 1), \\
\text{Str \& Aux}: \quad & \mathbb{H}_0^{(1)}: (\eta_2^*, \eta_3^*) = (1, 0), \\
& \mathbb{H}_0^{(2)}: (\eta_2^*, \eta_3^*) = (0, 1), \\
\text{VaR}: \quad & \mathbb{H}_0^{(1)}: (\beta_2^*, \beta_3^*) = (1, 0), \\
& \mathbb{H}_0^{(2)}: (\beta_2^*, \beta_3^*) = (0, 1).
\end{aligned} \tag{3.1}
$$

In the following, we describe two DGPs: for the first, both forecasting models stem from classical GARCH models; the second considers two joint GAS models for the VaR and ES of Patton et al. (2019). For both model classes, we simulate data as a convex combination of two distinct models with a flexible convex combination weight $\pi \in [0, 1]$. This implies that for $\pi = 0$, the first model encompasses the second, while for $\pi = 1$, the inverse holds. For all intermediate parameters $\pi \in (0, 1)$, the data stem from a linear combination, and both forecast encompassing null hypotheses should be rejected, which indicates that a forecast combination method is preferred.

The GARCH DGP

The two GARCH models, which are calibrated to daily IBM returns, are given by $\tilde{Y}_{j,t+1} = \hat{\sigma}_{j,t} u_{t+1}$, for $j = 1, 2$, where $u_{t+1} \overset{iid}{\sim} \mathcal{N}(0, 1)$ and the two distinct volatility specifications are given by

$$\hat{\sigma}_{1,t}^2 = 0.042 + 0.053 \tilde{Y}_{1,t}^2 + 0.925 \hat{\sigma}_{1,t-1}^2, \quad \text{and} \tag{3.2}$$

$$\hat{\sigma}_{2,t}^2 = 0.044 + \left(0.024 + 0.058 \cdot \mathbb{1}_{\{\tilde{Y}_{2,t} \leq 0\}}\right) \tilde{Y}_{2,t}^2 + 0.923 \hat{\sigma}_{2,t-1}^2. \tag{3.3}$$

For both models, we obtain VaR and ES forecasts by $\hat{q}_{j,t} = z_\alpha \hat{\sigma}_{j,t}$ and $\hat{e}_{j,t} = \xi_\alpha \hat{\sigma}_{j,t}$, for $j = 1, 2$, where z_α and ξ_α are the α-quantile and α-ES of the standard normal distribution. Notice that the time index t on $\hat{\sigma}_{j,t}$ indicates that it is a $\mathcal{F}_t$-measurable forecast for time $t + 1$. While the first specification in (3.2) is a classical GARCH(1,1) model (Bollerslev, 1986), the second specification in (3.3)

---

[7] We set the constant K such that it is sufficiently large to avoid restricting the parameter estimation in realistic settings while ensuring that the parameter space Θ is indeed compact.

follows the GJR-GARCH model of Glosten et al. (1993), which allows for a leverage effect. We simulate data from the convex combination of these processes, $Y_{t+1} = ((1 - \pi)\hat{\sigma}_{1,t} + \pi\hat{\sigma}_{2,t})u_{t+1}$ for 21 equally spaced values of $\pi \in [0, 1]$, where $u_{t+1} \overset{iid}{\sim} \mathcal{N}(0, 1)$.

The VaR/ES GAS DGP

In the second simulation setup, we implement the one-factor (1F) and two-factor (2F) GAS models for the VaR and ES of Patton et al. (2019). The 1F-GAS model evolves as

$$\hat{q}_{1,t} = -1.164 \exp(\hat{\kappa}_t) \qquad \text{and}$$
$$\hat{e}_{1,t} = -1.757 \exp(\hat{\kappa}_t), \qquad \text{where} \tag{3.4}$$

$$\hat{\kappa}_t = 0.995\hat{\kappa}_{t-1} + \frac{0.007}{\hat{e}_{1,t-1}} \left(\frac{\tilde{Y}_{1,t}}{\alpha} \mathbb{1}_{\{\tilde{Y}_{1,t} \leq \hat{q}_{1,t-1}\}} - \hat{e}_{1,t-1} \right). \tag{3.5}$$

The 2F-GAS model follows the specification

$$\begin{pmatrix} \hat{q}_{2,t} \\ \hat{e}_{2,t} \end{pmatrix} = \begin{pmatrix} -0.009 \\ -0.010 \end{pmatrix} + \begin{pmatrix} 0.993 & 0 \\ 0 & 0.994 \end{pmatrix} \begin{pmatrix} \hat{q}_{2,t-1} \\ \hat{e}_{2,t-1} \end{pmatrix}$$
$$+ \begin{pmatrix} -0.358 & -0.351 \\ -0.003 & -0.003 \end{pmatrix} \lambda_t, \tag{3.6}$$

where the forcing variable is given by $\lambda_t = \big(\hat{q}_{2,t-1}(\alpha - \mathbb{1}_{\{\tilde{Y}_{2,t} \leq \hat{q}_{2,t-1}\}}), \mathbb{1}_{\{\tilde{Y}_{2,t} \leq \hat{q}_{2,t-1}\}} \tilde{Y}_{2,t}/\alpha - \hat{e}_{2,t-1} \big)^{\top}$. For both models, $j = 1, 2$, we simulate $\tilde{Y}_{j,t+1} \sim \mathcal{N}(\hat{\mu}_{j,t}, \hat{\sigma}_{j,t}^2)$, where the conditional mean and standard deviations are given by $\hat{\mu}_{j,t} = \hat{q}_{j,t} - z_\alpha \frac{\hat{e}_{j,t} - \hat{q}_{j,t}}{\xi_\alpha - z_\alpha}$ and $\hat{\sigma}_{j,t} = \frac{\hat{e}_{j,t} - \hat{q}_{j,t}}{\xi_\alpha - z_\alpha}$, such that $Q_\alpha(\tilde{Y}_{j,t+1}|\mathcal{F}_t) = \hat{q}_{j,t}$ and $ES_\alpha(\tilde{Y}_{j,t+1}|\mathcal{F}_t) = \hat{e}_{j,t}$ almost surely. The parameter values for this model are obtained from Table 8 of Patton et al. (2019) and correspond to parameters calibrated to daily S&P 500 returns.

In order to simulate returns that follow a convex combination of these two conditional distributions, we simulate Bernoulli draws $\pi_{t+1} \sim \text{Bern}(\pi)$ for 21 equally spaced values of $\pi \in [0, 1]$, and let $Y_{t+1} = (1 - \pi_{t+1})\tilde{Y}_{1,t+1} + \pi_{t+1}\tilde{Y}_{2,t+1}$. Thus, for $\pi = 0$, Y_{t+1} follows the 1F-GAS model, for $\pi = 1$, Y_{t+1} follows the 2F-GAS model, and for $\pi \in (0, 1)$, Y_{t+1} follows some convex combination of these two models.[8]

Both models in the GARCH DGP generate data from a pure scale (volatility) process resulting in perfectly colinear VaR and ES forecasts. By contrast, the more general VaR/ES GAS models in the second DGP generate VaR and ES forecasts that are not colinear and consequently introduce misspecification in the quantile model of the strict ES encompassing test. As the utilized parameters are calibrated to daily financial returns, these models reflect a realistic degree of misspecification encountered in practical risk management.

---

[8] Whereas it is straightforward to generate returns stemming from convex combinations of GARCH-type volatility models by using convex combinations of the conditional volatilities, this is not as simple for the more general GAS models considered in this section. Consequently, we use this more involved approach based on Bernoulli draws in order to generate these convex model combinations.

3.2. Simulation results

Table 1 reports the empirical sizes of the three different ES encompassing tests introduced in Section 2 together with the VaR encompassing test of Giacomini and Komunjer (2005) at a 10% nominal significance level based on 2000 Monte Carlo replications. Tables S.1 and S.2 in the Supplementary Material present equivalent results for nominal sizes of 1% and 5%. The column panel $\mathbb{H}_0^{(1)}$ indicates that we test whether model 1 encompasses model 2, while the panel $\mathbb{H}_0^{(2)}$ indicates the reverse.

We find that the two ES encompassing tests (the strict and auxiliary test) are well sized, especially in large samples for both DGPs and for both null hypotheses. While the joint VaR and ES test is slightly oversized, the VaR test exhibits even larger sizes. This behavior is especially remarkable insofar as the ES is considerably further in the tail than the VaR at the same probability level and hence harder to estimate and test. This pattern can be explained by the fact that the asymptotic covariance of the two tests involving the VaR is subject to estimation of the density quantile function $h_t(g_t^q(\beta_n^*))$, which is hard to estimate for small probability levels (Dimitriadis & Bayer, 2019; Giacomini & Komunjer, 2005; Koenker & Bassett, 1978).

We further find that the strict and the auxiliary tests behave almost identically. This also holds for the VaR/ES GAS DGP, for which the regression model of the strict ES encompassing test is potentially misspecified.[9] This suggests that the approximation error induced by the misspecification in the strict ES test is negligible in realistic financial settings. Remarkably, in the vast majority of cases, the strict ES test exhibits better size properties than the correctly specified joint VaR and ES and the VaR encompassing tests.

We present power curves (empirical rejection rates) for both DGPs and different sample sizes in the individual plot panels in Fig. 1. In each plot, we depict the respective power curves of our three ES encompassing tests and the VaR encompassing test of Giacomini and Komunjer (2005) for both null hypotheses and for a nominal significance level of 10% based on 2000 Monte Carlo replications. We observe increasing power for both DGPs, both tested null hypotheses, and for all four encompassing tests for increasing (decreasing) values of the combination parameter π. We find that, while the VaR and joint VaR and ES tests are considerably oversized, they produce a similar test power compared to the strict and auxiliary ES encompassing tests, especially for larger (smaller) values

---

[9] Figure S.1 in the Supplementary Material plots the ratio of the VaR and ES forecasts for a simulated return series of the 2F-GAS model (and further misspecified models described in Section S.1 in the Supplementary Material). Following the discussion after Definition 2.6, this ratio mainly governs the degree of misspecification the regression model of the strict ES encompassing test is subject to. For the 2F-GAS model, the ratio of the VaR and ES forecasts fluctuates between approximately 0.7 and 0.85, while it equals 0.84 for the location-scale approaches under normality. This demonstrates that the VaR/ES GAS simulation designs, which are calibrated to real financial data, generate some moderate degree of model misspecification in the regression model.

Table 1
Empirical sizes of the forecast encompassing tests.

	$\mathbb{H}_0^{(1)}$				$\mathbb{H}_0^{(2)}$			
	Str ES	Aux ES	VaR ES	VaR	Str ES	Aux ES	VaR ES	VaR
n	GARCH							
500	15.25	15.20	18.35	22.75	14.40	14.65	18.80	22.50
1000	11.55	11.10	15.60	20.10	12.30	12.70	17.80	22.85
2500	11.45	11.55	16.35	18.80	11.00	11.25	14.60	17.55
5000	10.05	10.25	13.10	15.35	9.75	10.15	13.90	15.75
n	VaR/ES GAS							
500	29.35	29.75	24.15	27.85	21.70	21.15	23.10	27.80
1000	22.75	21.85	19.55	23.95	18.15	18.60	18.15	22.80
2500	16.05	15.80	16.20	18.35	12.65	13.50	15.65	19.65
5000	13.50	13.60	14.05	16.60	10.60	11.35	14.10	17.95

Notes: This table presents the empirical sizes (in %) of our three forecast encompassing tests for the ES together with a VaR encompassing test of Giacomini and Komunjer (2005) for a nominal size of 10%. The results are shown for the two DGPs described in Section 3.1 in the horizontal panels, for both tested hypotheses in the vertical panels and for different sample sizes.
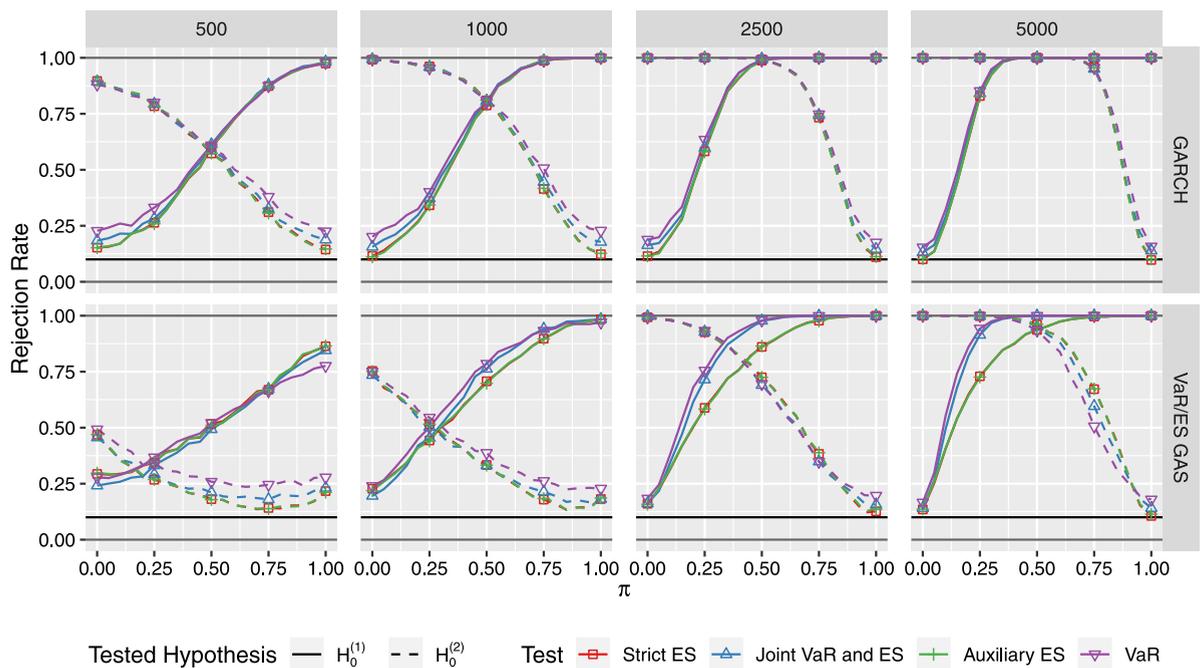

Fig. 1. This figure shows power curves (empirical rejection frequencies) for the encompassing tests with a nominal size of 10% and for the two DGPs described in Section 3.1 in the plot rows. The plot columns depict different sample sizes. The colors indicate the four different encompassing tests, and the line types refer to the two tested null hypotheses. An ideal test exhibits a rejection frequency of 10% for $\pi = 0$ and for $\mathbb{H}_0^{(1)}$ (and inversely for $\pi = 1$ and $\mathbb{H}_0^{(2)}$) and as sharply increasing rejection rates as possible for increasing (decreasing) values of π.

of π. Again, the strict and auxiliary ES encompassing tests are almost indistinguishable, which implies that the strict test is robust to misspecifications induced by the VaR/ES GAS models. Interestingly, we find that the power curves for the VaR/ES GAS specifications are slightly asymmetric, implying that the tests react differently to the specifications of different numbers of driving factors in the GAS models.

3.3. Extensions of the simulation setup

In this section, we consider three extensions of our simulation setup. First, in Section 3.3.1, we present two

additional DGPs. Second, in Section 3.3.2, we analyze the behavior of our tests under different strictly consistent loss functions for the VaR and ES pair. Third, in Section 3.3.3, we employ two additional link functions by testing for forecast encompassing for affine and nonlinear forecast combinations.

3.3.1. Forecast encompassing under different data generating processes

In this subsection, we consider two additional DGPs, namely the GAS-*t* model of Creal et al. (2013) and the ES-CAViaR models of Taylor (2019). These are described in detail in Section S.1 of the Supplementary Material.
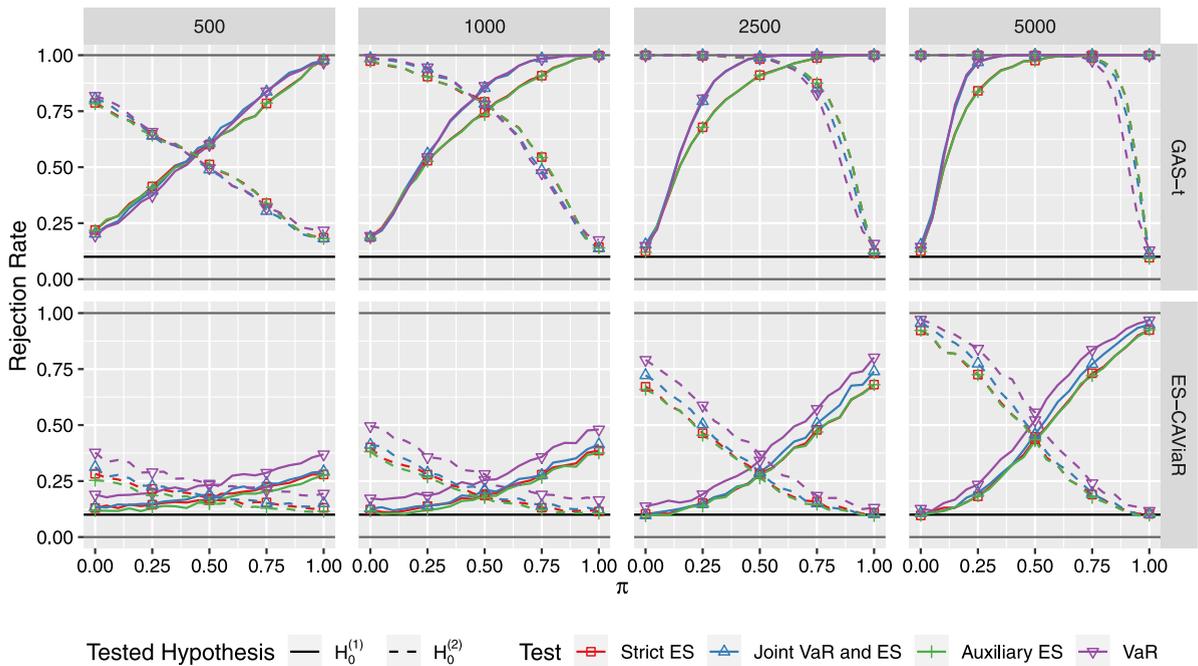
Fig. 2. This figure shows power curves (empirical rejection frequencies) of the encompassing tests with a nominal size of 10%. The plot rows correspond to the GAS-*t* and the ES-CAViaR DGPs described in Section S.1 of the Supplementary Material. The plot columns depict different sample sizes. The colors indicate the four different encompassing tests, and the line types refer to the two tested null hypotheses.

Both models go beyond the class of pure scale models and consequently generate model misspecifications in the quantile regression equation of the strict ES encompassing test. Fig. 2 presents power curves of these two DGPs, and Table S.3 in the Supplementary Material reports the corresponding test size.

The results of these two DGPs qualitatively confirm the simulation results in Section 3.2. The three ES specific tests exhibit accurate empirical test sizes, especially in large samples, and the strict and the auxiliary tests generally exhibit better size properties than the joint VaR and ES and the stand-alone VaR encompassing tests. Increasing the sample size results in increasing power for both DGPs and for all considered encompassing tests. It is noteworthy that all tests show considerably lower power for the ES-CAViaR DGP compared to the other DGPs. This result is comparable to the power results of Giacomini and Komunjer (2005), as this DGP is a slightly modified version of their DGP. In summary, these two additional DGPs demonstrate that the new ES encompassing tests perform well for a variety of realistic data generating processes.

3.3.2. Forecast encompassing under different loss functions

The three encompassing test specifications for the ES presented in Section 2.2 are built on the zero-homogeneous joint loss function for the VaR and ES given in (2.8). While this loss function is the most popular choice in recent literature on semiparametric ES models (Bayer & Dimitriadis, 2020; Patton et al., 2019; Taylor, 2019), there is an entire class of joint loss functions for

the VaR and ES, proposed by Fissler and Ziegel (2016) as

$$
\begin{aligned}
\rho(Y, q_\alpha, e_\alpha) = {} & \left(\mathbb{1}_{\{Y \leq q_\alpha\}} - \alpha \right) \mathfrak{g}(q_\alpha) - \mathbb{1}_{\{Y \leq q_\alpha\}} \mathfrak{g}(Y) \\
& + \phi'(e_\alpha) \left(e_\alpha - q_\alpha + \frac{(q_\alpha - Y) \mathbb{1}_{\{Y \leq q_\alpha\}}}{\alpha} \right) \quad (3.7) \\
& - \phi(e_\alpha) + a(Y),
\end{aligned}
$$

where the function $\mathfrak{g}$ is twice continuously differentiable and increasing, ϕ is three times continuously differentiable, strictly increasing, and strictly convex, and a and $\mathfrak{g}$ are integrable functions (Fissler & Ziegel, 2016). The loss function in (2.8) is a special case of (3.7) for the choices $\mathfrak{g}(z) = 0$, $\phi(z) = -\log(-z)$, and $a(z) = 0$.

The ES encompassing tests can generally be set up by using any choice of (3.7) (fulfilling certain further weak regularity conditions). We consider two additional specifications in the following. Following the theory of homogeneous loss functions (Nolde & Ziegel, 2017) and the numerical performance in linear regression settings (Dimitriadis & Bayer, 2019), we fix $\mathfrak{g}(z) = 0$ and, in addition to $\phi(z) = -\log(-z)$, we employ the choices $\phi(z) = 1/\sqrt{-z}$ and $\phi(z) = -1/z$.[10]

Fig. 3 shows rejection rates of the VaR/ES GAS DGP for different sample sizes and the three encompassing test specifications for the ES, where the different line colors

---

[10] Strictly speaking, the asymptotic theory in Theorem 2.10 only covers the M-estimator based on the loss function in (2.8). However, the proofs and the resulting asymptotic covariance matrices are easily extended to the general case by combining the methods of this paper with the extension of Dimitriadis and Bayer (2019) to the case of general loss functions, as given in (3.7).
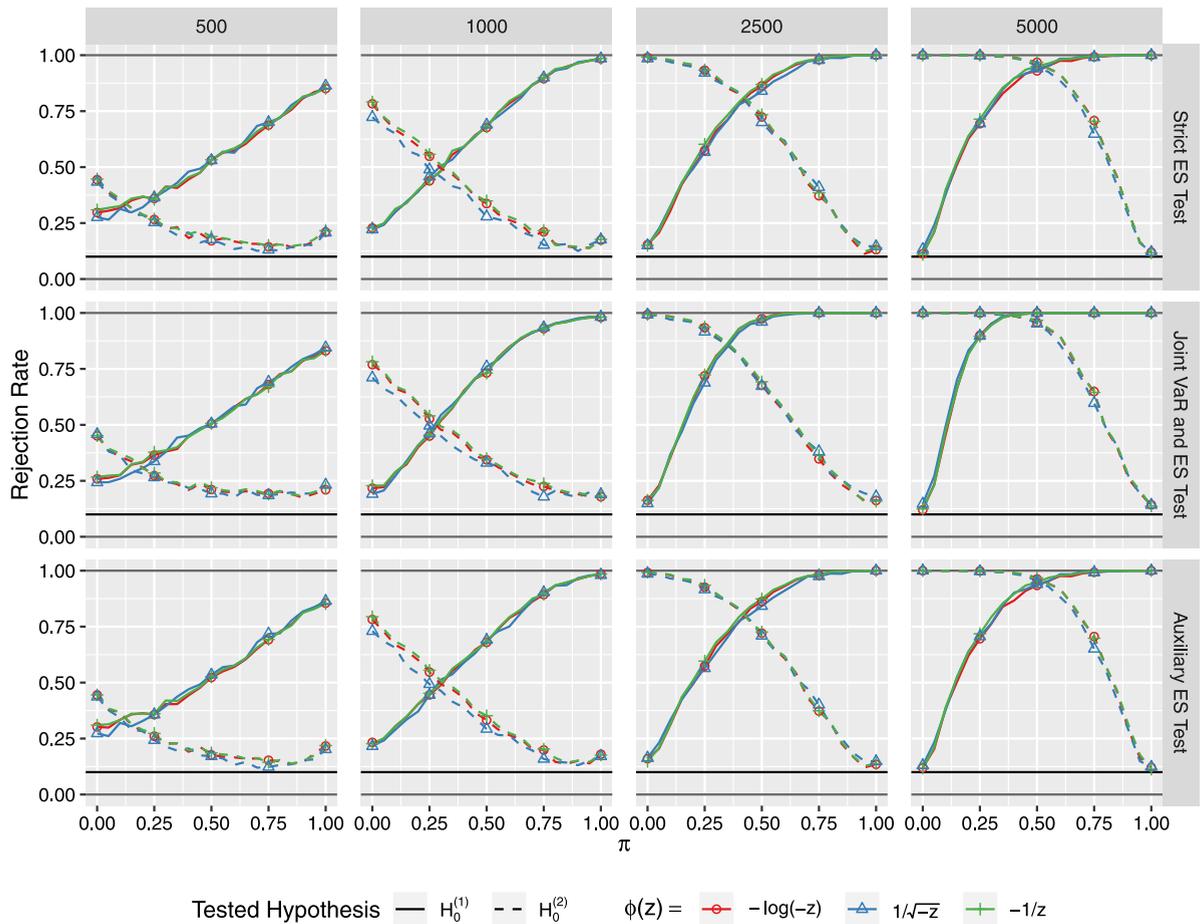
Fig. 3. This figure shows power curves (empirical rejection frequencies) of the three ES specific encompassing tests with a nominal size of 10% for the VaR/ES GAS DGP described in Section S.1 of the Supplementary Material, where the different colors represent the different loss functions employed in the tests, given in and below (3.7). The plot rows depict different sample sizes, the plot columns show the three different ES encompassing tests, and the line types refer to the tested null hypotheses.

represent the three different loss specifications. Table S.4 in the Supplementary Material reports the corresponding test sizes. We find that the tests based on the three different loss functions perform almost indistinguishably, especially in large samples. This result is not unexpected (especially in large samples where the expectation is well approximated by the sample mean) as Fissler and Ziegel (2016) showed that all loss functions in the class in (3.7) are uniquely minimized by the true VaR and ES.

For the potentially misspecified strict ES encompassing test, the pseudo-true parameter defined in (2.19) may theoretically depend on the underlying loss function. Notice that the VaR/ES GAS DGP used in Fig. 3 allows for such a model misspecification. However, we find that the rejection rates of the strict ES encompassing test and those of the two correctly specified encompassing tests in Fig. 3 are unaffected by employing different loss functions. This result implies that the potentially different pseudo-true parameters are almost entirely unaffected by the misspecification.

3.3.3. Forecast encompassing under different link functions

In this section, we employ two additional link function specifications. First, we consider an affine combination including an intercept[11]:

$$g^q(\hat{\boldsymbol{f}}_t, \beta) = \beta_1 + \beta_2 \hat{f}_{1,t} + (1 - \beta_2)\hat{f}_{2,t} \quad \text{and}$$
$$g^e(\hat{\boldsymbol{e}}_t, \eta) = \eta_1 + \eta_2 \hat{e}_{1,t} + (1 - \eta_2)\hat{e}_{2,t}, \tag{3.8}$$

where $\hat{\boldsymbol{f}}_t = \hat{\boldsymbol{q}}_t$ for the joint and auxiliary tests and $\hat{\boldsymbol{f}}_t = \hat{\boldsymbol{e}}_t$ for the strict test. For the joint test, the first null hypothesis is given by $\mathbb{H}_0^{(1)} : (\beta_2^*, \eta_2^*) = (1, 1)$. For the strict and auxiliary tests, it is given by $\mathbb{H}_0^{(1)} : \eta_2^* = 1$. The second, opposing null hypotheses, $\mathbb{H}_0^{(2)}$, are obtained by replacing the ones with zeros. For the affine link functions, we employ the same DGPs as for the encompassing tests based on linear link functions.

---

[11] We include an intercept, as this stabilizes the performance of the associated quantile regression. We do not include classical convex combinations (where $0 \le \beta_2 \le 1$), as our theoretical framework does not allow for testing on the boundary (see e.g. Andrews (1999) for details).
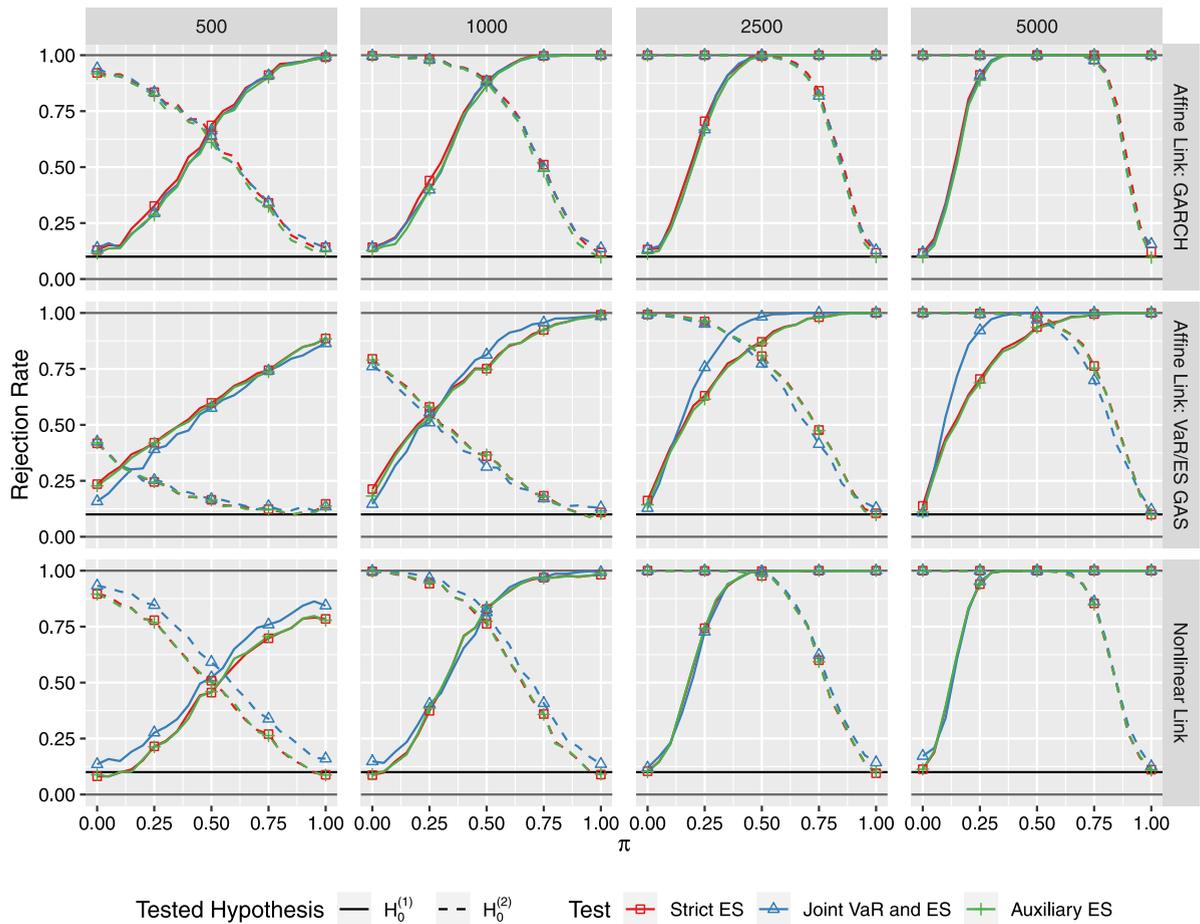
Fig. 4. This figure shows power curves (empirical rejection frequencies) of encompassing tests with a nominal size of 10% for two additional link functions. The first two rows of plots present results for the affine link functions given in (3.8) and for the GARCH and VaR/ES GAS DGPs. The third plot row presents results for the nonlinear link functions given in (3.9) and the respective nonlinear GARCH DGP. The plot columns depict different sample sizes. The colors indicate the three ES specific encompassing tests, and the line types refer to the two tested null hypotheses.

Furthermore, we employ nonlinear link functions, where $\hat{f}_t$ is given as in (3.8), and

$$g^q(\hat{f}_t, \beta) = \beta_1 - \exp(\beta_2 \log(-\hat{f}_{1,t}) + \beta_3 \log(-\hat{f}_{2,t})), \quad \text{and}$$
$$g^e(\hat{e}_t, \eta) = \eta_1 - \exp(\eta_2 \log(-\hat{e}_{1,t}) + \eta_3 \log(-\hat{e}_{2,t})),$$
$$(3.9)$$

and we test the same null hypotheses as for the linear link functions described in (3.1). For the nonlinear link function, we employ a slightly modified GARCH DGP. As in Section 3.1, let $\hat{\sigma}_{1,t}$ and $\hat{\sigma}_{2,t}$ denote the conditional volatilities of the GARCH and GJR-GARCH models. Then, we simulate data according to

$$Y_{t+1} = \exp\big((1 - \pi) \log(\hat{\sigma}_{1,t}) + \pi \log(\hat{\sigma}_{2,t})\big) \cdot u_{t+1} \quad (3.10)$$

for an equally spaced grid of 21 values for $\pi \in [0, 1]$ and where $u_{t+1} \overset{iid}{\sim} \mathcal{N}(0, 1)$. This ensures that for $\pi = 0$, $\hat{q}_{1,t}$ and $\hat{e}_{1,t}$ are the correct VaR and ES forecasts, and vice versa for $\pi = 1$. For any $\pi \in (0, 1)$, the true VaR and ES are given by a combination according to the nonlinear link functions in (3.9).

Fig. 4 presents power curves of these two additional link functions, where for the affine combination, we consider both the GARCH and the VaR/ES GAS DGPs. Table S.5 in the Supplementary Material presents the associated test sizes. We find that the power curves for the affine combinations in Fig. 4 are comparable to the linear specifications, as given in Fig. 1 for both considered DGPs. Moreover, the encompassing tests based on the nonlinear link function in the third row of plots perform similarly to the linear encompassing tests in the GARCH setting. This extension of the simulation setup shows that our ES encompassing tests can be applied based on a variety of different link functions.

4. Empirical application

We use close-to-close returns from the IBM stock, the S&P 500 index, and the DAX 30 index from June 1st, 2000 until May 31st, 2019, which amounts to a total of $T = 4779$ daily observations. We use a fixed forecasting scheme, i.e. the model parameters are estimated once on the first $m = 2000$ in-sample observations. These

parameter estimates are used to generate the VaR and ES forecasts in a rolling-window fashion for the remaining out-of-sample period of $n = 2779$ days. Following the suggestion of the Basel III Accords, we use the probability level $\alpha = 2.5\%$ for the VaR and the ES.

For the analysis, we consider the following competing forecasting models. First, we employ the Historical Simulation (HS) model, which generates VaR and ES forecasts by computing the empirical quantile and ES at level α of the past 250 trading days. The second model is the RiskMetrics (RM) model, which models the conditional volatility as an IGARCH equation with fixed parameter values, $\hat{\sigma}_t^2 = 0.94\hat{\sigma}_{t-1}^2 + 0.06Y_t^2$ and Gaussian residuals. Third, we use the GJR-GARCH(1,1)-t model of Glosten et al. (1993) with Student-t residuals. The fourth model is given by the Student-t-GAS model with time-varying variance and degrees of freedom, as introduced in Section S.1 of the Supplementary Material. The fifth and sixth models are the one- and two-factor GAS models for the VaR and ES of Patton et al. (2019) set out in Section 3.1 and estimated by minimizing the strictly consistent loss function for the VaR and ES given in (2.8). The last two models are the two dynamic ES-CAViaR models of Taylor (2019) described in Section S.1 of the Supplementary Material. Table S.6 in the Supplementary Material shows the correlations of the respective VaR and ES forecasts of these models. We find that no pair of forecasts is perfectly correlated, which is crucial for the applicability of the encompassing tests, as implied by condition (f) of Assumption 2.7.

We run pair-wise encompassing tests comparing all eight forecasting methods. Hence, for each model pair, we run encompassing tests for both hypotheses, i.e. that the first forecast encompasses the second, denoted by $\mathbb{H}_0^{(1)}$, and the inverse, denoted by $\mathbb{H}_0^{(2)}$. This results in four possible outcomes of these two tests: (1) *non-rejection* (NR) indicates that none of the null hypotheses is rejected and the tests are not helpful. (2) *encompassed* (E1) denotes the setting where the first model is encompassed by the competitor model but does not encompass it, i.e. $\mathbb{H}_0^{(1)}$ is rejected but $\mathbb{H}_0^{(2)}$ is not, which results in choosing the competitor model. (3) *encompassing* (E2) indicates that the first model encompasses the other but is not encompassed by it, i.e. $\mathbb{H}_0^{(1)}$ is not rejected but $\mathbb{H}_0^{(2)}$ is, which implies that we choose the first model. Finally, (4) *combination* (C) refers to a setting where both null hypotheses are rejected and we consequently opt for a forecast combination.

In Table 2, we report relative frequencies of test outcomes at the 10% significance level for the different encompassing tests for all three return time series. Tables S.7, S.8, and S.9 in the Supplementary Material report the individual p-values of the encompassing tests. The results can be summarized as follows: First, for the IBM stock returns, we find many cases of double rejections and hence empirical evidence for using forecast combinations. This implies that the individual models provide additional and exclusive information. Hence, a forecast combination is often superior to the stand-alone forecasting models. This finding supports the theoretical advantages of

forecast combinations, presented in a general setting, e.g., by Giacomini and Komunjer (2005), Halbleib and Pohlmeier (2012), and Timmermann (2006), and specifically for the pair VaR and ES by Taylor (2020). Second, for the S&P 500 index, and especially for the DAX 30 index, we overall observe fewer instances of double rejections of the ES encompassing tests. While the decrease in cases where the VaR encompassing test opts for a forecast combination is smaller, these rejections have to be considered carefully given that the VaR encompassing test is oversized in all simulation setups in Section 3, even in large samples. This result can be explained by the fact that the S&P 500 and DAX indices are well diversified, and the returns fluctuate to a lesser extent and exhibit less extreme outliers than single stock return series. Furthermore, the considered VaR and ES forecasts show larger correlations for the indices than for the single stock in Table S.6 in the Supplementary Material, negatively influencing the tests' power. Third, in terms of the frequencies of the cases E1 and E2, we observe recurring patterns over the different models for both time series. The ES-specific GAS and CAViaR type models in particular seem to exhibit a superior performance, while the HS, RM, GARCH, and GAS-t models tend to be encompassed more often. Finally, the two tests that only focus on testing the encompassing of ES forecasts perform almost identically. This supports the conclusion from the simulation study that potential misspecifications do not negatively influence the performance of the strict ES test in realistic financial settings. This is encouraging as the strict ES encompassing test can be applied in cases where one does not have VaR forecasts at hand, such as it is currently imposed by the Basel Committee on Banking Supervision (Basel Committee, 2016, 2017).

Tables S.10, S.11, and S.12 in the Supplementary Material report the joint VaR and ES losses for the zero-homogeneous loss function in (2.8) for forecasts stemming from the stand-alone models and the respective forecast combinations with estimated combination weights from the underlying regressions. These results qualitatively confirm the results of the encompassing tests: e.g., for the IBM stock in the first panel of Table 2, we find that forecast combinations are particularly preferred for the first four models (in the model ordering of the table). The average losses in Table S.10 show a similar pattern throughout all three panels. We observe that the optimal forecast combinations exhibit substantially smaller losses compared to the stand-alone models for the first four models. This decrease is of a considerably smaller magnitude for the last four models.

5. Conclusion

With the implementation of the Basel III Accords (Basel Committee, 2016, 2017), risk managers and regulators have shifted their attention towards the ES. As such, there is a need for forecast evaluation and comparison tools for the ES. In this paper, we introduced new forecast encompassing tests for the ES, based on a joint loss function and an associated joint regression framework for the ES together with the VaR (Dimitriadis & Bayer, 2019;

Table 2

Encompassing test results for the IBM stock, the DAX 30 Index, and the S&P 500 Index.

Model	Strict ES				Aux ES				Joint VaR ES				VaR			
	NR	E1	E2	C	NR	E1	E2	C	NR	E1	E2	C	NR	E1	E2	C
IBM																
HS		57		43		57		43		43		57		43		57
RM		57		43		57		43		43		57	14	43		43
GJR		57		43		57		43		43		57	14	29		57
GAS		43		57		43		57		29		71		29		71
G1F	14	29	43	14	14	29	43	14	14			86		14		86
G2F	14	29	57		14	29	57			29	57	14		29	43	29
ASES	14		86		14		86		14		57	29	14		57	29
SAVES	14		86		14		86		14		86		14		86	
S&P 500																
HS		86		14		86		14		71		29		86		14
RM		71	14	14		71	14	14		71	14	14	14	43	14	29
GJR	29		71		29		71				71	29			71	29
GAS	29	29	29	14	14	43	29	14	57	29		14	14	43	14	29
G1F		29		71		29		71	14			86	29	14	29	29
G2F	29	14	57		14	14	71				57	43	14		29	57
ASES	14		71	14	14		71	14			43	57			29	71
SAVES	14	43	29	14	14	43	29	14		43	43	14	14	29	29	29
DAX 30																
HS		86		14		100				86		14		86		14
RM		43	57			43	57			43	43	14		29	14	57
GJR		57	43			57	43			14	71	14	14	14	43	29
GAS	14	14	71		29		71		14	29	43	14		14	57	29
G1F		71	29			71	29			71	14	14		43	14	43
G2F		86		14		86	14			57		43		43		57
ASES	29		71		29		71				86	14	14		86	
SAVES	14		86		29		71		14		43	43	29	14	29	29

Notes: This table shows results of the pair-wise encompassing test outcomes for the three ES encompassing tests and the VaR encompassing test of Giacomini and Komunjer (2005). The eight considered models estimated an in-sample period of $m = 2000$ days, with a nominal test significance level of 10%, for the IBM stock, the S&P 500 index, and the DAX 30 index. The individual columns report the relative frequencies (in %) of how often the respective test outcome occurs for a model (given in the rows) when this model generates the *first* forecasts in the encompassing tests. The column "NR" (non-rejection) reports the frequency of cases where neither null hypothesis is rejected, while the column "C" (combination) indicates the frequency of cases where both null hypotheses are rejected. The column "E1" (encompassed) refers to the case where $\mathbb{H}_0^{(1)}$ is rejected but $\mathbb{H}_0^{(2)}$ is not, i.e. it gives the relative frequencies that this model is encompassed by its competitors. Column "E2" (encompassing) refers to the inverse case, i.e. it gives the relative frequencies that this model encompasses its competitors.

Fissler & Ziegel, 2016; Patton et al., 2019). We proposed three variants of the ES encompassing test that allow for flexible parametric forecast combination methods. As one test variant is potentially subject to model misspecification, we extended the existing asymptotic theory of Bayer and Dimitriadis (2020), Dimitriadis and Bayer (2019) and Patton et al. (2019) to cases of potential model misspecification for flexible parametric models. Potential future research on extending the setting presented in this paper includes encompassing tests for convex forecast combinations and forecast combinations that theoretically prevent crossings of the VaR and the ES. Both approaches require a non-standard asymptotic theory for tests on the boundary of the parameter space.

Tests for forecast encompassing establish a theoretical foundation for forecast combinations of two competing forecasts when both opposing hypotheses of forecast encompassing are rejected. This situation corresponds to a case where neither forecast encompasses its competitor. Generally, forecast combinations benefit from diversification gains stemming from combining different model specifications and underlying information sets. This benefit is especially pronounced for extreme risk measures such as the ES (Taylor, 2020), because stand-alone models are particularly sensitive to observations in the tails of the return distributions. Thus, combining forecasts can be seen as a robustification of the forecasts. Our application empirically validated this conjecture, especially in terms of the daily returns from the IBM stock.

Although we proposed encompassing tests suitable for cases where the ES is reported with and without its accompanying VaR forecast, encompassing tests for the VaR and ES together are more natural given their joint elicitability. Furthermore, this joint elicitability property clearly hints towards reporting ES forecasts jointly with their corresponding VaR forecasts by default. By contrast, the auxiliary test can be seen as the first forecast comparison procedure for the ES, insofar as it focuses almost entirely on the ES. Theoretically, it is best applied in cases where the primary focus of the researcher is on the ES, even though VaR forecasts are available. The strict ES test is specifically designed for scenarios where only competing ES forecasts are available. Furthermore, in cases where competing ES forecasts are built on the same VaR model (forecasts), the strict ES test is applicable—but the joint and auxiliary tests are infeasible, due to the collinearity of the (identical) VaR forecasts.

Acknowledgments

Appendix A. Proofs

Proof of Proposition 2.8. We check that the necessary conditions (i)–(iv) of the basic consistency theorem hold, as given in Theorem 2.1 in Newey and McFadden (1994), p. 2121. We consider the objective functions $Q_n(\theta)$ and $Q_n^0(\theta)$, as defined in (2.18) and (2.19). First, notice that condition (ii) holds by imposing condition (b) (in Assumption 2.7). The unique identification condition (i) holds by assumption (c). Next, we verify the uniform convergence condition (iv) by applying the uniform weak law of large numbers given in Theorem A.2.5. in White (1994). For that, we have to show that

1. the map $\theta \mapsto \rho\big(Y_{t+1}, g_t^q(\beta), g_t^e(\eta)\big)$ is Lipschitz-L_1 on Θ,[12]
2. For all $\theta^o \in \Theta$, there exists $\delta^o > 0$, such that for all $\delta, 0 < \delta \leq \delta^o$, the sequences

$$\bar{\rho}_t(\theta^o, \delta) \coloneqq \sup_{\theta \in \Theta} \big\{ \rho\big(Y_{t+1}, g_t^q(\beta), g_t^e(\eta)\big) \big| \|\theta - \theta^o\| < \delta \big\} \tag{A.1}$$

and

$$\underline{\rho}_t(\theta^o, \delta) \coloneqq \inf_{\theta \in \Theta} \big\{ \rho\big(Y_{t+1}, g_t^q(\beta), g_t^e(\eta)\big) \big| \|\theta - \theta^o\| < \delta \big\} \tag{A.2}$$

obey a weak law of large numbers.

Condition 1 follows directly from Lemma S.1 in the Supplementary Material and we turn to condition 2. The process Z_t is a strongly mixing process of size $-r/(r-2)$ for some $r > 2$ by condition (a), and the functions $\rho\big(Y_{t+1}, g_t^q(\beta), g_t^e(\eta)\big)$ and the supremum/infimum functions are $\mathcal{F}_t$-measurable for all $t \in \mathbb{N}$. Thus, we can conclude that the sequences $\bar{\rho}_t(\theta^o, \delta)$ and $\underline{\rho}_t(\theta^o, \delta)$ are also strong mixing sequences of the same size by applying the same theorem.

Furthermore, for $\tilde{r} > 1$ and for some sufficiently small $\delta > 0$, $r \geq \tilde{r} + \delta$ and thus $\mathbb{E}\Big[|\bar{\rho}_t(\theta^o, \delta)|^{\tilde{r}+\delta}\Big] \leq \sup_{1 \leq t \leq T} \mathbb{E}\Big[\sup_{\theta \in \Theta} \big|\rho\big(Y_{t+1}, g_t^q(\beta), g_t^e(\eta)\big)\big|^r\Big]$ for all $t, 1 \leq t \leq T, T \geq 1$. As Θ is compact, there exists some $c > 0$ such that $\sup_{\theta \in \Theta} \|\theta\| \leq c$. Thus, for all $t = 1, \ldots, T$, it holds that

$$\mathbb{E}\left[\sup_{\theta \in \Theta} \big|\rho\big(Y_{t+1}, g_t^q(\beta), g_t^e(\eta)\big)\big|^r \right] \tag{A.3}$$

$$\leq 4^{r-1}\left\{ 1 + \left(\frac{c}{K}\left(1 + \frac{1}{\alpha}\right)\right)\mathbb{E}\|g_t^q(\beta)\|^r \right.$$

$$\left. + \frac{1}{\alpha K}\mathbb{E}|Y_{t+1}|^r + \sup_{\theta \in \Theta}\mathbb{E}\| \log(g_t^e(\eta))\|^r \right\}, \tag{A.4}$$

which is bounded by condition (h) and as $\log(z) \leq z$ for a sufficiently large z. The same inequality holds for $|\underline{\rho}_t(\theta^o, \delta)|$. Thus, we can apply the weak law of large numbers for strong mixing sequences in Corollary 3.48 in White (2001), p. 49 in order to conclude that for all $\theta^o \in \Theta$ such that $\|\theta^o - \theta\| \leq \delta$, it holds that $\frac{1}{n}\sum_{t=m}^{T-1}\big(\bar{\rho}_t(\theta^o, \delta) - \mathbb{E}\left[\bar{\rho}_t(\theta^o, \delta)\right]\big) \xrightarrow{\mathbb{P}} 0$ and $\frac{1}{n}\sum_{t=m}^{T-1}\big(\underline{\rho}_t(\theta^o, \delta) - \mathbb{E}\left[\underline{\rho}_t(\theta^o, \delta)\right]\big) \xrightarrow{\mathbb{P}} 0$, which shows condition 2. Consequently, the uniform convergence condition (iv) holds by applying the uniform weak law of large numbers given in Theorem A.2.5. in White (1994).

As we have shown that the map $\theta \mapsto \rho\big(Y_{t+1}, g_t^q(\beta), g_t^e(\eta)\big)$ is Lipschitz-L_1 in Lemma S.1 in the Supplementary Material, the map $\theta \mapsto Q_n^0 = \frac{1}{n}\sum_{t=m}^{T-1}\mathbb{E}\left[\rho\big(Y_{t+1}, g_t^q(\beta), g_t^e(\eta)\big)\right]$ is also continuous, which shows condition (iii). Thus, we can apply Theorem 2.1. of Newey and McFadden (1994), which concludes the proof of this proposition. $\square$

Proof of Proposition 2.9. We define $\Psi_n(\theta) = \frac{1}{n}\sum_{t=m}^{T-1}\psi\big(Y_{t+1}, g_t^q(\beta), g_t^e(\eta)\big)$ and $\Psi_n^0(\theta) = \mathbb{E}[\Psi_n(\theta)]$. From the proof of Lemma S.2 in the Supplementary Material, we derive the mean value expansion (for $\hat{\theta}_n$ close to θ_n^*),

$$\Psi_n^0(\hat{\theta}_n) - \Psi_n^0(\theta_n^*) = \Delta_n(\tilde{\theta}_1, \ldots, \tilde{\theta}_k)(\hat{\theta}_n - \theta_n^*), \tag{A.5}$$

for (possibly different) values $\tilde{\theta}_1, \ldots, \tilde{\theta}_k$ somewhere on the line between $\hat{\theta}_n$ and θ_n^*, where the components of $\Delta_n(\tilde{\theta}_1, \ldots, \tilde{\theta}_k)$ are given in Lemma S.2 in the Supplementary Material, and where $\Psi_n^0(\theta_n^*) = 0$.[13]

Furthermore, it holds that $\Delta_n(\theta_n^*, \ldots, \theta_n^*) = \Lambda_n(\theta_n^*)$, and $\Delta_n(\tilde{\theta}_1, \ldots, \tilde{\theta}_k)$ is a continuous function in its arguments $\tilde{\theta}_1, \ldots, \tilde{\theta}_k$. Using that $\Lambda_n(\theta_n^*)$ has Eigenvalues bounded away from zero (for a sufficiently large n), we also get that $\Delta_n(\tilde{\theta}_1, \ldots, \tilde{\theta}_k)$ is non-singular in a neighborhood around θ_n^* (for all arguments) for a sufficiently large n, because the map that maps the matrix onto its Eigenvalues is continuous. As we also know that $\hat{\theta}_n - \theta_n^* \xrightarrow{\mathbb{P}} 0$ and $\|\tilde{\theta}_j - \theta_n^*\| \leq \|\hat{\theta}_n - \theta_n^*\|$ for all $j = 1, \ldots, k$, we get

---

[12] See Definition A.2.3 in White (1994) for a definition of Lipschitz-L_1. Notice that we do not have a double index. Thus, we suppress the n in the notation of White (1994). Furthermore, we apply the definition by using the identify function for a_t^o.

[13] The mean-value theorem cannot be generalized in a straightforward fashion to vector-valued functions. Thus, we have to consider the mean value expansion in each component separately, giving this more complicated expression.

from the continuous mapping theorem that

$$\Delta_n^{-1}(\tilde{\theta}_1, \ldots, \tilde{\theta}_k) - \Lambda_n^{-1}(\theta_n^*) \xrightarrow{\mathbb{P}} 0. \tag{A.6}$$

In the following, we apply Lemma A.1 in Weiss (1991) (by verifying its assumptions). Thus, we extend the iid results of Huber (1967) to strong mixing sequences. Assumption (N1) of Lemma A.1 in Weiss (1991) is satisfied as every almost surely continuous stochastic process is separable in the sense of Doob (Gikhman & Skorokhod, 2004), and the functions $\psi\left(Y_{t+1}, g_t^q(\beta), g_t^e(\eta)\right)$ are almost surely continuous for all $t \in \mathbb{N}$. Assumption (N2) is satisfied, as shown in the proof of Proposition 2.8. Assumption (N3)(i) is shown in Lemma S.2 in the Supplementary Material. The technical Assumptions (N3)(ii) and (N3)(iii) follow from Lemma 4 and Lemma 5 in the supplemental appendix of Patton et al. (2019). For this, notice that the moment conditions in Assumption 2(C) and (D) of Patton et al. (2019) are implied by the condition (h) in Assumption 2.7. Assumption (N4) follows from the moment conditions (h) in Assumption 2.7 and Assumption (N5) from the strong mixing condition (a). Furthermore, Lemma 2 in the supplemental appendix of Patton et al. (2019) implies that $\sqrt{n}\Psi_n(\hat{\theta}_n) \xrightarrow{\mathbb{P}} 0$. Thus, we can apply Lemma A.1 in Weiss (1991) and get that

$$\sqrt{n}\Psi_n^0(\hat{\theta}_n) - \sqrt{n}\Psi_n(\theta_n^*) \xrightarrow{\mathbb{P}} 0. \tag{A.7}$$

Combining (A.5), (A.6) and (A.7), we get that

$$\sqrt{n}(\hat{\theta}_n - \theta_n^*) = -\Delta_n(\tilde{\theta}_1, \ldots, \tilde{\theta}_k)^{-1}\sqrt{n}\,\Psi_n^0(\hat{\theta}_n) \tag{A.8}$$

$$= -\left(\Lambda_n^{-1}(\theta_n^*) + o_p(1)\right) \cdot \left(\sqrt{n}\Psi_n(\theta_n^*) + o_p(1)\right)$$

$$= -\Lambda_n^{-1}(\theta_n^*) \cdot \sqrt{n}\Psi_n(\theta_n^*) + o_p(1). \tag{A.9}$$

Furthermore, $\Sigma_n^{-1/2}(\theta_n^*)\sqrt{n}\Psi_n(\theta_n^*) \xrightarrow{d} \mathcal{N}(0, I_k)$ by Lemma S.3 in the Supplementary Material. Thus, $\Sigma_n^{-1/2}(\theta_n^*)\Lambda_n(\theta_n^*)$ $\sqrt{n}(\hat{\theta}_n - \theta_n^*) \xrightarrow{d} \mathcal{N}(0, I_k)$, which concludes the proof of this proposition. □

Proof of Theorem 2.10. We first notice that

$$\widehat{\Omega}_n^{-1/2}\sqrt{n}(\hat{\theta}_n - \theta_n^*) = \Omega_n^{-1/2}\sqrt{n}(\hat{\theta}_n - \theta_n^*)$$
$$+ \left(\widehat{\Omega}_n^{-1/2} - \Omega_n^{-1/2}\right)\sqrt{n}(\hat{\theta}_n - \theta_n^*). \tag{A.10}$$

From Proposition 2.9, we obtain that $\Omega_n^{-1/2}\sqrt{n}(\hat{\theta}_n - \theta_n^*) \xrightarrow{d} \mathcal{N}(0, I_k)$. Furthermore, as $\left(\widehat{\Omega}_n^{-1/2} - \Omega_n^{-1/2}\right) = o_P(1)$ by assumption, we apply Slutzky's theorem in order to get that $\left(\widehat{\Omega}_n^{-1/2} - \Omega_n^{-1/2}\right)\sqrt{n}(\hat{\theta}_n - \theta_n^*) = o_P(1)$. Thus, $\widehat{\Omega}_n^{-1/2}\sqrt{n}(\hat{\theta}_n - \theta_n^*) \xrightarrow{d} \mathcal{N}(0, I_k)$ and the result based on the selection matrices R follows by applying the continuous mapping theorem, which concludes the proof of this theorem. □

Appendix B. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ijforecast.2020.07.008.

References

Andrews, D. W. K. (1999). Estimation when a parameter is on a boundary. *Econometrica*, 67(6), 1341–1383.

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

Barendse, S. (2020). *Efficiently weighted estimation of tail and interquartile expectations, Working paper* (pp. 1–55). Available at https://drive.google.com/file/d/1nI0QAWbM_VchAZDVg79p2vJcKoCrQB8o/view.

Basel Committee (2013). *Fundamental review of the trading book: A revised market risk framework*: Technical report, Bank for International Settlements, Available at http://www.bis.org/publ/bcbs265.pdf.

Basel Committee (2016). *Minimum capital requirements for market risk*: Technical report, Bank for International Settlements, Available at http://www.bis.org/bcbs/publ/d352.pdf.

Basel Committee (2017). *Pillar 3 disclosure requirements – consolidated and enhanced framework*: Technical report, Basel Committee on Banking Supervision, Available at http://www.bis.org/bcbs/publ/d400.pdf.

Bayer, S., & Dimitriadis, T. (2020). Regression based expected shortfall backtesting. *Journal of Financial Econometrics* (in press), Available at https://arxiv.org/abs/1801.04112.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3), 307–327.

Chong, Y. Y., & Hendry, D. (1986). Econometric evaluation of linear macro-economic models. *Review of Economic Studies*, 53(4), 671–690.

Clark, T. E., & McCracken, M. W. (2001). Tests of equal forecast accuracy and encompassing for nested models. *Journal of Econometrics*, 105(1), 85–110.

Clements, M., & Harvey, D. (2009). Forecast combination and encompassing. In T. C. Mills, & K. Patterson (Eds.), *Palgrave Handbook of Econometrics: Volume 2: Applied Econometrics* (pp. 169–198). London: Palgrave Macmillan UK.

Clements, M., & Harvey, D. (2010). Forecast encompassing tests and probability forecasts. *Journal of Applied Econometrics*, 25(6), 1028–1062.

Cont, R., Deguest, R., & Scandolo, G. (2010). Robustness and sensitivity analysis of risk measurement procedures. *Quantitative Finance*, 10(6), 593–606.

Creal, D., Koopman, S. J., & Lucas, A. (2013). Generalized autoregressive score models with applications. *Journal of Applied Econometrics*, 28(5), 777–795.

Danielsson, J., Embrechts, P., Goodhart, C., Keating, C., Muennich, F., Renault, O., & Shin, H. S. (2001). *An academic response to Basel II*. Financial Markets Group, Financial Markets Group Special Papers, Available at https://EconPapers.repec.org/RePEc:fmg:fmgsps:sp130.

Diebold, F. X. (1989). Forecast combination and encompassing: Reconciling two divergent literatures. *International Journal of Forecasting*, 5(4), 589–592.

Diebold, F., & Mariano, R. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3), 253–263.

Dimitriadis, T., & Bayer, S. (2019). A joint quantile and expected shortfall regression framework. *Electronic Journal of Statistics*, 13(1), 1823–1871.

Elliott, G., Komunjer, I., & Timmermann, A. (2005). Estimation and testing of forecast rationality under flexible loss. *Review of Economic Studies*, 72(4), 1107–1125.

Embrechts, P., Liu, H., & Wang, R. (2018). Quantile-based risk sharing. *Operations Research*, 66(4), 936–949.

Engle, R., & Manganelli, S. (2004). CAViaR: Conditional autoregressive value at risk by regression quantiles. *Journal of Business & Economic Statistics*, 22(4), 367–381.

Ericsson, N. R. (1993). On the limitations of comparing mean square forecast errors: Clarifications and extensions. *Journal of Forecasting*, 12(8), 644–651.

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

Fissler, T., & Ziegel, J. F. (2019). Elicitability of range value at risk. Available at https://arxiv.org/abs/1902.04489.

Fissler, T., Ziegel, J. F., & Gneiting, T. (2016). Expected shortfall is jointly elicitable with value at risk - Implications for backtesting. *Risk*, January, 58–61.

Giacomini, R., & Komunjer, I. (2005). Evaluation and combination of conditional quantile forecasts. *Journal of Business & Economic Statistics*, 23, 416–431.

Giacomini, R., & White, H. (2006). Tests of conditional predictive ability. *Econometrica*, 74(6), 1545–1578.

Gikhman, I., & Skorokhod, A. (2004). *Classics in Mathematics*: *Vol. 210, The Theory of Stochastic Processes I*. Springer Berlin Heidelberg.

Glosten, L. R., Jagannathan, R., & Runkle, D. E. (1993). On the relation between the expected value and the volatility of the nominal excess return on stocks. *The Journal of Finance*, *48*(5), 1779–1801.

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

Halbleib, R., & Pohlmeier, W. (2012). Improving the value at risk forecasts: Theory and evidence from the financial crisis. *Journal of Economic Dynamics and Control*, *36*(8), 1212–1228.

Hall, A. R., & Inoue, A. (2003). The large sample behaviour of the generalized method of moments estimator in misspecified models. *Journal of Econometrics*, *114*(2), 361–394.

Hansen, B. E., & Lee, S. (2019). *Inference for iterated GMM under misspecification, Working paper*. Available at https://www.ssc.wisc.edu/~bhansen/papers/IteratedGMM.html.

Hansen, P. R., Lunde, A., & Nason, J. M. (2011). The model confidence set. *Econometrica*, *79*(2), 453–497.

Harvey, D., Leybourne, S., & Newbold, P. (1998). Tests for forecast encompassing. *Journal of Business & Economic Statistics*, *16*(2), 254–259.

Harvey, D., & Newbold, P. (2000). Tests for multiple forecast encompassing. *Journal of Applied Econometrics*, *15*(5), 471–482.

Heinrich, C. (2014). The mode functional is not elicitable. *Biometrika*, *101*(1), 245–251.

Hendricks, W., & Koenker, R. (1992). Hierarchical spline models for conditional quantiles and the demand for electricity. *Journal of the American Statistical Association*, *87*(417), 58–68.

Hendry, D., & Richard, J.-F. (1982). On the formulation of empirical models in dynamic econometrics. *Journal of Econometrics*, *20*(1), 3–33.

Holzmann, H., & Eulert, M. (2014). The role of the information set for forecasting—with applications to risk management. *The Annals of Applied Statistics*, *8*(1), 595–621.

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

Koenker, R. W., & Bassett, G. (1978). Regression quantiles. *Econometrica*, *46*(1), 33–50.

Mizon, G., & Richard, J.-F. (1986). The encompassing principle and its application to testing non-nested hypotheses. *Econometrica*, *54*(3), 657–678.

Newbold, P., & Harvey, D. (2007). A companion to economic forecasting, Chapter 12. In *Forecast Combination and Encompassing* (pp. 268–283). John Wiley and Sons, Ltd.

Newey, W., & McFadden, D. (1994). Large sample estimation and hypothesis testing. In R. Engle, & D. McFadden (Eds.), *Handbook of Econometrics, Vol. 4, Chapter 36* (pp. 2111–2245). Elsevier.

Nolde, N., & Ziegel, J. F. (2017). Elicitability and backtesting: Perspectives for banking regulation. *The Annals of Applied Statistics*, *11*(4), 1833–1874.

Patton, A. J. (2011). Data-based ranking of realised volatility estimators. *Journal of Econometrics*, *161*, 284–303.

Patton, A. J. (2019). Comparing possibly misspecified forecasts. *Journal of Business & Economic Statistics*, 1–14.

Patton, A. J., & Timmermann, A. (2007). Testing forecast optimality under unknown loss. *Journal of the American Statistical Association*, *102*(480), 1172–1184.

Patton, A. J., Ziegel, J. F., & Chen, R. (2019). Dynamic semiparametric models for expected shortfall (and Value-at-Risk). *Journal of Econometrics*, *211*(2), 388–413.

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

Taylor, J. W. (2020). Forecast combinations for value at risk and expected shortfall. *International Journal of Forecasting*, *36*(2), 428–441.

Timmermann, A. (2006). Forecast combinations. In G. Elliott, C. Granger, & A. Timmermann (Eds.), *Handbook of Economic Forecasting, Vol. 1, Chapter 04* (1st ed.). (pp. 135–196). Elsevier.

Weiss, A. A. (1991). Estimating nonlinear dynamic models using least absolute error estimation. *Econometric Theory*, *7*(01), 46–68.

West, K. (2006). Forecast evaluation. In G. Elliott, C. Granger, & A. Timmermann (Eds.), *Handbook of Economic Forecasting, Vol. 1, Chapter 03* (1st ed.). (pp. 99–134). Elsevier.

White, H. (1994). *EconomeTric Society Monographs, Estimation, Inference and Specification Analysis*. Cambridge University Press.

White, H. (2001). *AsymptotiC Theory for Econometricians*. San Diego: Academic Press.