

Backtesting Systemic Risk Forecasts using Multi-Objective Elicitability*

Tobias Fissler[†] Yannick Hoga[‡]

February 2, 2022

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 summarised as ‘backtesting’, impossible. The novel notion of *multi-objective elicibility* solves this problem. Specifically, we propose Diebold–Mariano type tests utilising two-dimensional scores equipped with the lexicographic order. We illustrate the test decisions by an easy-to-apply traffic-light approach. We apply our traffic-light approach to DAX 30 and S&P 500 returns, and infer some recommendations for regulators.

Keywords: Backtest; (Conditional) Elicitability; Forecasting; Identifiability; Lexicographic Order; Multi-objective Optimisation; Systemic Risk

JEL classification: C18 (Methodological Issues), C52 (Model Evaluation, Validation, and Selection), C58 (Financial Econometrics)

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

*We are grateful 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 Hlavinová, Kurt Hornik, Marc-Oliver Pohle, Birgit Rudloff and Johanna F. Ziegel for detailed comments and valuable discussions. Of course, all errors and opinions expressed in this article are solely the authors’ responsibility. Yannick Hoga gratefully acknowledges support of the German Research Foundation (DFG) through grant HO 6305/2-1.

[†]Vienna University of Economics and Business (WU), Department of Finance, Accounting and Statistics, Welthandelsplatz 1, 1020 Vienna, Austria, e-mail: tobias.fissler@wu.ac.at.

[‡]University of Duisburg-Essen, Faculty of Economics and Business Administration, Universitätsstraße 12, D-45117 Essen, Germany, tel. +49 201 1834365, e-mail: yannick.hoga@vwl.uni-due.de.

a pronounced comonotonic behaviour 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 literature on measuring systemic risk has emerged over the last decade (Giesecke and Kim, 2011; Chen *et al.*, 2013; Adrian and Brunnermeier, 2016; Acharya *et al.*, 2017; Brownlees and Engle, 2017; Feinstein *et al.*, 2017).

Systemic risk measures are important in various contexts. First, they are important in banking regulation under the Basel framework of the Bank for International Settlements (2019), where they are vital in determining which banks are among the globally systemically important banks (G-SIBs). Such G-SIBs are then subjected to higher capital requirements. Second, in finance, systemic risk measures may be used to study spillover effects in the financial system (Adrian and Brunnermeier, 2016) or the build-up of asset price bubbles (Brunnermeier *et al.*, 2020). Third, they may be used to study the linkage between the financial sector and the real economy. Among others, Giglio *et al.* (2016) and Brownlees and Engle (2017) show that an increase in systemic risk is predictive of future declines in real economic activity. All these examples underscore the importance of accurately measuring and predicting systemic risk.

In this paper, 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 are the losses of interest and X the losses of a reference position, $\text{CoVaR}_{\alpha|\beta}(Y|X)$ ($\text{CoES}_{\alpha|\beta}(Y|X)$) is the VaR (ES) of Y at level α , given that X is “in distress”. Here, we interpret the event that X is in distress as X being larger or equal than its β -quantile, i.e., $\{X \geq \text{VaR}_{\beta}(X)\}$. Finally, we consider Acharya *et al.*’s (2017) marginal expected shortfall, $\text{MES}_{\beta}(Y|X)$, as the conditional mean of Y given $\{X \geq \text{VaR}_{\beta}(X)\}$. Section 2 introduces the exact definitions.

Benoit *et al.* (2017) distinguish between the “source-specific approach” and the “global approach” to systemic risk measurement. The source-specific approach considers individual sources of systemic risk, such as contagion risk or liquidity crises. In contrast, global measures of systemic risk potentially incorporate all mechanisms studied in the source-specific approach. Benoit *et al.* (2017) categorize CoVaR, CoES and MES under the global approach.

In practice, forecasting systemic risk measures—such as CoVaR, CoES and MES—requires adequate models for the marginals X and Y , and for their dependence structure. The literature has developed numerous different modelling approaches for this; 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 outlined above, it is vital to develop statistical quality assessments of the various models’ predictive performances. It is the main aim of this paper 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 *et al.* (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 (possible *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 *et al.*, 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 minimised in expectation by the optimal forecast (see Definition 3.1).

Our contributions in this paper are the following: First, we show that $\text{CoVaR}_{\alpha|\beta}$, $\text{CoES}_{\alpha|\beta}$ and MES_β are not identifiable and 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 the adequacy of the forecasts can be determined nor can different systemic risk forecasts be sensibly compared (say to a regulatory standard model).

We provide a partial remedy for this drawback by giving joint 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 $(\text{VaR}_\beta(X), \text{MES}_\beta(Y|X))$ (Theorem 4.2). 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 4.3. 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 (Section G in the Supplement). Theoretically, our results are akin to the fact that $\text{ES}_\alpha(Y)$ is not identifiable on its own, but the pair $(\text{VaR}_\alpha(Y), \text{ES}_\alpha(Y))$ is identifiable (Fissler and Ziegel, 2016).

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

As a remedy to this negative result, we propose the novel concept of *multi-objective elicibility*, which works with *multivariate* scores S mapping to $\mathbb{R}^m$ equipped with a certain (partial) order relation $\preceq$. 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 $\leq$, which allows for straightforward comparisons of losses. Sub-

section 3.2 introduces *multi-objective scores* $\mathbf{S}$ and the corresponding concepts of *multi-objective consistency and elicibility*. The terminology stems from the field of *multi-objective optimisation*: According to Ehrgott (2005, p. v) it is “a mathematical theory of optimization under multiple objectives”, and can be encountered in various fields of science, economics, logistics and engineering. Since this novel concept to forecast evaluation may open up the avenue to a whole field of applications and research (which is underpinned by further instances; see Example B.1), we give a concise general outline of the theory, using partial orders on $\mathbb{R}^m$ or even infinite-dimensional real vector spaces.

For the systemic risk forecasts we consider here, scores mapping to $\mathbb{R}^2$ equipped with the lexicographic (total) order—described in Subsection 3.3—are sufficient. In particular, the performance of different systemic risk forecasts must be ranked with regard to the lexicographic order. Specifically, Theorem 4.4 shows 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, and it provides classes of strictly multi-objective consistent scores. We outline 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 elicibility (with scores mapping to $\mathbb{R}^2$ equipped with the lexicographic order) instead of the “standard” notion of elicibility (with scores mapping to $\mathbb{R}$ equipped with the canonical order $\leq$). In particular, some systemic risk forecast is now preferable to some other forecast when the $\mathbb{R}^2$ -valued score of the former is smaller (with regard to the lexicographic order) than that of the latter. Thus, financial institutions may build on this result to improve their prediction models for $(\text{VaR}_\beta, \text{CoVaR}_{\alpha|\beta})$, $(\text{VaR}_\beta, \text{CoVaR}_{\alpha|\beta}, \text{CoES}_{\alpha|\beta})$ and $(\text{VaR}_\beta, \text{MES}_\beta)$, which is crucial for an adequate assessment of the diverse risks faced by these institutions.

Since the multi-objective scores of Theorem 4.4 take values in $\mathbb{R}^2$ equipped with the lexicographic order, some particularities arise for statistical hypothesis tests. While simple “two-sided” null hypotheses of equal predictive performance can be tested with a classical Wald-test, particular caution must be taken when testing for superior predictive ability. Due to the particularities of the lexicographic order, a straightforward “one-sided” composite null hypothesis would be insensitive to the systemic risk measure forecast, ignoring the primary goal of the backtesting procedure. Therefore, we suggest to use “one and a half”-sided composite null hypotheses, testing for superior predictive ability in the systemic risk component and equal performance in the auxiliary $\text{VaR}_\beta(X)$ component. Section 5 provides details, including an adaptation of the Basel framework’s traffic-light approach to systemic risk backtests.

An empirical application in Section 6 demonstrates the viability of the comparative backtest. There, we consider daily log-losses of the DAX 30 with daily log-losses of the S&P 500 as a reference quantity. We compare systemic risk forecasts derived from a benchmark Gaussian copula model with those produced by a t -copula model, where in both models the correlation parameter of the copula is driven by generalised autoregressive score (GAS) dynamics (Creal *et al.*, 2013). We find that the predictive performance of the t -copula is superior with p -values close to 3%, which is consistent with its popularity in empirical work. One conclusion from

our empirical analysis is that fairly long samples are required to validly distinguish between different forecasts, because the effective sample sizes in comparing systemic risk forecasts are (almost by definition) reduced. Thus, the one year evaluation period for (univariate) VaR and ES forecasts in the Basel framework of the [Bank for International Settlements \(2019\)](#) is, in our view, insufficient for systemic risk forecasts.

The paper closes with a discussion and outlook (Section 7). Besides the parts already mentioned above, the Supplement provides proofs for the results of Section 3 (Section A) and further background material on multi-objective elicibility (Section B). All other proofs are relegated to Section E. Section F investigates the finite-sample properties of our comparative backtests in simulations. The R code to reproduce all numerical experiments is available online.

Throughout the paper, we indicate vectors with bold letters. We highlight the distinction between row and column vectors only when it is essential, and use the symbol $'$ to indicate the transpose of a vector or matrix.

2. Formal definition of CoVaR, CoES and MES

Fix some non-atomic probability space $(\Omega, \mathfrak{A}, P)$ 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 $(\Omega, \mathfrak{A}, P)$. 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 $\mathbf{X} \in L^0(\mathbb{R}^d)$ let $F_{\mathbf{X}}$ be its joint distribution. Then define for $p \in \{0\} \cup [1, \infty)$ the collection $\mathcal{F}^p(\mathbb{R}^d) := \{F_{\mathbf{X}} : \mathbf{X} \in L^p(\mathbb{R}^d)\}$. We overload notation and identify any $F \in \mathcal{F}^0(\mathbb{R}^d)$ with its cumulative distribution function (cdf) $\mathbb{R}^d \rightarrow [0, 1]$.

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 $\mathbb{R}^* := (-\infty, \infty]$. 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_{β} , $\beta \in [0, 1]$, is a law-determined map $L^0(\mathbb{R}) \rightarrow [-\infty, \infty]$. In the rest of the paper, 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 their marginals, respectively. Recall that for $\beta \in [0, 1]$ the β -quantile of F_X is the closed interval $q_{\beta}(F_X) = \{x \in \mathbb{R} : F_X(x-) \leq \beta \leq F_X(x)\}$, where $F_X(x-) := \lim_{t \uparrow x} F_X(t)$. Then, $\text{VaR}_{\beta}(X)$ is the lower β -quantile of F_X , i.e., $\text{VaR}_{\beta}(X) := \text{VaR}_{\beta}(F_X) := \inf q_{\beta}(F_X)$. For $\beta \in (0, 1)$, VaR_{β} is always finite. 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 β close to 1 for VaR_{β} , such as $\beta = 0.95$ or $\beta = 0.99$.

[Adrian and Brunnermeier \(2016\)](#) define CoVaR_β as the β -quantile of the conditional distribution function $F_Y(\cdot | X = \text{VaR}_\beta(X)) = \mathbb{P}\{Y \leq \cdot | X = \text{VaR}_\beta(X)\}$. The conditioning event in this definition is problematic for several reasons. First, it may have probability zero (which is the case when F_X is continuous). Second, it does not fully capture the tail-risk of F_X . Third, since the roles of Y and X are asymmetric by construction, one may want to consider different probability levels to specify the event of ‘being in distress’. Thus, we follow [Girardi and Tolga Ergün \(2013\)](#) and [Banulescu-Radu et al. \(2021\)](#) in redefining $\text{CoVaR}_{\alpha|\beta}: L^0(\mathbb{R}^2) \rightarrow \mathbb{R}$ for $\alpha \in (0, 1)$, $\beta \in [0, 1)$ as

$$\text{CoVaR}_{\alpha|\beta}(Y|X) := \text{CoVaR}_{\alpha|\beta}(F_{X,Y}) := \text{VaR}_\alpha(F_{Y|X \geq \text{VaR}_\beta(X)}), \quad (2.1)$$

where $F_{Y|X \geq \text{VaR}_\beta(X)} = \mathbb{P}\{Y \leq \cdot | X \geq \text{VaR}_\beta(X)\}$. For $\beta = 0$, we simply have $\text{CoVaR}_{\alpha|0}(Y|X) := \text{VaR}_\alpha(Y)$, 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 merely a quantile of the distribution $F_{Y|X \geq \text{VaR}_\beta(X)}$, it inherits the same defects as $\text{VaR}_\alpha(Y)$. That is, it ignores tail risks beyond the quantile level α and it fails to be coherent, particularly defying the rationale of advantageous diversification effects ([Artzner et al., 1999](#); [Mainik and Schaanning, 2014](#)). The well-known *Expected Shortfall* at level $\alpha \in (0, 1)$, $\text{ES}_\alpha(Y) := \text{ES}_\alpha(F_Y) := \frac{1}{1-\alpha} \int_\alpha^1 \text{VaR}_\gamma(Y) d\gamma \in \mathbb{R}^*$, does not suffer from these defects. This motivates [Adrian and Brunnermeier \(2016\)](#) to introduce the conditional Expected Shortfall (CoES). As for $\text{CoVaR}_{\alpha|\beta}$, we modify the conditioning event and formally introduce for $\alpha \in (0, 1)$, $\beta \in [0, 1)$, $\text{CoES}_{\alpha|\beta}: L^0(\mathbb{R}^2) \rightarrow \mathbb{R}^*$ via

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

If $F_{X,Y}$ is continuous, then $\text{CoES}_{\alpha|\beta}(Y|X) = \mathbb{E}[Y | Y \geq \text{CoVaR}_{\alpha|\beta}(Y|X), X \geq \text{VaR}_\beta(X)]$. Again, $\text{CoES}_{\alpha|0}(Y|X) = \text{ES}_\alpha(Y)$, and we write $\text{CoES}_\alpha(Y|X) := \text{CoES}_{\alpha|\alpha}(Y|X)$.

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

$$\text{MES}_\beta(Y|X) := \text{MES}_\beta(F_{X,Y}) := \text{CoES}_{0|\beta}(Y|X) = \int_0^1 \text{CoVaR}_{\gamma|\beta}(Y|X) d\gamma. \quad (2.3)$$

Again, $\text{MES}_0(Y|X) = \mathbb{E}[Y]$, and $\text{MES}_\beta(Y|X) = \mathbb{E}[Y | X \geq \text{VaR}_\beta(X)]$ if $F_{X,Y}$ is continuous. If F_X is discontinuous, Remark [C.1](#) proposes a novel correction term which generalises the three measures considered in this paper.

3. (Conditional) identifiability and multi-objective elicibility

We present the theory in this section in all generality to serve as a basis for future research on multi-objective scores. Therefore, we work with general functionals which do not necessarily

have the interpretation of risk functionals. All proofs are in Section A.

3.1. Notation, basic definitions and results

Adopting the decision-theoretic terminology of [Gneiting \(2011\)](#), we denote by $\mathbf{A}$ an action domain. This is the space of plausible forecasts, which can be finite for categorial forecasts, $\mathbb{R}$ or $\mathbb{R}^k$ for point forecasts, or a set of distributions for probabilistic forecasts. Moreover, let $\mathbf{O}$ be an observation domain—a set where verifying observations materialise—with $\mathbf{O} = \mathbb{R}^d$ as a leading example. Denote by $\mathcal{F}^0(\mathbf{O})$ the set of all probability distributions on $\mathbf{O}$. Let $\mathcal{F} \subseteq \mathcal{F}'$ be subclasses of $\mathcal{F}^0(\mathbf{O})$. We consider a general, possibly set-valued functional $\mathbf{T}: \mathcal{F}' \rightarrow \mathcal{P}(\mathbf{A})$ with $\mathcal{P}(\mathbf{A})$ the power set of $\mathbf{A}$. Later on, $\mathbf{T}$ will have the interpretation of a risk functional. Note that $\mathcal{F}'$ is the class of distributions, where our functional $\mathbf{T}$ is defined on, and $\mathcal{F} \subseteq \mathcal{F}'$ is the subclass on which $\mathbf{T}$ will be identifiable/elicitable (see Definitions 3.1 and 3.2). For instance, for $\mathbf{T}(F_{X,Y}) = (\text{VaR}_\beta(F_X), \text{CoVaR}_{\alpha|\beta}(F_{X,Y}))$, we have $\mathcal{F}' = \mathcal{F}^0(\mathbb{R}^2)$ and $\mathcal{F}$ is given in Theorem 4.2 (i)/Theorem 4.4 (ii).

We adopt the *selective* notion of forecasts discussed in [Fissler et al. \(2021\)](#) where one is content with correctly specifying a single element $\mathbf{t} \in \mathbf{T}(F) \subseteq \mathbf{A}$ as opposed to specifying the entire set $\mathbf{T}(F)$. (If one is interested in *exhaustive* forecasts—i.e., in forecasts for the whole set $\mathbf{T}(F)$ —one can change the action domain to $\mathcal{P}(\mathbf{A})$.) If $\mathbf{T}$ attains singletons only, we identify the value of $\mathbf{T}(F)$ with its unique element. This identification allows us to treat point-valued functionals as set-valued functionals without loss of generality. We mention that the risk functionals to be considered in Section 4 are all point-valued.

A function $\mathbf{g}: \mathbf{A} \times \mathbf{O} \rightarrow \mathbb{R}^{\mathcal{I}}$, where $\mathcal{I}$ is an index set, is called $\mathcal{F}$ -integrable if for all components g_i , $i \in \mathcal{I}$, it holds that $\int |g_i(\mathbf{r}, \mathbf{y})| dF(\mathbf{y}) < \infty$ for all $\mathbf{r} \in \mathbf{A}$, $F \in \mathcal{F}$. If $\mathbf{g}$ is $\mathcal{F}$ -integrable, we define the map $\bar{\mathbf{g}}: \mathbf{A} \times \mathcal{F} \rightarrow \mathbb{R}^{\mathcal{I}}$, $\bar{\mathbf{g}}(\mathbf{r}, F) := \int g_i(\mathbf{r}, \mathbf{y}) dF(\mathbf{y})$ for $\mathbf{r} \in \mathbf{A}$, $F \in \mathcal{F}$. A similar convention and notation is used for maps $\mathbf{a}: \mathbf{O} \rightarrow \mathbb{R}^{\mathcal{I}}$. We start with the classical definition of elicibility and consistent scoring functions, mapping to $\mathbb{R}$.

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

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

Suppose in a risk management context that $\mathbf{T}$ corresponds to the VaR_α -functional and y_t are the observed losses. Subject to mild conditions on $\mathcal{F}$, VaR_α is elicitable, where the ‘pinball loss’ $S(r, y) = (\mathbb{1}\{y > r\} - 1 + \alpha)(r - y)$ is strictly $\mathcal{F}$ -consistent. This allows to compare competing VaR forecasts $(r_{t,(1)})_{t=1,\dots,n}$ and $(r_{t,(2)})_{t=1,\dots,n}$ 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 $(r_{t,(1)})_{t=1,\dots,n}$ over $(r_{t,(2)})_{t=1,\dots,n}$. Identifiability, on the other hand, opens the way to test for calibration by checking, e.g., if the test statistic $\frac{1}{n} \sum_{t=1}^n V(r_{t,(i)}, y_t)$ is sufficiently close to $\mathbf{0}$ or not. E.g., when $\mathbf{T}$ corresponds to VaR_α checking calibration amounts to checking if the empirical VaR-violation rate is roughly $1 - \alpha$, which can be done in terms of $V(r, y) = \mathbb{1}\{y > r\} - (1 - \alpha)$. These examples demonstrate the importance of elicibility and identifiability for comparing and evaluating (risk) forecasts in practice.

Under regularity conditions, the notions of elicibility 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* elicibility and *conditional* identifiability can be helpful. Following the concept presented in Fissler and Ziegel (2016, Section 6), we slightly adapt Emmer *et al.*'s (2015) original definition of conditional elicibility. We also introduce the corresponding counterpart of conditional identifiability.

Definition 3.3. Consider two functionals $\mathbf{T}_j : \mathcal{F}' \rightarrow \mathcal{P}(A_j)$, $j = 1, 2$, and let $\mathcal{F} \subseteq \mathcal{F}'$.

- (i) $\mathbf{T}_2$ is *conditionally elicitable with $\mathbf{T}_1$* on $\mathcal{F}$, if $\mathbf{T}_1$ is elicitable on $\mathcal{F}$ and $\mathbf{T}_2$ is elicitable on $\mathcal{F}_{\mathbf{r}_1} := \{F \in \mathcal{F} : \mathbf{r}_1 \in \mathbf{T}_1(F)\}$ for any $\mathbf{r}_1 \in A_1$.
- (ii) $\mathbf{T}_2$ is *conditionally identifiable with $\mathbf{T}_1$* on $\mathcal{F}$, if $\mathbf{T}_1$ is identifiable on $\mathcal{F}$ and $\mathbf{T}_2$ is identifiable on $\mathcal{F}_{\mathbf{r}_1} := \{F \in \mathcal{F} : \mathbf{r}_1 \in \mathbf{T}_1(F)\}$ for any $\mathbf{r}_1 \in A_1$.

It is easy to see that the variance is conditionally elicitable and conditionally identifiable with the mean, and that ES_α is conditionally elicitable and conditionally identifiable with VaR_α on appropriate classes of distributions, respectively (Emmer *et al.*, 2015). The pairs (mean, variance) and $(\text{VaR}_\alpha, \text{ES}_\alpha)$ 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 $\mathbf{T}_2$ is conditionally identifiable with $\mathbf{T}_1$ on $\mathcal{F}$, then the pair $(\mathbf{T}_1, \mathbf{T}_2)$ is identifiable on $\mathcal{F}$.

Of course, an analogue to Proposition 3.4 for (conditional) elicibility 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 Section C establishes the conditional elicibility of $\text{CoVaR}_{\alpha|\beta}$, $(\text{CoVaR}_{\alpha|\beta}, \text{CoES}_{\alpha|\beta})$ and MES_β all with VaR_β , the corresponding pairs and the triplet generally fail to be elicitable; see Section D.

3.2. Multi-objective scores, consistency and elicibility

To overcome the structural drawback of elicibility 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 elicibility*. It is inspired by the fundamental observation that identification functions are generally multivariate.

To be more precise, the dimension m of the identification function $\mathbf{V}$ 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 D.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. This gap and its consequences on estimation are discussed in detail by Dimitriadis *et al.* (2020a).

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 $\mathbb{R}^m$, where usually $m = k$, or even more generally to some real vector space $\mathbb{R}^{\mathcal{I}}$, where the index set $\mathcal{I}$ may be finite, countable or even uncountable. In the application of the general theory developed here to the systemic risk measures, it suffices to consider $\mathbb{R}^2$ -valued scores; see Theorem 4.4 in Section 4. The motivation for defining a classical score as a univariate map to $\mathbb{R}$ (see Definition 3.1) is grounded in tradition on the one hand. On the other hand, $\mathbb{R}$ is equipped with the canonical (total) order relation $\leq$, which allows for straightforward comparisons of forecasts by checking whether the empirical average score differences satisfy $\bar{d}_n \geq 0$ or $\bar{d}_n \leq 0$. Consistency in Definition 3.1 ultimately relies on the existence of an order.

Hence, we need to equip $\mathbb{R}^{\mathcal{I}}$ with a vector partial order $\preceq$ and write $(\mathbb{R}^{\mathcal{I}}, \preceq)$. A binary relation $\preceq$ on $\mathbb{R}^{\mathcal{I}}$ is a *partial order* if it is reflexive ($\forall \mathbf{x} \in \mathbb{R}^{\mathcal{I}}, \mathbf{x} \preceq \mathbf{x}$), antisymmetric ($\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^{\mathcal{I}}$ if $\mathbf{x} \preceq \mathbf{y}$ and $\mathbf{y} \preceq \mathbf{x}$, then $\mathbf{x} = \mathbf{y}$), transitive ($\forall \mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^{\mathcal{I}}$ if $\mathbf{x} \preceq \mathbf{y}$ and $\mathbf{y} \preceq \mathbf{z}$, then $\mathbf{x} \preceq \mathbf{z}$). Two elements $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{\mathcal{I}}$ are *comparable* if $\mathbf{x} \preceq \mathbf{y}$ or $\mathbf{y} \preceq \mathbf{x}$. A *total order* is a partial order where all elements are comparable. A partial order $\preceq$ on $\mathbb{R}^{\mathcal{I}}$ induces a *strict* partial order $\prec$ on $\mathbb{R}^{\mathcal{I}}$ as follows. For all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{\mathcal{I}}$ it holds that $\mathbf{x} \prec \mathbf{y}$ if and only if $\mathbf{x} \preceq \mathbf{y}$ and $\mathbf{x} \neq \mathbf{y}$. A partial order $\preceq$ on $\mathbb{R}^{\mathcal{I}}$ is a *vector partial order* if it is compatible with addition and positive scaling. That is, if for all $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^{\mathcal{I}}$ and for all $\lambda \in (0, \infty)$ it holds that $\mathbf{x} \preceq \mathbf{y}$ implies that $\mathbf{x} + \mathbf{z} \preceq \mathbf{y} + \mathbf{z}$, and $\mathbf{x} \preceq \mathbf{y}$ implies that $\lambda \mathbf{x} \preceq \lambda \mathbf{y}$. (In the sequel, we always mean a vector partial order whenever we write “partial order”.) The canonical choice is the componentwise order defined for $\mathbf{x} = (x_i)_{i \in \mathcal{I}}, \mathbf{y} = (y_i)_{i \in \mathcal{I}} \in \mathbb{R}^{\mathcal{I}}$ as $\mathbf{x} \preceq \mathbf{y}$ if and only if $x_i \leq y_i$ for all $i \in \mathcal{I}$. The use of a partial order $\preceq$ on $\mathbb{R}^{\mathcal{I}}$ leads to two different generalisations of Definition 3.1. These are due to the fact that in a partial order $\mathbf{x} \preceq \mathbf{y}$ implies that $\mathbf{y} \not\prec \mathbf{x}$, but the reverse implication fails if $\mathbf{x}$ and $\mathbf{y}$ are not comparable.

Definition 3.5. (i) An $\mathcal{F}$ -integrable map $\mathbf{S}: \mathbf{A} \times \mathbf{O} \rightarrow (\mathbb{R}^{\mathcal{I}}, \preceq)$ is a *strongly* multi-objective

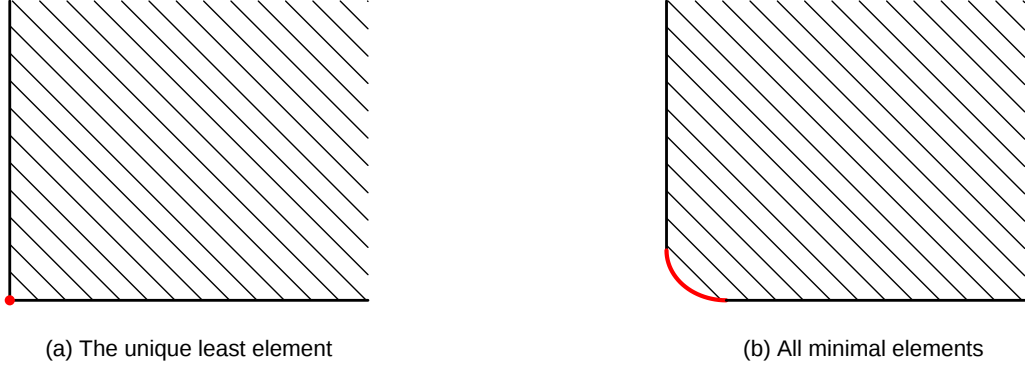


Figure 1: In both panels, the shaded areas correspond to $\bar{\mathbf{S}}(\mathbf{A}, F)$ of a multi-objective score $\mathbf{S}$ mapping to $\mathbb{R}^2$ equipped with the componentwise order. The red sets correspond to $\bar{\mathbf{S}}(\mathbf{T}(F), F)$. In panel (a), this is the unique least element, illustrating the situation of strict strong multi-objective consistency. In panel (b), this set corresponds to all minimal elements, depicting the situation of strict weak multi-objective elicibility.

$\mathcal{F}$ -consistent scoring function for $\mathbf{T}: \mathcal{F} \rightarrow \mathcal{P}(\mathbf{A})$ if $\bar{\mathbf{S}}(\mathbf{t}, F) \preceq \bar{\mathbf{S}}(\mathbf{r}, F)$ for all $\mathbf{t} \in \mathbf{T}(F)$, all $\mathbf{r} \in \mathbf{A}$ and all $F \in \mathcal{F}$. It is a strictly strongly multi-objective $\mathcal{F}$ -consistent scoring function for $\mathbf{T}$ if, additionally, $\bar{\mathbf{S}}(\mathbf{t}, F) = \bar{\mathbf{S}}(\mathbf{r}, F)$ implies that $\mathbf{r} \in \mathbf{T}(F)$. $\mathbf{T}$ is *strongly multi-objective-elicitable on $\mathcal{F}$* (with respect to the order $\preceq$ on $\mathbb{R}^{\mathcal{I}}$) if there is a strictly strongly multi-objective $\mathcal{F}$ -consistent scoring function for $\mathbf{T}$.

- (ii) An $\mathcal{F}$ -integrable map $\mathbf{S}: \mathbf{A} \times \mathcal{O} \rightarrow (\mathbb{R}^{\mathcal{I}}, \preceq)$ is a *weakly* multi-objective $\mathcal{F}$ -consistent scoring function for $\mathbf{T}: \mathcal{F} \rightarrow \mathcal{P}(\mathbf{A})$ if $\bar{\mathbf{S}}(\mathbf{r}, F) \not\preceq \bar{\mathbf{S}}(\mathbf{t}, F)$ for all $\mathbf{t} \in \mathbf{T}(F)$, all $\mathbf{r} \in \mathbf{A}$ and all $F \in \mathcal{F}$. It is a strictly weakly multi-objective $\mathcal{F}$ -consistent scoring function for $\mathbf{T}$ if, additionally, for all $F \in \mathcal{F}$ and for any $\mathbf{r}_0 \in \mathbf{A}$ it holds that if $\bar{\mathbf{S}}(\mathbf{r}, F) \not\preceq \bar{\mathbf{S}}(\mathbf{r}_0, F)$ for all $\mathbf{r} \in \mathbf{A}$, then $\mathbf{r}_0 \in \mathbf{T}(F)$. $\mathbf{T}$ is *weakly multi-objective-elicitable on $\mathcal{F}$* (with respect to the order $\preceq$ on $\mathbb{R}^{\mathcal{I}}$) if there is a strictly weakly multi-objective $\mathcal{F}$ -consistent scoring function for $\mathbf{T}$.

As discussed, a partial order on $\mathbb{R}^{\mathcal{I}}$ is sufficient to define multi-objective consistency. Using the notation $\bar{\mathbf{S}}(B, F) := \{\bar{\mathbf{S}}(\mathbf{r}, F): \mathbf{r} \in B\}$ for $F \in \mathcal{F}$ and $B \subseteq \mathbf{A}$, the definition of multi-objective consistency does not require comparability of all elements $\bar{\mathbf{S}}(\mathbf{A}, F)$ for some fixed $F \in \mathcal{F}$. Weak multi-objective consistency solely implies that all elements of $\bar{\mathbf{S}}(\mathbf{T}(F), F)$ are minimal in $\bar{\mathbf{S}}(\mathbf{A}, F)$. The strict version additionally ensures that all minimal elements of $\bar{\mathbf{S}}(\mathbf{A}, F)$ are in $\bar{\mathbf{S}}(\mathbf{T}(F), F)$. Strong multi-objective consistency additionally means that the elements of $\bar{\mathbf{S}}(\mathbf{T}(F), F)$ are not only minimal in $\bar{\mathbf{S}}(\mathbf{A}, F)$, but that they are the (unique) least element of $\bar{\mathbf{S}}(\mathbf{A}, F)$. Due to the uniqueness of a least element, $\bar{\mathbf{S}}(\mathbf{T}(F), F)$ is a singleton. The strict version of strong multi-objective consistency additionally means that $\bar{\mathbf{S}}(\mathbf{T}(F), F) \not\subseteq \bar{\mathbf{S}}(\mathbf{A} \setminus \mathbf{T}(F), F)$. See Figure 1 for an illustration of the two situations. Clearly, (strict) strong consistency implies (strict) weak consistency. If we omit the qualifiers “weak” or “strong”, we refer to the strong version.

Multi-objective consistency translates the essence of consistency as a ‘truth serum’, or being

incentive compatible, to the multivariate realm. No matter what forecast $\mathbf{r} \in \mathbf{A}$ an agent issues, they would be better off (or at least not worse off) in expectation when issuing a correctly specified functional value $\mathbf{t} \in \mathbf{T}(F)$. (In the weak version, if an agent issued some $\mathbf{t} \in \mathbf{T}(F)$, they would not be better off with some other forecast.) Strictness means that any action $\mathbf{r} \in \mathbf{A} \setminus \mathbf{T}(F)$ leads to a worse outcome in expectation than the truth $\mathbf{t} \in \mathbf{T}(F)$. So this minimal requirement of honouring truthful forecasts is preserved.

Remark 3.6. To the best of our knowledge, the notion of multi-objective scoring functions with the related concepts is novel to the forecast evaluation literature. However, it has some connections to the concept of *forecast dominance* introduced by Ehm *et al.* (2016, Section 3.2). Let $(S_i)_{i \in \mathcal{I}}$ be some class of univariate (strictly) $\mathcal{F}$ -consistent scoring functions $S_i: \mathbf{A} \times \mathbf{O} \rightarrow \mathbb{R}$ for some functional $\mathbf{T}: \mathcal{F} \rightarrow \mathcal{P}(\mathbf{A})$. Adapting their definition slightly, we say that a forecast $\mathbf{r} \in \mathbf{A}$ dominates $\tilde{\mathbf{r}} \in \mathbf{A}$ relative to $(S_i)_{i \in \mathcal{I}}$ if $\bar{S}_i(\mathbf{r}, F) \leq \bar{S}_i(\tilde{\mathbf{r}}, F)$ for all $F \in \mathcal{F}$ and for all $i \in \mathcal{I}$. In our terminology, we could phrase this forecast ranking in terms of a multivariate score $\mathbf{S}: \mathbf{A} \times \mathbf{O} \rightarrow (\mathbb{R}^{\mathcal{I}}, \preceq)$, $\mathbf{S}(\mathbf{r}, \mathbf{y}) = (S_i(\mathbf{r}, \mathbf{y}))_{i \in \mathcal{I}}$, where $\preceq$ is the usual componentwise partial order. By virtue of mixture representations, Ehm *et al.* (2016) impressively demonstrate that for quantiles and expectiles, one can rephrase forecast dominance with respect to (nearly) all consistent scoring functions equivalently in terms of extremal or elementary scores $(S_\theta)_{\theta \in \mathbb{R}} \subseteq (S_i)_{i \in \mathcal{I}}$. For this situation, we can easily construct a multi-objective score $\mathbf{S} = (S_\theta)_{\theta \in \mathbb{R}}$, mapping to the function space $\mathbb{R}^{\mathbb{R}}$ equipped with the componentwise order. Ehm *et al.* (2016) provide numerous instances of comparable and incomparable forecast rankings. This example provides an easy construction for a (strictly) multi-objective $\mathcal{F}$ -consistent score, $\mathbf{S} = (S_\theta)_{\theta \in \mathbb{R}}$, if $\mathbf{T}$ is elicitable.

For verifying observations $(\mathbf{y}_t)_{t=1, \dots, n}$, our Definition 3.5 suggests to compare two sequences of forecasts $(\mathbf{r}_{t,(i)})_{t=1, \dots, n}$ ($i = 1, 2$) via their average scores $\bar{\mathbf{S}}_{1n} = \frac{1}{n} \sum_{t=1}^n \mathbf{S}(\mathbf{r}_{t,(1)}, \mathbf{y}_t)$ and $\bar{\mathbf{S}}_{2n} = \frac{1}{n} \sum_{t=1}^n \mathbf{S}(\mathbf{r}_{t,(2)}, \mathbf{y}_t)$. However, while using classical univariate scores always leads to a conclusive forecast ranking (ignoring questions of statistical significance for a moment), the presence of a *partial* order may lead to inconclusive rankings, particularly if neither of the two forecasts is correctly specified. For instance, if $\bar{\mathbf{S}}_{1n} = (1, 4)$ and $\bar{\mathbf{S}}_{2n} = (4, 1)$ (e.g., for the bivariate systemic risk scores of Theorem 4.4), then neither $\bar{\mathbf{S}}_{1n} \preceq \bar{\mathbf{S}}_{2n}$ nor $\bar{\mathbf{S}}_{2n} \preceq \bar{\mathbf{S}}_{1n}$ in the componentwise order on $\mathbb{R}^2$.

3.3. Multi-objective elicitability with respect to the lexicographic order and conditional elicitability

To overcome the issue of inconclusive forecast rankings, we equip $\mathbb{R}^{\mathcal{I}}$ with a total order. In a total order, all elements are comparable. Hence, the notions of weak and strong multi-objective consistency / elicibility coincide and a distinction is redundant. A promising choice is the lexicographic order. For our purposes, and to ease the exposition, it suffices to consider $\mathbb{R}^2$ equipped with the lexicographic order $\preceq_{\text{lex}}$. On $\mathbb{R}^2$ it holds that $(x_1, x_2) \preceq_{\text{lex}} (y_1, y_2)$ if $x_1 < y_1$ or if $(x_1 = y_1 \text{ and } x_2 \leq y_2)$. The lexicographic order is widely used for preference relations in

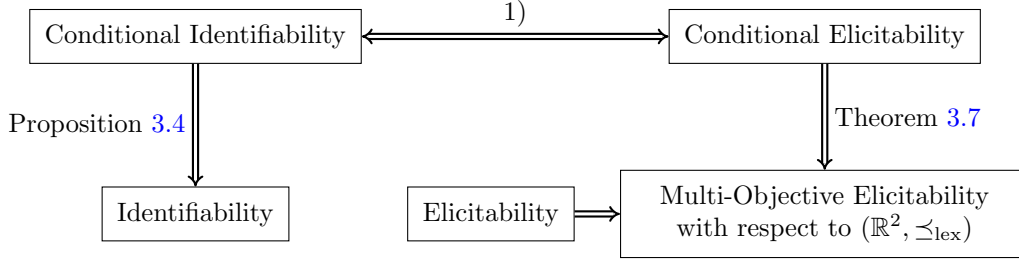


Figure 2: Illustration of the most important structural results for real-valued (risk) functionals $\mathbf{T}_1$ and $\mathbf{T}_2$. The equivalence 1) follows from Steinwart *et al.* (2014) under some regularity conditions.

microeconomics. In particular, it is well-known for the fact that it cannot be represented by a real-valued utility function (Mas-Colell *et al.*, 1995, Chapter 3.C).

There are at least two reasons for our choice of the lexicographic order. First, as a total order, it allows for conclusive forecast rankings, which we exploit in our comparative backtests in Section 5. For instance, if as above $\bar{\mathbf{S}}_{1n} = (1, 4)$ and $\bar{\mathbf{S}}_{2n} = (4, 1)$ for the systemic risk scores, then the inconclusive ranking is resolved because $\bar{\mathbf{S}}_{1n} \preceq_{\text{lex}} \bar{\mathbf{S}}_{2n}$. Second, the lexicographic order opens the way to the following analogue of Proposition 3.4.

Theorem 3.7. *If $\mathbf{T}_2$ is conditionally elicitable with $\mathbf{T}_1$ on $\mathcal{F}$ and $\mathbf{T}_1(F)$ is a singleton for all $F \in \mathcal{F}$, then the pair $(\mathbf{T}_1, \mathbf{T}_2)$ is multi-objective elicitable on $\mathcal{F}$ with respect to the lexicographic order $\preceq_{\text{lex}}$ on $\mathbb{R}^2$.*

Theorem 3.7 is almost a direct analogue of Proposition 3.4, with the intriguing exception that $\mathbf{T}_1$ is assumed to be a singleton on $\mathcal{F}$ in Theorem 3.7; see Remark A.1 for further details. 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 $(\text{VaR}_\alpha, \text{ES}_\alpha)$ in Subsection B.1. There, we also provide further examples of conditionally elicitable functionals failing to be elicitable in the traditional sense, but which are multi-objective elicitable with respect to $(\mathbb{R}^2, \preceq_{\text{lex}})$. Section B.2 elaborates on how the multi-objective scores can be considered a “generalised antiderivative” of a k -dimensional identification function with relaxed symmetry conditions. Section B.3 provides further details on comparing misspecified forecasts under multi-objective elicibility, and on the sensitivity of multi-objective scores with respect to increasing information sets. Finally, Section B.4 revisits a powerful necessary condition for identifiability and elicibility, namely the Convex Level Sets (CxLS) property, for multi-objective scores.

The proof of Theorem 3.7 explicitly exploits the asymmetric structure of the lexicographic order which fits well with the asymmetric notion of conditional elicibility, where the roles of $\mathbf{T}_1$ and $\mathbf{T}_2$ may not be changed. In particular, in the setup of Theorem 3.7, the pair $(\mathbf{T}_2, \mathbf{T}_1)$ is generally not multi-objective elicitable on $\mathcal{F}$ with respect to the lexicographic order $\preceq_{\text{lex}}$ on $\mathbb{R}^2$. More generally, we suspect that under the conditions of Theorem 3.7 the lexicographic order is the only order on $\mathbb{R}^2$ which renders $(\mathbf{T}_2, \mathbf{T}_1)$ multi-objective elicitable.

In the rest of the paper, we focus on multi-objective scores mapping to $(\mathbb{R}^2, \preceq_{\text{lex}})$. We use

the results of Section 3, summarised in Figure 2, extensively 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 risk measures $\text{CoVaR}_{\alpha|\beta}$, $\text{CoES}_{\alpha|\beta}$ and MES_β 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 the CxLS property (which remains necessary for multi-objective scores; see Proposition B.4) is violated in each case.

Proposition 4.1. *For $\alpha, \beta \in (0, 1)$, $\text{CoVaR}_{\alpha|\beta}$, $\text{CoES}_{\alpha|\beta}$ and MES_β 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 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.

Proposition 4.1 can also be shown on other classes $\mathcal{F} \subseteq \mathcal{F}^0(\mathbb{R}^2)$ that are sufficiently rich, e.g., the classes containing all measures with finite support (see Remark E.1). Importantly, $\mathcal{F}$ must be convex and must not only consist of distributions with independent marginals.

4.2. Joint identifiability results

Section C establishes the conditional identifiability and conditional elicibility of $\text{CoVaR}_{\alpha|\beta}(Y|X)$, $(\text{CoVaR}_{\alpha|\beta}(Y|X), \text{CoES}_{\alpha|\beta}(Y|X))$, and $\text{MES}_\beta(Y|X)$ with $\text{VaR}_\beta(X)$, respectively. This, in combination with Proposition 3.4, immediately yields the following joint identifiability results, where for $\alpha, \beta \in (0, 1)$ and $p \in \{0\} \cup [1, \infty]$ we use the notation

$$\begin{aligned}\mathcal{F}_\alpha^p(\mathbb{R}) &:= \{F \in \mathcal{F}^p(\mathbb{R}) : F(\text{VaR}_\alpha(F)) = \alpha\}, \\ \mathcal{F}_{(\alpha)}^p(\mathbb{R}) &:= \{F \in \mathcal{F}_\alpha^p(\mathbb{R}) : F(\text{VaR}_\alpha(F) + \varepsilon) > \alpha \text{ for all } \varepsilon > 0\}, \\ \mathcal{F}_{(\beta)}^0(\mathbb{R}^2) &:= \{F_{X,Y} \in \mathcal{F}^0(\mathbb{R}^2) : F_X \in \mathcal{F}_{(\beta)}^0(\mathbb{R})\}.\end{aligned}\tag{4.1}$$

Theorem 4.2. *Let $\alpha, \beta \in (0, 1)$. Consider the strict $\mathcal{F}_{(\beta)}^0(\mathbb{R}^2)$ -identification function for $F_{X,Y} \mapsto \text{VaR}_\beta(F_X)$, $V^{\text{VaR}}: \mathbb{R} \times \mathbb{R}^2 \rightarrow \mathbb{R}$, $V^{\text{VaR}}(v, (x, y)) = \mathbb{1}\{x \leq v\} - \beta$.*

- (i) *On $\{F_{X,Y} \in \mathcal{F}^0(\mathbb{R}^2) : F_X \in \mathcal{F}_{(\beta)}^0(\mathbb{R}), F_{Y|X \geq \text{VaR}_\beta(X)} \in \mathcal{F}_{(\alpha)}^0(\mathbb{R})\}$, the pair $F_{X,Y} \mapsto (\text{VaR}_\beta(F_X), \text{CoVaR}_{\alpha|\beta}(F_{X,Y}))$ is identifiable with a strict identification function $\mathbf{V}^{(\text{VaR}, \text{CoVaR})}: \mathbb{R}^2 \times \mathbb{R}^2 \rightarrow \mathbb{R}^2$,*

$$\mathbf{V}^{(\text{VaR}, \text{CoVaR})}((v, c), (x, y)) = \begin{pmatrix} \mathbb{1}\{x \leq v\} - \beta \\ \mathbb{1}\{x > v\} [\mathbb{1}\{y \leq c\} - \alpha] \end{pmatrix}.\tag{4.2}$$

- (ii) On $\{F_{X,Y} \in \mathcal{F}^0(\mathbb{R}^2): F_X \in \mathcal{F}_{(\beta)}^0(\mathbb{R}), F_{Y|X \geq \text{VaR}_\beta(X)} \in \mathcal{F}_{(\alpha)}^1(\mathbb{R})\}$, the triplet $F_{X,Y} \mapsto (\text{VaR}_\beta(F_X), \text{CoVaR}_{\alpha|\beta}(F_{X,Y}), \text{CoES}_{\alpha|\beta}(F_{X,Y}))$ is identifiable with a strict identification function $\mathbf{V}^{(\text{VaR}, \text{CoVaR}, \text{CoES})}: \mathbb{R}^3 \times \mathbb{R}^2 \rightarrow \mathbb{R}^3$,

$$((v, c, e), (x, y)) \mapsto \begin{pmatrix} \mathbb{1}\{x \leq v\} - \beta \\ \mathbb{1}\{x > v\} [\mathbb{1}\{y \leq c\} - \alpha] \\ \mathbb{1}\{x > v\} \left[e - \frac{1}{1-\alpha} (y \mathbb{1}\{y > c\} + c(\mathbb{1}\{y \leq c\} - \alpha)) \right] \end{pmatrix}. \quad (4.3)$$

- (iii) On $\{F_{X,Y} \in \mathcal{F}^0(\mathbb{R}^2): F_X \in \mathcal{F}_{(\beta)}^0(\mathbb{R}), F_{Y|X \geq \text{VaR}_\beta(X)} \in \mathcal{F}^1(\mathbb{R})\}$, the pair $F_{X,Y} \mapsto (\text{VaR}_\beta(F_X), \text{MES}_\beta(F_{X,Y}))$ is identifiable with a strict identification function $\mathbf{V}^{(\text{VaR}, \text{MES})}: \mathbb{R}^2 \times \mathbb{R}^2 \rightarrow \mathbb{R}^2$,

$$\mathbf{V}^{(\text{VaR}, \text{MES})}((v, \mu), (x, y)) = \begin{pmatrix} \mathbb{1}\{x \leq v\} - \beta \\ \mathbb{1}\{x > v\} [\mu - y] \end{pmatrix}. \quad (4.4)$$

Following [Nolde and Ziegel \(2017\)](#), Theorem 4.2 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, $\mathbb{E}[\mathbf{V}(\mathbf{r}_t, (X_t, Y_t))] = \mathbf{0}$ for all $t \in \mathbb{N}$, or the more informative null of *conditional* calibration, $\mathbb{E}[\mathbf{V}(\mathbf{r}_t, (X_t, Y_t)) | \mathfrak{F}_{t-1}] = \mathbf{0}$ for all $t \in \mathbb{N}$. Here, the σ -algebra $\mathfrak{F}_{t-1}$ represents the information available to the forecaster at time $t-1$. Recall that the null of conditional calibration is equivalent to $\mathbb{E}[\varphi_{t-1} \mathbf{V}(\mathbf{r}_t, (X_t, Y_t))] = \mathbf{0}$ for all $\mathfrak{F}_{t-1}$ -measurable random vectors φ_{t-1} for all $t \in \mathbb{N}$. Unless $\mathfrak{F}_{t-1}$ is particularly simple, this approach is statistically not feasible, because a finite selection $\varphi_{1,t-1}, \dots, \varphi_{\ell,t-1}$ of $\mathfrak{F}_{t-1}$ -measurable random vectors is required in practice. Note that using only a finite collection amounts to testing a broader null than the original one of conditional calibration.

Remark 4.3. It is worth pointing out the improvements upon the substantial work of [Banulescu-Radu et al. \(2021\)](#), who develop—inter alia—traditional backtests of systemic risk measures. Translating their notation into ours, they propose in their equation (4) a one-dimensional identification function for the pair $(\text{VaR}_\beta(X), \text{CoVaR}_{\alpha|\beta}(Y|X))$ of the form $V: \mathbb{R}^2 \times \mathbb{R}^2 \rightarrow \mathbb{R}$,

$$V((v, c), (x, y)) = \mathbb{1}\{x > v\} \mathbb{1}\{y > c\} - (1 - \alpha)(1 - \beta). \quad (4.5)$$

[Banulescu-Radu et al. \(2021\)](#) point out that this is an identification function since $\bar{V}((\text{VaR}_\beta(X), \text{CoVaR}_{\alpha|\beta}(X|Y)), F_{X,Y}) = 0$. However, it fails to be strict, since, so long as $(1 - \alpha')(1 - \beta') = (1 - \alpha)(1 - \beta)$, we have $\bar{V}((\text{VaR}_{\beta'}(X), \text{CoVaR}_{\alpha'|\beta'}(X|Y)), F_{X,Y}) = 0$. Thus, the specific null of ‘correct’ unconditional calibration $\mathbb{E}[V(\mathbf{r}_t, (X_t, Y_t))] = 0$ with V from (4.5) is too broad, since it is satisfied for all $\mathbf{r}_t = (\text{VaR}_{\beta'}(X), \text{CoVaR}_{\alpha'|\beta'}(X|Y))$ with $(1 - \alpha')(1 - \beta') = (1 - \alpha)(1 - \beta)$. This implies only trivial power against such alternatives for the corresponding Wald-test for calibration. This is in stark contrast to the Wald-test employing our two-dimensional identification function in (4.2), where $\mathbb{E}[\mathbf{V}^{(\text{VaR}, \text{CoVaR})}(\mathbf{r}_t, (X_t, Y_t))] = \mathbf{0}$ if and only

if $\mathbf{r}_t = (\text{VaR}_\beta(X), \text{CoVaR}_{\alpha|\beta}(X|Y))$. We refer to Section G for a simulation study illustrating the possible detrimental effects of the non-strictness of (4.5). We stress that the non-strictness is not only problematic in the uncountably many cases where $(1 - \alpha')(1 - \beta') = (1 - \alpha)(1 - \beta)$, but in *all* cases where VaR is overpredicted and CoVaR underpredicted or the other way around. This is because by overpredicting (underpredicting) VaR and underpredicting (overpredicting) CoVaR, the different biases cancel out in the one-dimensional identification function in (4.5), thus leading to a loss of power in identifying misspecified forecasts; see Figure 9. Intuitively speaking, two moment conditions via a two-dimensional identification function are necessary to uniquely identify the two-dimensional pair $(\text{VaR}_\beta(X), \text{CoVaR}_{\alpha|\beta}(Y|X))$. This is in line with other two-dimensional strict identification functions for two-dimensional functionals in the literature, e.g. for Value at Risk and Expected Shortfall, see Fissler and Ziegel (2016); Nolde and Ziegel (2017).

4.3. Multi-objective elicibility results

Recall that $\text{VaR}_\beta(Y)$, $(\text{VaR}_\beta(Y), \text{ES}_\beta(Y))$ and $E(Y)$ are all elicitable. Surprisingly, their conditional counterparts $(\text{VaR}_\beta(X), \text{CoVaR}_{\alpha|\beta}(Y|X))$, $(\text{VaR}_\beta(X), \text{CoVaR}_{\alpha|\beta}(Y|X), \text{CoES}_{\alpha|\beta}(Y|X))$ and $(\text{VaR}_\beta(X), \text{MES}_\beta(Y|X))$ fail to be elicitable despite being identifiable (Section D). 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 elicibility results of Section C can hardly be used for forecast comparisons, unless the $\text{VaR}_\beta(X)$ forecasts are the same and correctly specified. However, the conditional elicibility in combination with Theorem 3.7 immediately yields the following novel joint multi-objective elicibility results with respect to the lexicographic order $\preceq_{\text{lex}}$ on $\mathbb{R}^2$. These results can readily be used for comparing systemic risk forecasts as detailed in Section 5. We again use the notation introduced in (4.1).

Theorem 4.4. *Let $\alpha, \beta \in (0, 1)$.*

- (i) *On $\mathcal{F} \subseteq \mathcal{F}_\beta^0(\mathbb{R}^2)$, the score $S^{\text{VaR}}: \mathbb{R} \times \mathbb{R}^2 \rightarrow \mathbb{R}$,*

$$S^{\text{VaR}}(v, (x, y)) = (\mathbf{1}\{x \leq v\} - \beta)h(v) - \mathbf{1}\{x \leq v\}h(x) + a^{\text{VaR}}(x, y) \quad (4.6)$$

is strictly $\mathcal{F}$ -consistent for $\mathcal{F} \ni F_{X,Y} \mapsto \text{VaR}_\beta(F_X)$, if $h: \mathbb{R} \rightarrow \mathbb{R}$ is strictly increasing and for all $v \in \mathbb{R}$, the function $(x, y) \mapsto a^{\text{VaR}}(x, y) - \mathbf{1}\{x \leq v\}h(x)$ is $\mathcal{F}$ -integrable.

- (ii) *On $\mathcal{F} \subseteq \{F_{X,Y} \in \mathcal{F}^0(\mathbb{R}^2): F_X \in \mathcal{F}_{(\beta)}^0(\mathbb{R}), F_{Y|X \geq \text{VaR}_\beta(X)} \in \mathcal{F}_\alpha^0(\mathbb{R})\}$, the pair $\mathcal{F} \ni F_{X,Y} \mapsto (\text{VaR}_\beta(F_X), \text{CoVaR}_{\alpha|\beta}(F_{X,Y}))$ is multi-objective elicitable with respect to $(\mathbb{R}^2, \preceq_{\text{lex}})$. A strictly $\mathcal{F}$ -consistent multi-objective scoring function $\mathbf{S}^{(\text{VaR}, \text{CoVaR})}: \mathbb{R}^2 \times \mathbb{R}^2 \rightarrow (\mathbb{R}^2, \preceq_{\text{lex}})$*

is given by

$$\mathbf{S}^{(\text{VaR}, \text{CoVaR})}((v, c), (x, y)) = \begin{pmatrix} S^{\text{VaR}}(v, (x, y)) \\ S_v^{\text{CoVaR}}(c, (x, y)) \end{pmatrix}, \quad (4.7)$$

$$S_v^{\text{CoVaR}}(c, (x, y)) = \mathbb{1}\{x > v\} \left[(\mathbb{1}\{y \leq c\} - \alpha)g(c) - \mathbb{1}\{y \leq c\}g(y) + a(y) \right] + a^{\text{CoVaR}}(x, y),$$

where $g: \mathbb{R} \rightarrow \mathbb{R}$ is strictly increasing and S_v^{CoVaR} is $\mathcal{F}$ -integrable for all $v \in \mathbb{R}$.

- (iii) On $\mathcal{F} \subseteq \{F_{X,Y} \in \mathcal{F}^0(\mathbb{R}^2): F_X \in \mathcal{F}_{(\beta)}^0(\mathbb{R}), F_{Y|X \geq \text{VaR}_\beta(X)} \in \mathcal{F}_\alpha^1(\mathbb{R})\}$, the triplet $\mathcal{F} \ni F_{X,Y} \mapsto (\text{VaR}_\beta(F_X), \text{CoVaR}_{\alpha|\beta}(F_{X,Y}), \text{CoES}_{\alpha|\beta}(F_{X,Y}))$ is multi-objective elicitable with respect to $(\mathbb{R}^2, \preceq_{\text{lex}})$. A strictly $\mathcal{F}$ -consistent multi-objective scoring function $\mathbf{S}^{(\text{VaR}, \text{CoVaR}, \text{CoES})}: \mathbb{R}^3 \times \mathbb{R}^2 \rightarrow (\mathbb{R}^2, \preceq_{\text{lex}})$ is given by

$$\mathbf{S}^{(\text{VaR}, \text{CoVaR}, \text{CoES})}((v, c, e), (x, y)) = \begin{pmatrix} S^{\text{VaR}}(v, (x, y)) \\ S_v^{(\text{CoVaR}, \text{CoES})}((c, e), (x, y)) \end{pmatrix}, \quad (4.8)$$

$$S_v^{(\text{CoVaR}, \text{CoES})}((c, e), (x, y)) = \mathbb{1}\{x > v\} \left[(\mathbb{1}\{y \leq c\} - \alpha)g(c) - \mathbb{1}\{y \leq c\}g(y) \right. \\ \left. + \phi'(e) \left(e - \frac{1}{1-\alpha} (y \mathbb{1}\{y > c\} + c(\mathbb{1}\{y \leq c\} - \alpha)) \right) - \phi(e) + a(y) \right] + a^{\text{CoES}}(x, y),$$

where $g: \mathbb{R} \rightarrow \mathbb{R}$ is increasing, $\phi: \mathbb{R} \rightarrow \mathbb{R}$ is strictly convex with subgradient $\phi' < 0$, and $S_v^{(\text{CoVaR}, \text{CoES})}$ is $\mathcal{F}$ -integrable for all $v \in \mathbb{R}$.

- (iv) On $\mathcal{F} \subseteq \{F_{X,Y} \in \mathcal{F}^0(\mathbb{R}^2): F_X \in \mathcal{F}_{(\beta)}^0(\mathbb{R}), F_{Y|X \geq \text{VaR}_\beta(X)} \in \mathcal{F}^1(\mathbb{R})\}$, the pair $F_{X,Y} \mapsto (\text{VaR}_\beta(F_X), \text{MES}_\beta(F_{X,Y}))$ is multi-objective elicitable with respect to $(\mathbb{R}^2, \preceq_{\text{lex}})$. A strictly $\mathcal{F}$ -consistent multi-objective scoring function $\mathbf{S}^{(\text{VaR}, \text{MES})}: \mathbb{R}^2 \times \mathbb{R}^2 \rightarrow (\mathbb{R}^2, \preceq_{\text{lex}})$ is given by

$$\mathbf{S}^{(\text{VaR}, \text{MES})}((v, \mu), (x, y)) = \begin{pmatrix} S^{\text{VaR}}(v, (x, y)) \\ S_v^{\text{MES}}(\mu, (x, y)) \end{pmatrix}, \quad (4.9)$$

$$S_v^{\text{MES}}(\mu, (x, y)) = \mathbb{1}\{x > v\} \left[\phi'(\mu)(\mu - y) - \phi(\mu) + a(y) \right] + a^{\text{MES}}(x, y),$$

where $\phi: \mathbb{R} \rightarrow \mathbb{R}$ is strictly convex with subgradient ϕ' and S_v^{MES} is $\mathcal{F}$ -integrable for all $v \in \mathbb{R}$.

The choice of the functions $a, a^{\text{VaR}}, a^{\text{CoVaR}}, a^{\text{CoES}}, a^{\text{MES}}$ in Theorem 4.4 is inessential. It may only influence the integrability of the scoring functions on the one hand and it can control the sign of the scores on the other hand. For the remaining functions in Theorem 4.4, one could choose the standard choices, e.g., the identity for h and g in (4.6), (4.7) and (4.8), giving rise to the common ‘pinball loss’, or $\phi(y) = y^2$ in (4.9) leading to the square loss. For further possible

Table 1: Overview of properties of (systemic) risk measures. $\checkmark(\times)$ indicates that property does (does not) apply.

Risk Measure	Identifiability	Elicitability	Multi-objective Elicitability
$\text{CoVaR}_{\alpha \beta}(Y X)$	$\times$	$\times$	$\times$
$\text{CoES}_{\alpha \beta}(Y X)$	$\times$	$\times$	$\times$
$\text{MES}_{\alpha \beta}(Y X)$	$\times$	$\times$	$\times$
$(\text{VaR}_\beta(X), \text{CoVaR}_{\alpha \beta}(Y X))$	$\checkmark$	$\times$	$\checkmark$
$(\text{VaR}_\beta(X), \text{CoVaR}_{\alpha \beta}(Y X), \text{CoES}_{\alpha \beta}(Y X))$	$\checkmark$	$\times$	$\checkmark$
$(\text{VaR}_\beta(X), \text{MES}_{\alpha \beta}(Y X))$	$\checkmark$	$\times$	$\checkmark$
$\text{VaR}_\alpha(Y)$	$\checkmark$	$\checkmark$	$\checkmark$
$\text{ES}_\alpha(Y)$	$\times$	$\times$	$\times$
$(\text{VaR}_\alpha(Y), \text{ES}_\alpha(Y))$	$\checkmark$	$\checkmark$	$\checkmark$

choices, especially for ϕ in (4.8), we refer to Subsection F.3.

Table 1 summarises the results of Section 4. For purposes of comparison, the final three rows display the properties of $\text{VaR}_\beta(Y)$ and $\text{ES}_\beta(Y)$. The multi-objective elicibility of the pair $(\text{VaR}_\beta(Y), \text{ES}_\beta(Y))$ follows from Example B.1. Table 1 highlights the structural difference between $\text{VaR}_\beta(Y)$, $\text{E}(Y)$, $(\text{VaR}_\beta(Y), \text{ES}_\beta(Y))$ on the one hand and their conditional counterparts $(\text{VaR}_\beta(X), \text{CoVaR}_{\alpha|\beta}(Y|X))$, $(\text{VaR}_\beta(X), \text{MES}_{\alpha|\beta}(Y|X))$, and $(\text{VaR}_\beta(X), \text{CoVaR}_{\alpha|\beta}(Y|X), \text{CoES}_{\alpha|\beta}(Y|X))$ on the other hand. While elicibility in the usual sense holds for the former, it fails for the latter. This highlights the importance of the newly introduced concept of multi-objective elicibility, which allows for comparative backtests. Section 5 shows how to implement comparative backtests with the multi-objective scores of Theorem 4.4.

5. Diebold–Mariano tests for multi-objective scores

5.1. Two-sided tests

Diebold and Mariano (1995) propose to use formal hypothesis tests to account for sampling uncertainty in forecast comparisons. These so-called Diebold–Mariano (DM) tests are widely used in empirical forecast comparisons and continue to be studied in the theoretical literature. However, consistent with the extant notion of strict consistency, DM tests have hitherto relied on scalar scoring functions. Thus, here we show how to use our two-dimensional multi-objective scores from Theorem 4.4 in DM tests, with a special focus on the implications caused by the lexicographic order.

To that end, denote by $\mathbf{S} = (S_1, S_2)'$ one of the multi-objective scores of Theorem 4.4. Let $\{\mathbf{r}_{t,(1)}\}_{t=1,\dots,n}$ and $\{\mathbf{r}_{t,(2)}\}_{t=1,\dots,n}$ be the appertaining competing sequences of forecasts (e.g., if $\mathbf{S} = \mathbf{S}^{(\text{VaR}, \text{CoVaR})}$, then $\mathbf{r}_{t,(i)} = (\widehat{\text{VaR}}_{t,(i)}, \widehat{\text{CoVaR}}_{t,(i)})$ for $i = 1, 2$). The verifying observations are $\{(X_t, Y_t)\}_{t=1,\dots,n}$. We compare the two forecasts via the (bivariate) score differences $\mathbf{d}_t := (d_{1t}, d_{2t})' := \mathbf{S}(\mathbf{r}_{t,(1)}, (X_t, Y_t)) - \mathbf{S}(\mathbf{r}_{t,(2)}, (X_t, Y_t))$. The two-sided null hypothesis is that both forecasts predict equally well on average, i.e., $H_0^-: \mathbb{E}[\bar{\mathbf{d}}_n] = \mathbf{0}$ for all $n = 1, 2, \dots$, where

$\bar{\mathbf{d}}_n := (\bar{d}_{1n}, \bar{d}_{2n})' := \frac{1}{n} \sum_{t=1}^n \mathbf{d}_t$. (Along the lines of [Giacomini and White \(2006\)](#), one can also test the conditional null hypothesis $H_0^* =: \mathbb{E}[\mathbf{d}_t \mid \mathfrak{F}_{t-1}] = \mathbf{0}$ for all $t = 1, 2, \dots$, where the σ -algebra $\mathfrak{F}_{t-1}$ contains all information available at time $t-1$.) We test H_0^- using the Wald-type test statistic

$$\mathcal{T}_n = n \bar{\mathbf{d}}_n' \hat{\boldsymbol{\Omega}}_n^{-1} \bar{\mathbf{d}}_n, \quad (5.1)$$

where $\hat{\boldsymbol{\Omega}}_n$ is some consistent estimator of the variance-covariance matrix $\boldsymbol{\Omega}_n = \text{Var}(\sqrt{n} \bar{\mathbf{d}}_n)$ under the null hypothesis (i.e., in the componentwise norm, $\|\hat{\boldsymbol{\Omega}}_n - \boldsymbol{\Omega}_n\| \xrightarrow{P} 0$, as $n \rightarrow \infty$). To account for possible autocorrelation in the sequence $\{\mathbf{d}_t\}_{t=1, \dots, n}$, one can use

$$\begin{aligned} \hat{\boldsymbol{\Omega}}_n = \begin{pmatrix} \hat{\sigma}_{11,n} & \hat{\sigma}_{12,n} \\ \hat{\sigma}_{12,n} & \hat{\sigma}_{22,n} \end{pmatrix} &= \frac{1}{n} \sum_{t=1}^n (\mathbf{d}_t - \bar{\mathbf{d}}_n)(\mathbf{d}_t - \bar{\mathbf{d}}_n)' \\ &+ \frac{1}{n} \sum_{h=1}^{m_n} w_{n,h} \sum_{t=h+1}^n \left[(\mathbf{d}_t - \bar{\mathbf{d}}_n)(\mathbf{d}_{t-h} - \bar{\mathbf{d}}_n)' + (\mathbf{d}_{t-h} - \bar{\mathbf{d}}_n)(\mathbf{d}_t - \bar{\mathbf{d}}_n)' \right], \end{aligned} \quad (5.2)$$

where $m_n \rightarrow \infty$ is a sequence of integers satisfying $m_n = o(n^{1/4})$, and $w_{n,h}$ is a uniformly bounded scalar triangular array with $w_{n,h} \rightarrow 1$, as $n \rightarrow \infty$, for all $h = 1, \dots, m_n$; see [White \(2001\)](#) for detail. Under the assumption that $\{\mathbf{d}_t\}_{t=1, \dots, n}$ does not exhibit autocorrelation under the null, m_n can be set to 0 such that $\hat{\boldsymbol{\Omega}}_n$ is simply the sample variance-covariance matrix.

Theorem 5.1. *Suppose that $\boldsymbol{\Omega}_n \rightarrow \boldsymbol{\Omega}$, as $n \rightarrow \infty$, where $\boldsymbol{\Omega}$ is positive definite. Then, under technical Assumption [E.2](#) (see Supplement Section [E](#)), it holds under H_0^- that*

$$\sqrt{n} \hat{\boldsymbol{\Omega}}_n^{-1/2} \bar{\mathbf{d}}_n \xrightarrow{d} N(\mathbf{0}, \mathbf{I}_{2 \times 2}), \quad \text{as } n \rightarrow \infty,$$

where $\mathbf{I}_{2 \times 2}$ denotes the (2×2) -identity matrix. In particular, $\mathcal{T}_n \xrightarrow{d} \chi_2^2$, as $n \rightarrow \infty$, where χ_2^2 denotes a χ^2 -distribution with 2 degrees of freedom.

Thus, we reject H_0^- at significance level ν , if $\mathcal{T}_n > \chi_{2,1-\nu}^2$, where $\chi_{2,1-\nu}^2$ is the $(1-\nu)$ -quantile of the χ_2^2 -distribution. A typical non-rejection region in terms of $\bar{d}_{1n}$ and $\bar{d}_{2n}$ is sketched in [Figure 3](#) (a), and has the well-known ellipse shape. For brevity, we leave out a formal investigation of our test under the alternative. We mention, however, that consistency of our test can be established along the lines of [Giacomini and White \(2006\)](#).

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 $\{\mathbf{r}_{t,(2)}\}_{t=1, \dots, n}$ over some benchmark forecasts $\{\mathbf{r}_{t,(1)}\}_{t=1, \dots, n}$. These different forecasts could stem from two different internal 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 back-testing approach of [Fissler et al. \(2016\)](#)—the financial institution has the onus of proof to show

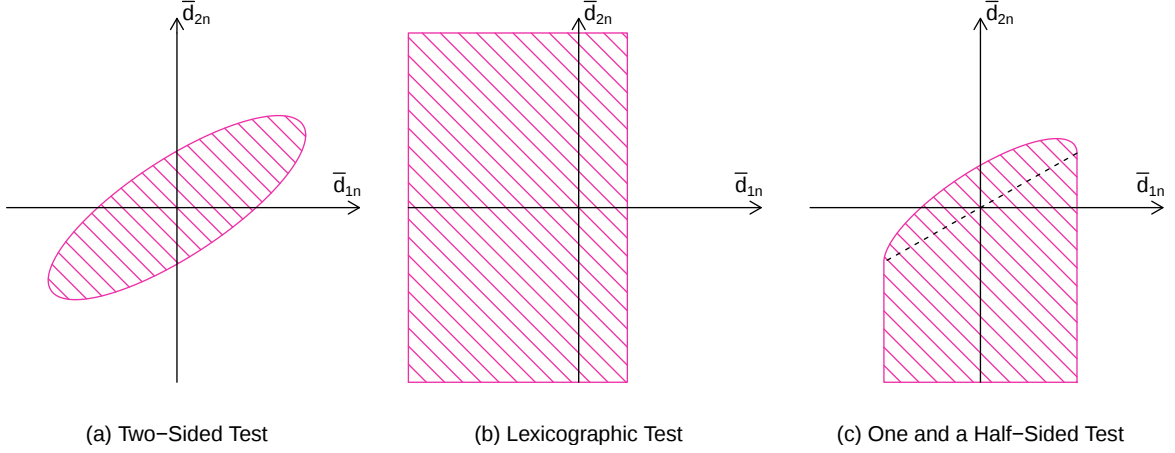


Figure 3: Non-rejection regions for two-sided DM test in (a), lexicographic DM test in (b), and one and a half-sided DM test in (c).

the superiority of its internal model over the standard model. In such situations, it is tempting to test the null hypothesis $E[\bar{\mathbf{d}}_n] \preceq_{\text{lex}} \mathbf{0}$, which is equivalent to

$$E[\bar{d}_{1n}] < 0 \quad \text{or} \quad \left(E[\bar{d}_{1n}] = 0 \quad \text{and} \quad E[\bar{d}_{2n}] \leq 0 \right). \quad (5.3)$$

Here, the goal would be to reject the null of better or, at least, equally good benchmark forecasts as evidence of the superiority of $\{\mathbf{r}_{t,(2)}\}_{t=1,\dots,n}$. However, as under standard conditions (see Theorem 5.1), $\sqrt{n}(\bar{d}_{1n}, \bar{d}_{2n})'$ has an asymptotic bivariate normal distribution, the probability that $\bar{d}_{1n} = 0$ and $\bar{d}_{2n} \leq 0$ vanishes for large sample sizes. Thus, testing (5.3) amounts to a test of the null $E[\bar{d}_{1n}] < 0$, i.e., that the benchmark VaR forecasts are superior. This null can be tested via $\mathcal{T}_{1n} = \sqrt{n}\bar{d}_{1n}/\hat{\sigma}_{11,n}^{1/2}$, where we reject (5.3) at significance level $\nu \in (0, 1)$ when $\mathcal{T}_{1n} > \Phi^{-1}(1 - \nu)$. The corresponding non-rejection region is sketched in Figure 3 (b). In particular, we would reject $E[\bar{\mathbf{d}}_n] \preceq_{\text{lex}} \mathbf{0}$ solely based on the predictive performance of the $\text{VaR}(X_t)$ component. In other words, once $E[\bar{d}_{1n}] < 0$ is rejected such that the internal VaR forecasts are superior, the internal forecasts $\mathbf{r}_{t,(1)}$ 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 $E[\bar{d}_{1n}] = 0$, our suggested null hypothesis takes the form

$$H_0^{\preceq_{\text{lex}}} : E[\bar{d}_{1n}] = 0 \quad \text{and} \quad E[\bar{d}_{2n}] \leq 0 \quad \text{for all } n = 1, 2, \dots$$

We demonstrate below how to interpret a rejection of this null. The region in $\mathbb{R}^2$ pertaining to $H_0^{\preceq_{\text{lex}}}$ is the lower part of the vertical axis in Figure 3 (c).

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

$$H_0^{(c)}: \mathbb{E}[\bar{d}_{1n}] = 0 \quad \text{and} \quad \mathbb{E}[\bar{d}_{2n}] = c \quad \text{for all } n = 1, 2, \dots$$

For each individual $c \leq 0$, this can be tested using the Wald-type test statistic $\mathcal{T}_n^{(c)}$, where $\mathcal{T}_n^{(c)}$ is defined similarly as $\mathcal{T}_n$, only with $\mathbf{d}_t$ replaced by $\mathbf{d}_t^{(c)} = (d_{1t}, d_{2t} - c)'$. (Note that this substitution leaves $\hat{\boldsymbol{\Omega}}_n$ unaffected.) Thus, we reject $H_0^{\leq \text{lex}}$ if and only if

$$\mathcal{T}_n^{(c)} > \chi_{2,1-\tilde{\nu}}^2 \quad \text{for all } c \leq 0, \quad (5.4)$$

where $\tilde{\nu} \in (0, 1)$. The area associated with the appertaining non-rejection region is shaded in pink in Figure 3 (c). The rejection condition (5.4) is of course equivalent to $\mathcal{T}_n^{\text{OS}} := \inf_{c \leq 0} \mathcal{T}_n^{(c)} = \mathcal{T}_n^{(c^*)} > \chi_{2,1-\tilde{\nu}}^2$, where the solution $c^* := \min \{0, \bar{d}_{2n} - (\hat{\sigma}_{12,n}/\hat{\sigma}_{11,n})\bar{d}_{1n}\}$ follows from a simple quadratic minimisation problem. Hence,

$$\mathcal{T}_n^{\text{OS}} = n \left(\bar{d}_{1n}, \max \{ \bar{d}_{2n}, (\hat{\sigma}_{12,n}/\hat{\sigma}_{11,n})\bar{d}_{1n} \} \right) \hat{\boldsymbol{\Omega}}_n^{-1} \begin{pmatrix} \bar{d}_{1n} \\ \max \{ \bar{d}_{2n}, (\hat{\sigma}_{12,n}/\hat{\sigma}_{11,n})\bar{d}_{1n} \} \end{pmatrix}. \quad (5.5)$$

To illustrate this graphically, note that the line defined by $z_2 = (\hat{\sigma}_{12,n}/\hat{\sigma}_{11,n})z_1$ —indicated by the dashed line in Figure 3 (c)—passes through the extremal (negative and positive) horizontal points of the ellipse. Thus, if $\bar{d}_{2n} > (\hat{\sigma}_{12,n}/\hat{\sigma}_{11,n})\bar{d}_{1n}$, (5.5) parametrizes the upper half of the tilted ellipse, and if $\bar{d}_{2n} \leq (\hat{\sigma}_{12,n}/\hat{\sigma}_{11,n})\bar{d}_{1n}$ the area below. In our numerical experiments, we use $\mathcal{T}_n^{\text{OS}}$ to test $H_0^{\leq \text{lex}}$.

The next proposition shows that rejecting $H_0^{\leq \text{lex}}$ when $\mathcal{T}_n^{\text{OS}} > \chi_{2,1-\tilde{\nu}}^2$ leads to a test of level $\nu = 1/2[1 + \tilde{\nu} - F_{\chi_1^2}(\chi_{2,1-\tilde{\nu}}^2)]$, where $F_{\chi_1^2}$ denotes the cdf of a χ_1^2 -distribution.

Proposition 5.2. *Under the conditions of Theorem 5.1, rejecting $H_0^{\leq \text{lex}}$ if $\mathcal{T}_n^{\text{OS}} > \chi_{2,1-\tilde{\nu}}^2$ leads to an asymptotic size ν -test with $\nu = 1/2[1 + \tilde{\nu} - F_{\chi_1^2}(\chi_{2,1-\tilde{\nu}}^2)]$. That is,*

$$\sup_{c \leq 0} \lim_{n \rightarrow \infty} \mathbb{P} \left\{ H_0^{\leq \text{lex}} \text{ is rejected based on } \mathcal{T}_n^{\text{OS}} \mid H_0^{(c)} \text{ holds} \right\} = \nu.$$

Remark 5.3. For a test of $H_0^{\leq \text{lex}}$ with desired significance level of ν , one can determine the required level $\tilde{\nu}$ from $\nu = 1/2[1 + \tilde{\nu} - F_{\chi_1^2}(\chi_{2,1-\tilde{\nu}}^2)]$ using standard root-finding algorithms. E.g., if $\nu = 1\%$ / $\nu = 5\%$ / $\nu = 10\%$, then $\tilde{\nu} = 1.60\%$ / $\tilde{\nu} = 7.66\%$ / $\tilde{\nu} = 14.9\%$.

Remark 5.4. To compare systemic risk forecasts on an equal footing, it may be desirable to compare them based on the same marginal model being used and, hence, based on the same VaR forecasts $\widehat{\text{VaR}}_t = \widehat{\text{VaR}}_{t,(1)} = \widehat{\text{VaR}}_{t,(2)}$. In this case, differences in predictive ability can be attributed solely to the different dependence models; see, e.g., [Nolde and Zhang \(2020\)](#) and [Hoga \(2021\)](#). (Note that Theorem 5.1 no longer applies for identical VaR forecasts, since the limit of the covariance matrix $\boldsymbol{\Omega}$ is only positive semi-definite.) When $\widehat{\text{VaR}}_t = \widehat{\text{VaR}}_{t,(1)} = \widehat{\text{VaR}}_{t,(2)}$, testing for $\mathbb{E}[\bar{d}_{1n}] = 0$ is redundant, and H_0^- and $H_0^{\leq \text{lex}}$ reduce to $\mathbb{E}[\bar{d}_{2n}] = 0$ and $\mathbb{E}[\bar{d}_{2n}] \leq 0$,

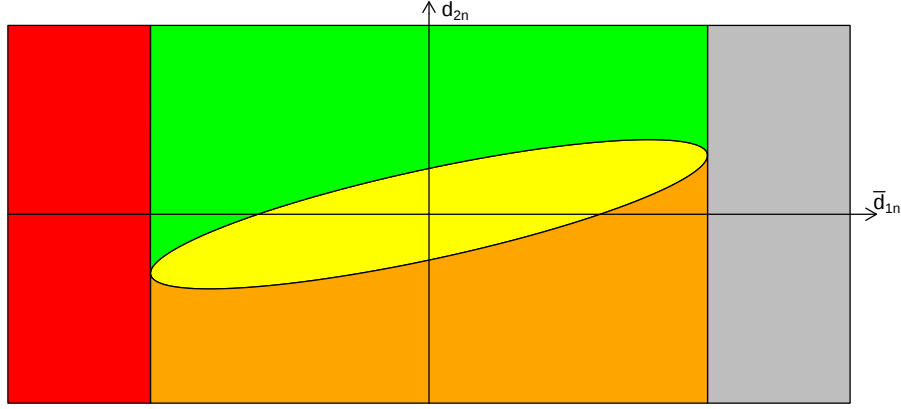


Figure 4: Interpretation of test decisions. Yellow zone corresponds to ellipse. Red (grey) zone corresponds to the rectangle to the left (right) of the ellipse. Green (orange) zone corresponds to area immediately above (below) the ellipse.

respectively. These hypotheses can be tested using $\mathcal{T}_{2n} = \sqrt{n}\bar{d}_{2n}/\hat{\sigma}_{22,n}^{1/2}$. Under $E[\bar{d}_{2n}] = 0$, $\mathcal{T}_{2n}$ converges to an $N(0, 1)$ -limit when restricting the assumptions of Theorem 5.1 to the sequence d_{2t} . Hence, we would reject $E[\bar{d}_{2n}] = 0$ (or $E[\bar{d}_{2n}] \leq 0$) at significance level $\nu \in (0, 1)$, if $|\mathcal{T}_{2n}| > \Phi^{-1}(1 - \nu/2)$ (or $\mathcal{T}_{2n} > \Phi^{-1}(1 - \nu)$).

Figure 4 presents a schematic for interpreting test decisions on $H_0^{\leq \text{lex}}$ depending on the values of $\bar{d}_{1n}$ and $\bar{d}_{2n}$. In a regulatory context, it extends the three-zone traffic-light classification of the [Bank for International Settlements \(2019, Paragraph 99.39\)](#):

“The green zone corresponds to backtesting results that do not themselves suggest a problem with the quality or accuracy of a bank’s model. The yellow zone encompasses results that do raise questions in this regard, but where such a conclusion is not definitive. The red zone indicates a backtesting result that almost certainly indicates a problem with a bank’s risk model.”

The union of the yellow and the orange areas in Figure 4 corresponds to the non-rejection region depicted in Figure 3 (c). By symmetry, the union of the yellow and the green areas is the non-rejection region associated with the null

$$H_0^{\geq \text{lex}}: E[\bar{d}_{1n}] = 0 \quad \text{and} \quad E[\bar{d}_{2n}] \geq 0 \quad \text{for all } n = 1, 2, \dots$$

Thus, for $(\bar{d}_{1n}, \bar{d}_{2n})'$ in the yellow area we can neither reject $H_0^{\leq \text{lex}}$ nor $H_0^{\geq \text{lex}}$ (at significance level $\nu = 1/2[1 + \tilde{\nu} - F_{\chi_1^2}(\chi_{2,1-\tilde{\nu}}^2)]$, respectively). This implies that there is no evidence of differences in predictive ability between the internal and the benchmark model (at level $\tilde{\nu}$). From a regulatory perspective, the bank’s internal model is ‘at the boundary’ of what can be deemed acceptable. Hence, we suggest heightened attention and close monitoring by the regulator, as indicated by the yellow colour.

When $(\bar{d}_{1n}, \bar{d}_{2n})'$ falls into the orange area, then—while the VaR forecasts are of comparable

quality—the benchmark model provides superior systemic risk forecasts. Here, as indicated by the orange colour, a revision of the internal systemic risk forecasts is called for, while the internal marginal model is in order.

In contrast, the green area indicates superior systemic risk forecasts of the internal model, with VaR forecasts being comparably accurate. In this case, the financial institution should be allowed to use its internal model for risk forecasting purposes. The green colour indicates a pass of the backtest.

In the red region, the VaR forecasts of the internal model are deemed inferior to the benchmark predictions, whence there is no basis to compare the systemic risk forecasts. (The red region corresponds to the rejection region of the null $E[\bar{d}_{1n}] \geq 0$ at level $\nu' = \frac{1}{2} - \frac{1}{2}F_{\chi_1^2}(\chi_{2,1-\tilde{\nu}}^2)$. E.g., for $\nu = 5\%$ we get $\tilde{\nu} = 7.66\%$ and $\nu' = 1.17\%$.) 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 $\mathcal{T}_{2n}$; see Remark 5.4. For the comparison via $\mathcal{T}_{2n}$, we suggest to adopt a similar decision heuristic as in Fissler *et al.* (2016): Neither $H_0^{\leq \text{lex}}$ nor $H_0^{\geq \text{lex}}$ can be rejected when $|\mathcal{T}_{2n}| \leq \Phi^{-1}(1 - \nu)$ (corresponding to our yellow zone). When $\mathcal{T}_{2n} > \Phi^{-1}(1 - \nu)$ ($\mathcal{T}_{2n} < \Phi^{-1}(\nu)$) such that $H_0^{\leq \text{lex}}$ ($H_0^{\geq \text{lex}}$) can be rejected, the internal systemic risk forecasts are 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 $E[\bar{d}_{1n}] \leq 0$ at level $\nu' = \frac{1}{2} - \frac{1}{2}F_{\chi_1^2}(\chi_{2,1-\tilde{\nu}}^2)$.) In this case, there is no cause for action on the end of the bank (hence the grey colour), but the regulator should rather adopt the bank's VaR model as a basis for comparing the systemic risk forecasts. As before, the subsequent comparison of systemic risk forecasts should be carried out by using $\mathcal{T}_{2n}$, since the VaR forecasts are identical.

Section F investigates the finite-sample properties of $\mathcal{T}_n$, $\mathcal{T}_n^{\text{OS}}$, and $\mathcal{T}_{2n}$ under $H_0^=$ and $H_0^{\leq \text{lex}}$ 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 risk forecast comparisons are small. Second, power increases markedly in n . Third, comparisons for (CoVaR, 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 $H_0^{\leq \text{lex}}$ and $H_0^=$ are easier to detect for the former. Fifth, it is in general easier to detect differences in predictive ability of the systemic risk component, when the VaR 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.

While the schematic in Figure 4 is motivated by a regulatory framework, we stress that it can be used in the context of any comparative backtest between different models. Such a general example is provided in Section 6.

6. Empirical application

Consider daily log-losses $X_{-r+1}, \dots, X_n$ on the S&P 500 and log-losses $Y_{-r+1}, \dots, Y_n$ on the DAX 30 from 2000–2020, where the data are taken from www.wsj.com/market-data/quotes (ticker symbols: SPX and DX:DAX). So if $P_{Z,t}$ 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 $n + r = 5193$ observations. Here, for $\alpha = \beta = 0.95$, we compare rolling-window (VaR, CoVaR, CoES) forecasts for the series $\{(X_t, Y_t)\}_{t=1, \dots, n}$, 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 Section F 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, $F_{(X_t, Y_t)|\mathfrak{F}_{t-1}}(x, y) = \mathbb{P}\{X_t \leq x, Y_t \leq y \mid \mathfrak{F}_{t-1}\} =: \mathbb{P}_{t-1}\{X_t \leq x, Y_t \leq y\}$ for $x, y \in \mathbb{R}$. Here, the filtration $\mathfrak{F}_{t-1}$ is generated by the information available to a forecaster at time $t-1$. These are usually past observations $(X_{t-1}, Y_{t-1}), (X_{t-2}, Y_{t-2}), \dots$, and possibly additional exogenous information. Here, we assume $\mathfrak{F}_{t-1} = \sigma((X_{t-1}, Y_{t-1}), (X_{t-2}, Y_{t-2}), \dots)$, such that we forecast the *conditional* risk measures $\text{VaR}_t(X_t) = \text{VaR}_\beta(F_{X_t|\mathfrak{F}_{t-1}})$, $\text{CoVaR}_t(Y_t|X_t) = \text{CoVaR}_{\alpha|\beta}(F_{(X_t, Y_t)|\mathfrak{F}_{t-1}})$, and $\text{CoES}_t(Y_t|X_t) = \text{CoES}_{\alpha|\beta}(F_{(X_t, Y_t)|\mathfrak{F}_{t-1}})$. For notational brevity, we suppress the dependence of the risk measures on the risk levels α and β , 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 X_t and Y_t and—as a dependence model for the respective innovations $\varepsilon_{x,t}$ and $\varepsilon_{y,t}$ —a Gaussian copula driven by GAS dynamics. That is, we assume $(\varepsilon_{x,t}, \varepsilon_{y,t}) \mid \mathfrak{F}_{t-1}$ to have 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 Subsection F.1, where the same model (with $\vartheta = \infty$ in Equation (F.2)) 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 et al. \(1993\)](#). 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 Equation (F.2). 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 robust Gaussian quasi-maximum likelihood estimator for the GARCH-type marginal models) and the risk forecasting stage (by using their empirical cdfs $\hat{F}_x$ and $\hat{F}_y$ in computing the risk measures). For details on how the risk predictions are calculated, we refer to Subsection F.2.

We now compare two sequences of rolling-window predictions. For each rolling window of length $r = 1000$, we re-fit the two models on a daily basis. This gives us $n = 4193$ forecasts

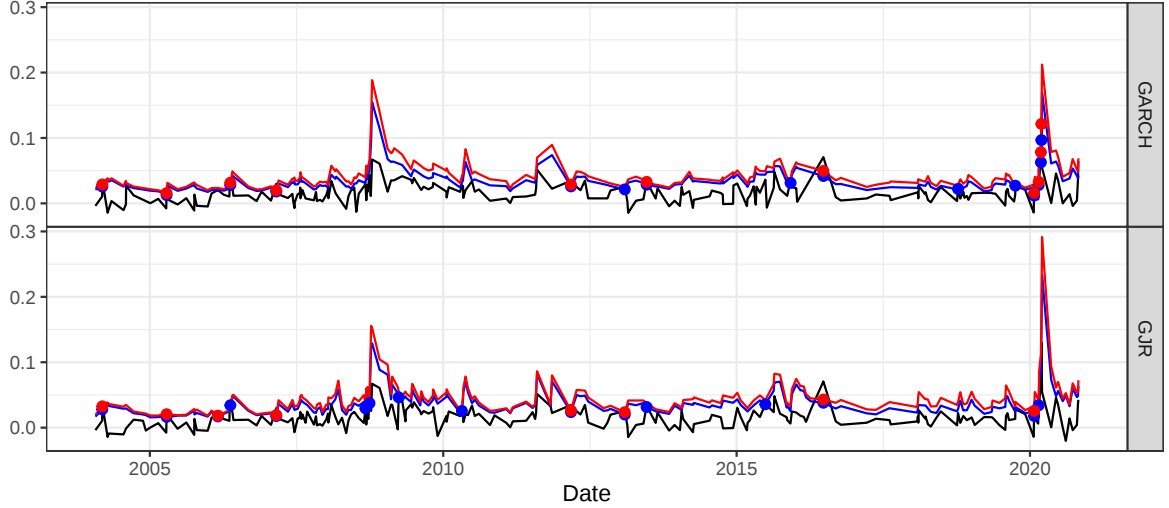


Figure 5: Top: DAX log-losses on days of VaR violation of S&P 500 (black). CoVaR (CoES) forecasts are shown as the blue (red) line. Violation of CoVaR (CoES) forecast is indicated by a blue (red) dot. All forecasts from GARCH with Gaussian copula. Bottom: Same as top, only with forecasts from GJR–GARCH with t -copula.

$\{\mathbf{r}_{t,(1)}\}_{t=1,\dots,n}$ from the GARCH model with Gaussian copula, and $\{\mathbf{r}_{t,(2)}\}_{t=1,\dots,n}$ from the GJR–GARCH with t -copula. We interpret the $\mathbf{r}_{t,(1)}$ as generic benchmark forecasts, which are to be improved upon by the $\mathbf{r}_{t,(2)}$. In an oversight context, the $\mathbf{r}_{t,(1)}$ may be some regulatory benchmark forecasts and the $\mathbf{r}_{t,(2)}$ forecasts from the bank’s internal model. Or from a banking perspective, the $\mathbf{r}_{t,(1)}$ may be forecasts issued from a trading desk’s legacy model and the $\mathbf{r}_{t,(2)}$ are 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 of 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 $\{(X_t, Y_t)\}_{t=1,\dots,n}$.

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. By definition, the S&P 500 should only exceed its VaR forecast on 5% of all trading days, i.e., 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 non-exceedance 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

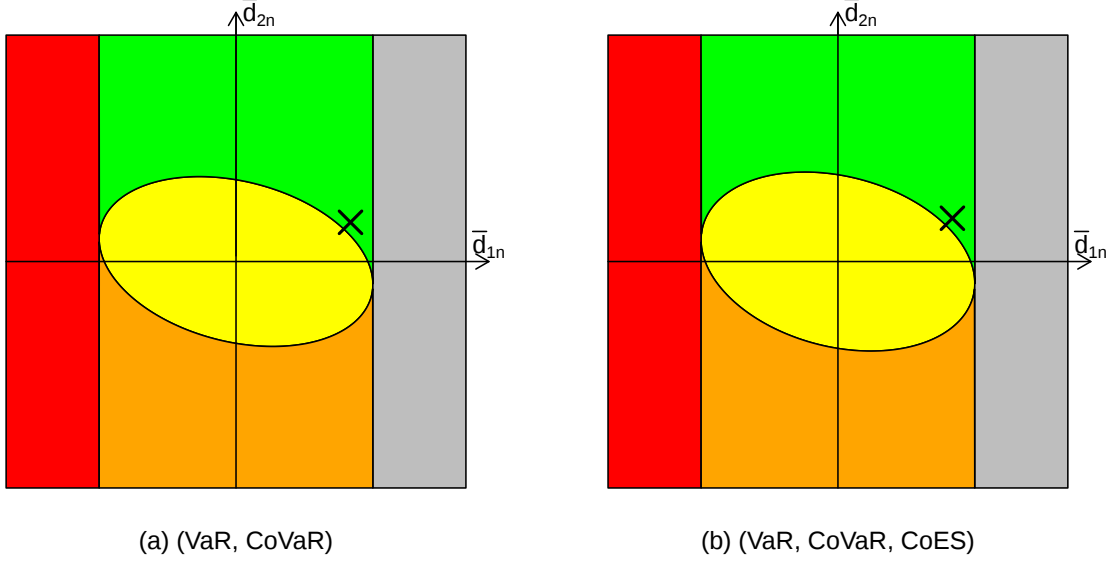


Figure 6: Panel (a): Values of $\bar{d}_{1n}$ and $\bar{d}_{2n}$ for (VaR, CoVaR) forecasts, indicated by $\times$. Panel (b): Values of $\bar{d}_{1n}$ and $\bar{d}_{2n}$ for (VaR, CoVaR, CoES) forecasts, indicated by $\times$.

undesirable from a risk management perspective, providing some informal evidence in favour of the t -copula model. We investigate this more formally in the following. Note 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 2010’s, and—most markedly—in 2020 as a consequence of the Covid-19 pandemic.

As pointed out above, we regard the GARCH model with Gaussian copula as our benchmark model. So we now want to test $H_0^{\leq \text{lex}}$ for (VaR, CoVaR) and (VaR, CoVaR, CoES). Due to the less clustered CoVaR exceedances and the compelling empirical evidence in favour of GJR–GARCH models (Glosten *et al.*, 1993; Brownlees *et al.*, 2011) and GAS- t -copula models (Creal *et al.*, 2013; Bernardi and Catania, 2019), we expect the GJR–GARCH model with t -copula to produce lower scores, i.e., better risk forecasts, possibly leading to a rejection of $H_0^{\leq \text{lex}}$. We carry out the tests using $\mathcal{T}_n^{\text{OS}}$ with the scores of Equations (F.4) and (F.5) having 0-homogeneous score differences, and with $\hat{\boldsymbol{\Omega}}_n$ from (5.2) (with $m_n = 0$); see Supplement Subsection F.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 (Nolde and Ziegel, 2017). We confirm the latter in our simulations for systemic risk forecasts as well, thus justifying our choice. Let $\bar{\mathbf{d}}_n = (\bar{d}_{1n}, \bar{d}_{2n})'$ be defined as in Subsection 5.1. Indeed, computing $\bar{\mathbf{d}}_n$, we find that the GJR–GARCH with t -copula produces lower scores for both the (VaR, CoVaR) and the (VaR, CoVaR, CoES) forecasts. The score differences are even statistically significant at the 5%-level: The p -values for the $\mathcal{T}_n^{\text{OS}}$ -based Wald test are 2.9% for (VaR, CoVaR) and 3.0% for (VaR, CoVaR, CoES).

Figure 6 illustrates the test decisions. Panel (a) shows the results for the (VaR, CoVaR)

comparison. Both $\bar{d}_{1n}$ and $\bar{d}_{2n}$ are positive, favouring the forecasts $\mathbf{r}_{t,(2)}$ from the GJR–GARCH model with t -copula. Additionally, the pair $\bar{\mathbf{d}}_n = (\bar{d}_{1n}, \bar{d}_{2n})'$ 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 (VaR, CoVaR, 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 $\beta = 0.95$. This is as expected, because by only considering observations with one extreme component, the effective sample size is massively reduced to roughly $n(1 - \beta)$. Indeed, for both models, 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 , e.g., 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 et al., 2020b](#)), 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 paper 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 modelling attempts of these quantities in financial institutions. Moreover, the fact that our notion of multi-objective elicibility serves as a ‘truth serum’ implies that the regulatory framework can be improved by enticing financial institutions to accurately model systemic risk. In terms of calibration tests, we are only aware of one more paper: [Banulescu-Radu et al. \(2021\)](#) introduce such tests, which unfortunately hinge on non-strict identification functions. This may lead to a severe loss of power under the alternative as demonstrated in Remark 4.3 and Section G. Due to the strictness of our identification functions, such a phenomenon cannot happen in our context.

The novel concept of multi-objective elicibility is likely to be fruitful also in applications beyond the proposed DM-backtests for systemic risk measures. In Subsection B.1, we provide more examples of interesting situations where conditional elicibility holds, yet classical joint elicibility fails. By virtue of Theorem 3.7, one can now construct incentive compatible elicitation mechanisms for these functionals, or can come up with M -estimation procedures in a regression context.

Beyond the confines of finance, we anticipate many other interesting applications of our backtests. For instance, in economics, [Adrian et al. \(2019\)](#) and [Adrian et al. \(2021\)](#) have recently drawn attention to tail risks and their interconnections by popularizing the Growth-at-Risk, which is simply the VaR of GDP growth. This literature has led to an increase in the use of risk forecast evaluation methods in economics (e.g., [Brownlees and Souza, 2021](#)). Hence, the backtests developed in this paper should be relevant for future macroeconomic applications as well.

References

- Acharya VV, Pedersen LH, Philippon T, Richardson M. 2017. Measuring systemic risk. *The Review of Financial Studies* **30**: 2–47.
- Adrian T, Boyarchenko N, Giannone D. 2019. Vulnerable growth. *The American Economic Review* **109**: 1263–1289.
- Adrian T, Brunnermeier MK. 2016. CoVaR. *The American Economic Review* **106**: 1705–1741.
- Adrian T, Grinberg F, Liang N, Malik S, Yu J. 2021. The term structure of Growth-at-Risk. Forthcoming in *American Economic Journal: Macroeconomics* <https://www.aeaweb.org/articles/pdf/doi/10.1257/mac.20180428>.
- Artzner P, Delbaen F, Eber JM, Heath D. 1999. Coherent measures of risk. *Mathematical Finance* **9**: 203–228.
- Bank for International Settlements. 2019. *Basel Framework*. Basel, http://www.bis.org/basel_framework/index.htm?export=pdf.
- Banulescu-Radu D, Hurlin C, Leymarie J, Scaillet O. 2021. Backtesting marginal expected shortfall and related systemic risk measures. *Management Science* **67**: 5730–5754.
- Benoit S, Colliard JE, Hurlin C, Pérignon C. 2017. Where the risks lie: A survey on systemic risk. *Review of Finance* **21**: 109–152.
- Bernardi M, Catania L. 2019. Switching generalized autoregressive score copula models with application to systemic risk. *Journal of Applied Econometrics* **34**: 43–65.
- Brownlees C, Engle R, Kelly B. 2011. A practical guide to volatility forecasting through calm and storm. *Journal of Risk* **14**: 3–22.
- Brownlees C, Engle RF. 2017. SRISK: A conditional capital shortfall measure of systemic risk. *The Review of Financial Studies* **30**: 48–79.
- Brownlees C, Souza ABM. 2021. Backtesting global Growth-at-Risk. *Journal of Monetary Economics* **118**: 312–330.
- Brunnermeier M, Rother S, Schnabel I. 2020. Asset price bubbles and systemic risk. *The Review of Financial Studies* **33**: 4272–4317.
- Chen C, Iyengar G, Moallemi CC. 2013. An Axiomatic Approach to Systemic Risk. *Management Science* **59**: 1373–1388.
- Creal D, Koopman SJ, Lucas A. 2013. Generalized autoregressive score models with applications. *Journal of Applied Econometrics* **28**: 777–795.
- Diebold FX, Mariano RS. 1995. Comparing predictive accuracy. *Journal of Business & Economic Statistics* **13**: 253–263.

- Dimitriadis T, Fissler T, Ziegel JF. 2020a. The Efficiency Gap. *Preprint*. <https://arxiv.org/abs/2010.14146>.
- Dimitriadis T, Liu X, Schnaitmann J. 2020b. Encompassing tests for value at risk and expected shortfall multi-step forecasts based on inference on the boundary. *Preprint*. <https://arxiv.org/abs/2009.07341>.
- Eckernkemper T. 2018. Modeling systemic risk: Time-varying tail dependence when forecasting marginal expected shortfall. *Journal of Financial Econometrics* **16**: 63–117.
- Ehm W, Gneiting T, Jordan A, Krüger F. 2016. Of quantiles and expectiles: consistent scoring functions, Choquet representations and forecast rankings. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **78**: 505–562.
- Ehrgott M. 2005. *Multicriteria Optimization*. Berlin, Heidelberg: Springer.
- Emmer S, Kratz M, Tasche D. 2015. What is the best risk measure in practice? a comparison of standard measures. *Journal of Risk* **18**: 31–60.
- Feinstein Z, Rudloff B, Weber S. 2017. Measures of systemic risk. *SIAM Journal on Financial Mathematics* **8**: 672–708.
- Fissler T, Frongillo R, Hlavínová J, Rudloff B. 2021. Forecast evaluation of quantiles, prediction intervals, and other set-valued functionals. *Electronic Journal of Statistics* **15**: 1034–1084.
- Fissler T, Ziegel JF. 2016. Higher order elicibility and Osband’s principle. *The Annals of Statistics* **44**: 1680–1707.
- Fissler T, Ziegel JF, Gneiting T. 2016. Expected shortfall is jointly elicitable with value-at-risk: Implications for backtesting. *Risk Magazine* : 58–61.
- Giacomini R, White H. 2006. Tests of conditional predictive ability. *Econometrica* **74**: 1545–1578.
- Giesecke K, Kim B. 2011. Systemic risk: What defaults are telling us. *Management Science* **57**: 1387–1405.
- Giglio S, Kelly B, Pruitt S. 2016. Systemic risk and the macroeconomy: An empirical evaluation. *Journal of Financial Economics* **119**: 457–471.
- Girardi G, Tolga Ergün A. 2013. Systemic risk measurement: Multivariate GARCH estimation of CoVaR. *Journal of Banking & Finance* **37**: 3169–3180.
- Glosten LR, Jagannathan R, Runkle DE. 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.
- Gneiting T. 2011. Making and evaluating point forecasts. *Journal of the American Statistical Association* **106**: 746–762.
- Hoga Y. 2021. Modeling time-varying tail dependence, with application to systemic risk forecasting. Forthcoming in *Journal of Financial Econometrics* : 1–31.
- Mainik G, Schaanning E. 2014. On dependence consistency of CoVaR and some other systemic risk measures. *Statistics & Risk Modeling* **31**: 49–77.
- Mas-Colell A, Whinston MD, Green JR. 1995. *Microeconomic Theory*. Oxford University Press.
- Nolde N, Zhang J. 2020. Conditional extremes in asymmetric financial markets. *Journal of Business & Economic Statistics* **38**: 201–213.
- Nolde N, Ziegel JF. 2017. Elicibility and backtesting: Perspectives for banking regulation. *The Annals of Applied Statistics* **11**: 1833–1874.
- Steinwart I, Pasin C, Williamson R, Zhang S. 2014. Elicitation and Identification of Properties. *JMLR Workshop Conf. Proc.* **35**: 1–45.
- White H. 2001. *Asymptotic Theory for Econometricians*. San Diego: Academic Press, First edn.

Supplement

A. Proofs for Section 3

Proof of Proposition 3.4: Let $V_1: A_1 \times \mathcal{O} \rightarrow \mathbb{R}^{m_1}$ be a strict $\mathcal{F}$ -identification function for T_1 and, for any $r_1 \in A_1$, let $V_{2,r_1}: A_2 \times \mathcal{O} \rightarrow \mathbb{R}^{m_2}$ be a strict $\mathcal{F}_{r_1}$ -identification function for T_2 . We claim that

$$V: (A_1 \times A_2) \times \mathcal{O} \rightarrow \mathbb{R}^{m_1+m_2}, \quad ((r_1, r_2), y) \mapsto (V_1(r_1, y), V_{2,r_1}(r_2, y))$$

is a strict $\mathcal{F}$ -identification function for (T_1, T_2) . Consider some $F \in \mathcal{F}$ and $(r_1, r_2) \in A_1 \times A_2$. First, assume $(r_1, r_2) \in T_1(F) \times T_2(F)$. Then $\bar{V}_1(r_1, F) = \mathbf{0}$ and $\bar{V}_{2,r_1}(r_2, F) = \mathbf{0}$. Second, assume $\bar{V}_1(r_1, F) = \mathbf{0}$ and $\bar{V}_{2,r_1}(r_2, F) = \mathbf{0}$. Due to the former equality and the strictness of V_1 , it holds that $r_1 \in T_1(F)$. Now we can invoke the strictness of V_{2,r_1} on $\mathcal{F}_{r_1}$ to conclude that $r_2 \in T_2(F)$. ■

Proof of Theorem 3.7: Let $S_1: A_1 \times \mathcal{O} \rightarrow \mathbb{R}$ be a strictly $\mathcal{F}$ -consistent scoring function for T_1 . For any $r_1 \in A_1$ let $S_{2,r_1}: A_2 \times \mathcal{O} \rightarrow \mathbb{R}$ be a strictly $\mathcal{F}_{r_1}$ -consistent scoring function for T_2 . We claim that

$$S: (A_1 \times A_2) \times \mathbb{R}^d \rightarrow (\mathbb{R}^2, \preceq_{\text{lex}}), \quad ((r_1, r_2), y) \mapsto (S_1(r_1, y), S_{2,r_1}(r_2, y)) \quad (\text{A.1})$$

is a strictly multi-objective $\mathcal{F}$ -consistent scoring function for (T_1, T_2) with respect to $\preceq_{\text{lex}}$. To see this, consider some $F \in \mathcal{F}$ with $(t_1, t_2) \in T_1(F) \times T_2(F)$ and $(r_1, r_2) \in A_1 \times A_2$. From the strict $\mathcal{F}$ -consistency of S_1 it follows that

$$\bar{S}_1(t_1, F) \leq \bar{S}_1(r_1, F), \quad (\text{A.2})$$

where the inequality is strict if $r_1 \notin T_1(F)$. If the inequality is strict, we can already conclude that $\bar{S}(t_1, t_2, F) \prec_{\text{lex}} \bar{S}(r_1, r_2, F)$. Otherwise, it must be the case that $r_1 \in T_1(F)$. Since by assumption $T_1(F)$ is a singleton, $r_1 = t_1$. Invoking the strict $\mathcal{F}_{t_1}$ -consistency of S_{2,t_1} , we obtain that

$$\bar{S}_{2,t_1}(t_2, F) \leq \bar{S}_{2,t_1}(r_2, F), \quad (\text{A.3})$$

where the inequality is strict if $r_2 \notin T_2(F)$. If the inequality is strict, this implies that $\bar{S}((t_1, t_2), F) \prec_{\text{lex}} \bar{S}((r_1, r_2), F)$. Otherwise, equality implies that $r_2 \in T_2(F)$. ■

Remark A.1. Theorem 3.7 is almost a direct analogue of Proposition 3.4, with the intriguing exception that T_1 is assumed to be a singleton on $\mathcal{F}$ in Theorem 3.7. An inspection of the proof reveals that this is needed to show the inequality in (A.3). If at this stage of the proof one has that $t_1, r_1 \in T_1(F)$ and $t_1 \neq r_1$, it is *per se* not clear (and in general also not the case) that $S_{2,t_1} = S_{2,r_1}$ from which the inequality in (A.3) would follow. What would be sufficient, however, to dispense with the assumption on T_1 and to still show (A.3) is the following: It is possible to

choose a family of strictly $\mathcal{F}_{r_1}$ -consistent scores S_{2,r_1} , $r_1 \in A_1$, for T_2 such that for all $F \in \mathcal{F}$ and for all $t_1, r_1 \in T_1(F)$ we have equality for the expected scores, $\bar{S}_{2,t_1}(\cdot, F) = \bar{S}_{2,r_1}(\cdot, F)$. In many practical examples, this can be achieved; see, e.g., Example B.1 (b).

B. More background on multi-objective elicibility

B.1. Further examples of multi-objective elicitable functionals

We provide more examples of functionals that are conditionally elicitable, yet fail to be elicitable in the traditional sense. Examples (c) and (d) are concerned with α -prediction intervals. In line with Fissler *et al.* (2021, Section 4) for some random variable Y with distribution F , this is any interval $[a, b] \subseteq \mathbb{R}$ such that $P(Y \in [a, b]) = F(b) - F(a-) \geq \alpha$. Of course, each distribution F has a multitude of different α -prediction intervals. We consider two possible attempts of coming up with *interpretable* α -prediction intervals.

Two univariate scores $S, \tilde{S}: A \times O \rightarrow \mathbb{R}$ are *equivalent* if there exists a positive constant $c > 0$ and a function $a: O \rightarrow \mathbb{R}$ such that $\tilde{S}(r, y) = cS(r, y) + a(y)$. Ignoring integrability issues for a moment, it is clear that equivalence preserves (strict) consistency. If S and a are $\mathcal{F}$ -integrable, so is $\tilde{S}$. However, by a convenient choice of a , the integrability conditions of $\tilde{S}$ can be milder than those of S . E.g., while $S(r, y) = (r - y)^2$, $r, y \in \mathbb{R}$, is a strictly $\mathcal{F}^2(\mathbb{R})$ -consistent scoring function for the mean, the equivalent score $\tilde{S}(r, y) = r^2 - 2ry$ is strictly $\mathcal{F}^1(\mathbb{R})$ -consistent score for the mean. A straightforward generalisation is obvious for multi-objective scores of the form (A.1), i.e.,

$$\mathbf{S}: (A_1 \times A_2) \times \mathbb{R}^d \rightarrow (\mathbb{R}^2, \preceq_{\text{lex}}), \quad ((r_1, r_2), y) \mapsto (S_1(r_1, y), S_{2,r_1}(r_2, y))'. \quad (\text{B.1})$$

Example B.1. (a) The variance functional is conditionally elicitable with the mean functional on $\mathcal{F}^4(\mathbb{R})$. A strictly multi-objective $\mathcal{F}^4(\mathbb{R})$ -consistent score $\mathbf{S}: \mathbb{R}^2 \times \mathbb{R} \rightarrow (\mathbb{R}^2, \preceq_{\text{lex}})$ is given by (B.1), where $S_1(r_1, y) = (r_1 - y)^2$ and $S_{2,r_1}(r_2, y) = [r_2 - (y - r_1)^2]^2$. Alternatively, to achieve a strictly multi-objective $\mathcal{F}^2(\mathbb{R})$ -consistent score, one can consider the equivalent version $\tilde{S}_{2,r_1}(r_2, y) = r_2^2 - 2r_2(y - r_1)^2$, or more generally $\tilde{S}_{2,r_1}(r_2, y) = -\phi(r_2) + \phi'(r_2)[r_2 - (y - r_1)]$, where ϕ is strictly convex with subgradient ϕ' .

(b) For $\alpha \in (0, 1)$, ES_α is conditionally elicitable with the α -quantile, q_α , on $\mathcal{F}^2(\mathbb{R})$. A strictly multi-objective $\mathcal{F}^2(\mathbb{R})$ -consistent score $\mathbf{S}: \mathbb{R}^2 \times \mathbb{R} \rightarrow (\mathbb{R}^2, \preceq_{\text{lex}})$ is given via (B.1), where $S_1(r_1, y) = (\mathbb{1}\{y \leq r_1\} - \alpha)(r_1 - y)$ and $S_{2,r_1}(r_2, y) = \left\{ r_2 - \frac{1}{1-\alpha} [\mathbb{1}\{y > r_1\}y + r_1(\mathbb{1}\{y \leq r_1\} - \alpha)] \right\}^2$. It is straightforward to verify that this family of scores satisfies the condition discussed in Remark A.1. Similar to (a), we can achieve strict multi-objective $\mathcal{F}^1(\mathbb{R})$ -consistency upon replacing S_{2,r_1} by $\tilde{S}_{2,r_1}(r_2, y) = r_2^2 - \frac{2r_2}{1-\alpha} [\mathbb{1}\{y > r_1\}y + r_1(\mathbb{1}\{y \leq r_1\} - \alpha)]$, or more generally by $\tilde{S}_{2,r_1}(r_2, y) = -\phi(r_2) + \phi'(r_2) \left\{ r_2 - \frac{1}{1-\alpha} [\mathbb{1}\{y > r_1\}y + r_1(\mathbb{1}\{y \leq r_1\} - \alpha)] \right\}$, where ϕ is strictly convex with subgradient ϕ' .

- (c) Fix $\alpha \in (0, 1)$. Motivated by prediction intervals induced by quantiles of the form $[q_\beta^-(F), q_{\alpha+\beta}^-(F)]$ for $\beta \in [0, 1 - \alpha]$ we consider the following form of α -prediction intervals. Let $T_1: \mathcal{F} \rightarrow \mathbb{R}$ be elicitable on $\mathcal{F} \subseteq \mathcal{F}_{\text{inc}} \subset \mathcal{F}^0(\mathbb{R})$, where all $F \in \mathcal{F}_{\text{inc}}$ are strictly increasing cdfs, and let $T_2: \mathcal{F} \rightarrow (-\infty, \infty]$ be the right endpoint of the shortest α -prediction interval $[T_1(F), T_2(F)]$ whose left endpoint is given via T_1 . [Fissler et al. \(2021, Proposition 4.10\)](#) asserts that (T_1, T_2) is not elicitable, subject to smoothness conditions, but [Fissler et al. \(2021, Proposition A.1\)](#) yields that T_2 is conditionally elicitable with T_1 on $\mathcal{F}$. A strictly multi-objective $\mathcal{F}$ -consistent scoring function $\mathbf{S}: \mathbb{R}^2 \times \mathbb{R} \rightarrow (\mathbb{R}^2, \preceq_{\text{lex}})$ is given by (B.1), where $S_1(r_1, y)$ is a strictly $\mathcal{F}$ -consistent score for T_1 and

$$S_{2,r_1}(r_2, y) = \mathbb{1}\{y \geq r_1\} \left[(\mathbb{1}\{y \leq r_2\} - \alpha)(g(x) - g(y)) \right] - \mathbb{1}\{y < r_1\} \alpha g(r_2),$$

where $g: (-\infty, \infty] \rightarrow \mathbb{R}$ is strictly increasing and $\mathcal{F}$ -integrable.

- (d) A different approach to constructing interpretable α -prediction intervals is to specify the midpoint in terms of an elicitable functional, such as the mean or the median. Again, let $\alpha \in (0, 1)$ and let $T_1: \mathcal{F} \rightarrow \mathbb{R}$ be elicitable on $\mathcal{F} \subseteq \mathcal{F}_{\text{inc}}$ and let $T_2: \mathcal{F} \rightarrow [0, \infty)$ be half of the length of the shortest α -prediction interval $[T_1(F) - T_2(F), T_1(F) + T_2(F)]$ with midpoint T_1 . [Fissler et al. \(2021, Proposition 4.12\)](#) asserts that (T_1, T_2) is not elicitable, subject to smoothness conditions, but [Fissler et al. \(2021, Proposition A.3\)](#) yields that T_2 is conditionally elicitable with T_1 on $\mathcal{F}$. A strictly multi-objective $\mathcal{F}$ -consistent scoring function $\mathbf{S}: \mathbb{R}^2 \times \mathbb{R} \rightarrow (\mathbb{R}^2, \preceq_{\text{lex}})$ is given via (B.1), where $S_1(r_1, y)$ is a strictly $\mathcal{F}$ -consistent score for T_1 and

$$S_{2,r_1}(r_2, y) = (\mathbb{1}\{|y - r_1| \leq r_2\} - \alpha) [g(r_2) - g(|y - r_1|)],$$

where $g: [0, \infty) \rightarrow \mathbb{R}$ is strictly increasing with $g(|\cdot - r_1|)$ being $\mathcal{F}$ -integrable for all $r_1 \in \mathbb{R}$.

- (e) For $\alpha \in (0, 1)$, let $\mathcal{F}_{(\alpha)}^0(\mathbb{R})$ as in (4.1). Define the right $(1 - \alpha)$ -tail of a distribution $F \in \mathcal{F}_{(\alpha)}^0(\mathbb{R})$ as $T_\alpha(F) \in \mathcal{F}^0(\mathbb{R})$ where $T_\alpha(F)(x) = \mathbb{1}\{x > \text{VaR}_\alpha(F)\}(F(x) - \alpha)/(1 - \alpha)$. Then, T_α is conditionally elicitable with VaR_α on $\mathcal{F}_{(\alpha)}^0(\mathbb{R})$. A strictly multi-objective $\mathcal{F}_{(\alpha)}^0(\mathbb{R})$ -consistent score $\mathbf{S}: \mathbb{R} \times \mathcal{F}_{(\alpha)}^0(\mathbb{R}) \times \mathbb{R} \rightarrow (\mathbb{R}^2, \preceq_{\text{lex}})$ is given via (B.1), where $S_1(r, y) = (\mathbb{1}\{y \leq r\} - \alpha)[g(r) - g(y)]$ with strictly increasing and bounded $g: \mathbb{R} \rightarrow \mathbb{R}$, and the second component of $\mathbf{S}$ is given by $S_{2,r}(F, y) = \mathbb{1}\{y > r\}R(F, y)$. Here, $R: \mathcal{F}^0(\mathbb{R}) \times \mathbb{R} \rightarrow \mathbb{R}$ is a strictly proper scoring rule. This result follows from [Gneiting \(2011b, Theorem 5\)](#); see also Theorem C.2 and [Holzmann and Klar \(2017\)](#) for related approaches.

B.2. Dimensionality considerations

Consider an elicitable functional $\mathbf{T}: \mathcal{F} \rightarrow \mathcal{P}(\mathbf{A})$ with strictly $\mathcal{F}$ -consistent score $S: \mathbf{A} \times \mathbf{O} \rightarrow \mathbb{R}$, where $\mathbf{A} \subseteq \mathbb{R}^k$. Under sufficient smoothness conditions on the score $S(\cdot, \mathbf{y})$ and the expected score $\bar{S}(\cdot, F)$, first order conditions yield that the k -dimensional gradient of S is an

$\mathcal{F}$ -identification function for $\mathbf{T}$ (which is not necessarily strict). Under the conditions of Theorem 3.7, first order conditions of the corresponding strictly multi-objective $\mathcal{F}$ -consistent score (B.1) also induce an $\mathcal{F}$ -identification function whose dimension coincides with $\mathbf{T}$. To illustrate this, suppose that $\mathbf{T}_1$ and $\mathbf{T}_2$ are univariate, such that $k_1 + k_2 = k = 2$. We have the scores $\mathbf{S}((r_1, r_2), \mathbf{y}) = (S_1(r_1, \mathbf{y}), S_{2,r_1}(r_2, \mathbf{y}))' \in (\mathbb{R}^2, \preceq_{\text{lex}})$. At the optimum $(t_1, t_2) \in T_1(F) \times T_2(F)$ we obtain the first order condition $\partial \bar{S}_1(r_1, F) = 0$ and $\partial \bar{S}_{2,r_1}(r_2, F) = 0$. Hence, if the derivatives exist and tacitly assuming that integration and differentiation commute, $\mathbf{V}((r_1, r_2), \mathbf{y}) = (\partial S_1(r_1, \mathbf{y}), \partial S_{2,r_1}(r_2, \mathbf{y}))'$ constitutes a two-dimensional $\mathcal{F}$ -identification function for (T_1, T_2) . This means we can consider the two-dimensional score (B.1) as a generalised antiderivative of a k -dimensional identification function where the symmetry conditions imposed by the Hessian are massively relaxed, which renders the existence of such an object possible. Thus, multi-objective scores provide a means to close the gap between identification and scoring functions for multivariate functionals.

B.3. More on multi-objective elicitability with respect to the lexicographic order

Remark B.2. Resuming with the discussion right before Theorem 3.7, the use of the lexicographic order on $\mathbb{R}^2$ allows to compare any forecasts, also misspecified ones. As for the classical univariate concept of strict consistency, strong 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 generalisations of order-sensitivity (see, e.g., Lambert *et al.*, 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 B.3. On the level of the prediction space setting (Gneiting and Ranjan, 2013), 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. This is an instance of the more general calibration–resolution principle of Pohle (2020), showing that minimising a consistent scoring function amounts to “jointly maximizing information content and minimizing systematic mistakes”. Using the same arguments as in Holzmann and Eulert (2014) and Pohle (2020) (basically exploiting the tower property of conditional expectations), one can establish a similar principle for multi-objective consistent scores mapping to $(\mathbb{R}^2, \preceq_{\text{lex}})$.

B.4. The necessity of the Convex Level Sets property

Here, we revisit a powerful necessary condition for identifiability and elicibility, namely the Convex Level Sets (CxLS) property, for multi-objective scores. We use the same notation as in Section 3 of the main paper, in particular we let $\mathcal{F} \subseteq \mathcal{F}' \subseteq \mathcal{F}^0(\mathcal{O})$. Following the terminology of Fissler *et al.* (2021, Section 3) we say that a functional $\mathbf{T}: \mathcal{F}' \rightarrow \mathcal{P}(\mathcal{A})$, satisfies the *selective CxLS property* on $\mathcal{F}$ if for all $F_0, F_1 \in \mathcal{F}$ and for all $\lambda \in (0, 1)$ such that $F_\lambda := (1-\lambda)F_0 + \lambda F_1 \in \mathcal{F}$ we have $\mathbf{T}(F_0) \cap \mathbf{T}(F_1) \subseteq \mathbf{T}(F_\lambda)$. $\mathbf{T}$ satisfies the *selective CxLS* property* on $\mathcal{F}$ if for all $F_0, F_1 \in \mathcal{F}$ and for all $\lambda \in (0, 1)$ such that $F_\lambda := (1-\lambda)F_0 + \lambda F_1 \in \mathcal{F}$ $\mathbf{T}(F_0) \cap \mathbf{T}(F_1) \neq \emptyset$ implies that $\mathbf{T}(F_0) \cap \mathbf{T}(F_1) = \mathbf{T}(F_\lambda)$. We shall frequently omit the term “selective” and will just speak of the CxLS and the CxLS* property. Clearly, the CxLS* property implies the CxLS property. If $\mathbf{T}$ attains singletons only on $\mathcal{F}$, then the two properties coincide. They then take the familiar form that $\mathbf{T}(F_0) = \mathbf{T}(F_1)$ implies $\mathbf{T}(F_0) = \mathbf{T}(F_\lambda)$. It is well known that the CxLS property is necessary for identifiability (which we state for the sake of completeness). Proposition 3.4 in Fissler *et al.* (2021) establishes that the CxLS* property is necessary for univariate elicibility. The following generalises this result to strong multi-objective elicibility.

Proposition B.4. *Consider a functional $\mathbf{T}: \mathcal{F}' \rightarrow \mathcal{P}(\mathcal{A})$.*

- (i) *If $\mathbf{T}$ is identifiable on $\mathcal{F} \subseteq \mathcal{F}'$, it satisfies the CxLS property on $\mathcal{F}$.*
- (ii) *If $\mathbf{T}$ is strongly multi-objective elicitable on $\mathcal{F} \subseteq \mathcal{F}'$, it satisfies the CxLS* property on $\mathcal{F}$.*

Proof of Proposition B.4: For (i), let $\mathbf{V}: \mathcal{A} \times \mathcal{O} \rightarrow \mathbb{R}^m$ be a strict identification function. Let $F_0, F_1 \in \mathcal{F}$ such that $F_\lambda \in \mathcal{F}$ for some $\lambda \in (0, 1)$. For $\mathbf{t} \in \mathbf{T}(F_0) \cap \mathbf{T}(F_1)$ it holds that $\mathbf{0} = (1-\lambda)\bar{\mathbf{V}}(\mathbf{t}, F_0) + \lambda\bar{\mathbf{V}}(\mathbf{t}, F_1) = \bar{\mathbf{V}}(\mathbf{t}, F_\lambda)$. Due to the strictness of $\mathbf{V}$, we have $\mathbf{t} \in \mathbf{T}(F_\lambda)$. For (ii), the proof follows along the lines of the proof of Proposition 3.4 in Fissler *et al.* (2021). With the same set-up as above, consider some strictly strongly multi-objective $\mathcal{F}$ -consistent scoring function $\mathbf{S}: \mathcal{A} \times \mathcal{O} \rightarrow (\mathbb{R}^I, \preceq)$. For $i \in \{0, 1\}$, the strict strong multi-objective $\mathcal{F}$ -consistency of $\mathbf{S}$ implies that for any $\mathbf{r} \in \mathcal{A}$ we have $\bar{\mathbf{S}}(\mathbf{t}, F_i) - \bar{\mathbf{S}}(\mathbf{r}, F_i) = \mathbf{0}$ for $\mathbf{r} \in \mathbf{T}(F_i)$ and for $\mathbf{r} \notin \mathbf{T}(F_i)$ the score difference is $\prec \mathbf{0}$. Hence, $\bar{\mathbf{S}}(\mathbf{t}, F_\lambda) - \bar{\mathbf{S}}(\mathbf{r}, F_\lambda)$ equals $(1-\lambda)(\bar{\mathbf{S}}(\mathbf{t}, F_0) - \bar{\mathbf{S}}(\mathbf{r}, F_0)) + \lambda(\bar{\mathbf{S}}(\mathbf{t}, F_1) - \bar{\mathbf{S}}(\mathbf{r}, F_1))$, which is $\mathbf{0}$ for $\mathbf{r} \in \mathbf{T}(F_0) \cap \mathbf{T}(F_1)$ and otherwise $\prec \mathbf{0}$. ■

Proposition B.4 implies that strong multi-objective elicibility shares an important feature with the usual univariate notion of elicibility. Showing that the CxLS* is violated is the standard procedure to rule out elicibility. Consequently, functionals violating the CxLS* property such as variance or ES also fail to be strongly multi-objective elicitable. This applies in particular to scoring functions mapping to $(\mathbb{R}^2, \preceq_{\text{lex}})$, since $\preceq_{\text{lex}}$ is a total order.

On the other hand, weak multi-objective elicibility appears to be much more flexible than its strong counterpart. Indeed, the following consideration shows that the CxLS property is not necessary for weak multi-objective elicibility: Suppose we have a strictly weakly multi-objective $\mathcal{F}$ -consistent scoring function $\mathbf{S}$ for $\mathbf{T}$ mapping to $(\mathbb{R}^2, \preceq)$, where $\preceq$ is the component-wise order. Suppose further that there are $F_0, F_1 \in \mathcal{F}$ such that $F_\lambda = (1-\lambda)F_0 + \lambda F_1 \in \mathcal{F}$

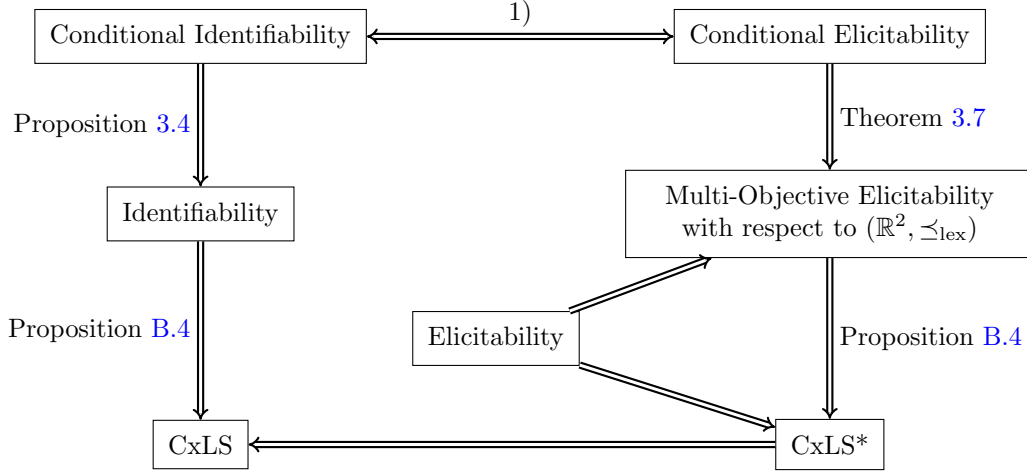


Figure 7: Illustration of the most important structural results for functionals $T_1 : \mathcal{F} \rightarrow \mathcal{P}(A_1)$ and $T_2 : \mathcal{F} \rightarrow \mathcal{P}(A_2)$. The equivalence 1) holds under some regularity conditions if T_1 and T_2 are point-valued mappings to $\mathbb{R}$ (Steinwart *et al.*, 2014).

for all $\lambda \in (0, 1)$. Assume there is some $t \in T(F_0) \cap T(F_1)$ and some $r \notin T(F_0) \cup T(F_1)$ such that $\bar{S}(r, F_0) - \bar{S}(t, F_0) = (-5, 1)' \not\prec (0, 0)'$ and $\bar{S}(r, F_1) - \bar{S}(t, F_1) = (1, -5)' \not\prec (0, 0)'$. Then, invoking the linearity of the integral, $\bar{S}(r, F_{1/2}) - \bar{S}(t, F_{1/2}) = (-2, -2)' \prec (0, 0)'$. Therefore, $t \notin T(F_{1/2})$ and the CxLS property is violated (and thus also the CxLS* property).

A further investigation of weak multi-objective elicibility therefore seems to be an interesting field. Such an investigation will be deferred to future work.

As a complete summary of the results of this section and Section 3, Figure 7 complements Figure 2 in the main paper by including the implications of Proposition B.4.

C. Conditional identifiability and conditional elicibility results

Here, we show that $\text{CoVaR}_{\alpha|\beta}(Y|X)$, $\text{MES}_{\beta}(Y|X)$ and $(\text{CoVaR}_{\alpha|\beta}(Y|X), \text{CoES}_{\alpha|\beta}(Y|X))$ are all conditionally elicitable and conditionally identifiable with $\text{VaR}_{\beta}(X)$, subject to mild assumptions on the corresponding class of bivariate distributions.

Remark C.1. Our definitions of the systemic risk measures in Section 2 rely on the conditional cdf $F_{Y|X \geq \text{VaR}_{\beta}(X)}$; see (2.1). If F_X is continuous, the event $\{X \geq \text{VaR}_{\beta}(X)\}$ has probability $1 - \beta$, such that conditioning on the right tail is akin to the classical definition of expected shortfall at level β , as $\text{ES}_{\beta}(X) = \mathbb{E}[X | X \geq \text{VaR}_{\beta}(X)]$. However, if $X = \text{VaR}_{\beta}(X)$ with positive probability, then $\mathbb{P}\{X \geq \text{VaR}_{\beta}(X)\} > 1 - \beta$, such that $\{X \geq \text{VaR}_{\beta}(X)\}$ loses its interpretation as a tail event. (E.g., for X with $\mathbb{P}\{X = 0\} = 1 - \mathbb{P}\{X = 1\} = p < 1$ it holds that $\text{VaR}_{\beta}(X) = 1$ for all $\beta \in [p, 1]$. Hence, $\mathbb{P}\{X \geq \text{VaR}_{\beta}(X)\} = 1 - p > 1 - \beta$ if $p < \beta$.) To allow for point masses of F_X at $\text{VaR}_{\beta}(X)$, one may use a similar correction term as for ES_{β} ; see McNeil *et al.* (2015, Proposition 8.15). Specifically, one may adapt the definition of

$F_{Y|X \geq \text{VaR}_\beta(X)}$ as follows. For $\beta \in [0, 1)$ and $y \in \mathbb{R}$ we define the cdf

$$F_{Y|X \geq \text{VaR}_\beta(X)}(y) := \frac{1}{1-\beta} \left[\mathbb{P}\{Y \leq y, X > \text{VaR}_\beta(X)\} + \mathbb{P}\{Y \leq y \mid X = \text{VaR}_\beta(X)\}(1 - \beta - \mathbb{P}\{X > \text{VaR}_\beta(X)\}) \right]. \quad (\text{C.1})$$

If $\mathbb{P}\{X > \text{VaR}_\beta(X)\} = 1 - \beta$, the second summand vanishes and the first summand simplifies to $F_{Y|X \geq \text{VaR}_\beta(X)}(y)$. If $\mathbb{P}\{X > \text{VaR}_\beta(X)\} < 1 - \beta$, then $\mathbb{P}\{X = \text{VaR}_\beta(X)\} > 0$, such that the conditional probability $\mathbb{P}\{Y \leq y \mid X = \text{VaR}_\beta(X)\}$ can be defined elementarily. Note that $F_{Y|X \geq \text{VaR}_0(X)} = F_Y$. In particular, one may replace $F_{Y|X \geq \text{VaR}_\beta(X)}$ in (2.1) with $F_{Y|X \geq \text{VaR}_\beta(X)}(y)$ to obtain alternative definitions of $\text{CoVaR}_{\alpha|\beta}(Y|X)$, $\text{CoES}_{\alpha|\beta}(Y|X)$ and $\text{MES}_\beta(Y|X)$. All results of Section C, Section D and Section 4 go through unaltered with this change. In fact, we shall prove all these results with the systemic risk measures defined in terms of $F_{Y|X \geq \text{VaR}_\beta(X)}(y)$. We mention that the definitions coincide when F_X is continuous, because then $F_{Y|X \geq \text{VaR}_\beta(X)}(\cdot) \equiv F_{Y|X \geq \text{VaR}_\beta(X)}(\cdot)$ for $\beta \in [0, 1)$.

We start with a description of the core rationale behind the conditional identifiability and elicibility results. To this end, first recall a powerful result about weighted scoring and identification functions.

Theorem C.2 (Gneiting (2011b) Theorem 5; Fissler (2017) Proposition 2.3.5). *Let $\mathcal{F} \subseteq \mathcal{F}' \subseteq \mathcal{F}^0(\mathbb{R}^d)$ and let $\mathbf{V}: \mathbf{A} \times \mathbb{R}^d \rightarrow \mathbb{R}^m$ and $S: \mathbf{A} \times \mathbb{R}^d \rightarrow \mathbb{R}$ be a strict $\mathcal{F}$ -identification function and a strictly $\mathcal{F}$ -consistent scoring function for $\mathbf{T}: \mathcal{F}' \rightarrow \mathcal{P}(\mathbf{A})$. Let $w: \mathbb{R}^d \rightarrow [0, \infty)$ be a non-negative measurable weight function. For any $F \in \mathcal{F}^0(\mathbb{R}^d)$ such that $\bar{w}(F) \in (0, \infty)$ define the probability measure*

$$F^{(w)}(d\mathbf{x}) := \frac{w(\mathbf{x})F(d\mathbf{x})}{\bar{w}(F)}, \quad \mathbf{x} \in \mathbb{R}^d.$$

Define the class $\mathcal{F}^{(w)} = \{F \in \mathcal{F}^0(\mathbb{R}^d): F^{(w)} \in \mathcal{F}\}$ and the functional $\mathbf{T}^{(w)}: \mathcal{F}^{(w)} \rightarrow \mathcal{P}(\mathbf{A})$, $F \mapsto \mathbf{T}^{(w)}(F) := \mathbf{T}(F^{(w)})$. Then $\mathbf{V}^{(w)}(\mathbf{x}, \mathbf{y}) := w(\mathbf{x})\mathbf{V}(\mathbf{x}, \mathbf{y})$ a strict $\mathcal{F}^{(w)}$ -identification function and $S^{(w)}(\mathbf{x}, \mathbf{y}) := w(\mathbf{x})S(\mathbf{x}, \mathbf{y})$ is a strictly $\mathcal{F}^{(w)}$ -consistent scoring function for $\mathbf{T}^{(w)}$.

Secondly, consider a class of bivariate distributions $F_{X,Y} \in \mathcal{F}^0(\mathbb{R}^2)$, where $F_{X,Y} \mapsto \text{VaR}_\beta(F_X)$ is elicitable and identifiable. This is, e.g., the case for $\mathcal{F}_{(\beta)}^0(\mathbb{R}^2)$ defined in (4.1). Conveniently, for any $F_{X,Y} \in \mathcal{F}_{(\beta)}^0(\mathbb{R}^2)$, the distribution $F_{Y|X \geq \text{VaR}_\beta(X)}$ defined in (C.1) reduces to $F_{Y|X > \text{VaR}_\beta(X)}$. Crucially, this conditional distribution can be retrieved as the marginal distribution

$$F_Y^{(w)}(dy) := \int_{\mathbb{R}} F_{X,Y}^{(w)}(dx, dy), \quad y \in \mathbb{R}$$

with weight function $w(x, y) = \mathbb{1}\{x > \text{VaR}_\beta(F_X)\}$. Hence, Theorem C.2 along with well-known results about the elicibility and identifiability of VaR_α , the mean (Gneiting, 2011a), and those for $(\text{VaR}_\alpha, \text{ES}_\alpha)$ (Fissler and Ziegel, 2016) implies that $\text{CoVaR}_{\alpha|\beta}(Y|X)$, $\text{MES}_\beta(Y|X)$ and $(\text{CoVaR}_{\alpha|\beta}(Y|X), \text{CoES}_{\alpha|\beta}(Y|X))$ are identifiable and elicitable on a suitable restriction of the class $\mathcal{F}_{(\beta),v} := \{F_{X,Y} \in \mathcal{F}_{(\beta)}^0(\mathbb{R}^2): \text{VaR}_\beta(F_X) = v\}$, where $v \in \mathbb{R}$.

D. Negative results about joint elicibility

This section shows that the pairs $(\text{VaR}_\beta(X), \text{CoVaR}_{\alpha|\beta}(Y|X))$, $(\text{VaR}_\beta(X), \text{MES}_\beta(Y|X))$ and the triplet $(\text{VaR}_\beta(X), \text{CoVaR}_{\alpha|\beta}(Y|X), \text{CoES}_{\alpha|\beta}(Y|X))$ fail to be elicitable, subject to regularity conditions, despite the fact that they are identifiable and thus also satisfy the CxLS property. This negative result is in stark contrast to the fact that the unconditional counterparts, $\text{VaR}_\alpha(Y)$, $E(Y)$ and $(\text{VaR}_\alpha(Y), \text{ES}_\alpha(Y))$, are all elicitable.

The reason for this intriguing finding lies in integrability conditions: If the functionals were elicitable and admitted strictly consistent scoring functions, which are twice continuously differentiable in expectation for sufficiently smooth distributions, their Hessian would need to be symmetric. Osband's Principle (Fissler and Ziegel, 2016, Theorem 3.2) asserts that, under some richness conditions on the underlying class of distributions $\mathcal{F}$, for any consistent score $S(\mathbf{r}, y)$ and any strict identification function $\mathbf{V}(\mathbf{r}, y)$, there exists a matrix-valued function $\mathbf{h}(\mathbf{r})$ such that $\nabla \bar{S}(\mathbf{r}, F) = \mathbf{h}(\mathbf{r}) \bar{\mathbf{V}}(\mathbf{r}, F)$ for all $\mathbf{r}$ and for all $F \in \mathcal{F}$. Using the strict identification functions from Theorem 4.2 it is possible to show that such a function $\mathbf{h}(\mathbf{r})$ cannot exist for the three functionals of interest here.

This is an instance of a more general phenomenon: Dimitriadis *et al.* (2020) argue that there is generally a *gap* between the classes of strictly consistent scoring functions and strict identification functions for multivariate functionals, which is due to these integrability conditions. The following example illustrates this.

Example D.1. Consider the double quantile functional $\mathbf{T} = (q_\alpha, q_\beta)$, $0 < \alpha < \beta < 1$ with the standard identification function $\mathbf{V}((r_1, r_2), y) = (\mathbb{1}\{y \leq r_1\} - \alpha, \mathbb{1}\{y \leq r_2\} - \beta)'$. Then the matrix-valued function $\mathbf{h}$ induced by a consistent scoring function is necessarily diagonal with non-negative diagonal entries such that h_{11} is only a function of r_1 and h_{22} is only a function of r_2 . Put differently, any consistent scoring function for the double quantile is necessarily the sum of two consistent scoring functions for the individual quantiles. On the other hand, it is easy to see that for any matrix-valued function $\mathbf{h}$ such that $\mathbf{h}(\mathbf{r})$ has full rank everywhere, the product $\mathbf{h}(\mathbf{r})\mathbf{V}(\mathbf{r}, y)$ is a strict identification function for $\mathbf{T}$. Clearly, the latter class is considerably larger.

In the case at hand, this gap is extreme: While there exist strict identification functions, the respective classes of strict scoring functions are empty. We are only aware of one more instance in the literature of such an extreme gap (Fissler *et al.*, 2021, Subsection 4.3).

As mentioned, we need to impose certain richness conditions along with smoothness conditions on the class of bivariate distributions $\mathcal{F} \subset \mathcal{F}^0(\mathbb{R}^2)$. Then, the negative results hold on any class containing such an $\mathcal{F}$. We state the particular conditions needed for $\mathcal{F}$ explicitly, noting that the class of bivariate normal distributions (excluding perfect correlation) along with all mixtures always satisfies these conditions.

Assumption D.2. (a) *For any $F_{X,Y} \in \mathcal{F}$ with marginals F_X and F_Y , the distribution $F_X F_Y$ is in $\mathcal{F}$, having the same marginal distributions, but independent components.*

(b) $\mathcal{F}$ is convex.

(c) All $F_{X,Y} \in \mathcal{F}$ are continuously differentiable and admit a strictly positive continuous density $f_{X,Y}$.

In the following, let $\text{conv}(M)$ denote the convex hull of some set M , and $\text{int}(M)$ its interior.

Proposition D.3. Let $\alpha, \beta \in (0, 1)$ and denote by $\mathbf{T}$ the functional mapping $F_{X,Y}$ to $(\text{VaR}_\beta(F_X), \text{CoVaR}_{\alpha|\beta}(F_{X,Y}))$. Let $\mathcal{F} \subseteq \mathcal{F}^0(\mathbb{R}^2)$ satisfy Assumption D.2 and suppose that

- (i) $\mathbf{A} := \mathbf{T}(\mathcal{F}) = \{\mathbf{T}(F) : F \in \mathcal{F}\} \subseteq \mathbb{R}^2$ has a non-empty and simply connected interior $\text{int}(\mathbf{A})$.
- (ii) For $\mathbf{V} = \mathbf{V}^{(\text{VaR}, \text{CoVaR})}$ defined in (4.2) it holds that for all $(v, c) \in \text{int}(\mathbf{A})$ there are $F_1, F_2, F_3 \in \mathcal{F}$ such that $\mathbf{0} \in \text{int}(\text{conv}\{\bar{\mathbf{V}}((v, c), F_1), \bar{\mathbf{V}}((v, c), F_2), \bar{\mathbf{V}}((v, c), F_3)\})$.
- (iii) For all $F_{X,Y} \in \mathcal{F}$ and for all $(v, c) \in \text{int}(\mathbf{A})$ there is some $\tilde{F}_{X,Y} \in \mathcal{F}$ such that $F_Y(c) = \tilde{F}_Y(c)$, $F_X(v) = \tilde{F}_X(v)$, $f_Y(c) = \tilde{f}_Y(c)$, $f_X(v) \neq \tilde{f}_X(v)$.

Then, for any $\mathcal{F} \subseteq \mathcal{F}' \subseteq \mathcal{F}^0(\mathbb{R}^2)$, there is no strictly $\mathcal{F}$ -consistent scoring function $S : \mathbf{A} \times \mathbb{R}^2 \rightarrow \mathbb{R}$ for $\mathbf{T}$, such that the expected score $\bar{S}(\cdot, F) : \mathbf{A} \rightarrow \mathbb{R}$ is twice partially differentiable with continuous second order derivatives for all $F \in \mathcal{F}$.

It is worth emphasising that it is the *expected* score in Proposition D.3 which is assumed to be twice differentiable with continuous second order derivatives, and not necessarily the score itself. E.g., while the usual score for VaR_α , $S(x, y) = (\mathbb{1}\{y \leq x\} - \alpha)(x - y)$ clearly fails to even be differentiable everywhere, the expected score $\bar{S}(x, F)$ is twice continuously differentiable if F is continuously differentiable, with $\bar{S}'(x, F) = F(x) - \alpha$ and $\bar{S}''(x, F) = F'(x) = f(x)$. Therefore, requiring this smoothness condition on the expected score seems to be reasonable on $\mathcal{F}$ of the form considered in the proposition. The other two negative results are of a similar form.

Proposition D.4. Let $\alpha, \beta \in (0, 1)$ and denote by $\mathbf{T}$ the functional mapping $F_{X,Y}$ to $(\text{VaR}_\beta(F_X), \text{CoVaR}_{\alpha|\beta}(F_{X,Y}), \text{CoES}_{\alpha|\beta}(F_{X,Y}))$. Let $\mathcal{F} \subseteq \mathcal{F}^0(\mathbb{R}^2)$ satisfy Assumption D.2 and suppose that

- (i) $\mathbf{A} := \mathbf{T}(\mathcal{F}) = \{\mathbf{T}(F) : F \in \mathcal{F}\} \subseteq \mathbb{R}^3$ has a non-empty and simply connected interior $\text{int}(\mathbf{A})$.
- (ii) For $\mathbf{V} = \mathbf{V}^{(\text{VaR}, \text{CoVaR}, \text{CoES})}$ defined in (4.3) it holds that for all $(v, c, \mu) \in \text{int}(\mathbf{A})$ there are $F_1, \dots, F_4 \in \mathcal{F}$ such that $\mathbf{0} \in \text{int}(\text{conv}\{\bar{\mathbf{V}}((v, c, \mu), F_1), \dots, \bar{\mathbf{V}}((v, c, \mu), F_4)\})$.
- (iii) For all $F_{X,Y} \in \mathcal{F}$ and for all $(v, c, \mu) \in \text{int}(\mathbf{A})$ there are some $\tilde{F}_{X,Y}, \hat{F}_{X,Y} \in \mathcal{F}$ such that $F_Y(c) = \tilde{F}_Y(c) = \hat{F}_Y(c)$, $\int_c^\infty y f_Y(y) dy = \int_c^\infty y \tilde{f}_Y(y) dy = \int_c^\infty y \hat{f}_Y(y) dy$, $F_X(v) = \tilde{F}_X(v) = \hat{F}_X(v)$, $f_Y(c) = \tilde{f}_Y(c) \neq \hat{f}_Y(c)$, $f_X(v) \neq \tilde{f}_X(v) = \hat{f}_X(v)$.

Then, for any $\mathcal{F} \subseteq \mathcal{F}' \subseteq \mathcal{F}^0(\mathbb{R}^2)$, there is no strictly $\mathcal{F}'$ -consistent scoring function $S : \mathbf{A} \times \mathbb{R}^2 \rightarrow \mathbb{R}$ for $\mathbf{T}$, such that the expected score $\bar{S}(\cdot, F) : \mathbf{A} \rightarrow \mathbb{R}$ is twice partially differentiable with continuous second order derivatives for all $F \in \mathcal{F}$.

Proposition D.5. Let $\alpha, \beta \in (0, 1)$ and denote by $\mathbf{T}$ the functional mapping $F_{X,Y}$ to $(\text{VaR}_\beta(F_X), \text{MES}_\beta(F_{X,Y}))$. Let $\mathcal{F} \subseteq \mathcal{F}^0(\mathbb{R}^2)$ satisfy Assumption D.2 and suppose that

- (i) $\mathbf{A} := \mathbf{T}(\mathcal{F}) = \{\mathbf{T}(F) : F \in \mathcal{F}\} \subseteq \mathbb{R}^2$ has a non-empty and simply connected interior $\text{int}(\mathbf{A})$.
- (ii) For $\mathbf{V} = \mathbf{V}^{(\text{VaR}, \text{MES})}$ defined in (4.4) it holds that for all $(v, \mu) \in \text{int}(\mathbf{A})$ there are $F_1, F_2, F_3 \in \mathcal{F}$ such that $\mathbf{0} \in \text{int}(\text{conv}\{\bar{\mathbf{V}}((v, \mu), F_1), \bar{\mathbf{V}}((v, \mu), F_2), \bar{\mathbf{V}}((v, \mu), F_3)\})$.
- (iii) For all $F_{X,Y} \in \mathcal{F}$ and for all $(v, \mu) \in \text{int}(\mathbf{A})$ there is some $\tilde{F}_{X,Y} \in \mathcal{F}$ such that $E(F_Y) = E(\tilde{F}_Y)$, $F_X(v) = \tilde{F}_X(v)$, $f_X(v) \neq \tilde{f}_X(v)$.

Then, for any $\mathcal{F} \subseteq \mathcal{F}' \subseteq \mathcal{F}^0(\mathbb{R}^2)$, there is no strictly $\mathcal{F}$ -consistent scoring function $S : \mathbf{A} \times \mathbb{R}^2 \rightarrow \mathbb{R}$ for $\mathbf{T}$, such that the expected score $\bar{S}(\cdot, F) : \mathbf{A} \rightarrow \mathbb{R}$ is twice partially differentiable with continuous second order derivatives for all $F \in \mathcal{F}$.

E. Proofs

We first prove Proposition 4.1 via Proposition B.4. Specifically, we show that $\text{CoVaR}_{\alpha|\beta}$, $\text{CoES}_{\alpha|\beta}$ and MES_{β} violate the CxLS property on any class $\mathcal{F} \subseteq \mathcal{F}^0(\mathbb{R}^2)$ containing all bivariate normal distributions along with their mixtures. Recall from Section C that we take the systemic risk measures to be defined in terms of $F_{Y|X \succcurlyeq \text{VaR}_{\beta}(X)}$ from (C.1) here.

Remark E.1. Before proving Proposition 4.1, we outline the core rationale why the systemic risk measures of interest violate the CxLS property on a class $\mathcal{F}$ under very mild conditions. Suppose $\mathcal{F}$ contains two bivariate distributions F^0, F^1 of the form $F^0(x, y) = C(F_1^0(x), F_2(y))$, $F^1(x, y) = C(F_1^1(x), F_2(y))$. Here, $C : [0, 1]^2 \rightarrow [0, 1]$ is a copula and F_1^0, F_1^1, F_2 are univariate marginal distributions. Let $(X^0, Y^0) \sim F^0$ and $(X^1, Y^1) \sim F^1$. If F_1^0 and F_1^1 are continuous, then

$$F_{Y^0|X^0 \succcurlyeq \text{VaR}_{\beta}(X^0)}(y) = F_{Y^1|X^1 \succcurlyeq \text{VaR}_{\beta}(X^1)}(y) = \left[F_2(y) - C(\beta, F_2(y)) \right] / (1 - \beta).$$

Hence, clearly all functionals of $F_{Y^0|X^0 \succcurlyeq \text{VaR}_{\beta}(X^0)}$ and $F_{Y^1|X^1 \succcurlyeq \text{VaR}_{\beta}(X^1)}$ coincide. Next, consider the convex combination $F^{\lambda}(x, y) := (1 - \lambda)F^0(x, y) + \lambda F^1(x, y)$ and let $(X^{\lambda}, Y^{\lambda}) \sim F^{\lambda}$ where $\lambda \in (0, 1)$. To simplify the argument, suppose that F_1^0 and F_1^1 are strictly increasing (otherwise, the argument can be adapted). If $\text{VaR}_{\beta}(X^0) < \text{VaR}_{\beta}(X^1)$, then $\text{VaR}_{\beta}(X^0) < \text{VaR}_{\beta}(X^{\lambda}) < \text{VaR}_{\beta}(X^1)$. Hence, $\beta^0 := F_1^0(\text{VaR}_{\beta}(X^{\lambda})) > \beta > F_1^1(\text{VaR}_{\beta}(X^{\lambda})) =: \beta^1$. We obtain

$$F_{Y^{\lambda}|X^{\lambda} \succcurlyeq \text{VaR}_{\beta}(X^{\lambda})}(y) = (1 - \lambda) \frac{F_2(y) - C(\beta^0, F_2(y))}{1 - \beta^0} + \lambda \frac{F_2(y) - C(\beta^1, F_2(y))}{1 - \beta^1}.$$

If C is not the independence copula, then different choices of $u_1 \in (0, 1)$ generally result in different mappings $[0, 1] \ni u_2 \mapsto [u_2 - C(u_1, u_2)] / (1 - u_1)$. Therefore, $F_{Y^{\lambda}|X^{\lambda} \succcurlyeq \text{VaR}_{\beta}(X^{\lambda})}$ will in general be different from $F_{Y^0|X^0 \succcurlyeq \text{VaR}_{\beta}(X^0)}$. Hence, a violation of the CxLS property can be shown upon choosing appropriate F_1^0, F_1^1 and $\lambda \in (0, 1)$.

Proof of Proposition 4.1: By Proposition B.4, it suffices to show that $\text{CoVaR}_{\alpha|\beta}$, $\text{CoES}_{\alpha|\beta}$ and MES_{β} violate the CxLS property on any class $\mathcal{F} \subseteq \mathcal{F}^0(\mathbb{R}^2)$ containing all bivariate normal distributions along with their mixtures.

To task this, consider $X, Y \sim \mathcal{N}(0, 1)$ which are jointly Gaussian and have a correlation of $\rho \neq 0$. Let $(X^0, Y^0) = (X - 1, Y)$ and $(X^1, Y^1) = (X + 1, Y)$ with distribution functions F^0 and F^1 . Moreover, let $F^{0.5} = 0.5F^0 + 0.5F^1$ and let $(X^{0.5}, Y^{0.5})$ have distribution $F^{0.5}$. Clearly $\text{VaR}_\beta(X^0) = \Phi^{-1}(\beta) - 1$ and $\text{VaR}_\beta(X^1) = \Phi^{-1}(\beta) + 1$, where Φ is the distribution function of a standard normal and ϕ is its density function. For $\text{VaR}_\beta(X^{0.5})$, there is generally no closed form solution. We therefore illustrate the result numerically for $\beta = 0.95$. Here, $\text{VaR}_{0.95}(X^{0.5}) \approx 2.28$, while $\Phi^{-1}(0.95) \approx 1.64$. It holds that

$$\begin{aligned} F_{Y^0|X^0 \succ \text{VaR}_\beta(X^0)}(y) &= F_{Y^1|X^1 \succ \text{VaR}_\beta(X^1)}(y) = \mathbb{P}(Y \leq y \mid X > \Phi^{-1}(\beta)), \\ F_{Y^{0.5}|X^{0.5} \succ \text{VaR}_\beta(X^{0.5})}(y) &= 0.5 \mathbb{P}\{Y \leq y \mid X - 1 > \text{VaR}_\beta(X^{0.5})\} \\ &\quad + 0.5 \mathbb{P}\{Y \leq y \mid X + 1 > \text{VaR}_\beta(X^{0.5})\}. \end{aligned}$$

We exploit the fact that $\mathbb{E}[Y \mid X > z] = \rho\phi(z)/[1 - \Phi(z)]$. Therefore, $\text{MES}_\beta(Y^0|X^0) = \text{MES}_\beta(Y^1|X^1) = \mathbb{E}[Y \mid X > \Phi^{-1}(\beta)] = \rho\phi(\Phi^{-1}(\beta))/[1 - \beta]$. On the other hand,

$$\text{MES}_\beta(Y^{0.5}|X^{0.5}) = 0.5\rho \left[\frac{\phi(\text{VaR}_\beta(X^{0.5}) + 1)}{1 - \Phi(\text{VaR}_\beta(X^{0.5}) + 1)} + \frac{\phi(\text{VaR}_\beta(X^{0.5}) - 1)}{1 - \Phi(\text{VaR}_\beta(X^{0.5}) - 1)} \right].$$

These values do not coincide for $\beta \in (0, 1)$ and $\rho \neq 0$. E.g., $\text{MES}_{0.95}(Y^0|X^0) = 2.06\rho$ and $\text{MES}_{0.95}(Y^{0.5}|X^{0.5}) = 2.65\rho$.

For $\text{CoVaR}_{\alpha|\beta}$, it is sufficient to check that $\text{VaR}_\alpha(F_{Y^{0.5}|X^{0.5} \succ \text{VaR}_\beta(X^{0.5})})$ does not coincide with $\text{VaR}_\alpha(F_{Y^0|X^0 \succ \text{VaR}_\beta(X^0)})$. This can only be solved numerically and we illustrate again the case for $\alpha = \beta = 0.95$, setting $\rho = 0.8$, emphasising that other choices for $\alpha, \beta \in (0, 1)$ and $\rho \neq 0$ also hold similarly. We get $\text{VaR}_{0.95}(F_{Y^0|X^0 \succ \text{VaR}_{0.95}(X^0)}) \approx 2.77$ and $\text{VaR}_{0.95}(F_{Y^{0.5}|X^{0.5} \succ \text{VaR}_{0.95}(X^{0.5})}) \approx 3.20$.

The results for $\text{CoES}_{\alpha|\beta}$ follow from the fact that ES_α itself violates the CxLS-property. We mention that $(\text{CoVaR}_{\alpha|\beta}, \text{CoES}_{\alpha|\beta})$ also violates the CxLS-property, which can be shown similarly as for $\text{CoVaR}_{\alpha|\beta}$. ■

Next, we prove Theorem 5.1. The technical assumptions of Theorem 5.1 are only required to ensure a multivariate central limit theorem for the score differences $\{\mathbf{d}_t\}$

$$\boldsymbol{\Omega}_n^{-1/2} \sqrt{n} \mathbf{d}_n \xrightarrow{d} N(\mathbf{0}, \mathbf{I}_{2 \times 2}), \quad \text{as } n \rightarrow \infty, \quad (\text{E.1})$$

and consistent long-run variance estimation

$$\|\widehat{\boldsymbol{\Omega}}_n - \boldsymbol{\Omega}_n\| \xrightarrow{\text{P}} \mathbf{0}, \quad \text{as } n \rightarrow \infty. \quad (\text{E.2})$$

Thus, Assumption E.2 may be replaced by any other conditions ensuring (E.1) and (E.2). Some of the most general conditions can be found in Davidson (1994, Section 24.4) for (E.1) and De Jong and Davidson (2000, Theorem 2.1) for (E.2). While our conditions are slightly less general, they are much simpler.

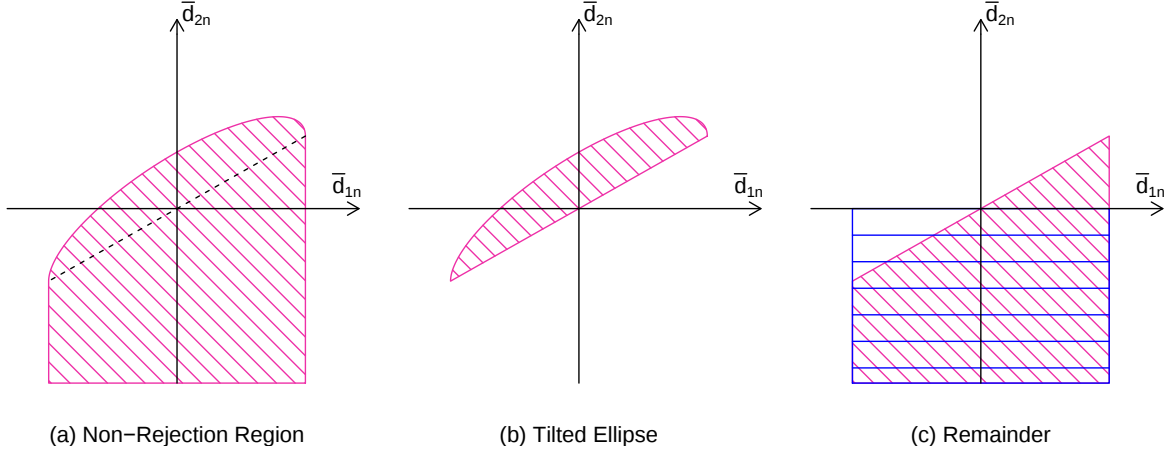


Figure 8: Decomposition of Non-Rejection Region.

Assumption E.2. *B1: There is some $\Delta_d < \infty$ such that $E|\mathbf{d}'_t \mathbf{d}_t|^r \leq \Delta_d$ for all $t \geq 1$, where $r > 2$.*

B2: $\{\mathbf{d}_t\}$ is α -mixing of size $-2r/(r-2)$ or ϕ -mixing of size $-r/(r-1)$.

B3: The sequence of integers m_n satisfies $m_n \rightarrow \infty$ and $m_n = o(n^{1/4})$, as $n \rightarrow \infty$.

B4: There is some $\Delta_w < \infty$ such that $|w_{n,h}| \leq \Delta_w$ for all $n \in \mathbb{N}$ and $h \in \{1, \dots, m_n\}$, and $w_{n,h} \rightarrow 1$, as $n \rightarrow \infty$, for all $h = 1, \dots, m_n$.

Proof of Theorem 5.1: The proof is similar to that of Theorem 4 in [Giacomini and White \(2006\)](#), so we only sketch it here. A Cramér–Wold device and Theorem 5.20 in [White \(2001\)](#) ensure that $\boldsymbol{\Omega}_n^{-1/2} \sqrt{n} \bar{\mathbf{d}}_n \xrightarrow{d} N(\mathbf{0}, \mathbf{I}_{2 \times 2})$, as $n \rightarrow \infty$, under B1 and B2. The fact that $\|\hat{\boldsymbol{\Omega}}_n - \boldsymbol{\Omega}_n\| \xrightarrow{P} \mathbf{0}$, as $n \rightarrow \infty$, follows from Theorem 6.20 in [White \(2001\)](#) under B1–B4. ■

Proof of Proposition 5.2: By construction of the non-rejection region (displayed in Figure 8 (a)), we have

$$\begin{aligned} \sup_{c \leq 0} \lim_{n \rightarrow \infty} P \left\{ \mathcal{T}_n^{\text{OS}} > \chi_{2,1-\tilde{\nu}}^2 \mid H_0^{(c)} \text{ holds} \right\} &= \lim_{n \rightarrow \infty} P \left\{ \mathcal{T}_n^{\text{OS}} > \chi_{2,1-\tilde{\nu}}^2 \mid H_0^{(0)} \text{ holds} \right\} \\ &= \lim_{n \rightarrow \infty} P \left\{ \mathcal{T}_n^{\text{OS}} > \chi_{2,1-\tilde{\nu}}^2 \mid E[\bar{d}_{1n}] = 0 \text{ and } E[\bar{d}_{2n}] = 0 \right\} = 1 - P \left\{ \mathcal{T}^{\text{OS}} \leq \chi_{2,1-\tilde{\nu}}^2 \right\}, \end{aligned} \quad (\text{E.3})$$

where, for $\boldsymbol{\Omega} = ((\sigma_{11}, \sigma_{12})', (\sigma_{12}, \sigma_{22})')$,

$$\mathcal{T}^{\text{OS}} = \left(Z_1, \max \{ Z_2, (\sigma_{12}/\sigma_{11}) Z_1 \} \right) \boldsymbol{\Omega}^{-1} \begin{pmatrix} Z_1 \\ \max \{ Z_2, (\sigma_{12}/\sigma_{11}) Z_1 \} \end{pmatrix},$$

with $(Z_1, Z_2)' \sim N(\mathbf{0}, \boldsymbol{\Omega})$, which is the distributional limit of $\sqrt{n}(\bar{d}_{1n}, \bar{d}_{2n})'$. By construction, $P\{\mathcal{T}^{\text{OS}} \leq \chi_{2,1-\tilde{\nu}}^2\}$ corresponds to the probability that $(Z_1, Z_2)'$ falls into the non-rejection region

sketched in Figure 8 (a). This region admits a disjoint decomposition into the upper half of the tilted ellipse (displayed in panel (b)), corresponding to $Z_2 > (\sigma_{12}/\sigma_{11})Z_1$ and the remainder (displayed in panel (c)), corresponding to $Z_2 \leq (\sigma_{12}/\sigma_{11})Z_1$. Since the full ellipse has probability content $1 - \tilde{\nu}$, we get

$$\mathbb{P} \left\{ (Z_1, Z_2)' \text{ falls into upper half of tilted ellipse} \right\} = \frac{1}{2}(1 - \tilde{\nu}).$$

By symmetry of the normal distribution, the probability content of the remainder equals that of the area shaded in blue in panel (c). Thus,

$$\mathbb{P} \left\{ (Z_1, Z_2)' \text{ falls into remainder} \right\} = \mathbb{P} \left\{ |Z_1| \leq e_{1,\max}, Z_2 \leq 0 \right\} = \frac{1}{2} \mathbb{P} \left\{ |Z_1| \leq e_{1,\max} \right\},$$

where $e_{1,\max}$ is the maximal extension of the tilted ellipse in the horizontal direction. For the ellipse $\{z \in \mathbb{R}^2: z' \Omega^{-1} z \leq \chi_{2,1-\tilde{\nu}}^2\}$, this is $e_{1,\max} = \sqrt{\sigma_{11}\chi_{2,1-\tilde{\nu}}^2}$. Thus,

$$\mathbb{P} \left\{ (Z_1, Z_2)' \text{ falls into remainder} \right\} = \frac{1}{2} \mathbb{P} \left\{ \sigma_{11}^{-1/2} |Z_1| \leq \sqrt{\chi_{2,1-\tilde{\nu}}^2} \right\} = \frac{1}{2} F_{\chi_1^2}(\chi_{2,1-\tilde{\nu}}^2).$$

We end up with $\mathbb{P} \left\{ \mathcal{T}^{\text{OS}} \leq \chi_{2,1-\tilde{\nu}}^2 \right\} = \frac{1}{2}(1 - \tilde{\nu} + F_{\chi_1^2}(\chi_{2,1-\tilde{\nu}}^2))$. Combining this with (E.3), the conclusion follows. $\blacksquare$

Proof of Proposition D.3: The proof exploits Osband's Principle in Fissler and Ziegel (2016, Theorem 3.2), which originates from the seminal work of Osband (1985). Suppose there is an $\mathcal{F}$ -consistent scoring function $S: \mathbf{A} \times \mathbb{R}^2 \rightarrow \mathbb{R}$, whose expectation $\bar{S}: \mathbf{A} \rightarrow \mathbb{R}$ is twice partially differentiable with continuous second-order derivatives for all $F \in \mathcal{F}$. Using Assumptions D.2 and conditions (i) and (ii) of Proposition D.3, Fissler and Ziegel (2016, Theorem 3.2) shows that there is a differentiable matrix-valued function $\mathbf{h}: \text{int}(\mathbf{A}) \rightarrow \mathbb{R}^{2 \times 2}$ with components h_{ij} , $i, j = 1, 2$, such that for all $(v, c) \in \text{int}(\mathbf{A})$ and for all $F_{X,Y} \in \mathcal{F}$

$$\nabla \bar{S}((v, c), F_{X,Y}) = \mathbf{h}(v, c) \bar{\mathbf{V}}((v, c), F_{X,Y}). \quad (\text{E.4})$$

The symmetry of the Hessian $\nabla^2 \bar{S}((v, c), F_{X,Y})$ (Fissler and Ziegel, 2016, Corollary 3.3) then yields that

$$\partial_c \partial_v \bar{S}((v, c), F_{X,Y}) = \partial_v \partial_c \bar{S}((v, c), F_{X,Y}) \quad (\text{E.5})$$

for all $(v, c) \in \text{int}(\mathbf{A})$ and for all $F_{X,Y} \in \mathcal{F}$. Denote by $\mathcal{F}^0(\mathbb{R})^{\otimes 2}$ the family of independent bivariate distributions. Then, straightforward calculations yield that for $F_{X,Y} \in \mathcal{F} \cap \mathcal{F}^0(\mathbb{R})^{\otimes 2}$, it holds that

$$\begin{aligned} \bar{V}_1((v, c), F_{X,Y}) &= F_X(v) - \beta, \\ \bar{V}_2((v, c), F_{X,Y}) &= (F_Y(c) - \alpha)(1 - F_X(v)), \\ \partial_c \bar{V}_1((v, c), F_{X,Y}) &= 0, \\ \partial_v \bar{V}_1((v, c), F_{X,Y}) &= f_X(v), \end{aligned}$$

$$\begin{aligned}\partial_c \bar{V}_2((v, c), F_{X,Y}) &= f_Y(c)(1 - F_X(v)), \\ \partial_v \bar{V}_2((v, c), F_{X,Y}) &= -f_X(v)(F_Y(c) - \alpha).\end{aligned}$$

Let $(v^*, c^*) = \mathbf{T}(F_{X,Y})$. Using (E.4), the chain rule, and the fact that the expected identification function vanishes at the true report, we get

$$h_{12}(v^*, c^*) \partial_c \bar{V}_2((v^*, c^*), F_{X,Y}) = \partial_c \partial_v \bar{S}((v^*, c^*), F_{X,Y}).$$

Similarly,

$$\partial_v \partial_c \bar{S}((v^*, c^*), F_{X,Y}) = h_{21}(v^*, c^*) \partial_v \bar{V}_1((v^*, c^*), F_{X,Y}) + h_{22}(v^*, c^*) \partial_v \bar{V}_2((v^*, c^*), F_{X,Y}).$$

For $F_{X,Y} \in \mathcal{F} \cap \mathcal{F}^0(\mathbb{R})^{\otimes 2}$, (E.5) yields that

$$h_{12}(v^*, c^*) f_Y(c^*)(1 - \beta) = h_{21}(v^*, c^*) f_X(v^*). \quad (\text{E.6})$$

Exploiting part (iii) of Proposition D.3 we get that $h_{21}(v^*, c^*) = 0$. Since $f_Y(c^*) > 0$, also $h_{12}(v^*, c^*) = 0$. Due to the surjectivity of $\mathbf{T}$ onto $\mathbf{A}$, we obtain that for all $(v, c) \in \text{int}(\mathbf{A})$, $h_{12}(v, c) = h_{21}(v, c) = 0$. Now we exploit (E.5) for a general point $(v, c) \in \text{int}(\mathbf{A})$, such that $F_X(v) \neq \beta$ and $F_Y(c) \neq \alpha$. For $F_{X,Y} \in \mathcal{F} \cap \mathcal{F}^0(\mathbb{R})^{\otimes 2}$ we obtain $\partial_c h_{11}(v, c)(F_X(v) - \beta) = (F_Y(c) - \alpha)[\partial_v h_{22}(v, c)(1 - F_X(v)) - h_{22}(v, c)f_X(v)]$. Again invoking condition (iii), we get that $h_{22}(v, c) = 0$. Since we can perform the argument for any $(v, c) \in \text{int}(\mathbf{A})$, making use of the surjectivity, it holds that $h_{22} \equiv \partial_v h_{22} \equiv 0$. Therefore, also $\partial_c h_{11} \equiv 0$. In conclusion, the matrix function $\mathbf{h}$ is a diagonal matrix and the only non vanishing entry is $h_{11}(v, c) = g(v)$ for some function $g: \mathbb{R} \rightarrow \mathbb{R}$. Hence, using (E.4) we obtain that $\partial_c \bar{S}((v, c), F_{X,Y}) = 0$ for all $(v, c) \in \text{int}(\mathbf{A})$. Since $\text{int}(\mathbf{A})$ is open, non-empty and simply connected, the expected score is constant in c . This shows that S cannot be strictly $\mathcal{F}$ -consistent. ■

Proof of Proposition D.4: The proof basically follows along the lines of the proof of Proposition D.3. So we get the counterpart of (E.4)

$$\nabla \bar{S}((v, c, e), F_{X,Y}) = \mathbf{h}(v, c, e) \bar{\mathbf{V}}((v, c, e), F_{X,Y}), \quad (\text{E.7})$$

where $\mathbf{h}(v, c, e) \in \mathbb{R}^{3 \times 3}$. For $F_{X,Y} \in \mathcal{F} \cap \mathcal{F}^0(\mathbb{R})^{\otimes 2}$ we get that

$$\begin{aligned}\bar{V}_1((v, c, e), F_{X,Y}) &= F_X(v) - \beta, \\ \bar{V}_2((v, c, e), F_{X,Y}) &= (1 - F_X(v))(F_Y(c) - \alpha), \\ \bar{V}_3((v, c, e), F_{X,Y}) &= (1 - F_X(v)) \left[e - \frac{1}{1 - \alpha} \left(\int_c^\infty y f_Y(y) dy + c(F_Y(c) - \alpha) \right) \right],\end{aligned}$$

and

$$\partial_v \bar{V}_1((v, c, e), F_{X,Y}) = f_X(v),$$

$$\begin{aligned}
\partial_c \bar{V}_1((v, c, e), F_{X,Y}) &= \partial_e \bar{V}_1((v, c, e), F_{X,Y}) = 0, \\
\partial_v \bar{V}_2((v, c, e), F_{X,Y}) &= -f_X(v)(F_Y(c) - \alpha), \\
\partial_c \bar{V}_2((v, c, e), F_{X,Y}) &= f_Y(c)(1 - F_X(v)), \\
\partial_e \bar{V}_2((v, c, e), F_{X,Y}) &= 0, \\
\partial_v \bar{V}_3((v, c, e), F_{X,Y}) &= f_X(v) \left[e - \frac{1}{1-\alpha} \left(\int_c^\infty y f_Y(y) dy + c(F_Y(c) - \alpha) \right) \right], \\
\partial_c \bar{V}_3((v, c, e), F_{X,Y}) &= -(1 - F_X(v))(F_Y(c) - \alpha)/(1 - \alpha), \\
\partial_e \bar{V}_3((v, c, e), F_{X,Y}) &= (1 - F_X(v)).
\end{aligned}$$

The symmetry of the Hessian implies that for all $(v, c, \mu) \in \text{int}(\mathbf{A})$ and for all $F_{X,Y} \in \mathcal{F} \cap \mathcal{F}^0(\mathbb{R})^{\otimes 2}$,

$$\partial_c \partial_v \bar{S}((v, c, e), F_{X,Y}) = \partial_v \partial_c \bar{S}((v, c, e), F_{X,Y}), \quad (\text{E.8})$$

$$\partial_e \partial_v \bar{S}((v, c, e), F_{X,Y}) = \partial_v \partial_e \bar{S}((v, c, e), F_{X,Y}), \quad (\text{E.9})$$

$$\partial_e \partial_c \bar{S}((v, c, e), F_{X,Y}) = \partial_c \partial_e \bar{S}((v, c, e), F_{X,Y}). \quad (\text{E.10})$$

Using (E.7) to evaluate (E.8) at $(v^*, c^*, e^*) = \mathbf{T}(F_{X,Y}) \in \text{int}(\mathbf{A})$ implies the identity

$$h_{12}(v^*, c^*, e^*) f_Y(c^*)(1 - \beta) = h_{21}(v^*, c^*, e^*) f_X(v^*).$$

Using condition (iii) of Proposition D.4 together with the positivity of f_Y and with the surjectivity of $\mathbf{T}$ implies that $h_{12} \equiv h_{21} \equiv 0$. Using (E.7) to evaluate (E.9) at $(v^*, c^*, e^*) = \mathbf{T}(F_{X,Y}) \in \text{int}(\mathbf{A})$ implies that $h_{13}(v^*, c^*, e^*)(1 - \beta) = h_{31}(v^*, c^*, e^*) f_X(v^*)$. Using the same arguments, we obtain that $h_{13} \equiv h_{31} \equiv 0$. Finally, evaluating (E.10) with (E.7) at $(v^*, c^*, e^*) = \mathbf{T}(F_{X,Y}) \in \text{int}(\mathbf{A})$ yields that $h_{23}(v^*, c^*, e^*)(1 - \beta) = h_{32}(v^*, c^*, e^*) f_Y(c^*)(1 - \beta)$. Hence, invoking the surjectivity of $\mathbf{T}$ and condition (iii), we get that $h_{23} \equiv h_{32} \equiv 0$. We already know that $\mathbf{h}$ is a diagonal matrix. (E.8) therefore simplifies to $\partial_c h_{11}(v, c, e)(F_X(v) - \beta) = (F_Y(c) - \alpha)[\partial_v h_{22}(v, c, e)(1 - F_X(v)) - h_{22}(v, c, e)f_X(v)]$. Evaluating this for points such that $F_X(v) \neq \beta$ and $F_Y(c) \neq \alpha$, we can again exploit condition (iii) of Proposition D.4 and the surjectivity of $\mathbf{T}$ to obtain $h_{22} \equiv \partial_v h_{22} \equiv \partial_c h_{11} \equiv 0$. Now, we can evaluate (E.10) for some point $(v, c, e) \in \text{int}(\mathbf{A})$ such that $F_Y(c) \neq \alpha$ and $e \neq \frac{1}{1-\alpha} \left(\int_c^\infty y f_Y(y) dy + c(F_Y(c) - \alpha) \right)$ to obtain $0 = \left[e - \frac{1}{1-\alpha} \left(\int_c^\infty y f_Y(y) dy + c(F_Y(c) - \alpha) \right) \right] \times \left[(1 - F_X(v)) \partial_v h_{33}(v, c, e) + f_X(v) h_{33}(v, c, e) \right]$.

Again invoking condition (iii) and the surjectivity, we get that $h_{33} \equiv 0$. Finally, (E.9) simplifies to $\partial_e h_{11}(v, c, e)(F_X(v) - \beta) = 0$, which implies that $\partial_e h_{11} \equiv 0$. In conclusion, the only non-vanishing component of $\mathbf{h}$ is $h_{11}(v, c, e) = g(v)$ for some function $g: \mathbb{R} \rightarrow \mathbb{R}$. With the same arguments as in the proof of Proposition D.3 we deduce that $(v, c, e) \mapsto \bar{S}((v, c, e), F_{X,Y})$ is constant in c and e . Therefore, S cannot be strictly $\mathcal{F}$ -consistent. ■

Proof of Proposition D.5: The proof first follows along the lines of the proof of Proposi-

tion D.3 up to equation (E.5), *mutatis mutandis*. For $F_{X,Y} \in \mathcal{F} \cap \mathcal{F}^0(\mathbb{R})^{\otimes 2}$, we get

$$\begin{aligned}\bar{V}_1((v, \mu), F_{X,Y}) &= F_X(v) - \beta, \\ \bar{V}_2((v, \mu), F_{X,Y}) &= (1 - F_X(v))(\mu - \mathbb{E}(F_Y)), \\ \partial_v \bar{V}_1((v, \mu), F_{X,Y}) &= f_X(v), \\ \partial_\mu \bar{V}_1((v, \mu), F_{X,Y}) &= 0, \\ \partial_v \bar{V}_2((v, \mu), F_{X,Y}) &= -f_X(v)(\mu - \mathbb{E}(F_Y)), \\ \partial_\mu \bar{V}_2((v, \mu), F_{X,Y}) &= (1 - F_X(v)).\end{aligned}$$

Let $(v^*, \mu^*) = \mathbf{T}(F_{X,Y})$. Since $\partial_\mu \partial_v \bar{S}((v^*, \mu^*), F_{X,Y}) = \partial_v \partial_\mu \bar{S}((v^*, \mu^*), F_{X,Y})$ by the symmetry of the Hessian, we obtain $h_{12}(v^*, \mu^*)(1 - \beta) = h_{21}(v^*, \mu^*)f_X(v^*)$. Exploiting condition (iii) of Proposition D.5 and the surjectivity of $\mathbf{T}$ we get that $h_{12} \equiv h_{21} \equiv 0$. Now, we exploit the symmetry of the Hessian at a general point $(v, \mu) \in \text{int}(\mathbf{A})$ such that $F_X(v) \neq \beta$ and $\mathbb{E}(F_Y) \neq \mu$. We obtain

$$\begin{aligned}\partial_\mu \partial_v \bar{S}((v, \mu), F_{X,Y}) &= \partial_\mu h_{11}(v, \mu)(F_X(v) - \beta), \\ \partial_v \partial_\mu \bar{S}((v, \mu), F_{X,Y}) &= (\mu - \mathbb{E}(F_Y))[\partial_v h_{22}(v, \mu)(1 - F_X(v)) - h_{22}(v, \mu)f_X(v)].\end{aligned}$$

Again invoking (iii) of Proposition D.5 and the surjectivity of $\mathbf{T}$ we get that $h_{22} \equiv \partial_v h_{22} \equiv \partial_c h_{11} \equiv 0$. Therefore, the only non-vanishing component of $\mathbf{h}$ is $h_{11}(v, \mu) = g(v)$ for some function $g: \mathbb{R} \rightarrow \mathbb{R}$. Concluding as in the proof of Proposition D.3, S fails to be strictly $\mathcal{F}$ -consistent for $\mathbf{T}$. $\blacksquare$

F. Monte Carlo simulations

Here, we investigate the finite-sample performance of our tests of H_0^- and $H_0^{\prec \text{lex}}$. For distinct VaR forecasts, we use $\mathcal{T}_n$ (defined in (5.1)) to test H_0^- , and $\mathcal{T}_n^{\text{OS}}$ (defined in (5.5)) to test $H_0^{\prec \text{lex}}$. For identical VaR forecasts, we use $\mathcal{T}_{2n} = \sqrt{n} \bar{d}_{2n} / \hat{\sigma}_{22,n}^{1/2}$ to test both hypotheses. We do so for (VaR, CoVaR) and (VaR, CoVaR, CoES) forecasts generated from a model of asset returns that may reasonably approximate actual stock market dynamics. We leave out the pair (VaR, MES) for brevity, because the results are qualitatively similar. Throughout, we consider one-step-ahead forecasts. For one-step-ahead forecasts, Diebold and Mariano (1995), Giacomini and White (2006) and others recommend to use $m_n = 0$ for $\hat{\Omega}_n$ in (5.2) (where an empty sum is defined to be zero), implicitly assuming that $\{\mathbf{d}_t\}$ is uncorrelated. We follow their suggestion here.

F.1. Data generating process

We simulate $\{(X_t, Y_t)\}_{t=1, \dots, n}$ from a bivariate GARCH model with GAS-driven t -copula. In particular, we use GARCH(1,1) marginals

$$\begin{aligned} X_t &= \sigma_{x,t} \varepsilon_{x,t}, & \sigma_{x,t}^2 &= \omega_x + \alpha_x X_{t-1}^2 + \beta_x \sigma_{x,t-1}^2, \\ Y_t &= \sigma_{y,t} \varepsilon_{y,t}, & \sigma_{y,t}^2 &= \omega_y + \alpha_y Y_{t-1}^2 + \beta_y \sigma_{y,t-1}^2, \end{aligned} \quad (\text{F.1})$$

where $\omega_z > 0$, $\alpha_z \geq 0$, $\beta_z \geq 0$ ($z \in \{x, y\}$). The innovations $(\varepsilon_{x,t}, \varepsilon_{y,t})$ are identically distributed with $\varepsilon_{x,t} \sim N(0, 1)$ and $\varepsilon_{y,t}$ having a standardised Student's t_5 -distribution.

For the dependence structure, consider the probability integral transforms (PITs) $\mathbf{U}_t = (F_x(\varepsilon_{x,t}), F_y(\varepsilon_{y,t}))$, where F_x and F_y denote the cdfs of $\varepsilon_{x,t}$ and $\varepsilon_{y,t}$, respectively. Define $\mathfrak{F}_t = \sigma((X_t, Y_t), (X_{t-1}, Y_{t-1}), \dots)$. Then, we model the conditional distribution of $\mathbf{U}_t \mid \mathfrak{F}_{t-1}$ with a t -copula with density $c(\cdot; \vartheta, \rho_t): [0, 1]^2 \rightarrow [0, \infty)$, where $\vartheta > 0$ is the (constant) degrees-of-freedom parameter, and $\rho_t \in (-1, 1)$ is the time-varying correlation parameter. Letting the correlation vary over time while keeping the degrees of freedom constant is standard in the literature (De Lira Salvatierra and Patton, 2015; Bernardi and Catania, 2019). To restrict ρ_t to $(-1, 1)$, we model the real-valued parameter f_t with a GAS dynamic and then set $\rho_t := \Delta(f_t)$ where $\Delta(x) = [1 - \exp(-x)]/[1 + \exp(x)] \in (-1, 1)$. In particular, following Creal *et al.* (2013), we set

$$f_t = \omega^\dagger + \alpha^\dagger s_{t-1} + \beta^\dagger f_{t-1}, \quad s_{t-1} := \frac{d}{df_{t-1}} \log c(\mathbf{U}_{t-1}; \vartheta, \Delta(f_{t-1})). \quad (\text{F.2})$$

We choose the empirically plausible values $(\omega^\dagger, \alpha^\dagger, \beta^\dagger, \vartheta) = (0.001, 0.1, 0.99, 5)$ for the dependence parameters, and $(\omega_x, \alpha_x, \beta_x) = (\omega_y, \alpha_y, \beta_y) = (0.001, 0.2, 0.79)$ for the marginal parameters (see, e.g., De Lira Salvatierra and Patton, 2015; Hoga, 2021).

F.2. Risk forecasts

In practice, the most popular approach to compute (risk) forecasts is a moving-window approach. In this case, a window that is rolled through the sample is used as the basis for frequent (often daily) re-estimation of the model producing the risk forecasts (e.g., a bivariate GARCH with GAS-driven t -copula, or a simple white noise model). We adopt such a rolling-window scheme in the empirical application in Section 6. However, since here we perform 10 000 replications for the simulation results to have low standard errors, such an approach would be computationally infeasible. Instead, we use a fixed-window approach here. That is, we obtain model parameter estimates from some ‘in-sample’ period $\{(X_t, Y_t)\}_{t=-r+1, \dots, 0}$, and use these estimates to produce ‘out-of-sample’ risk forecasts for $\{(X_t, Y_t)\}_{t=1, \dots, n}$.

We describe how we do this next. First, we generate a trajectory $\{(X_t, Y_t)\}_{t=-r+1, \dots, n}$ from the bivariate GARCH in (F.1) with GAS-driven t -copula. Then, based on the ‘in-sample’ data $\{(X_t, Y_t)\}_{t=-r+1, \dots, 0}$, we estimate the marginal parameters $(\omega_x, \alpha_x, \beta_x)$ and $(\omega_y, \alpha_y, \beta_y)$ via standard Gaussian quasi-maximum likelihood estimation (Francq and Zakoïan, 2004). In particular, we obtain the estimated conditional variances $\hat{\sigma}_{x,t}^2 = \hat{\omega}_x + \hat{\alpha}_x X_{t-1}^2 + \hat{\beta}_x \hat{\sigma}_{x,t-1}^2$ (and

similarly for $\hat{\sigma}_{y,t}^2$) and the standardized residuals $\hat{\varepsilon}_{x,t} = X_t/\hat{\sigma}_{x,t}$ and $\hat{\varepsilon}_{y,t} = Y_t/\hat{\sigma}_{y,t}$. By doing so, we also compute the estimated PITs $\hat{U}_t = (\hat{F}_x(\hat{\varepsilon}_{x,t}), \hat{F}_y(\hat{\varepsilon}_{y,t}))$, where $\hat{F}_x$ and $\hat{F}_y$ are the empirical cdfs of the $\hat{\varepsilon}_{x,t}$ and $\hat{\varepsilon}_{y,t}$, respectively. Based on the $\hat{U}_t$, we estimate $(\omega^\dagger, \alpha^\dagger, \beta^\dagger, \vartheta)'$ via maximum likelihood, as proposed by [Creal et al. \(2013\)](#).

In line with the empirical application in the main paper, we forecast the conditional risk measures $\text{VaR}_t(X_t) = \text{VaR}_\beta(F_{X_t|\mathfrak{F}_{t-1}})$, $\text{CoVaR}_t(Y_t|X_t) = \text{CoVaR}_{\alpha|\beta}(F_{(X_t,Y_t)|\mathfrak{F}_{t-1}})$, and $\text{CoES}_t(Y_t|X_t) = \text{CoES}_{\alpha|\beta}(F_{(X_t,Y_t)|\mathfrak{F}_{t-1}})$. Here, $F_{(X_t,Y_t)|\mathfrak{F}_{t-1}}(x,y) = \text{P}\{X_t \leq x, Y_t \leq y \mid \mathfrak{F}_{t-1}\} =: \text{P}_{t-1}\{X_t \leq x, Y_t \leq y\}$ for $x, y \in \mathbb{R}$, and $\mathfrak{F}_{t-1} = \sigma((X_{t-1}, Y_{t-1}), (X_{t-2}, Y_{t-2}), \dots)$. To ease notation, we suppress the dependence of the risk measures on the risk levels α and β , which we set to be $\alpha = \beta = 0.95$.

Now, we describe how we calculate the ‘out-of-sample’ (conditional) risk forecasts for the observations $\{(X_t, Y_t)\}_{t=1,\dots,n}$. We use the estimated parameters from the ‘in-sample’ period to compute the (one-step-ahead) Value-at-Risk for the X_t via

$$\widehat{\text{VaR}}_t := \widehat{\text{VaR}}_t(X_t) = \hat{\sigma}_{x,t} \widehat{\text{VaR}}_{x,\varepsilon}, \quad t = 1, \dots, n,$$

where $\widehat{\text{VaR}}_{x,\varepsilon}$ is the empirical β -quantile of the $\{\hat{\varepsilon}_{x,t}\}_{t=-r+1,\dots,0}$. In terms of the conditional distribution, CoVaR solves $\text{P}_{t-1}\{Y_t > \text{CoVaR}_t(Y_t|X_t) \mid X_t \geq \text{VaR}_t(X_t)\} = 1 - \alpha$. Using this and the fact that $\text{P}_{t-1}\{X_t \geq \text{VaR}_t(X_t)\} = 1 - \beta$, elementary calculations yield that $\text{CoVaR}_t(Y_t|X_t)$ is implicitly defined by the equation

$$(1 - \alpha)(1 - \beta) = \int_{F_y(\text{CoVaR}_t(Y_t|X_t)/\sigma_{y,t})}^1 \int_{\beta}^1 c((u_1, u_2), \vartheta, \rho_t) du_1 du_2. \quad (\text{F.3})$$

Plugging in estimates (e.g., replacing ϑ with the ‘in-sample’ estimate $\hat{\vartheta}$, replacing F_y with the ‘in-sample’ empirical cdf $\hat{F}_y$, etc.) and numerically solving (F.3) for $\text{CoVaR}_t(Y_t|X_t)$, yields the forecasts $\widehat{\text{CoVaR}}_t$ for $t = 1, \dots, n$. From a discretization of the integral in formula (2.2), we also obtain CoES predictions $\widehat{\text{CoES}}_t$ for $t = 1, \dots, n$. We collect these (correctly specified) VaR and systemic risk forecasts in the vector $\hat{\mathbf{r}}_t = (\hat{r}_t^{\text{VaR}}, \hat{r}_t^{\text{SR}}) = (\widehat{\text{VaR}}_t, \widehat{\text{CoVaR}}_t, \widehat{\text{CoES}}_t)$.

Clearly, it is very challenging to come up with two forecasts from distinct models, such that H_0^- is satisfied for our data-generating process; see also [Zhu and Timmermann \(2020\)](#). Hence, for simplicity, we confound the correctly specified forecasts $\hat{\mathbf{r}}_t$ by two multiplicative noises of equal magnitude to simulate under H_0^- . We use (strictly positive) multiplicative noise instead of additive noise, because this ensures that positive risk forecasts remain positive after confounding them. This is important, since the 0-homogeneous loss functions we use require positive (systemic) risk forecasts; see also Section F.3. The noises may be thought of as uninformative (multiplicative) predictors. Specifically, consider $\hat{\mathbf{r}}_{t,(i)} = (\hat{r}_{t,(i)}^{\text{VaR}}, \hat{r}_{t,(i)}^{\text{SR}}) = (\widehat{\text{VaR}}_{t,(i)}, \widehat{\text{CoVaR}}_{t,(i)}, \widehat{\text{CoES}}_{t,(i)})$ ($i = 1, 2$), where $\widehat{\text{VaR}}_{t,(i)} = \widehat{\text{VaR}}_t \cdot \epsilon_{t,(i)}^{\text{VaR}}$, $\widehat{\text{CoVaR}}_{t,(i)} = \widehat{\text{CoVaR}}_t \cdot \epsilon_{t,(i)}^{\text{CoVaR}}$ and $\widehat{\text{CoES}}_{t,(i)} = \widehat{\text{CoES}}_t \cdot \epsilon_{t,(i)}^{\text{CoES}}$. Thus, the forecasts from the correctly specified model are contaminated by some mutually independent, positive multiplicative noises $\{\epsilon_{t,(i)}^{\text{VaR}}\}_{t=1,\dots,n}$, $\{\epsilon_{t,(i)}^{\text{CoVaR}}\}_{t=1,\dots,n}$ and $\{\epsilon_{t,(i)}^{\text{CoES}}\}_{t=1,\dots,n}$, each assumed to be serially independent and Weibull-

distributed. Hence, they have common density $f(x) = (k/\lambda)(x/\lambda)^{k-1}e^{-(x/\lambda)^k}$ for $x > 0$, where we choose $k = 10$ for the shape parameter and $\lambda = 0.3$ for the scale parameter. This results in a mean of $\lambda\Gamma(1 + 1/k) \approx 0.285$, so that the $\hat{\mathbf{r}}_{t,(i)}$ are seriously misspecified on average. Clearly, H_0^- (and also $H_0^{\leq \text{lex}}$) holds for $\hat{\mathbf{r}}_{t,(1)}$ and $\hat{\mathbf{r}}_{t,(2)}$.

F.3. Description of tests

In evaluating the risk forecasts, we choose multi-objective scores with components that render the score differences 0-homogeneous. That is, the score differences remain unchanged when scaling all forecasts and observations with a positive constant. This leads to ‘unit-consistent’ comparisons and, quite often, higher power of DM tests (Nolde and Ziegel, 2017; Taylor, 2019; Patton *et al.*, 2019). We also found in unreported simulations that 0-homogeneous score differences lead to DM tests with higher power than other degrees of homogeneity, e.g., 1-homogeneous scores. This may be explained as follows. Our data and forecasts are conditionally heteroscedastic. Invoking the 0-homogeneity, the loss differences $\mathbf{d}_t$ then are homoscedastic. Thus, we obtain smaller ‘standard errors’ in $\hat{\boldsymbol{\Omega}}_n$ by using 0-homogenous instead of, say, 1-homogeneous score differences. This effect is likely not outweighed by the increased average score differences of 1-homogenous vis-à-vis 0-homogenous scores. (Recall that 1-homogeneous scores are more sensitive to outliers.) If this is indeed the case, then 0-homogenous scores give tests higher power. When using 0-homogeneous score differences, one has to assume additionally that risk forecasts are positive. However, this assumption is typically innocuous.

To achieve 0-homogeneity in the first component of our multi-objective score, we set $h(z) = \log(z)$ and $a^{\text{VaR}}(x, y) = \log(x)$ in the S^{VaR} -function from Theorem 4.4 (Nolde and Ziegel, 2017, Example 4). When comparing (VaR, CoVaR) forecasts, we choose $g(z) = \log(z)$, $a(y) = \log(y)$ and $a^{\text{CoVaR}}(x, y) = 0$ in S_v^{CoVaR} from (4.7), leading to

$$S_v^{\text{CoVaR}}(c, (x, y)) = \mathbb{1}\{x > v\} \left[(\mathbb{1}\{y \leq c\} - \alpha) \log(c) + \mathbb{1}\{y \geq c\} \log(y) \right]. \quad (\text{F.4})$$

When also CoES forecasts are of interest, we use $S_v^{(\text{CoVaR}, \text{CoES})}$ from (4.8) with $g(z) = 0$, $a(y) = a^{\text{CoES}}(x, y) = 0$, and $\phi(z) = -\log(z)$ ($z > 0$). This choice leads to

$$S_v^{(\text{CoVaR}, \text{CoES})}((c, e), (x, y)) = \frac{\mathbb{1}\{x > v\}}{1 - \alpha} \left[\mathbb{1}\{y > c\} \frac{y - c}{e} + (1 - \alpha) \left(\frac{c}{e} - 1 + \log(e) \right) \right]. \quad (\text{F.5})$$

To test H_0^- and $H_0^{\leq \text{lex}}$, we simulate $R = 10\,000$ trajectories $\{(X_t, Y_t)\}_{t=-r+1, \dots, n}$ from the bivariate GARCH model (F.1) with GAS-driven t -copula. We fix $r = 1000$ and let $n \in \{500, 1000\}$. For each trajectory, we fit the model, forecast VaR and systemic risk, and compute the score differences. Then, depending on the forecasts, we carry out the formal tests based on $\mathcal{T}_n$ (for a test of H_0^-) and $\mathcal{T}_n^{\text{OS}}$ (for a test of $H_0^{\leq \text{lex}}$), or—when VaR forecasts are identical—based on $\mathcal{T}_{2n}$. In the latter case, H_0^- ($H_0^{\leq \text{lex}}$) boils down to $\mathbb{E}[\bar{d}_{2n}] = 0$ ($\mathbb{E}[\bar{d}_{2n}] \leq 0$) and is rejected if $|\mathcal{T}_{2n}| > \Phi^{-1}(1 - \nu/2)$ ($\mathcal{T}_{2n} > \Phi^{-1}(1 - \nu)$); see Remark 5.4.

Table 2: Rejection frequencies (in %) of H_0^- and $H_0^{\leq \text{lex}}$ using Remark 5.4 (column ' $\widehat{\text{VaR}}_{t,(1)}$ ') and Theorem 5.1 (column ' $\widehat{\text{VaR}}_{t,(i)}$ ').

n	Null	Forecasts	(VaR, CoVaR)		(VaR, CoVaR, CoES)	
			$r_{t,(i)}^{\text{VaR}}$		$r_{t,(i)}^{\text{VaR}}$	
			$\widehat{\text{VaR}}_{t,(1)}$	$\widehat{\text{VaR}}_{t,(i)}$	$\widehat{\text{VaR}}_{t,(1)}$	$\widehat{\text{VaR}}_{t,(i)}$
500	H_0^-	$\mathbf{r}_{t,(1)}^{\text{SR}} = \widehat{\mathbf{r}}_{t,(1)}^{\text{SR}}$	5.1	4.7	4.0	4.5
		$\mathbf{r}_{t,(2)}^{\text{SR}} = \widehat{\mathbf{r}}_{t,(2)}^{\text{SR}}$				
		$\mathbf{r}_{t,(1)}^{\text{SR}} = \widehat{\mathbf{r}}_{t,(1)}^{\text{SR}}$	83.6	76.0	85.0	76.5
		$\mathbf{r}_{t,(2)}^{\text{SR}} = \widehat{\mathbf{r}}_t^{\text{SR}}$				
	$H_0^{\leq \text{lex}}$	$\mathbf{r}_{t,(1)}^{\text{SR}} = \widehat{\mathbf{r}}_{t,(1)}^{\text{SR}}$	5.3	4.7	5.1	4.9
		$\mathbf{r}_{t,(2)}^{\text{SR}} = \widehat{\mathbf{r}}_{t,(2)}^{\text{SR}}$				
1000	H_0^-	$\mathbf{r}_{t,(1)}^{\text{SR}} = \widehat{\mathbf{r}}_{t,(1)}^{\text{SR}}$	5.2	4.4	4.3	4.6
		$\mathbf{r}_{t,(2)}^{\text{SR}} = \widehat{\mathbf{r}}_{t,(2)}^{\text{SR}}$				
		$\mathbf{r}_{t,(1)}^{\text{SR}} = \widehat{\mathbf{r}}_{t,(1)}^{\text{SR}}$	95.3	92.8	96.2	94.2
		$\mathbf{r}_{t,(2)}^{\text{SR}} = \widehat{\mathbf{r}}_t^{\text{SR}}$				
	$H_0^{\leq \text{lex}}$	$\mathbf{r}_{t,(1)}^{\text{SR}} = \widehat{\mathbf{r}}_{t,(1)}^{\text{SR}}$	5.0	4.5	4.8	4.5
		$\mathbf{r}_{t,(2)}^{\text{SR}} = \widehat{\mathbf{r}}_{t,(2)}^{\text{SR}}$				
		$\mathbf{r}_{t,(1)}^{\text{SR}} = \widehat{\mathbf{r}}_{t,(1)}^{\text{SR}}$	96.5	93.8	96.6	94.9
		$\mathbf{r}_{t,(2)}^{\text{SR}} = \widehat{\mathbf{r}}_t^{\text{SR}}$				

For distinct VaR forecasts, we consider $\mathbf{r}_{t,(1)} = \widehat{\mathbf{r}}_{t,(1)}$ and $\mathbf{r}_{t,(2)} = \widehat{\mathbf{r}}_{t,(2)}$ under both H_0^- and $H_0^{\leq \text{lex}}$. Since both forecasts are equally misspecified, we have $E[\bar{\mathbf{d}}_n] = \mathbf{0}$ when we evaluate either (VaR, CoVaR) or (VaR, CoVaR, CoES). Under the alternative, we compare the forecasts $\mathbf{r}_{t,(1)} = \widehat{\mathbf{r}}_{t,(1)} = (\widehat{r}_{t,(1)}^{\text{VaR}}, \widehat{\mathbf{r}}_{t,(1)}^{\text{SR}})$ and $\mathbf{r}_{t,(2)} = (\widehat{r}_{t,(2)}^{\text{VaR}}, \widehat{\mathbf{r}}_t^{\text{SR}}) = (\widehat{\text{VaR}}_{t,(2)}, \widehat{\text{CoVaR}}_t, \widehat{\text{CoES}}_t)$. The VaR forecasts are comparable, giving $E[\bar{\mathbf{d}}_{1n}] = 0$. Yet, the systemic risk forecasts of $\mathbf{r}_{t,(1)}$ are inferior to those of $\mathbf{r}_{t,(2)}$, implying $E[\bar{\mathbf{d}}_{2n}] > 0$. In particular, we simulate under the alternative of both H_0^- and $H_0^{\leq \text{lex}}$.

For identical VaR forecasts, we consider $\mathbf{r}_{t,(1)} = \widehat{\mathbf{r}}_{t,(1)} = (\widehat{r}_{t,(1)}^{\text{VaR}}, \widehat{\mathbf{r}}_{t,(1)}^{\text{SR}})$ and $\mathbf{r}_{t,(2)} = (\widehat{r}_{t,(1)}^{\text{VaR}}, \widehat{\mathbf{r}}_{t,(2)}^{\text{SR}})$ under H_0^- and $H_0^{\leq \text{lex}}$. Under the alternative, we compare $\mathbf{r}_{t,(1)} = (\widehat{r}_{t,(1)}^{\text{VaR}}, \widehat{\mathbf{r}}_{t,(1)}^{\text{SR}})$ and $\mathbf{r}_{t,(2)} = (\widehat{r}_{t,(1)}^{\text{VaR}}, \widehat{\mathbf{r}}_t^{\text{SR}})$, such that again $E[\bar{\mathbf{d}}_{2n}] > 0$.

F.4. Simulation results

We test H_0^- and $H_0^{\leq \text{lex}}$ at significance level $\nu = 5\%$. The columns ' $\widehat{\text{VaR}}_{t,(i)}$ ' (' $\widehat{\text{VaR}}_{t,(1)}$ ') in Table 2 present the results for distinct (identical) VaR forecasts. We draw the following conclusions:

1. Size is adequate in all cases, even for $n = 500$. This is encouraging, since effective sample

sizes in risk forecast comparisons are small, and are smaller still when systemic risk is concerned.

2. Comparing the results for $n = 500$ and $n = 1000$, we see the expected increase in power.
3. When comparing systemic risk forecasts, it is slightly easier to distinguish between two different (CoVaR, CoES) forecasts than if CoVaR alone is evaluated. For instance, under H_0^- and for $n = 500$ there is a statically significant difference between $\mathbf{r}_{t,(1)} = (\widehat{\text{VaR}}_{t,(1)}, \widehat{\text{CoVaR}}_{t,(1)})$ and $\mathbf{r}_{t,(2)} = (\widehat{\text{VaR}}_{t,(2)}, \widehat{\text{CoVaR}}_t)$ in 76.0% of all cases, and when CoES is added to the evaluation (so that $\mathbf{r}_{t,(1)} = (\widehat{\text{VaR}}_{t,(1)}, \widehat{\text{CoVaR}}_{t,(1)}, \widehat{\text{CoES}}_{t,(1)})$ and $\mathbf{r}_{t,(2)} = (\widehat{\text{VaR}}_{t,(2)}, \widehat{\text{CoVaR}}_t, \widehat{\text{CoES}}_t)$) this number is 76.5%. One reason for this may be the increase in informational content of the forecasts, such that differences are easier to identify.
4. Under both distinct and identical VaR forecasts, departures from $H_0^{\leq \text{lex}}$ are detected more often than from H_0^- . E.g., for $\mathbf{r}_{t,(1)} = (\widehat{\text{VaR}}_{t,(1)}, \widehat{\text{CoVaR}}_{t,(1)})$ and $\mathbf{r}_{t,(2)} = (\widehat{\text{VaR}}_{t,(2)}, \widehat{\text{CoVaR}}_t)$, we reject H_0^- with a percentage of 76.0% for $n = 500$. However, $H_0^{\leq \text{lex}}$ is rejected in 81.2% of all cases for the same forecasts. This is as expected for one-sided tests.
5. Generally, it is easier to detect differences in the systemic risk forecasts $\mathbf{r}_{t,(1)}^{\text{SR}}$ and $\mathbf{r}_{t,(2)}^{\text{SR}}$, when the same VaR forecasts are used compared to when VaR forecasts are merely comparable. E.g., for $\mathbf{r}_{t,(1)} = (\widehat{\text{VaR}}_{t,(1)}, \widehat{\text{CoVaR}}_{t,(1)})$ and $\mathbf{r}_{t,(2)} = (\widehat{\text{VaR}}_{t,(1)}, \widehat{\text{CoVaR}}_t)$, we reject H_0^- with probability 83.6% for $n = 500$. Yet, when comparing $\mathbf{r}_{t,(1)} = (\widehat{\text{VaR}}_{t,(1)}, \widehat{\text{CoVaR}}_{t,(1)})$ and $\mathbf{r}_{t,(2)} = (\widehat{\text{VaR}}_{t,(2)}, \widehat{\text{CoVaR}}_t)$, H_0^- is rejected in only 76.0% of all replications. Even when taking into account the slightly lower size in the latter case, there appears to be a power difference. Intuitively, when VaR forecasts are identical, instead of merely comparable, the test can exclusively focus on differences in the systemic risk component, thus increasing power.

Overall, the simulations show that our tests work quite well in that they keep size and have good power in detecting significant differences in forecast ability.

G. The perils of using non-strict identification functions for backtesting

For $t = 1, \dots, n$, consider $(X_t, Y_t) \stackrel{\text{i.i.d.}}{\sim} N(\mathbf{0}, \mathbf{\Omega})$ with $\mathbf{\Omega} = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 2 \end{pmatrix}$. Suppose the goal is to forecast $\text{VaR}_\beta(X_t)$ and $\text{CoVaR}_{\alpha|\beta}(Y_t|X_t)$ for $\alpha = \beta = 0.95$. Then, the correctly specified forecasts (conditional or unconditional—due to the i.i.d. nature this is the same) are $\widehat{\text{VaR}}_{\beta,t} = \text{VaR}_\beta(X_t) \approx 1.64$ and $\widehat{\text{CoVaR}}_{\alpha|\beta,t} = \text{CoVaR}_{\alpha|\beta}(Y_t|X_t) \approx 3.23$. We also consider the misspecified forecasts $\widehat{\text{VaR}}_{\beta',t} = \text{VaR}_{\beta'}(X_t) \approx 2.33$ and $\widehat{\text{CoVaR}}_{\alpha'|\beta',t} = \text{CoVaR}_{\alpha'|\beta'}(Y_t|X_t) \approx 2.23$, where $\alpha' = 0.75$ and $\beta' = 0.99$. Note that in this case $(1 - \alpha')(1 - \beta') = (1 - \alpha)(1 - \beta)$, such that the (non-strict) identification function of Banulescu-Radu *et al.* (2021) in (4.5) vanishes in

Table 3: Rejection frequencies (in %) of H_0 based on $\overline{\mathbf{V}}^{(\text{VaR}, \text{CoVaR})}$ and $\overline{V}$ for $n \in \{500, 1000\}$. Results are displayed for correctly specified forecasts $(\widehat{\text{VaR}}_{\beta,t}, \widehat{\text{CoVaR}}_{\alpha|\beta,t})$ and misspecified forecasts $(\widehat{\text{VaR}}_{\beta',t}, \widehat{\text{CoVaR}}_{\alpha'|\beta',t})$.

n	$\overline{\mathbf{V}}^{(\text{VaR}, \text{CoVaR})}$		$\overline{V}$	
	$\begin{pmatrix} \widehat{\text{VaR}}_{\beta,t} \\ \widehat{\text{CoVaR}}_{\alpha \beta,t} \end{pmatrix}$	$\begin{pmatrix} \widehat{\text{VaR}}_{\beta',t} \\ \widehat{\text{CoVaR}}_{\alpha' \beta',t} \end{pmatrix}$	$\begin{pmatrix} \widehat{\text{VaR}}_{\beta,t} \\ \widehat{\text{CoVaR}}_{\alpha \beta,t} \end{pmatrix}$	$\begin{pmatrix} \widehat{\text{VaR}}_{\beta',t} \\ \widehat{\text{CoVaR}}_{\alpha' \beta',t} \end{pmatrix}$
500	6.8	99.9	28.9	28.1
1000	6.4	100	8.1	8.3

expectation, yet our strict one $\mathbf{V}^{(\text{VaR}, \text{CoVaR})}$ from (4.2) does not. We test the null hypothesis of correct unconditional calibration, which—in terms of our strict identification function—reads as

$$H_0: \mathbb{E} \left[\mathbf{V}^{(\text{VaR}, \text{CoVaR})}((\widehat{\text{VaR}}_t, \widehat{\text{CoVaR}}_t), (X_t, Y_t)) \right] = \mathbf{0} \quad \text{for all } t = 1, 2, \dots$$

for $(\widehat{\text{VaR}}_t, \widehat{\text{CoVaR}}_t) \in \{(\widehat{\text{VaR}}_{\beta,t}, \widehat{\text{CoVaR}}_{\alpha|\beta,t}), (\widehat{\text{VaR}}_{\beta',t}, \widehat{\text{CoVaR}}_{\alpha'|\beta',t})\}$. Following [Nolde and Ziegel \(2017\)](#), we test H_0 using a standard Wald-test based on

$$\begin{aligned} \overline{\mathbf{V}}^{(\text{VaR}, \text{CoVaR})} &= \frac{1}{n} \sum_{t=1}^n \mathbf{V}^{(\text{VaR}, \text{CoVaR})}((\widehat{\text{VaR}}_t, \widehat{\text{CoVaR}}_t), (X_t, Y_t)) \quad \text{and} \\ \overline{V} &= \frac{1}{n} \sum_{t=1}^n V((\widehat{\text{VaR}}_t, \widehat{\text{CoVaR}}_t), (X_t, Y_t)), \end{aligned}$$

respectively.

Table 3 displays the rejection frequencies—calculated from 10 000 replications—for the tests (at a 5%-level) for $n \in \{500, 1000\}$. For our test statistic $\overline{\mathbf{V}}^{(\text{VaR}, \text{CoVaR})}$ size is close to the nominal level, and the misspecified forecasts are identified as such almost with certainty. However, with [Banulescu-Radu et al.'s \(2021\)](#) test statistic $\overline{V}$, the null is rejected with about the same frequency for the correctly specified *and* the misspecified forecasts. This is because the null that is actually tested with $\overline{V}$ is

$$H_0^*: \mathbb{E} \left[V((\widehat{\text{VaR}}_t, \widehat{\text{CoVaR}}_t), (X_t, Y_t)) \right] = 0 \quad \text{for all } t = 1, 2, \dots$$

where V is given in (4.5). Obviously, H_0^* does not amount to a null of correct calibration, because it is also satisfied for specific misspecified forecasts, such as $(\widehat{\text{VaR}}_{\beta',t}, \widehat{\text{CoVaR}}_{\alpha'|\beta',t})$. This highlights that traditional backtests should not be carried out using non-strict identification functions, because H_0^* is too broad in that it accommodates possibly misspecified forecasts.

Observe that particularly for $n = 500$, size for the test based on $\overline{V}$ is very far from the nominal level of 5%. This may be explained as follows: From (4.5) we see that the identification function is constant except when the X component exceeds the VaR forecast *and* the Y component

exceeds the CoVaR forecast. Under the null this only occurs with probability $(1 - \alpha)(1 - \beta) = (1 - 0.95)(1 - 0.95) = 0.25\%$, such that the effective sample size is reduced from $n = 500$ to $n \cdot 0.25\% = 1.25$. Thus, size distortions can be expected. Note that this issue is somewhat alleviated for our two-dimensional score $\mathbf{V}^{(\text{VaR}, \text{CoVaR})}$, where the effective sample size—at least in the first component of the score—is $n(1 - \beta) = 25$, leading to much better size for $n = 500$.

We emphasize that the non-strictness of Banulescu-Radu *et al.*'s (2021) V at (4.5) is not only problematic when $(1 - \alpha)(1 - \beta) = (1 - \alpha')(1 - \beta')$ (which is already the case for uncountably many configurations), but in all cases where VaR is overpredicted (underpredicted) and CoVaR is underpredicted (overpredicted). This is because the different biases cancel each other out such that V remains close to zero on average, leading to a loss of power. To illustrate this, we fix the probability levels of the misspecified forecasts at $\beta' = 0.99$ (as before) and let α' vary between $(0.75, 0.95)$. Then, $(1 - \alpha)(1 - \beta) \neq (1 - \alpha')(1 - \beta')$, but VaR is still overpredicted and CoVaR is underpredicted. We again test H_0 using $\bar{\mathbf{V}}^{(\text{VaR}, \text{CoVaR})}$ and $\bar{V}$ for $(\widehat{\text{VaR}}_t, \widehat{\text{CoVaR}}_t) \in \{(\widehat{\text{VaR}}_{\beta, t}, \widehat{\text{CoVaR}}_{\alpha|\beta, t}), (\widehat{\text{VaR}}_{\beta', t}, \widehat{\text{CoVaR}}_{\alpha'|\beta', t})\}$. These forecasts are misspecified for all $\alpha' \in (0.75, 0.95)$, such that we expect high rejection frequencies.

Figure 9 plots the rejection probabilities as a function of α' for $n = 1000$. We see that $\bar{\mathbf{V}}^{(\text{VaR}, \text{CoVaR})}$ always identifies the forecasts as misspecified. As could be expected from Table 3, for $\alpha' = 0.75$ the test based on $\bar{V}$ has only close to trivial power of 8.3%. However, the power increases only slowly for increasing α' . Even for $\alpha' \in (0.95, 0.99)$, where both VaR and CoVaR are overpredicted, our proposal based on the strict identification function has markedly higher power than the test of Banulescu-Radu *et al.* (2021).

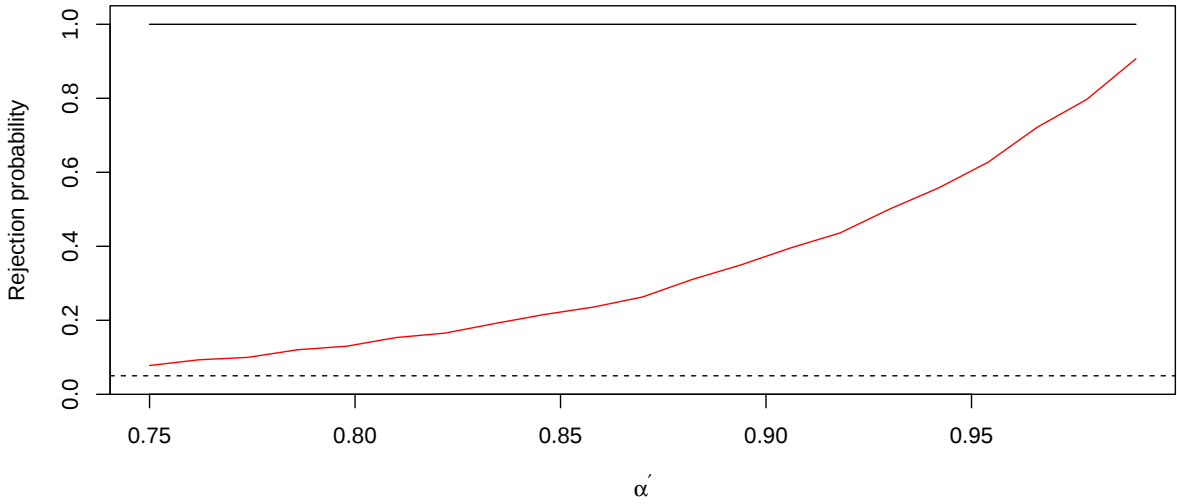


Figure 9: Rejection probabilities of H_0 as a function of $\alpha' \in (0.75, 0.99)$ for our $\bar{\mathbf{V}}^{(\text{VaR}, \text{CoVaR})}$ (black) and Banulescu-Radu *et al.*'s (2021) $\bar{V}$ (red) based on $n = 1000$. Nominal level of 5% indicated by the dashed horizontal line.

References

- Banulescu-Radu D, Hurlin C, Leymarie J, Scaillet O. 2021. Backtesting marginal expected shortfall and related systemic risk measures. *Management Science* **67**: 5730–5754.
- Bellini F, Bignozzi V. 2015. On elicitable risk measures. *Quantitative Finance* **15**: 725–733.
- Bernardi M, Catania L. 2019. Switching generalized autoregressive score copula models with application to systemic risk. *Journal of Applied Econometrics* **34**: 43–65.
- Creal D, Koopman SJ, Lucas A. 2013. Generalized autoregressive score models with applications. *Journal of Applied Econometrics* **28**: 777–795.
- Davidson J. 1994. *Stochastic Limit Theory*. Oxford: Oxford University Press.
- De Jong RM, Davidson J. 2000. Consistency of kernel estimators of heteroscedastic and autocorrelated covariance matrices. *Econometrica* **68**: 407–424.
- De Lira Salvatierra I, Patton AJ. 2015. Dynamic copula models and high frequency data. *Journal of Empirical Finance* **30**: 120–135.
- Diebold FX, Mariano RS. 1995. Comparing predictive accuracy. *Journal of Business & Economic Statistics* **13**: 253–263.
- Dimitriadis T, Fissler T, Ziegel JF. 2020. The Efficiency Gap. *Preprint*. <https://arxiv.org/abs/2010.14146>.
- Fissler T. 2017. *On Higher Order Elicitability and Some Limit Theorems on the Poisson and Wiener Space*. Ph.D. thesis, University of Bern.
- Fissler T, Frongillo R, Hlavínová J, Rudloff B. 2021. Forecast evaluation of quantiles, prediction intervals, and other set-valued functionals. *Electronic Journal of Statistics* **15**: 1034–1084.
- Fissler T, Ziegel JF. 2016. Higher order elicibility and Osband’s principle. *The Annals of Statistics* **44**: 1680–1707.
- Fissler T, Ziegel JF. 2019. Order-sensitivity and equivariance of scoring functions. *Electronic Journal of Statistics* **13**: 1166–1211.
- Francq C, Zakoïan JM. 2004. Maximum likelihood estimation of pure GARCH and ARMA–GARCH processes. *Bernoulli* **10**: 605–637.
- Giacomini R, White H. 2006. Tests of conditional predictive ability. *Econometrica* **74**: 1545–1578.
- Gneiting T. 2011a. Making and evaluating point forecasts. *Journal of the American Statistical Association* **106**: 746–762.
- Gneiting T. 2011b. Quantiles as optimal point forecasts. *International Journal of Forecasting* **27**: 197–207.
- Gneiting T, Ranjan R. 2013. Combining predictive distributions. *Electronic Journal of Statistics* **7**: 1747–1782.
- Hoga Y. 2021. Modeling time-varying tail dependence, with application to systemic risk forecasting. Forthcoming in *Journal of Financial Econometrics* : 1–31.
- Holzmah H, Eulert M. 2014. The role of the information set for forecasting – with applications to risk management. *The Annals of Applied Statistics* **8**: 79–83.
- Holzmah H, Klar B. 2017. Focusing on regions of interest in forecast evaluation. *The Annals of Applied Statistics* **11**: 2404–2431.
- Lambert N. 2019. Elicitation and Evaluation of Statistical Forecasts. *Preprint*. .
- Lambert N, Pennock DM, Shoham Y. 2008. Eliciting properties of probability distributions. In *Proceedings of the 9th ACM Conference on Electronic Commerce*, Chicago, IL, USA: ACM, pages 129–138.

- McNeil AJ, Frey R, Embrechts P. 2015. *Quantitative Risk Management: Concepts, Techniques and Tools*. Princeton: Princeton University Press, revised edn.
- Nolde N, Ziegel JF. 2017. Elicitability and backtesting: Perspectives for banking regulation. *The Annals of Applied Statistics* **11**: 1833–1874.
- Osband KH. 1985. *Providing Incentives for Better Cost Forecasting*. Ph.D. thesis, University of California, Berkeley.
- Patton AJ. 2020. Comparing possibly misspecified forecasts. *Journal of Business & Economic Statistics* **38**: 796–809.
- Patton AJ, Ziegel JF, Chen R. 2019. Dynamic semiparametric models for expected shortfall (and value-at-risk). *Journal of Econometrics* **211**: 388–413.
- Pohle MO. 2020. The Murphy decomposition and the calibration-resolution principle: A new perspective on forecast evaluation. *Preprint*. .
- Steinwart I, Pasin C, Williamson R, Zhang S. 2014. Elicitation and Identification of Properties. *JMLR Workshop Conf. Proc.* **35**: 1–45.
- Taylor JW. 2019. Forecasting value at risk and expected shortfall using a semiparametric approach based on the asymmetric Laplace distribution. *Journal of Business & Economic Statistics* **37**: 121–133.
- White H. 2001. *Asymptotic Theory for Econometricians*. San Diego: Academic Press, First edn.
- Zhu Y, Timmermann A. 2020. Can two forecasts have the same conditional expected accuracy? *Preprint*. .