Backtesting Systemic Risk Forecasts Using Multi-Objective Elicitability

Tobias Fissler and Yannick Hoga

ABSTRACT

Systemic risk measures such as CoVaR, CoES, and MES are widely-used in finance, macroeconomics and by regulatory bodies. Despite their importance, we show that they fail to be elicitable and identifiable. This renders forecast comparison and validation, commonly summarized as “backtesting,” impossible. The novel notion of multi-objective elicitability solves this problem by relying on bivariate scores equipped with the lexicographic order. Based on this concept, we propose Diebold–Mariano type tests with suitable bivariate scores to compare systemic risk forecasts. We illustrate the test decisions by an easy-to-apply traffic-light approach. Finally, we apply our traffic-light approach to DAX 30 and S&P 500 returns, and infer some recommendations for regulators.

1. Motivation

Regulating financial institutions in isolation is often not sufficient to prevent financial crises due to the interdependent risks these institutions face. In particular, their losses commonly exhibit a pronounced comonotonic behavior in the extreme tails: When one financial institution, or the market as a whole, is in distress, other institutions are much more prone to being at risk as well. The U.S. subprime mortgage crisis of 2008–2009, the European sovereign debt crisis of 2010–2011 and the Covid-19 crash of 2020 have forcefully demonstrated this fact and also the need to assess the systemic nature of risk. As a consequence of these crises, a huge strand of literature on measuring systemic risk has emerged over the last decade (Giesecke and Kim 2011; Adrian and Brunnermeier 2016; Acharya et al. 2017; Oh and Patton 2018). In this article, we revisit three influential systemic risk measures.

First, we consider Adrian and Brunnermeier’s (2016) conditional value-at-risk (CoVaR) and conditional expected shortfall (CoES) as extensions of the well-known value-at-risk (VaR) and expected shortfall (ES) to the realm of systemic risk. If $Y$ denotes the losses of interest and $X$ the losses of a reference position, $\text{CoVaR}_{\alpha|\beta}(Y|X)$ is the VaR (ES) of $Y$ at level $\alpha$, given that $X$ is “in distress.” Here, we interpret the event that $X$ is in distress as $X$ being larger or equal than its $\beta$-quantile, that is, $\{X \geq \text{VaR}_\beta(X)\}$. Finally, we consider Acharya et al’s (2017) marginal expected shortfall (MES), $\text{MES}_\beta(Y|X)$, as the conditional mean of $Y$ given $\{X \geq \text{VaR}_\beta(X)\}$. Section 2 introduces the exact definitions.

In practice, numerous models for forecasting CoVaR, CoES, and MES are available; see Girardi and Tolga Ergün (2013) and Bernardi and Catania (2019) for forecasting models for CoVaR and CoES, and Brownlees and Engle (2017) and Eckernkemper (2018) for MES models. Due to the importance of systemic risk measures, it is vital to develop statistical quality assessments of the various models’ predictive performances. It is the main aim of this article to provide such tools, which are referred to as “backtests” in finance.

Backtests have two main goals. On the one hand, one may wish to assess the absolute quality of forecasting models, also called the calibration, akin to model validation in statistics. Following the terminology of Fissler, Ziegel, and Gneiting (2016), we call such procedures “traditional backtests.” Roughly speaking, they check how well a sequence of risk measure forecasts aligns with corresponding observations of losses. Traditional backtests rely on the identifiability of the underlying risk measure, which ensures the existence of a (possibly multivariate) function $V$ that uniquely “identifies” the true report (see Definition 3.2). On the other hand, the presence of several alternative prediction models for a risk measure necessitates “comparative backtests” (Fissler, Ziegel, and Gneiting 2016) to assess their predictive accuracy relative to each other. This is akin to statistical model selection procedures. Comparative backtests exploit the elicitability of the underlying risk measure. This implies the existence of a real-valued loss (or also: scoring) function $S$, which is minimized in expectation by the optimal forecast (see Definition 3.1).

Our contributions in this article are the following: First, we show that $\text{CoVaR}_{\alpha|\beta}$, $\text{CoES}_{\alpha|\beta}$, and $\text{MES}_\beta$ are neither identifiable nor elicitable as standalone risk measures (Proposition 4.1). The practical implication of this is that neither traditional nor comparative backtests can be carried out. In particular, any regulation based solely on these systemic risk measures is pointless, because neither can the adequacy of the forecasts be determined nor can different systemic risk forecasts be sensibly compared (say to a regulatory standard model).
Second, we provide a partial remedy for this drawback by coming up with joint multivariate identification functions for \((\text{VaR}_\beta(X), \text{CoVaR}_{\alpha|\beta}(Y|X)), (\text{VaR}_\beta(X), \text{CoVaR}_{\alpha|\beta}(Y|X), \text{CoES}_{\alpha|\beta}(Y|X)), \) and \((\text{VaR}_\beta(X), \text{MES}_\beta(Y|X))\) in Theorem 3.1. (In the rest of the article, all references starting with an “S.” refer to the Supplement.) These identification functions can be used for (conditional) calibration tests in the spirit of Nolde and Ziegel (2017). To the best of our knowledge, this entails the first traditional backtest for these systemic risk measures apart from Banulescu-Radu et al. (2021). We contrast our approach with theirs in detail in Remark 3.2. In particular, they use one-dimensional identification functions for \((\text{VaR}_\beta(X), \text{CoVaR}_{\alpha|\beta}(Y|X)), \) and \((\text{VaR}_\beta(X), \text{MES}_\beta(Y|X))\), which fail to be strict in contrast to our two-dimensional identification functions. For the backtest of Banulescu-Radu et al. (2021), this non-strictness leads to a complete loss of power in identifying certain misspecified systemic risk forecasts (Supplement S.8). Theoretically, our results are akin to the fact that \(\text{ES}_\beta(Y)\) is not identifiable on its own, but the pair \((\text{VaR}_\beta(Y), \text{ES}_\beta(Y))\) is identifiable (Fissler and Ziegel 2016).

Third, in stark contrast to the joint elicitability of the pair \((\text{VaR}_\beta, \text{ES}_\beta)\), however, we show that the pairs \((\text{VaR}_\beta, \text{CoVaR}_{\alpha|\beta}), (\text{VaR}_\beta, \text{MES}_\beta)\) and the triplet \((\text{VaR}_\beta, \text{CoVaR}_{\alpha|\beta}, \text{CoES}_{\alpha|\beta}), (\text{VaR}_\beta, \text{MES}_\beta))\) fail to be elicitable (Supplement S.4). So while traditional backtests for the above pairs and the triplet may be constructed by virtue of their identifiability, classical comparative backtests exploiting elicitation are not feasible.

Our fifth contribution outlines in Section 5 how these scores can be used for comparative backtests of Diebold–Mariano type. These comparative backtests are different from—in our case infeasible—“standard” comparative backtests in that they build on the newly introduced notion of multi-objective elicitability (with scores mapping to \(\mathbb{R}^2\) equipped with the lexicographic order \(\preceq_{\text{lex}}\)). This contrasts sharply with classical \(\mathbb{R}\)-valued losses \(S\). Their prevalence to date is grounded in tradition (Gneiting 2011) and the fact that \(\mathbb{R}\) is equipped with the canonical (total) order relation \(\preceq\), which allows for straightforward comparisons of losses. We show that \((\text{VaR}_\beta, \text{CoVaR}_{\alpha|\beta}), (\text{VaR}_\beta, \text{CoVaR}_{\alpha|\beta}, \text{CoES}_{\alpha|\beta}), \) and \((\text{VaR}_\beta, \text{MES}_\beta))\) are multi-objective elicitable with respect to \(\preceq_{\text{lex}}\). Theorem 4.2 provides rich classes of strictly multi-objective consistent scores.

The article closes with a discussion and outlook (Section 7). A summary of the supplementary material is given at the end of the article. The R code to reproduce all numerical experiments is also available online. R implementations of the comparative backtests developed in this paper are also available in the SystemicRisk package of Dimitriadis and Hoga (2023b).

### 2. Formal Definition of CoVaR, CoES, and MES

Throughout the article, we indicate vectors with bold letters. We highlight the distinction between row and column vectors only when it is essential, and use the symbol ‘t’ to indicate the transpose of a vector or matrix.

Fix some nonatomic probability space \((\Omega, \mathcal{F}, P)\) where all random objects are defined. Using standard notation, let \(L^0(\mathbb{R}^d)\), \(d = 1, 2\), be the space of all \(\mathbb{R}^d\)-valued random vectors on \((\Omega, \mathcal{F}, P)\). Furthermore, for \(p \in [1, \infty)\), let \(L^p(\mathbb{R}^d) \subseteq L^0(\mathbb{R}^d)\) be the collection of random vectors whose components possess a finite \(p\)th moment. For \(X \in L^0(\mathbb{R}^d)\) let \(F_X\) be its joint distribution function. Then define for \(p \in [0, 1]\) the collection \(\mathcal{F}_p(\mathbb{R}^d) := \{F_X : X \in L^p(\mathbb{R}^d)\}\). The class of distribution functions \(F \in \mathcal{F}(\mathbb{R}^d)\) with a strictly positive Lebesgue density for all values \(x \in \mathbb{R}^d\) such that \(F(x) \in (0, 1)\) is denoted by \(\mathcal{F}_1(\mathbb{R}^d)\). Similarly, we write \(X \in L^p(\mathbb{R}^d)\) if \(F_X \in \mathcal{F}_p(\mathbb{R}^d)\) for \(p \in [0, 1]\).

Our systemic risk measures of interest—CoVaR, CoES, and MES—are maps from \(L^0(\mathbb{R}^2)\) (or \(L^1(\mathbb{R}^2)\) for MES) to \(\mathbb{R}^2 := (\infty, \infty]\). They are law-determined, meaning that their values for \((X, Y)\) and \((\tilde{X}, \tilde{Y})\) coincide if \(F_{X,Y} = F_{\tilde{X},\tilde{Y}}\). Hence, we
can consider them as risk-functionals on $\mathcal{F}^0(\mathbb{R}^2)$ (or $\mathcal{F}^1(\mathbb{R}^2)$). Similarly, the popular univariate risk measure VaR$_{\beta}$, $\beta \in [0,1]$, is a law-determined map $L^0(\mathbb{R}) \rightarrow [-\infty, \infty]$. In the rest of the article, we will frequently overload notation and identify these law-determined risk measures with their induced risk functionals. As such, we will use the terms “risk measure” and “risk functional” interchangeably.

Let $(X,Y) \in L^0(\mathbb{R}^2)$ be a two-dimensional random vector. Here, $Y$ stands for the losses of a position of interest (with the sign convention that positive values are losses and negative values are gains) and $X$ is a univariate reference position or aggregate of a reference system, having the same sign convention. Denote by $F_{X,Y}$ their joint distribution function, and by $F_X$ and $F_Y$ the respective marginals. To ease the exposition, we assume that $F_{X,Y} \in \mathcal{F}^0(\mathbb{R}^2)$, but we mention that most of the results can be generalized beyond this assumption; see Supplement S.2.

Recall that for $\beta \in [0,1]$ and $F_X \in \mathcal{F}^0(\mathbb{R})$, $\text{VaR}_{\beta}(F_X) := \text{VaR}_{\beta}(X) = \inf\{x \in \mathbb{R}: F_X(x) \geq \beta\}$, which is the $\beta$-quantile of $F_X$. For $\beta = (0,1)$ $\text{VaR}_{\beta}(X)$ is always finite. For $\beta \in [0,1]$ we define the expected shortfall as $\text{ES}_\beta(X) := \text{ES}_\beta(F_X) := 1/(1-\beta) \int_0^1 \text{VaR}_\gamma(X) d\gamma$. Our sign convention is such that the larger the risk measure of a position, the riskier it is deemed. Hence, we typically choose a probability level of $\beta$ close to 1 for $\text{VaR}_\beta$ and $\text{ES}_\beta$, such as $\beta = 0.95$ or $\beta = 0.99$.

The original definition of CoVaR is due to Adrian and Brunnermeier (2016). Here, we follow Girardi and Tolga Ergün (2013) and Nolde and Zhang (2020) in generalizing their definition to $\text{CoVaR}_{\alpha|\beta}: L^0(\mathbb{R}^2) \rightarrow \mathbb{R}$ for $\alpha \in (0,1)$, $\beta \in [0,1]$ via

$$\text{CoVaR}_{\alpha|\beta}(Y|X) := \text{CoVaR}_{\alpha|\beta}(F_{X,Y}) := \text{VaR}_\alpha(F_{Y|X \geq \text{VaR}_{\beta}(X)}),$$

where $F_{Y|X \geq \text{VaR}_{\beta}(X)} = \mathbb{P}[Y \leq \cdot | X \geq \text{VaR}_{\beta}(X)]$. For $\beta = 0$, $\text{CoVaR}_{0|\beta}(Y|X) := \text{VaR}_\alpha(Y)$, and if $\beta = \alpha$ we simply write $\text{CoVaR}_\alpha(Y|X) := \text{CoVaR}_{\alpha|\alpha}(Y|X)$.

Since $\text{CoVaR}_{\alpha|\beta}(Y|X)$ is the $\alpha$-quantile of the distribution $F_{Y|X \geq \text{VaR}_{\beta}(X)}$ and, as such, is blind against losses exceeding $\text{CoVaR}_{\alpha|\beta}(Y|X)$, Adrian and Brunnermeier (2016) also consider the conditional Expected Shortfall (CoES) as a more conservative measure on the bivariate tail. As for CoVaR, we consider the following generalized definition. For $\alpha \in (0,1)$, $\beta \in [0,1]$, define $\text{CoES}_{\alpha|\beta}: L^0(\mathbb{R}^2) \rightarrow \mathbb{R}$ via

$$\text{CoES}_{\alpha|\beta}(Y|X) := \text{CoES}_{\alpha|\beta}(F_{X,Y}) \equiv \frac{1}{1-\alpha} \int_0^1 \text{CoVaR}_{\gamma|\beta}(Y|X) d\gamma. \quad (2.2)$$

The assumed continuity of $F_{X,Y}$ implies that $\text{CoES}_{\alpha|\beta}(Y|X)$ coincides with the conditional expectation $\mathbb{E}[Y|Y \geq \text{CoVaR}_{\alpha|\beta}(Y|X), X \geq \text{VaR}_{\beta}(X)]$. Again, $\text{CoES}_{\alpha|\beta}(Y|X) = \text{ES}_\alpha(Y)$, and we write $\text{CoES}_{\alpha|\beta}(Y|X) := \text{CoES}_{\alpha|\alpha}(Y|X)$.

Finally, we consider the marginal Expected Shortfall (MES) of Acharya et al. (2017), which measures the expectation of $Y$ when $X$ is in distress, that is, when $X$ is in its right tail. Specifically, we introduce for $\beta \in [0,1)$ the map $\text{MES}_\beta: L^0(\mathbb{R}^2) \rightarrow \mathbb{R}$,

$$\text{MES}_\beta(Y|X) := \text{MES}_\beta(F_{X,Y}) := \text{CoES}_{\beta|\beta}(Y|X) = \int_0^1 \text{CoVaR}_{\gamma|\beta}(Y|X) d\gamma. \quad (2.3)$$

Again, $\text{MES}_0(Y|X) = \mathbb{E}[Y]$, and $\text{MES}_\beta(Y|X) = \mathbb{E}[Y|X \geq \text{VaR}_{\beta}(X)]$ since $F_{X,Y}$ is continuous.

## 3. (Conditional) Identifiability and Multi-Objective Elicitability

### 3.1. Notation, Basic Definitions and Results

Adopting the decision-theoretic terminology of Gneiting (2011) and Fissler and Ziegel (2016), we denote by $\mathcal{A} \subseteq \mathbb{R}^d$ the “action domain,” the space of all possible forecasts. The observation domain is $\mathbb{R}^d$. We consider a functional $T$ mapping from some class of distributions $\mathcal{F} \subseteq \mathcal{F}^0(\mathbb{R}^d)$ to $\mathcal{A}$. To simplify the exposition, we only consider point-valued functionals.

A function $a : \mathbb{R}^d \rightarrow \mathbb{R}$ is called $\mathcal{F}$-integrable if $\int [a(y)] dF(y) < \infty$ for all $F \in \mathcal{F}$. If $a$ is $\mathcal{F}$-integrable, we define the map $\bar{a} : \mathcal{F} \rightarrow \mathbb{R}$, $\bar{a}(F) := \int [a(y)] dF(y)$. Similarly, a function $g : \mathbb{A} 	imes \mathbb{R}^d \rightarrow \mathbb{R}$, is called $\mathcal{F}$-integrable if $g(r, \cdot)$ is $\mathcal{F}$-integrable for all $r \in \mathbb{A}$. For multivariate functions $a$ and $g$, similar conventions and notation apply, which should be understood componentwise. We start with the classical definition of elicitation and consistent scoring functions mapping to $\mathbb{R}$.

**Definition 3.1.** An $\mathcal{F}$-integrable map $S : \mathbb{A} \times \mathbb{R}^d \rightarrow \mathbb{R}$ is an $\mathcal{F}$-consistent scoring function for $T$ if $\bar{S}(T(F), F) \leq \bar{S}(r, F)$ for all $r \in \mathbb{A}$ and for all $F \in \mathcal{F}$. It is a strictly $\mathcal{F}$-consistent scoring function for $T$ if, additionally, equality only holds for $r = T(F)$. $T$ is elicitable on $\mathcal{F}$ if there is a strictly $\mathcal{F}$-consistent scoring function for $T$.

**Definition 3.2.** An $\mathcal{F}$-integrable map $V : \mathbb{A} \times \mathbb{R}^d \rightarrow \mathbb{R}$ is an $\mathcal{F}$-identification function for $T$ if $\bar{V}(T(F), F) = 0$ for all $F \in \mathcal{F}$. It is a strict $\mathcal{F}$-identification function for $T$ if, additionally, for all $F \in \mathcal{F}$ and for all $r \in \mathbb{A}$, $V(r, F) = 0$ implies that $r = T(F)$. $T$ is identifiable on $\mathcal{F}$ if there is a strict $\mathcal{F}$-identification function for $T$.

Suppose in a risk management context that $T$ corresponds to the $\text{VaR}_\beta$-functional and $y_t$ are the observed losses. Subject to mild conditions on $\mathcal{F}$, $\text{VaR}_\beta$ is elicitable, where the “pinball loss” $S(r,y) = \mathbb{I}[y > r] - 1 + \beta(r-y)$ is strictly $\mathcal{F}$-consistent. This allows to compare competing VaR forecasts $(r_{(t,1)}), t=1, \ldots, n$ and $(r_{(t,2)}), t=1, \ldots, n$ via their empirical average score differences $\bar{d}_n = \bar{S}_{1n} - \bar{S}_{2n} = \frac{1}{n} \sum_{t=1}^n S(r_{(t,1)}, y_t) - S(r_{(t,2)}, y_t)$. A negative (positive) sign of $\bar{d}_n$ indicates superiority (inferiority) of $(r_{(t,1)}), t=1, \ldots, n$ compared to $(r_{(t,2)}), t=1, \ldots, n$. Identifiability, on the other hand, opens the way to test for calibration by checking, for example, if the test statistic $\frac{1}{n} \sum_{t=1}^n V(r_{(t,0)}, y_t)$ is sufficiently close to $0$ or not. For example, when $T$ corresponds to $\text{VaR}_\beta$ checking calibration amounts to checking if the empirical VaR-violation rate is roughly $1-\beta$, which can be done in terms of $V(r, y) = \mathbb{I}[y > r] - (1-\beta)$. These examples demonstrate the importance of elicitation and identifiability for comparing and evaluating (risk) forecasts in practice.

Under regularity conditions, the notions of elicitation and identifiability are equivalent for point-valued functionals mapping to $\mathbb{R}$ (Steinwart et al. 2014). There are also important functionals which fail to be elicitable and identifiable, most prominently the variance and expected
shortfall (Gneiting 2011). In such situations, the notions of conditional elicitability and conditional identifiability can be helpful. Following the concept presented in Fissler and Ziegel (2016, sec. 6), we slightly adapt Emmer, Kratz, and Tasche’s (2015) original definition of conditional elicitability. We also introduce the corresponding counterpart of conditional identifiability.

**Definition 3.3.** Consider two functionals $T_j : F \to A_j$, $j = 1, 2$.

(i) $T_2$ is conditionally elicitable with $T_1$ on $F$, if $T_1$ is elicitable on $F$ and $T_2$ is elicitable on $F_{r_1} := \{ F \in F : r_1 = T_1(F) \}$ for any $r_1 \in A_1$.

(ii) $T_2$ is conditionally identifiable with $T_1$ on $F$, if $T_1$ is identifiable on $F$ and $T_2$ is identifiable on $F_{r_1} := \{ F \in F : r_1 = T_1(F) \}$ for any $r_1 \in A_1$.

It is easy to see that the variance is conditionally elicitable and conditionally identifiable with the mean, and that $ES_\beta$ is conditionally elicitable and conditionally identifiable with VaR$_\beta$ on appropriate classes of distributions, respectively (Emmer, Kratz, and Tasche 2015). The pairs (mean, variance) and (VaR$_\beta$, ES$_\beta$) even turn out to be elicitable and identifiable (Fissler and Ziegel 2016). For identifiability, this is an instance of the following proposition, which is already stated without a proof in the discussion of Fissler and Ziegel (2016).

**Proposition 3.4.** If $T_2$ is conditionally identifiable with $T_1$ on $F$, then the pair $(T_1, T_2)$ is identifiable on $F$.

**Remark 3.5.** The reverse implication of Proposition 3.4 does not hold. To see this, let $A_1 = A_2 = \mathbb{O} = \mathbb{R}$ and $F = F^2(\mathbb{R})$. Let $T_1'$ and $T_2'$ be the first and second moment. Then $(T_1, T_2) = (T_1' + T_2', T_1' - T_2')$ is elicitable and identifiable, invoking the revelation principle (Gneiting 2011, Theorem 4). But on their own, $T_1$ and $T_2$ are neither elicitable nor identifiable.

Of course, an analogue to Proposition 3.4 for (conditional) elicitation would be desirable, and it has been stated as an open problem in the discussion of Fissler and Ziegel (2016). Unfortunately, the answer is negative: While Supplement S.2 establishes the conditional elicitability of CoVaR$_{a|\beta}$, (CoVaR$_{a|\beta}$, CoES$_{\beta|a}$), and MES$_{a|\beta}$ all with VaR$_\beta$, the corresponding pairs and the triplet generally fail to be elicitable; see Supplement S.4.

### 3.2. Multi-Objective Elicitability with Respect to the Lexicographic Order and Conditional Elicitability

To overcome the structural drawback of elicitability in comparison to identifiability, in particular the lack of an analogue to Proposition 3.4, we introduce the novel notion of multi-objective scoring functions and the corresponding concepts of multi-objective consistency and multi-objective elicitability. It is inspired by the fundamental observation that identification functions are generally multivariate.

To be more precise, the dimension $m$ of the identification function $V$ usually coincides with the dimension $k$ of the functional. If $k = m = 1$, an identification function is often induced by the derivative of a consistent scoring function $S$. Also, the antiderivative of an (oriented) identification function yields a consistent score, thus, roughly establishing a one-to-one correspondence between the class of identification functions and the class of consistent scoring functions. For a $k$-dimensional functional, the gradient of a consistent score is $\mathbb{R}^k$-valued and naturally induces an identification function, subject to smoothness conditions. However, not every $k$-dimensional identification function possesses an antiderivative for $k \geq 2$. This is due to integrability conditions asserting that if it was integrable, the corresponding Hessian of the stipulated antiderivative would need to be symmetric (see also Example S.4.1 for an illustration). This rules out the one-to-one relation in the multivariate setting, giving rise to a gap between the class of consistent scoring functions and the one of identification functions.

We sidestep this integrability constraint by introducing the concept of multivariate scoring functions. Indeed, it is this multivariate structure of identification functions which facilitates the straightforward proof of Proposition 3.4. Therefore, we mimic this multi-dimensionality for scores, letting them map to $\mathbb{R}^2$. We equip $\mathbb{R}^2$ with the lexicographic order $\preceq_{\text{lex}}$. Recall that $(x_1, x_2) \preceq_{\text{lex}} (y_1, y_2)$ if $x_1 < y_1$ or if $(x_1 = y_1$ and $x_2 \leq y_2$).

**Definition 3.6.** An $F$-integrable map $S : A \times \mathbb{R}^d \to \mathbb{R}^2$ is a multi-objective $F$-consistent scoring function for $T$ with respect to the lexicographic order $\preceq_{\text{lex}}$ if $S(T(F), F) \preceq_{\text{lex}} S(r, F)$ for all $r \in A$ and for all $F \in F$. It is a strictly multi-objective $F$-consistent scoring function for $T$ with respect to $\preceq_{\text{lex}}$ if, additionally, equality only holds for $r = T(F)$.

Multi-objective elicitability with respect to the lexicographic order opens the way to the following analogue of Proposition 3.4.

**Theorem 3.7.** If $T_2$ is conditionally elicitable with $T_1$ on $F$, then the pair $(T_1, T_2)$ is multi-objective elicitable on $F$ with respect to the lexicographic order $\preceq_{\text{lex}}$.

Since elicitability in the usual sense obviously induces multi-objective elicitability with respect to $(\mathbb{R}^2, \preceq_{\text{lex}})$, Remark 3.5 also shows that the reverse implication of Theorem 3.7 does not hold.

**Remark 3.8.** The use of the lexicographic order on $\mathbb{R}^2$ allows to compare any forecasts, also misspecified ones. Note that here and in the rest of the article, the term "misspecified" refers to forecasts that are not calibrated due to (e.g.,) model misspecification and/or estimation error. As for the classical univariate concept of strict consistency, strict multi-objective consistency stays silent about the ranking of possibly misspecified forecasts, which is, however, the more realistic scenario (Patton 2020). For univariate functionals, consistency implies order-sensitivity under mild conditions (Bellini and Bignozzi 2015; Lambert 2019): If two forecasts are both smaller or larger than the true functional value, the one closer to the true value...
achieves an expected score at most as large as the other forecast. For multivariate forecasts, there are various generalizations of order-sensitivity (see, e.g., Lambert, Pennock, and Shoham 2008; Fissler and Ziegel 2019). For multi-objective scores similar order-sensitivity results would be desirable. We suspect that the componentwise order-sensitivity concept would be particularly promising in that regard.

**Remark 3.9.** Holzmann and Eulert (2014) establish that consistent scoring functions are sensitive with respect to increasing information sets. That is, when comparing two ideal forecasts based on nested information sets, the more informed forecast outperforms the less informed one on average. Using the same arguments as in Holzmann and Eulert (2014) (basically exploiting the tower property of conditional expectations), one can establish a similar principle for multi-objective consistent scores mapping to \((\mathbb{R}^2, \preceq_{\text{lex}})\).

The proof of Theorem 3.7 explicitly exploits the asymmetric structure of the lexicographic order which fits well with the asymmetric notion of conditionalelicability, where the roles of \(T_1\) and \(T_2\) may not be interchanged. In particular, in the setup of Theorem 3.7, the pair \((T_2, T_1)\) is generally not multi-objective elicitable on \(\mathcal{F}\) with respect to the lexicographic order \(\preceq_{\text{lex}}\) on \(\mathbb{R}^2\). We illustrate how the construction of Theorem 3.7 leads to strictly consistent multi-objective scores mapping to \((\mathbb{R}^2, \preceq_{\text{lex}})\) for the cases of (mean, variance) and \((\text{VaR}_\beta, \text{ES}_\beta)\) in Supplement S.1.

We end this section by showing that the convex level sets (CxLS) property of a functional, which is known to be necessary for elicitability (Osband 1985; Gneiting 2011), is also necessary for identifiability and for multi-objective elicibility with respect to \(\preceq_{\text{lex}}\). We say that a functional \(T: \mathcal{F} \rightarrow A\) satisfies the CxLS property on \(\mathcal{F}\) if for any \(F_0, F_1 \in \mathcal{F}\) such that \(T(F_0) = T(F_1) \iff t\) it holds that \(T(F_\lambda) = t\) for all \(\lambda \in (0, 1)\) with \(F_\lambda := (1 - \lambda)F_0 + \lambda F_1 \in \mathcal{F}\).

**Proposition 3.10.** If \(T: \mathcal{F} \rightarrow A\) is identifiable on \(\mathcal{F}\) or multi-objective elicitable on \(\mathcal{F}\) with respect to \(\preceq_{\text{lex}}\), it satisfies the CxLS property on \(\mathcal{F}\).

We extensively use the results of Section 3, summarized in Figure 1, to prove our structural results for the systemic risk measures in the next section.

### 4. Structural Results for CoVaR, CoES, and MES

#### 4.1. CoVaR, CoES, and MES Fail to be Identifiable or Elicitable

The following proposition shows that the three systemic risk measures \(\text{CoVaR}_{\alpha\beta}, \text{CoES}_{\alpha\beta},\) and \(\text{MES}_\beta\) generally fail to be identifiable or elicitable on sufficiently rich classes of bivariate distributions \(\mathcal{F} \subseteq \mathcal{F}^0(\mathbb{R}^2)\). It is proven by showing that these functionals do not satisfy the CxLS property, exploiting Proposition 3.10.

**Proposition 4.1.** For \(\alpha, \beta \in (0, 1)\), \(\text{CoVaR}_{\alpha\beta}, \text{CoES}_{\alpha\beta},\) and \(\text{MES}_\beta\) are neither identifiable nor elicitable on any class \(\mathcal{F} \subseteq \mathcal{F}^0(\mathbb{R}^2)\) containing all bivariate normal distributions along with their finite mixtures.

**Proposition 4.1** casts doubt on traditional and comparative backtesting approaches for CoVaR, CoES, and MES as standalone systemic risk measures. Thus, these measures should not be used for regulatory purposes on their own, because forecasts for them can neither be verified for their adequacy nor can they be sensibly compared to improve their modeling.

#### 4.2. Joint Identifiability Results

Supplement S.2 establishes the conditional identifiability and conditional elicibility of \(\text{CoVaR}_{\alpha\beta}(Y|X), \text{CoES}_{\alpha\beta}(Y|X),\) and \(\text{MES}_\beta(Y|X)\) with \(\text{VaR}_\beta(X)\), respectively. This, in combination with Proposition 3.4, immediately yields the joint identifiability results of Theorem S.3.1.

Following Nolde and Ziegel (2017), the strict identification functions \(V\) of Theorem S.3.1 can readily be used to assess the absolute forecast quality via joint (Wald-type) calibration tests. These either test the null hypothesis of unconditional calibration, \(E[V(r_t, (X_t, Y_t))] = 0\) for all \(t \in \mathbb{N}\), or the more informative null of conditional calibration, \(E[V(r_t, (X_t, Y_t)) | \mathcal{F}_{t-1}] = 0\) for all \(t \in \mathbb{N}\). Here, the \(\sigma\)-algebra \(\mathcal{F}_{t-1}\) represents the information available to the forecaster at time \(t - 1\). Details are in Supplement S.3.

There, in Remark S.3.2, we also discuss the potential perils of testing calibration with nonstrict identification functions. Banulescu-Radu et al. (2021) recently pioneered calibration tests for systemic risk forecasts by relying on a (real-valued) identification function which is, however, not strict. This implies

---

**Figure 1.** Illustration of the most important structural results for real-valued (risk) functionals \(T_1\) and \(T_2\). The equivalence 1) follows from Steinwart et al. (2014) under some regularity conditions.
that their identification function equals zero in expectation not only for the correct forecasts but also for certain misspecified forecasts, leading to a loss of power in identifying such misspecifications; see Supplement S.8 for some simulation evidence.

We mention that our strict identification functions may not only be used for one-shot calibration testing, but also for online monitoring of calibration, where adequacy of the systemic risk forecasts is assessed daily as new observations become available. Such monitoring schemes could be constructed along the same lines as the VaR and ES surveillance procedures of Hoga and Demetrescu (2023).

### 4.3. Multi-Objective Elicitability Results

Recall that VaRβ(Y), (VaRβ(Y), ESB(Y)), and E(Y) are all elicitable. Surprisingly, their conditional counterparts (VaRβ(X), CoVaRββ(Y|X)), (VaRβ(X), CoVaRββ(Y|X), CoESββ(Y|X)), and (VaRβ(X), MESβ(Y|X)) fail to be elicitable despite being identifiable (Supplement S.4). This is due to integrability conditions, causing an extreme gap between the class of strict identification functions and the class of strictly consistent scoring functions, which turns out to be empty. Thus, comparative backtests cannot be implemented using a scalar strictly consistent scoring function. Furthermore, the conditional elicitability results of Supplement S.2 can hardly be used for forecast comparisons, unless the VaRβ(X) forecasts are the same and correctly specified. However, the conditional elicitability in combination with Theorem 3.7 immediately yields the following novel joint multi-objective elicitability results with respect to the lexicographic order ≤lex on R². These results can readily be used for comparing systemic risk forecasts as detailed in Section 5.

**Theorem 4.2.** Let α, β ∈ (0, 1).

(i) On $F \subseteq F^0_ε(R^2)$, the score $S^\text{VaR}$: $R \times R^2 \rightarrow R$, $S^\text{VaR}(v, (x, y)) = \left\{ \begin{array}{ll} (1 - h(\nu)) & \text{if } x \leq \nu \\ h(x) + a^\text{VaR}(x) & \text{if } x > \nu \end{array} \right.$ is strictly $F$-consistent for $F \ni F_{XY} \mapsto \text{VaR}_β(F_X)$, if $h: R \rightarrow R$ is strictly increasing and the function $(x, y) \mapsto a^\text{VaR}(x) - I[x \leq \nu]h(x)$ is $F$-integrable for all $v \in R$.

(ii) On $F \subseteq F^0_ε(R^2)$, the pair $F \ni F_{XY} \mapsto (\text{VaR}_β(F_X), \text{CoVaR}_ββ(F_{XY}), \text{CoES}_ββ(F_{XY}))$ is multi-objective elicitable with respect to $R^2$, $≤lex$. A strictly $F$-consistent multi-objective scoring function $S^\text{VaR,CoVaR}: R^2 \times R^2 \rightarrow (R^2, ≤lex)$ is given by

\[
S^\text{VaR,CoVaR}(v, (x, y)) = \begin{pmatrix} S^\text{VaR}(v, (x, y)) \\ S^\text{CoVaR}(v, (x, y)) \end{pmatrix},
\]

\[
S^\text{CoVaR}(v, (x, y)) = \left\{ \begin{array}{ll} 1 \{y \leq \nu\} - a^\text{CoVaR}(x) & \text{if } x \leq \nu \\ a^\text{CoVaR}(x) & \text{if } x > \nu \end{array} \right.,
\]

where $\text{VaR}_β(Y), \text{ES}_β(Y), \text{VaR}_β(Y)$ from Fissler and Ziegel (2016) is the also the case that for a misspecified VaR-forecast, the expected score is generally not minimized by the correctly specified ES-forecast.

**Example 4.4.** The theoretical reason why CoVaR forecasts can only be compared when VaR forecasts are comparable lies in the
To that end, denote by $S = (S_1, S_2)$ one of the multi-objective scores of Theorem 4.2. Let $\{r_{i(1)}\}_{i=1}^n$ and $\{r_{i(2)}\}_{i=1}^n$ be the appertaining competing sequences of forecasts (e.g., if $S = S^{(\text{VAR,CoVaR})}$, then $r_{i(1)} = (\text{VAR}_t, \text{CoVar}_t)$ for $i = 1, 2$). The verifying observations are $(X_t, Y_t)_{t=1}^n$. We compare the two forecasts via the (bivariate) score differences $d_i := (d_{1i}, d_{2i})' := S(r_{i(1)}, (X_t, Y_t)) - S(r_{i(2)}, (X_t, Y_t))$. The two-sided null hypothesis is that both forecasts predict equally well on average, that is, $H_0^n := \mathbb{E}[\tilde{d}_n] = 0$ for all $n = 1, 2, \ldots$ where $\tilde{d}_n := (\tilde{d}_{1n}, \tilde{d}_{2n})' := \frac{1}{n} \sum_{i=1}^n d_i$. (Along the lines of Giacomini and White (2006), one can also test the conditional null hypothesis $H_0^n := \mathbb{E}[d_i | \tilde{y}_{i-1}] = 0$ for all $t = 1, 2, \ldots$ where the $\sigma$-algebra $\tilde{y}_{i-1}$ contains information available at time $t-1$.)

We test $H_0^n$ using the Wald-type test statistic

$$T_n = n \hat{\alpha}_n^{-1} \hat{\sigma}_n,$$

where $\hat{\sigma}_n$ is some consistent estimator of the variance-covariance matrix $\sigma_n = \text{var}(\sqrt{n} \hat{d}_n)$ under the null hypothesis (i.e., in the componentwise norm, $\mathbb{E}[\hat{\sigma}_n - \sigma_n] \rightarrow^p 0$, as $n \rightarrow \infty$). To account for possible autocorrelation in the sequence $\{d_i\}_{i=1}^n$, one can use

$$\hat{\sigma}_n = \left( \hat{\sigma}_{11n}, \hat{\sigma}_{12n}, \hat{\sigma}_{21n}, \hat{\sigma}_{22n} \right) = \frac{1}{n} \sum_{i=1}^n (d_i - \bar{d}_n)(d_i - \bar{d}_n)' + \frac{1}{n} \sum_{h=1}^{\lfloor n/2 \rfloor} w_{nh} \sum_{t=h+1}^n \left[ (d_i - \bar{d}_n)(d_{i-t+h} - \bar{d}_n)' + (d_{i-t+h} - \bar{d}_n)(d_i - \bar{d}_n)' \right],$$

where $m_n \rightarrow \infty$ is a sequence of integers satisfying $m_n = o(n^{1/4})$, and $w_{nh}$ is a uniformly bounded scalar triangular array with $w_{nh} \rightarrow 1$, as $n \rightarrow \infty$, for all $h = 1, \ldots, m_n$ see White (2001) for details. Under the assumption that $\{d_i\}_{i=1}^n$ does not exhibit autocorrelation under the null, $m_n$ can be set to 0 such that $\hat{\sigma}_n$ is simply the sample variance of the score differences.

**Theorem 5.1.** Suppose that $\Omega_n \rightarrow \Omega$, as $n \rightarrow \infty$, where $\Omega$ is positive definite. Then, under technical Assumption S.5.3 (see Supplement S.5), it holds under $H_0^n$ that

$$\sqrt{n} \hat{\Omega}_n^{-1/2} \hat{d}_n \rightarrow^d N(0, I_{2\times 2}),$$

as $n \rightarrow \infty$, where $I_{2\times 2}$ denotes the $(2 \times 2)$-identity matrix. In particular, $T_n \rightarrow^d \chi^2_2$, as $n \rightarrow \infty$, where $\chi^2_2$ denotes a $\chi^2$-distribution with 2 degrees of freedom.

Thus, we reject $H_0^n$ at significance level $\nu$, if $T_n > \chi^2_{2,1-\nu}$, where $\chi^2_{2,1-\nu}$ is the $(1 - \nu)$-quantile of the $\chi^2_2$-distribution.

A typical nonrejection region in terms of $\tilde{d}_{1n}$ and $\tilde{d}_{2n}$ is sketched in Figure 2(a), and has the well-known ellipse shape. For brevity, we defer a formal investigation of our test under local alternatives to Supplement S.6.

One of the more restrictive assumptions of Theorem 5.1 is that the serial dependence in the loss differences is restricted to satisfy a certain mixing condition; see item B2 of Assumption S.5.3. This can be justified when limited-memory estimators (i.e., estimators based on at most a finite number of observations) are used in computing the forecasts, as pointed out by Giacomini and White (2006). While a weakening of this assumption seems desirable, it is beyond the scope of the present article.
5.2. One and a Half-Sided Tests

In several contexts, it may be desirable to perform a one-sided comparative backtest to establish the superiority of risk forecasts \( \{r_{t,1}\}_{t=1,...,n} \) over some benchmark forecasts \( \{r_{t,2}\}_{t=1,...,n} \). These different forecasts could stem from different risk models of a financial institution (say, a legacy model and an extension thereof). It could also be the case that the benchmark forecasts originate from a regulatory standard model, and—in line with the conservative backtesting approach of Fissler, Ziegel, and Gneiting (2016)—the financial institution has the onus of proof to show the superiority of its internal model over the standard model.

In such situations, it is tempting to test the null hypothesis \( E[\tilde{d}_n] \not\preceq_{\text{lex}} 0 \), which is equivalent to

\[
E[\tilde{d}_n] < 0 \quad \text{or} \quad \left( E[\tilde{d}_n] = 0 \text{ and } E[\tilde{d}_{2n}] \leq 0 \right). \tag{5.3}
\]

Here, the goal would be to reject the null of better or, at least, equally good benchmark forecasts as evidence of the superiority of \( \{r_{t,1}\}_{t=1,...,n} \). However, as under standard conditions (see Theorem 5.1), \( \sqrt{n}[\tilde{d}_n, \tilde{d}_{2n}] \) has an asymptotic bivariate normal distribution, the probability that \( \tilde{d}_n = 0 \) and \( \tilde{d}_{2n} \leq 0 \) vanishes for large sample sizes. Thus, testing (5.3) amounts to a test of the null \( E[\tilde{d}_n] < 0 \), that is, that the benchmark VaR forecasts are superior. This null can be tested via \( T_{1n} := \frac{\sqrt{n}[\tilde{d}_n, \tilde{d}_{2n}]}{\tilde{\sigma}_n} \), where we reject (5.3) at significance level \( \nu \in (0,1) \) when \( T_{1n} > \Phi^{-1}(1 - \nu) \). The corresponding nonrejection region is sketched in Figure 2(b). In particular, we would reject \( E[\tilde{d}_n] \not\preceq_{\text{lex}} 0 \) solely based on the predictive performance of the VaR(X_t) component. In other words, once \( E[\tilde{d}_n] < 0 \) is rejected such that the internal VaR forecasts are superior, the internal forecasts \( r_{t,1} \) are preferable in the lexicographic order irrespective of the quality of the systemic risk component. Since this would entirely ignore the motivation of the backtest, we disregard this one-sided backtesting approach.

As a compromise, we suggest to test the following one and a half-sided null hypothesis. Since the two systemic risk forecasts only play a role for the ranking in the lexicographic order when \( E[\tilde{d}_n] = 0 \), our suggested null hypothesis takes the form

\[
H_0^{\text{lex}}: E[\tilde{d}_n] = 0 \quad \text{and} \quad E[\tilde{d}_{2n}] \leq 0 \quad \text{for all } n = 1, 2, \ldots.
\]

We demonstrate below how to interpret a rejection of this null. The region in \( \mathbb{R}^2 \) pertaining to \( H_0^{\text{lex}} \) is the lower part of the vertical axis in Figure 2(c).

Obviously, \( H_0^{\text{lex}} \) is the union of all \( H_0^{(c)} \) with \( c \leq 0 \), where

\[
H_0^{(c)}: E[\tilde{d}_n] = 0 \quad \text{and} \quad E[\tilde{d}_{2n}] = c \quad \text{for all } n = 1, 2, \ldots
\]

For each individual \( c \leq 0 \), this can be tested using the Wald-type test statistic \( T_n^{(c)} \), where \( T_n^{(c)} \) is defined similarly as \( T_n \) only with \( \tilde{d}_n \) replaced by \( \tilde{d}_n^{(c)} = (\tilde{d}_n, \tilde{d}_{2n} - c)' \). (Note that this substitution leaves \( \tilde{\Omega}_n \) unaffected.) Thus, we reject \( H_0^{\text{lex}} \) if and only if

\[
T_n^{(c)} > \chi^2_{2,1−\nu} \quad \text{for all } c \leq 0, \tag{5.4}
\]

where \( \nu \in (0,1) \). The area associated with the appertaining nonrejection region is shaded in pink in Figure 2(c). The rejection condition (5.4) is of course equivalent to \( T_n^{\text{OS}} := \inf_{c \leq 0} T_n^{(c)} = T_n^{(c^*)} > \chi^2_{2,1−\nu} \), where the solution \( c^* \) is \( c^* = \min \left\{ 0, \tilde{d}_n - (\tilde{\sigma}_{1,2n}/\tilde{\sigma}_{11,n})\tilde{d}_{1n} \right\} \) follows from a simple quadratic minimization problem. Hence,

\[
T_n^{\text{OS}} = n\left[\tilde{d}_n, \max \left\{ \tilde{d}_{2n}, (\tilde{\sigma}_{1,2n}/\tilde{\sigma}_{11,n})\tilde{d}_{1n} \right\}\right]\tilde{\Omega}_n^{-1}\left(\max \left\{ \tilde{d}_{2n}, (\tilde{\sigma}_{1,2n}/\tilde{\sigma}_{11,n})\tilde{d}_{1n} \right\}\right) . \tag{5.5}
\]

In our numerical experiments, we use \( T_n^{\text{OS}} \) to test \( H_0^{\text{lex}} \).

The next proposition shows that rejecting \( H_0^{\text{lex}} \) when \( T_n^{\text{OS}} > \chi^2_{2,1−\nu} \) leads to a test of level \( \nu = 1/2[1 + \tilde{\nu} - F_{\chi^2_{2,1−\nu}}(\chi^2_{2,1−\nu})] \), where \( F_{\chi^2_{2,1−\nu}} \) denotes the cdf of a \( \chi^2_{2,1−\nu} \)-distribution.

**Proposition 5.2.** Under the conditions of Theorem 5.1, rejecting \( H_0^{\text{lex}} \) if \( T_n^{\text{OS}} > \chi^2_{2,1−\nu} \) leads to an asymptotic size \( \nu \)-test with

\[
\nu = 1/2[1 + \tilde{\nu} - F_{\chi^2_{2,1−\nu}}(\chi^2_{2,1−\nu})] .
\]

That is,

\[
\sup_{c \leq 0} \lim_{n \to \infty} \Pr\left( H_0^{\text{lex}} \text{ is rejected based on } T_n^{\text{OS}} \middle| H_0^{(c)} \text{ holds} \right) = \nu.
\]

**Remark 5.3.** For a test of \( H_0^{\text{lex}} \) with desired significance level of \( \nu \), one can determine the corresponding level \( \tilde{\nu} \) in the critical value \( \chi^2_{2,1−\nu} \) from \( \nu = 1/2[1 + \tilde{\nu} - F_{\chi^2_{2,1−\nu}}(\chi^2_{2,1−\nu})] \) using standard root-finding algorithms. For example, if \( \nu = 1% \) / \( \nu = 5% \) / \( \nu = 10% \), then \( \tilde{\nu} = 1.60% / \tilde{\nu} = 7.66% / \tilde{\nu} = 14.9% \) with corresponding critical values \( \chi^2_{2,1−\nu} \) equalling 8.27 / 5.14 / 3.81.
Remark 5.4. In practice, it will most often be the case that the VaR and systemic risk forecasts, which are to be compared, differ in their VaR component (e.g., when comparing a bank’s internal model to a supervisory benchmark). Yet sometimes the VaR forecasts may be identical, such that $\text{VaR}_t = \text{VaR}_{t-1}$ for $t = 2, 3, \ldots$. Thus, one can focus solely on the systemic risk component when comparing $H_0$ and $H_0^\text{syst}$ for the marginals to ensure a fair assessment of the systemic risk forecasts. For this comparison, where the VaR forecasts are identical, one can focus solely on the systemic risk component by using $\mathcal{T}_n$; see Remark 5.4. For the comparison via $\mathcal{T}_n$, we suggest a similar decision heuristic as in Fissler, Ziegel, and Gneiting (2016): Neither $H_0^\text{syst}$ nor $H_0^\text{syst}$ can be rejected when $|\mathcal{T}_n| \leq \Phi^{-1}(1 - \nu)$ (corresponding to our yellow zone). When $\mathcal{T}_n > \Phi^{-1}(1 - \nu)$, the internal systemic risk forecast is superior (inferior), corresponding to our green (red) zone.

Similarly as for the red region, there are no grounds for meaningful systemic risk forecast comparisons in the grey area, since the internal model’s VaR forecasts are superior. (Formally, it corresponds to the rejection region of the null hypothesis that $\mathcal{T}_n$ is bounded by $\Phi^{-1}(1 - \nu)$.) Here, the bank should not be allowed to use its own marginal model (the traffic light is red), but instead should be required to use the benchmark model for the marginals to ensure a fair assessment of the systemic risk forecasts. For this comparison, where the VaR forecasts are identical, one can focus solely on the systemic risk component by using $\mathcal{T}_n$; see Remark 5.4. For the comparison via $\mathcal{T}_n$, we suggest a similar decision heuristic as in Fissler, Ziegel, and Gneiting (2016): Neither $H_0^\text{syst}$ nor $H_0^\text{syst}$ can be rejected when $|\mathcal{T}_n| \leq \Phi^{-1}(1 - \nu)$ (corresponding to our yellow zone). When $\mathcal{T}_n > \Phi^{-1}(1 - \nu)$, the internal systemic risk forecast is superior (inferior), corresponding to our green (red) zone.

5.3. Simulation Evidence

Supplement S.7 investigates the finite-sample properties of $\mathcal{T}_n$, $\mathcal{T}_n^\text{OS}$, and $\mathcal{T}_n$ under $H_0$ and $H_0^\text{syst}$ in detail. Here, we only summarize the main findings. First, size is adequate already for $n = 500$, which is encouraging since effective sample sizes in systemic risk forecast comparisons are small. Second, power increases markedly in $n$. Third, comparisons for (Co)VaR, CoES are slightly more powerful than those for CoVaR alone, most likely due to the increased informational content of the CoES component. Fourth, as expected for one-sided tests, departures from $H_0^\text{syst}$ and $H_0^\text{syst}$ are easier to detect for the former than for the latter. Fifth, it is in general easier to detect differences in predictive ability of the systemic risk component, when the VaR com-
component is identical (instead of only comparable) across forecasts. Intuitively, the inclusion of comparable VaR forecasts dilutes the power of the test in the systemic risk component. Sixth, it is easiest to reject $H_0^{\text{systemic}}$ and $H_0$ if—in addition to the systemic risk forecasts—the VaR forecasts are also of different quality.

### 5.4. An Alternative Two-Step Approach

The tests $T_n$ (for $H_0^+$) and $T_{n,\text{OS}}$ (for $H_0^{\text{systemic}}$) are based on one-step procedures. Alternatively, one may rely on the following two-step test for $H_0^+$ and $H_0^{\text{systemic}}$. First, $E(\bar{a}_{1n}) = 0$ is tested via $T_{1n} = \sqrt{n}d_{1n}/\sigma_{11,n}^{1/2}$. Then, if there was no rejection in the first step, $E(\bar{a}_{2n}) = 0$ (for $H_0^+$) and $E(\bar{a}_{2n}) \leq 0$ (for $H_0^{\text{systemic}}$) are tested via $T_{2n} = \sqrt{n}d_{2n}/\sigma_{22,n}^{1/2}$. Therefore, size is controlled asymptotically under $H_0^+$ if for critical values $c_1 > 0$ and $c_2 > 0$

$$\alpha \leftarrow P_{H_0^+} \left( \{ |T_{1n}| > c_1 \} \cup \{ |T_{1n}| \leq c_1, |T_{2n}| > c_2 \} \right)$$

$$= P_{H_0^+} \{ |T_{1n}| > c_1 \} + P_{H_0^+} \{ |T_{1n}| \leq c_1, |T_{2n}| > c_2 \}$$

(5.6)

for some pre-specified $\alpha \in (0,1)$; see Figure 4(a) for the resulting nonrejection region. (A similar statement as (5.6) holds for testing $H_0^{\text{systemic}}$; see also Figure 4(b).) By virtue of Theorem 5.1, $(T_{1n}, T_{2n})'$ is asymptotically normal with variance-covariance matrix that can be estimated consistently. Hence, it is possible to choose $c_1$ and $c_2$ such that (5.6) holds. Apart from the specific choices of the critical values $c_1$ and $c_2$, the two-step approach amounts to a sequential application of standard Diebold and Mariano (1995) tests. We stress that while the actual testing becomes somewhat standard, the two-step test (as its one-step counterpart) builds on the novel notion of multi-objective elicitation, where the lexicographic order first prescribes a comparison of the VaR scores before comparing the systemic risk scores.

The appeal of the above two-step approach is that it allows for a straightforward interpretation of a rejection of the composite null $H_0^+$: When $|T_{1n}| > c_1$, a violation of equal VaR accuracy, $E(\bar{a}_{1n}) = 0$, is implied; yet when $|T_{1n}| \leq c_1$ and $|T_{2n}| > c_2$ the rejection is due to $E(\bar{a}_{2n}) = 0$ being violated (such that systemic risk forecasts are of different quality).

However, the one-step test has three advantages relative to the two-step approach. First, with readily available critical values it is simple to implement. This contrasts with the two-step test, where obtaining appropriate critical values $c_1$ and $c_2$ involves the computation of bivariate normal rectangular probabilities in (5.6), which is notoriously demanding (Genz 2004). We mention that this problem may be side-stepped by opting for a Bonferroni-type correction in the two-step test (Romano, Shaikh, and Wolf 2010). Given that such a procedure would only consist of two steps, the resulting test would not necessarily be very conservative, although this cannot be guaranteed of course. Second, from well-known results in classical statistics (e.g., Engle 1984) the one-step Wald test $T_n$ possesses certain optimality properties. In particular, this implies that the better forecasts are identified with higher likelihood than for other tests, such as the two-step test. Third, the choice of the critical values $c_1$ and $c_2$ in the two-step test is not unique. Specifically, it is unclear how to weight the two probabilities in (5.6), introducing some degree of arbitrariness in the two-step test. For all these reasons, we prefer our one-step test.

### 6. Empirical Application

While the schematic in Figure 3 is motivated by a regulatory framework, we stress that it can be used in the context of any comparative backtest between different models. We provide such a general example now. Consider daily log-losses $X_{-r+1}, \ldots, X_n$ on the S&P 500 and log-losses $Y_{-r+1}, \ldots, Y_n$ on the DAX 30 from 2000 to 2020, where the data are taken from www.wsj.com/market-data/quotes (ticker symbols: SPX and DX:DAX). So if $Z_{t-1}$ denotes the stock index value at time $t$, then $Z_t = -\log(P_{Z_t}/P_{Z_{t-1}}) (Z \in \{X,Y\})$. We only keep those observations where data on both indexes are available, giving us $n + r = 5,193$ observations. Here, for $\alpha = \beta = 0.95$, we compare rolling-window (VaR, CoVaR, CoES) forecasts for the series $(X_t, Y_t)_{t=1,\ldots,n}$, where $r = 1000$ denotes the moving window length. The choice of $X$ and $Y$ amounts to considering the risk for large losses of the DAX 30, given that the world’s leading stock index—the S&P 500—is in distress. To promote flow in this section, we often refer to the simulation setup in Supplement S.7 for details on the time series models and the risk forecast computation.

For short-term risk management purposes, conditional (systemic) risk measure forecasts are more informative than unconditional ones. Conditional risk measures are based on the conditional distribution of $(X_t, Y_t)$, that is, $F_{(X_t,Y_t)\mid \bar{Y}_{t-1}}(x,y) = P(X_t \leq x, Y_t \leq y \mid \bar{Y}_{t-1}) = P_{t-1}(X_t \leq x, Y_t \leq y)$ for $x, y \in \mathbb{R}$. Here, the filtration $\bar{Y}_{t-1}$ is generated by the

![Figure 4](https://example.com/figure4.png)

Figure 4. Nonrejection regions of two-step tests for two-sided null in (a), and one and a half-sided null in (b).
information available to a forecaster at time \( t = 1 \). These are usually past observations \((X_{t-1}, Y_{t-1}), (X_{t-2}, Y_{t-2}), \ldots\), and possibly additional exogenous information. Here, we assume \( \tilde{F}_{t-1} = \sigma((X_{t-1}, Y_{t-1}), (X_{t-2}, Y_{t-2}), \ldots) \), such that we forecast the conditional risk measures \( \text{VaR}_t(X_t) = \text{VaR}_\theta(F_{X|\tilde{F}_{t-1}}), \text{CoVaR}_t(Y_t|X_t) = \text{CoVaR}_\alpha(\beta(F_{X|\tilde{F}_{t-1}}, Y_t)|\tilde{F}_{t-1}), \) and \( \text{CoES}_t(Y_t|X_t) = \text{CoES}_\alpha(\beta(F_{X|\tilde{F}_{t-1}}, Y_t)|\tilde{F}_{t-1}) \). For notational brevity, we suppress the dependence of the risk measures on the risk levels \( \alpha \) and \( \beta \), which we fix at \( \alpha = \beta = 0.95 \).

We consider two different methods for \((\text{VaR}_t(X_t), \text{CoVaR}_t(Y_t|X_t), \text{CoES}_t(Y_t|X_t))\) forecasting. The first method uses a simple GARCH(1,1) model for the respective innovations \( \varepsilon_{x,t} \) and \( \varepsilon_{y,t} \)—a Gaussian copula driven by GAS dynamics. That is, we assume \((\varepsilon_{x,t}, \varepsilon_{y,t}) | \tilde{F}_{t-1} \) to have a Gaussian copula density \( c(\cdot ; \rho_t) \) with time-varying correlation parameter \( \rho_t \in (-1, 1) \) following GAS dynamics. Details on the precise specification are in Supplement S.7.1, where the same model (with \( \theta = \infty \) in (S.25)) is used in the simulations. Both the marginal and the dependence model are regularly used as benchmark models in forecast comparisons.

The second method uses the GJR–GARCH(1,1) model of Glosten, Jagannathan, and Runkle (1993). The GJR–GARCH model possesses an additional parameter that allows positive and negative shocks of equal magnitude to have a different effect on volatility. As a dependence model, we now use a \( t \)-copula driven by GAS dynamics, similarly as in (S.25). For both models we remain agnostic regarding the specific distribution of the \( \varepsilon_{x,t} \) and \( \varepsilon_{y,t} \) both in the model estimation (by using the Gaussian quasi-maximum likelihood estimator for the GARCH-type marginal models, which is robust to other error distributions) and the risk forecasting stage (by using their empirical cdfs \( \tilde{F}_x \) and \( \tilde{F}_y \) in computing the risk measures). For details on how the risk predictions are calculated, we refer to Supplement S.7.2.

We now compare two sequences of rolling-window predictions. For each rolling window of length \( r = 1000 \), we refit the two models on a daily basis. This gives us \( n = 4193 \) forecasts \( \{r_{t(1)}|t = 1, \ldots, n\} \) from the GARCH model with Gaussian copula, and \( \{r_{t(2)}|t = 1, \ldots, n\} \) from the GJR–GARCH with \( t \)-copula. We interpret the \( r_{t(1)} \) as generic benchmark forecasts, which are to be improved upon by the \( r_{t(2)} \). In an oversight context, the \( r_{t(1)} \) may be some regulatory benchmark forecasts and the \( r_{t(2)} \) forecasts from the bank's internal model. Or from a banking perspective, the \( r_{t(1)} \) may be forecasts issued from a trading desk's legacy model and the \( r_{t(2)} \) forecasts from a refined version thereof. The latter context is more realistic here, because in the regulatory framework the focus is more often on the relation between an individual bank's returns and the market as a whole. In any case, the methodology remains the same. We compare forecasts based on the verifying observations \( \{(X_t, Y_t)\}_{t = 1, \ldots, n} \).

Figure 5 shows CoVaR and CoES forecasts from both models, where the top panel corresponds to the GARCH model with Gaussian copula and the bottom panel to the GJR–GARCH with \( t \)-copula. Specifically, the panels show the CoVaR (blue) and CoES (red) forecasts for the DAX log-losses (black) on days where the S&P 500 exceeds its VaR forecast. Note that due to the different marginal models (and, hence, the different VaR forecasts), the black lines differ slightly in the upper and lower panel of Figure 5. Observe that both panels of Figure 5 indicate marked spikes in systemic risk during the financial crisis of 2008–2009, during the European sovereign debt crisis in the first half of the 2010s, and—most markedly—in 2020 as a consequence of the Covid-19 pandemic.

By definition, the S&P 500 should only exceed its VaR forecast on 5% of all trading days, that is, on \( 0.05 \cdot 4193 = 209.65 \) days in our out-of-sample period. With 213 (218) VaR violations, our marginal GARCH(1,1) model (GJR–GARCH(1,1) model) is close to the ideal frequency. By definition, we expect our CoVaR forecasts to be not exceeded on 95% of these 213 days (218 days) with a VaR violation. With 15 and 16 exceedances (blue dots), which correspond to nonexceedance frequencies of 93.0% and 92.7%, the Gaussian copula and the \( t \)-copula model are reasonably close to the 95%-benchmark. However, for the Gaussian copula, the CoVaR exceedances seem to cluster more, such as during the beginning of the Covid-19 pandemic in early 2020 (top panel of Figure 5). Such violation clusters are undesirable from a risk management perspective, providing some informal evidence in favor of the \( t \)-copula model. We investigate this more formally in the following.

---

**Figure 5.** Top: DAX log-losses on days of VaR violation of S&P 500 (black). CoVaR (CoES) forecasts are shown as the blue (red) line. Violation of CoVaR (CoES) forecast is indicated by a blue (red) dot. All forecasts from GARCH with Gaussian copula. Bottom: Same as top, only with forecasts from GJR–GARCH with \( t \)-copula.
As pointed out above, we regard the GARCH model with Gaussian copula as our benchmark model. So we now want to test $H_{0}^{\text{clus}}$ for $(\text{VaR}, \text{CoVaR})$ and $(\text{VaR}, \text{CoVaR}, \text{CoES})$. Due to the less clustered CoVaR exceedances and the compelling empirical evidence in favor of GJR–GARCH models (Glosten, Jagannathan, and Runkle 1993; Brownlees, Engle, and Kelly 2011) and GAS-t-copula models (Creal, Koopman, and Lucas 2013; Bernardi and Catania 2019), we expect the GJR–GARCH model with t-copula to produce lower scores, that is, better risk forecasts, possibly leading to a rejection of $H_{0}^{\text{clus}}$. We carry out the tests using $\tau_{n}^{\text{OS}}$ with the scores of (5.27) and (5.28) having 0-homogeneous score differences, and with $\Omega_{n}$ from (5.2) (with $m_{n} = 0$); see Supplement S.7.3 for details. For VaR and ES forecasts, scoring functions giving 0-homogeneous score differences are typically recommended, since they allow for ‘unit-consistent’ and powerful comparisons (Patton 2011; Nolde and Ziegel 2017; Patton, Ziegel, and Chen 2019). We confirm the latter in our simulations for systemic risk forecasts as well, thus, justifying our choice. Let $\tilde{d}_{n} = (\tilde{d}_{1n}, \tilde{d}_{2n})'$ be defined as in Section 5.1. Indeed, computing $\tilde{d}_{n}$, we find that the GJR–GARCH with t-copula produces lower scores for both the $(\text{VaR}, \text{CoVaR})$ and the $(\text{VaR}, \text{CoVaR}, \text{CoES})$ forecasts. The score differences are even statistically significant at the 5%-level: The $p$-values for the $\tau_{n}^{\text{OS}}$-based Wald test are 2.9% for $(\text{VaR}, \text{CoVaR})$ and 3.0% for $(\text{VaR}, \text{CoVaR}, \text{CoES})$.

Figure 6 illustrates the test decisions. Panel (a) shows the results for the $(\text{VaR}, \text{CoVaR})$ comparison. Both $\tilde{d}_{1n} = \tilde{d}_{2n}$ are positive, favoring the forecasts $r_{1,2}$ from the GJR–GARCH model with t-copula. Additionally, the pair $\tilde{d}_{n} = (\tilde{d}_{1n}, \tilde{d}_{2n})'$ lies above the yellow ellipse, which traces the contour of a bivariate normal distribution with probability content $(100 - 7.66\%) = 92.34\%$. This confirms the significance of the score difference at the 5%-level; recall Remark 5.3. The results for the $(\text{VaR}, \text{CoVaR}, \text{CoES})$ forecasts in panel (b) are qualitatively similar. Adopting our traffic-light interpretation, the GJR–GARCH model with t-copula would be deemed an adequate risk forecasting model.

Nonetheless, the borderline significance of this example shows that discriminating between (systemic) risk forecasts requires long samples for the given parameter choice of $\alpha = 0.95$. This is as expected, because by only considering observations with one extreme component, the effective sample size is massively reduced to roughly $n(1 - \beta)$. Therefore, the effective out-of-sample period for comparing systemic risk forecasts is reduced from a length of 4193 to just over 200. The practical implication is that one should allow for sufficiently large samples to prove the superiority of the internal model over the benchmark. Alternatively, one could adapt a slightly looser definition of distress in the reference asset $X$, for example, by setting $\beta = 0.9$. The current evaluation period for VaR forecasts specified in the Basel framework by the Bank for International Settlements (2019) is one year, amounting to roughly 250 daily returns. Even for evaluating VaR and ES forecasts, the horizon of 250 days has been called into question for being too short (Dimitriadis, Liu, and Schnaitmann 2023), and this is only magnified for systemic risk forecasts. So whatever evaluation period for systemic risk measures is eventually settled on in a regulatory context, it likely needs to be far in excess of one year.

7. Discussion and Outlook

To our knowledge, this is the first article to come up with comparative backtests for the systemic risk measures CoVaR, CoES, and MES, which are crucial inputs in financial, macroeconomic and regulatory applications. Model selection procedures based on our results may enhance modeling attempts of these quantities in financial institutions. Moreover, the fact that our notion of multi-objective elicitability serves as a “truth serum” implies that the regulatory framework can be improved by enticing financial institutions to accurately model systemic risk.

The novel concept of multi-objective elicitability is likely to be valuable also in applications beyond the proposed DM-backtests for systemic risk measures, such as macroeconomics, where—with the Growth-at-Risk popularized by Adrian, Boyarchenko, and Giannone (2019)—increasing attention is being paid to economic tail risks and their interconnections. Furthermore, Supplement S.1 provides additional examples of interesting situations where conditional elicitability holds, yet classical joint elicitability fails. By virtue of Theorem 3.7, one can now construct incentive compatible elicitation mechanisms for these functionals.

![Figure 6](image-url)
Even for multivariate functionals that are elicitable in the classical sense, most prominently the pair (VaR, ES), we think that applying two dimensional multi-objective scores equipped with the lexicographic order can be beneficial. Forecast comparisons using multi-objective scores control for the predictive performance of the VaR-component. Hence, multi-objective scores allow to infer if a different predictive performance can be attributed to superiority in the ES-forecast or in the VaR-forecast. Such an attribution is currently impossible when using the classical univariate scores due to Fissler and Ziegel (2016). Hence, multi-objective scores are more informative, safeguarding against possibly false conclusions in backtesting.

Finally, next to forecast comparisons, the classical concept of elicitation is also strong enough to support estimation theory. For instance, Patton, Ziegel, and Chen (2019) use the (VaR, ES) scores of Fissler and Ziegel (2016) to estimate their dynamic (VaR, ES) models. Multi-objective elicibility (with respect to \((\mathbb{R}^2, \preceq_{\text{lex}})\)) is likewise a strong enough concept to be used in estimation. Specifically, Dimitriadis and Hoga (2023a) estimate their dynamic (VaR, CoVaR) models by a two-step minimization of the multi-objective scores for (VaR, CoVaR) from Theorem 4.2. As prescribed by the lexicographic order, first the VaR parameters are chosen to minimize the VaR score and, in a second step, the CoVaR scores are minimized.

There are, of course, many more open and interesting questions to be answered in the context of multi-objective elicitation, which are beyond the scope of the present article. We therefore see the need for further research to address these issues and to apply the concept to further situations in economics, finance, data science, and beyond.

Supplementary Materials

A supplementary appendix contains Section S.1, which provides further examples of multi-objective elicitable functionals. Section S.2 presents results concerning the conditional identifiability and conditional elicitation of the systemic risk measures under consideration, while Section S.3 features joint identifiability results and Section S.4 non-elicitation results. All proofs are deferred to Section S.5. In Section S.6, we show the power of the tests of Section 5 under local alternatives. Monte Carlo simulations are presented in Section S.7. Finally, Section S.8 illustrates the perils of using non-strict identification functions for calibration tests. We also provide the \(R\) code to reproduce all numerical results in the article and the supplementary appendix.

Acknowledgments

We are grateful to two referees whose comments significantly improved the quality of the article. We are indebted to Immanuel Bomze for suggesting to consider multivariate scoring functions equipped with general orders. Furthermore, we would like to thank Timo Dimitriadis, Rüdiger Frey, Christoph Hanck, Jana Hlavínová, Kurt Hornik, Marc-Oliver Pohle, Birgit Rudloff and Johanna Ziegel for detailed comments and valuable discussions. Of course, all errors and opinions expressed in this article are solely the authors’ responsibility.

Disclosure Statement

The authors report that there are no competing interests to declare.

Funding

Tobias Fissler gratefully acknowledges support from the OeNB anniversary fund, project number 17793. Yannick Hoga is thankful for support of the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through project 460479886.
Girardi, G., and Tolga Ergün, A. (2013), "Systemic Risk Measurement: Multivariate GARCH Estimation of CoVaR," *Journal of Banking & Finance*, 37, 3169–3180. [485,487]

Glosten, L. R., Jagannathan, R., and Runkle, D. E. (1993), "On the Relation between the Expected Value and the Volatility of the Nominal Excess Return on Stocks," *The Journal of Finance*, 48, 1779–1801. [495,496]

Gneiting, T. (2011), "Making and Evaluating Point Forecasts," *Journal of the American Statistical Association*, 106, 746–762. [486,487,488,489]

Hoga, Y. (2022), "Modeling Time-Varying Tail Dependence, With Application to Systemic Risk Forecasting," *Journal of Financial Econometrics*, 20, 1007–1037. [493]

Hoga, Y., and Demetrescu, M. (2023), "Monitoring Value-at-Risk and Expected Shortfall Forecasts," *Management Science*, 1–18. DOI:10.1287/mnsc.2022.4460 [490]

Holzmann, H., and Eulert, M. (2014), "The Role of the Information Set for Forecasting – With Applications to Risk Management," *The Annals of Applied Statistics*, 8, 79–83. [489]

Lambert, N. (2019), "Elicitation and Evaluation of Statistical Forecasts," *Preprint. Available at http://web.stanford.edu/~nlambert/papers/elicitation.pdf*. [488]

Lambert, N., Pennock, D. M., and Shoham, Y. (2008), "Eliciting Properties of Probability Distributions," in *Proceedings of the 9th ACM Conference on Electronic Commerce*, pp. 129–138, Chicago, IL: ACM. [489]

Nolde, N., and Zhang, J. (2020), "Conditional Extremes in Asymmetric Financial Markets," *Journal of Business & Economic Statistics*, 38, 201–213. [487,493]

Nolde, N., and Ziegel, J. F. (2017), "Elicitability and Backtesting: Perspectives for Banking Regulation," *The Annals of Applied Statistics*, 11, 1833–1874. [486,489,496]

Oh, D. H., Patton, A. J. (2018), "Time-Varying Systemic Risk: Evidence from a Dynamic Copula Model of CDS Spreads," *Journal of Business & Economic Statistics*, 36, 181–195. [485]

Osband, K. H. (1985), "Providing Incentives for Better Cost Forecasting," Ph.D. thesis, University of California, Berkeley. [489]

Patton, A. J. (2011), "Volatility Forecast Comparison Using Imperfect Volatility Proxies," *Journal of Econometrics*, 160, 246–256. [496]

Patton, A. J., Ziegel, J. F., Chen, R. (2019), "Dynamic Semiparametric Models for Expected Shortfall (and Value-at-Risk)," *Journal of Econometrics*, 211, 388–413. [496,497]

Romano, J. P., Shaikh, A. M., and Wolf, M. (2010), "Hypothesis Testing in Econometrics," *Annual Review of Economics*, 2, 75–104. [494]

Steinwart, I., Pasin, C., Williamson, R., and Zhang, S. (2014), "Elicitation and Identification of Properties," in *JMLR Workshop Conf. Proc.* (35), pp. 1–45. [487,489]

White, H. (2001), *Asymptotic Theory for Econometricians* (1st ed.), San Diego: Academic Press. [491]