Joint Backtesting Procedure for Systemic Risk Measures

Sylvain Benoit[*] Yujian Chen[†] Jérémy Leymarie[‡]

Preliminary and Incomplete Draft

Please do not cite or circulate without permission

Abstract

We develop a comprehensive joint backtesting framework for systemic risk measures, including the Marginal Expected Shortfall (MES), Conditional Value-at-Risk (CoVaR), and ΔCoVaR. Our framework makes two principal contributions. First, building on strict joint identification, we construct moment-based calibration backtests for (VaR, CoVaR), (VaR, MES), and the stressed and median components of ΔCoVaR. In particular, CoVaR relies on moment conditions that jointly enforce market VaR calibration and the conditional quantile restriction, and MES is tested directly using the joint (VaR, MES) identification function rather than integrated CoVaR violations. Second, we develop the first forecast-based backtests for systemic risk measures with valid inference under parameter estimation uncertainty: test statistics are computed solely from reported forecasts and realized returns, without requiring the full predictive distribution or access to the forecasting model. We derive the asymptotic distribution of the tests under standard estimation schemes with explicit covariance corrections. Monte Carlo simulations show that ignoring estimation risk causes severe size distortions, while our robust tests maintain accurate size and competitive size-corrected power across the designs considered.

JEL classification: C12, C52, G01, G17, G28

Keywords: Systemic risk, backtesting tests, estimation risk, identification functions, model-free tests

[*] Université Paris Dauphine - PSL, UMR CNRS 8007, LEDa-SDFi, 75016 Paris, France. E-mail: sylvain.benoit@dauphine.psl.eu.

[†] Ecole Polytechnique, ENSAE, IP Paris, 5 Avenue Henry Le Chatelier, 91120 Palaiseau, France. E-mail: yujian.chen@polytechnique.edu.

[‡] Clermont School of Business, 4 Bd Trudaine, 63000 Clermont-Ferrand, E-mail: jeremy.leymarie@clermont-sb.fr.

1 Introduction

The 2008 global financial crisis fundamentally transformed our understanding of financial risk. While traditional market risk measures focused on individual portfolio losses, the crisis revealed that the real threat to financial stability lies in the interconnectedness of financial institutions and their collective contribution to systemic fragility. This recognition prompted regulators worldwide to shift their attention from firm-level risk assessment toward measuring and monitoring systemic risk contributions. The assessment of systemic risk measures is thus crucial for financial stability monitoring and regulatory oversight. Among the systemic risk measures that emerged in the aftermath of the crisis, the Marginal Expected Shortfall (MES), the Conditional Value-at-Risk (CoVaR), and related indicators such as SRISK and the ΔCoVaR have become central tools for identifying systemically important financial institutions (SIFIs) and monitoring at a high frequency individual systemic footprint.

The MES, introduced by Acharya, Pedersen, Philippon, and Richardson (2017), captures an institution's expected equity loss conditional on the market experiencing severe distress. Unlike traditional portfolio risk measures that examine losses in isolation, the MES explicitly quantifies how an institution's performance deteriorates when the financial system as a whole is under stress. This conditional perspective is crucial because systemic risk materializes precisely when multiple institutions face simultaneous difficulties, creating cascading failures and contagion effects. The CoVaR, proposed by Adrian and Brunnermeier (2016), provides a complementary view by measuring the value-at-risk of the financial system conditional on a particular institution being in distress. The ΔCoVaR extends this concept by comparing an institution's contribution to systemic risk in stressed versus normal market conditions, thereby identifying institutions whose failure would have particularly severe spillover effects.

Despite their widespread adoption by regulators and risk managers to evaluate the vulnerability of financial institutions and their contributions to systemic risk, the statistical validation of these systemic risk measures remains challenging. Unlike portfolio-level risk measures where violations can be directly observed by comparing forecasts to realized losses, systemic risk indicators involve bivariate conditional expectations that cannot be directly verified from market data. This fundamental difficulty has led to a notable gap in the literature. While extensive backtesting frameworks exist for univariate risk measures such as Value-at-Risk (Christoffersen, 1998) and Expected Shortfall (Du and Escanciano, 2017), rigorous statistical tests for systemic risk forecasts are still in their infancy. Existing validation approaches have primarily relied on indirect methods, such as examining whether institutions with high ex-ante systemic risk scores experienced greater difficulties during actual crises, or investigating whether these measures successfully predict which institutions required government bailouts. While informative, these approaches do not constitute formal statistical tests of forecast validity in the spirit of traditional backtesting procedures used for market risk measures.

Recent theoretical advances have opened new possibilities for rigorous backtesting of systemic risk measures. Fissler and Ziegel (2016) demonstrated that Expected Shortfall, despite not being individually elicitable, becomes jointly elicitable when paired with Value-at-Risk, meaning that strictly consistent scoring functions exist for the bivariate functional (see Fissler, Ziegel, and Gneiting, 2016, for an empirical application). This insight, extended by Fissler and Hoga (2021) to the systemic risk context, establishes that the MES and CoVaR of a firm are respectively jointly identifiable with the VaR of the market. Building on this theoretical foundation, Barendse, Kole, and van Dijk (2023) developed a comprehensive framework for backtesting VaR and ES forecasts that properly accounts for parameter estimation uncertainty. Their approach, based on joint identification functions rather than direct violations, provides the methodological blueprint for extending backtesting procedures to the bivariate setting required for systemic risk assessment.

Our paper leverages these theoretical developments to construct a comprehensive backtesting methodology for systemic risk measures. Following the approach of Banulescu-Radu, Hurlin, Leymarie, and Scaillet (2021), who first proposed backtests for the MES based on cumulative joint violations, we develop a unified testing framework that addresses three critical challenges. First, we must define an appropriate violation concept that captures the joint behavior of firm and market returns in tail regions. Unlike univariate backtests where a violation simply occurs when the realized loss exceeds the forecast, systemic risk backtesting requires tracking when both the institution experiences significant losses and the market is in distress, integrating this information across all possible risk levels. Second, since systemic risk forecasts are necessarily based on estimated model parameters, we must rigorously account for estimation uncertainty to avoid spurious rejections due to parameter estimation error rather than genuine model misspecification. Third, we must develop tests that maintain correct size and adequate power across realistic sample sizes encountered in regulatory practice, where estimation periods may be relatively short and evaluation periods potentially long.

The contribution of our paper is threefold. First, we develop unconditional coverage (UC) and independence (IND) tests for systemic risk measures that explicitly incorporate corrections for estimation risk. We derive the precise adjustments to the asymptotic covariance matrix of our test statistics that result from parameter uncertainty, extending the estimation risk corrections of Banulescu-Radu, Hurlin, Leymarie, and Scaillet (2021) and Barendse, Kole, and van Dijk (2023) to our joint testing framework. Through extensive Monte Carlo simulations, we demonstrate that ignoring estimation risk leads to severe size distortions, with empirical rejection frequencies reaching up to three times the nominal level when the ratio of out-of-sample to in-sample observations is large. Our robust test statistics, by contrast, maintain accurate size across all configurations.

Second, we provide a unified framework applicable to multiple systemic risk measures including the MES, CoVaR, and ΔCoVaR. A particularly valuable feature of our methodology is its ability to decompose ΔCoVaR backtesting results into separate assessments of stressed and median CoVaR components, thereby providing diagnostic insights into the sources of model misspecification.

Third, we conduct a comprehensive simulation study examining how the finite-sample performance of our tests depends on the relative sizes of the estimation and evaluation periods. By systematically varying the in-sample size from 250 to 1000 observations and the out-of-sample size from 250 to 1000 observations, we provide practitioners with clear guidance on when estimation risk corrections are essential versus when they have negligible impact. This analysis is particularly relevant for regulatory applications, where models estimated on pre-crisis data must be evaluated over extended post-crisis periods, creating precisely the conditions where estimation risk effects are most pronounced.

The practical implications of our methodology extend beyond academic interest. Regulatory frameworks for macroprudential oversight increasingly rely on systemic risk measures to identify global systemically important banks, calibrate capital surcharges, and design stress testing scenarios. The validity of these regulatory applications fundamentally depends on whether the underlying forecasting models accurately capture tail dependencies and systemic linkages. Our backtesting framework provides regulators with statistically sound tools to assess model adequacy, identify periods when forecasts become unreliable, and detect early warning signals of potential forecast breakdowns that may herald changes in systemic risk dynamics. Moreover, by documenting substantial size distortions when estimation risk is ignored, our findings underscore the importance of incorporating these corrections into regulatory backtesting protocols to avoid erroneous rejections that could unnecessarily burden financial institutions with modeling restrictions, and worry regulators with unforeseen systemic risk.

The remainder of this paper is organized as follows. Section 2 defines the systemic risk measures considered in this paper (MES, CoVaR, and ΔCoVaR) and their model-based forecasts. Section 3 develops the joint backtesting framework, derives the asymptotic distribution of the test statistics under parameter estimation uncertainty, and provides the corresponding covariance matrix corrections. Section 4 presents Monte Carlo simulation evidence on the finite-sample size properties of the standard and robust tests. Section 5 investigates the size-corrected power properties of the proposed joint backtests under a range of forecast-misspecification scenarios.

2 Systemic Risk Measures

This section defines MES, CoVaR, and ΔCoVaR. We first present their theoretical definitions. We then explain how they are constructed to produce predictions, distinguishing the theoretical measures from those obtained through a parametric risk model.

2.1 Theoretical Definition of Systemic Risk Measures

Let $\boldsymbol{y}_t = (y_{m,t}, y_{i,t})'$ be the vector of stock returns of two assets at time t, where $y_{m,t}$ corresponds to the market return and $y_{i,t}$ corresponds to the stock return of the i-th financial institution. Let $\mathcal{F}_{t-1}$ be the information set available at time $t-1$, with $\left(\boldsymbol{y}_{t-1}, \boldsymbol{y}_{t-2}, ...\right) \subseteq \mathcal{F}_{t-1}$, and let $F\left(.; \mathcal{F}_{t-1}\right)$ denote the joint cumulative distribution function (cdf) of $\boldsymbol{y}_t$ given $\mathcal{F}_{t-1}$, such that $F\left(\boldsymbol{y}; \mathcal{F}_{t-1}\right) \equiv \Pr\left(y_{m,t} \leq y_m, y_{i,t} \leq y_i \right| \mathcal{F}_{t-1})$ for any $\boldsymbol{y} = (y_m, y_i)' \in \mathbb{R}^2$.

Marginal Expected Shortfall. According to Acharya, Pedersen, Philippon, and Richardson (2017) and Brownlees and Engle (2017), the MES of a financial firm corresponds to its short-run expected equity loss conditional on the market taking a loss greater than its VaR. Formally, the α-level MES of the financial institution i at time t given $\mathcal{F}_{t-1}$ is defined as,

$$MES_{i,t}\left(\alpha\right) = \mathbb{E}(y_{i,t}|y_{m,t} \leq VaR_{m,t}(\alpha); \mathcal{F}_{t-1}), \tag{1}$$

where $VaR_{m,t}\left(\alpha\right)$ is the α-level VaR of $y_{m,t}$ with

$$VaR_{m,t}\left(\alpha\right) = F_{y_m}^{-1}\left(\alpha; \mathcal{F}_{t-1}\right), \tag{2}$$

with F_{y_m} denoting the cdf of y_m, and $\alpha \in [0,1]$. If the market return $y_{m,t}$ is defined as the value-weighted average of the returns of all firms in the financial system, then MES can be interpreted as the sensitivity of the market ES to a marginal increase in the institution's market share (Scaillet, 2004; Acharya, Pedersen, Philippon, and Richardson, 2017). This interpretation explains the use of the term "marginal" and highlights that MES captures the contribution of a given institution to the tail risk of the financial system as a whole. Note that MES serves as the basis for various further measures, which, in addition to MES, incorporate bank's balance sheet data and market capitalization to capture the institution's leverage and size. Notable examples are the systemic risk measure SRISK (Acharya, Engle, and Richardson, 2012; Brownlees and Engle, 2017) and the SES (Acharya, Pedersen, Philippon, and Richardson, 2017) (see e.g., the New York University V-Lab website for real-time computation of these measures). As these depend solely on MES and extra-financial variables that belong to $\mathcal{F}_{t-1}$, their backtesting naturally reduces to the backtesting of MES.

Conditional Value at Risk. The Conditional Value at Risk (CoVaR) of Adrian and Brunnermeier (2016) corresponds to the VaR of the firm return $y_{i,t}$ conditionally on some adverse event observed for $y_{m,t}$.[1] This state is represented by a situation in which the market return $y_{m,t}$ equals its $VaR_{m,t}(\alpha)$, and the CoVaR is then estimated using quantile regression. A more general definition considers the system to be under stress whenever $y_{m,t} \leq VaR_{m,t}(\alpha)$, as adopted in Girardi and Ergün (2013), Banulescu-Radu, Hurlin, Leymarie, and Scaillet (2021), or Francq and Zakoïan (2025). Formally, the (β, α)-CoVaR at time t is

$$CoVaR_{i,t}(\beta, \alpha) = F^{-1}_{y_{i,t}|y_{m,t} \leq VaR_{m,t}(\alpha)}(\beta; \mathcal{F}_{t-1}). \tag{3}$$

This definition provides a conditional risk measure that isolates the institution's loss level specifically in periods of systemic distress.

Delta Conditional Value at Risk. The ΔCoVaR is constructed by contrasting two CoVaR measures: one computed under systemic distress (see Eq. (3)) and another evaluated in a normal or median state of the financial system. The normal regime corresponds to situations in which the market return lies between two quantiles $VaR_{m,t}(\alpha_{\text{inf}})$ and $VaR_{m,t}(\alpha_{\text{sup}})$, with $\alpha < \alpha_{\text{inf}} < \alpha_{\text{sup}}$. A common choice in empirical application is $\alpha_{\text{inf}} = 0.25$ and $\alpha_{\text{sup}} = 0.75$ (see e.g. Francq and Zakoïan, 2025). The associated

[1] Adrian and Brunnermeier propose several variants of CoVaR depending on how the conditioning event is specified. For consistency with the conditioning scheme adopted for MES, we focus on the version of CoVaR that underlies the construction of ΔCoVaR, commonly referred to as Exposure-ΔCoVaR. This indicator captures how vulnerable an individual institution is to periods of systemic distress.

$(\beta, \alpha_{\text{inf}}, \alpha_{\text{sup}})$-CoVaR at time t is thus defined as

$$CoVaR_{i,t}(\beta, \alpha_{\text{inf}}, \alpha_{\text{sup}}) = F^{-1}_{y_{i,t}|VaR_{m,t}(\alpha_{\text{inf}}) \leq y_{m,t} \leq VaR_{m,t}(\alpha_{\text{sup}})}(\beta; \mathcal{F}_{t-1}). \tag{4}$$

Using these two conditioning events, the $(\beta, \alpha, \alpha_{\text{inf}}, \alpha_{\text{sup}})$-ΔCoVaR of institution i at time t is given by

$$\Delta CoVaR_{i,t}(\beta, \alpha, \alpha_{\text{inf}}, \alpha_{\text{sup}}) = CoVaR_{i,t}(\beta, \alpha) - CoVaR_{i,t}(\beta, \alpha_{\text{inf}}, \alpha_{\text{sup}}), \tag{5}$$

where $CoVaR_{i,t}(\beta, \alpha)$ is such that $\Pr(y_{i,t} \leq CoVaR_{i,t}(\beta, \alpha)|y_{m,t} \leq VaR_{m,t}(\alpha); \mathcal{F}_{t-1}) = \beta$ and $CoVaR_{i,t}(\beta, \alpha_{\text{inf}}, \alpha_{\text{sup}})$ verifies $\Pr(y_{i,t} \leq CoVaR_{i,t}(\beta, \alpha_{\text{inf}}, \alpha_{\text{sup}})|VaR_{m,t}(\alpha_{\text{inf}}) \leq y_{m,t} \leq VaR_{m,t}(\alpha_{\text{sup}}); \mathcal{F}_{t-1}) = \beta$. The first term captures the institution's loss level under systemic distress (i.e., when the market return falls below $VaR_{m,t}(\alpha)$). The second term reflects the same institution's risk in a normal or median state of the system (that is, when $VaR_{m,t}(\alpha_{\text{inf}}) \leq y_{m,t} \leq VaR_{m,t}(\alpha_{\text{sup}})$). Thus, ΔCoVaR measures how much additional risk an institution is exposed to when the market is in distress relative to normal times.

2.2 From Theoretical Measures to Model-Based Measures

In general, MES, CoVaR, and ΔCoVaR forecasts are generated from a parametric model specified by the researcher, the risk manager, or the regulatory authority. For example, Brownlees and Engle (2017) and Acharya, Engle, and Richardson (2012) rely on a bivariate DCC model to compute MES and SRISK. In practice, the cdf $F(\,\cdot\,; \mathcal{F}_{t-1}, \boldsymbol{\theta}_0)$ of the joint distribution of $\boldsymbol{y}_t$, the marginal cdf $F_{y_m}(\,\cdot\,; \mathcal{F}_{t-1}, \boldsymbol{\theta}_0)$, and the truncated cdf $F_{y_i|y_m \leq v_{m,t}(\alpha, \boldsymbol{\theta}_0)}(\,\cdot\,; \mathcal{F}_{t-1}, \boldsymbol{\theta}_0)$ all depend on an unknown parameter vector $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta} \subset \mathbb{R}^{d_\theta}$. Hence, these parameters must be estimated before computing the forecasts. We refer to this specification as the "risk model", in analogy with the internal risk models banks use to produce VaR or ES forecasts. We denote respectively by $v_{m,t}(\alpha, \boldsymbol{\theta}_0)$, $\mu_{i,t}(\alpha, \boldsymbol{\theta}_0)$, $c_{i,t}(\beta, \alpha, \boldsymbol{\theta}_0)$, and $c_{i,t}(\beta, \alpha_{\text{inf}}, \alpha_{\text{sup}}, \boldsymbol{\theta}_0)$ the risk model-based representations of $VaR_{m,t}(\alpha)$, $MES_{i,t}(\alpha)$, $CoVaR_{i,t}(\beta, \alpha)$, and $CoVaR_{i,t}(\beta, \alpha_{\text{inf}}, \alpha_{\text{sup}})$.

Backtesting procedures in the literature follow two main strategies. A first class of methods evaluates the validity of the risk model itself, using it as the primary input of the testing procedure. For instance, Banulescu-Radu, Hurlin, Leymarie, and Scaillet (2021) adopt such a perspective when backtesting MES. A second class directly assesses the correctness of the model-based measures, treating them as the terminal objects to be tested. Our approach belongs to the latter category. In our framework, these quantities enter the backtesting procedure directly: we test the model-based measures themselves, rather than the risk model used to generate them.

3 Backtesting Systemic Risk Measures

In this section, we first introduce the time-series processes used to backtest MES, CoVaR, and ΔCoVaR, and we discuss the conditions that characterize their correct specification. We then present the test statistic employed to assess these conditions.

3.1 Processes for Backtesting Systemic Risk Measures and Moment Conditions

We rely on identification functions for MES, VaR, and CoVaR introduced by Fissler and Hoga (2024). Identification of a functional relies on the existence of a first-order moment condition, referred to as an identification function, that uniquely characterizes the functional. This notion is linked to elicitability, since the first-order derivative of an expected scoring function yields precisely such a condition (Osband, 1985). Identification functions coincide with moment functions in the sense of the (generalized) method of moments of Newey and McFadden (1994). Their structure makes them suited for backtesting.

Process for MES. We define the process at time t for backtesting MES as the vector $\boldsymbol{g}_t^{\mu}(\alpha, \boldsymbol{\theta}_0) = \Big(g_{1,t}^{\mu}(\alpha, \boldsymbol{\theta}_0)\ ,\ g_{2,t}^{\mu}(\alpha, \boldsymbol{\theta}_0)\Big)'$, with components,

$$\begin{aligned} g_{1,t}^{\mu}(\alpha, \boldsymbol{\theta}_0) &= \mathbf{1}\{y_{m,t} \leq v_{m,t}(\alpha, \boldsymbol{\theta}_0)\} - \alpha, \\ g_{2,t}^{\mu}(\alpha, \boldsymbol{\theta}_0) &= \mathbf{1}\{y_{m,t} \leq v_{m,t}(\alpha, \boldsymbol{\theta}_0)\}\Big(\mu_{i,t}(\alpha, \boldsymbol{\theta}_0) - y_{i,t}\Big), \end{aligned} \tag{6}$$

where $v_{m,t}(\alpha, \boldsymbol{\theta}_0)$ is the α-VaR forecast of the market return at time t, and $\mu_{i,t}(\alpha, \boldsymbol{\theta}_0)$ is the α-MES forecast of institution i at time t, both produced by the bank risk model.

Equation 6 defines the process for backtesting MES. The first component $g_{1,t}^{\mu}$ records a VaR exceedance of the market return, that is, it equals $1-\alpha$ when $y_{m,t}$ falls below the predicted VaR level and equals $-\alpha$ otherwise. The second component $g_{2,t}^{\mu}$ refines this exceedance information by capturing the deviation between the MES forecast and the realized loss of institution i, but only on the subset of times when a market VaR exceedance occurs.

The process $\boldsymbol{g}_t^{\mu}$ differs fundamentally from the process H_t introduced in Banulescu-Radu, Hurlin, Leymarie, and Scaillet (2021) for backtesting MES. First, $\boldsymbol{g}_t^{\mu}$ is explicit in the quantities $\mu_{i,t}(\alpha, \boldsymbol{\theta}_0)$ and $v_{m,t}(\alpha, \boldsymbol{\theta}_0)$ whereas Banulescu-Radu, Hurlin, Leymarie, and Scaillet (2021) formulate their violation in terms of the cdf of the risk model and therefore assess the risk model itself rather than the risk measures. As a result, our computation of $\boldsymbol{g}_t^{\mu}$ requires only the MES and VaR forecasts of the risk model, while the approach of Banulescu-Radu, Hurlin, Leymarie, and Scaillet (2021) relies on the full predictive cdf. This implies that their procedure necessitates reimplementing the entire risk model during backtesting, even though in practical applications only the forecasts are available to the analyst or regulator. Second, $\boldsymbol{g}_t^{\mu}$ is bivariate, which is required because MES cannot be identified without the market VaR. Backtesting MES necessitates verifying both quantities jointly. Using a single component as in Banulescu-Radu, Hurlin, Leymarie, and Scaillet (2021) would leave the system of moment conditions incomplete and can result in substantial power losses (see the discussion in Fissler and Hoga, 2024).

Process for CoVaR. We define the process at time t for backtesting CoVaR as the vector $\boldsymbol{g}_t^{\mathrm{c}}(\beta, \alpha, \boldsymbol{\theta}_0) = \left(g_{1,t}^{\mathrm{c}}(\alpha, \boldsymbol{\theta}_0)\ ,\ g_{2,t}^{\mathrm{c}}(\beta, \alpha, \boldsymbol{\theta}_0)\right)'$, with components

$$\begin{aligned} g_{1,t}^{\mathrm{c}}(\alpha, \boldsymbol{\theta}_0) &= \mathbf{1}\{y_{m,t} \leq v_{m,t}(\alpha, \boldsymbol{\theta}_0)\} - \alpha, \\ g_{2,t}^{\mathrm{c}}(\beta, \alpha, \boldsymbol{\theta}_0) &= \mathbf{1}\{y_{m,t} \leq v_{m,t}(\alpha, \boldsymbol{\theta}_0)\}\Big(\mathbf{1}\{y_{i,t} \leq c_{i,t}(\beta, \alpha, \boldsymbol{\theta}_0)\} - \beta\Big), \end{aligned} \tag{7}$$

where $v_{m,t}(\alpha, \boldsymbol{\theta}_0)$ is the α-VaR forecast of the market return at time t, and $c_{i,t}(\beta, \alpha, \boldsymbol{\theta}_0)$ is the (β, α)-CoVaR forecast of institution i at time t, both produced by the bank risk model.

Equation 7 introduces the process used for backtesting CoVaR. The first component $g^c_{1,t}$ is identical to $g^\mu_{1,t}$ in Equation 6. The second component $g^c_{2,t}$ refines this information by tracking whether the institution return $y_{i,t}$ also falls below its predicted (β,α)-CoVaR, but only on dates when a market VaR exceedance occurs. Hence, it takes the value $1-\beta$ when both a market VaR exceedance and an institution CoVaR exceedance occur, and $-\beta$ when the market VaR is violated but the institution does not exceed its CoVaR.

The structure of Equation 7 is closely related to that of Banulescu-Radu, Hurlin, Leymarie, and Scaillet (2021), in the sense that their statistic h_t related to CoVaR corresponds to $g^c_{2,t}$. In contrast, we additionally include $g^c_{1,t}$ for the VaR. As discussed in Fissler and Hoga (2024), one-dimensional identification functions for CoVaR fail to be strict, unlike the two-dimensional identification functions used here, which results in a substantial loss of power against certain alternatives.

Process for ΔCoVaR. We define the process at time t for ΔCoVaR as the vector $\boldsymbol{g}^\Delta_t(\beta,\alpha,\alpha_{\text{inf}},\alpha_{\text{sup}},\boldsymbol{\theta}_0) =$ $\Big(g^\Delta_{1,t}(\alpha,\boldsymbol{\theta}_0)\ ,\ g^\Delta_{2,t}(\beta,\alpha,\boldsymbol{\theta}_0)\ ,\ g^\Delta_{3,t}(\alpha_{\text{inf}},\alpha_{\text{sup}},\boldsymbol{\theta}_0)\ ,\ g^\Delta_{4,t}(\beta,\alpha_{\text{inf}},\alpha_{\text{sup}},\boldsymbol{\theta}_0)\Big)'$, with components

$$\begin{aligned}
&g^\Delta_{1,t}(\alpha,\boldsymbol{\theta}_0) = \mathbf{1}\{y_{m,t} \le v_{m,t}(\alpha,\boldsymbol{\theta}_0)\} - \alpha,\\
&g^\Delta_{2,t}(\beta,\alpha,\boldsymbol{\theta}_0) = \mathbf{1}\{y_{m,t} \le v_{m,t}(\alpha,\boldsymbol{\theta}_0)\}\Big(\mathbf{1}\{y_{i,t} \le c_{i,t}(\beta,\alpha,\boldsymbol{\theta}_0)\} - \beta\Big),\\
&g^\Delta_{3,t}(\alpha_{\text{inf}},\alpha_{\text{sup}},\boldsymbol{\theta}_0) = \mathbf{1}\{v_{m,t}(\alpha_{\text{inf}},\boldsymbol{\theta}_0) \le y_{m,t} \le v_{m,t}(\alpha_{\text{sup}},\boldsymbol{\theta}_0)\} - (\alpha_{\text{sup}} - \alpha_{\text{inf}}),\\
&g^\Delta_{4,t}(\beta,\alpha_{\text{inf}},\alpha_{\text{sup}},\boldsymbol{\theta}_0) = \mathbf{1}\{v_{m,t}(\alpha_{\text{inf}},\boldsymbol{\theta}_0) \le y_{m,t} \le v_{m,t}(\alpha_{\text{sup}},\boldsymbol{\theta}_0)\}\Big(\mathbf{1}\{y_{i,t} \le c_{i,t}(\beta,\alpha_{\text{inf}},\alpha_{\text{sup}},\boldsymbol{\theta}_0)\} - \beta\Big),
\end{aligned} \tag{8}$$

where $v_{m,t}(\alpha,\boldsymbol{\theta}_0)$, $v_{m,t}(\alpha_{\text{inf}},\boldsymbol{\theta}_0)$, and $v_{m,t}(\alpha_{\text{sup}},\boldsymbol{\theta}_0)$ denote the α-VaR forecast, the α_{inf}-VaR forecast, and the α_{sup}-VaR forecast of the market return at time t, with $\alpha < \alpha_{\text{inf}} < \alpha_{\text{sup}}$, and $c_{i,t}(\beta,\alpha,\boldsymbol{\theta}_0)$ and $c_{i,t}(\beta,\alpha_{\text{inf}},\alpha_{\text{sup}},\boldsymbol{\theta}_0)$ denote the (β,α)- and $(\beta,\alpha_{\text{inf}},\alpha_{\text{sup}})$-CoVaR forecasts of institution i at time t, produced by the bank risk model.

Equation 8 introduces the 4×1 vector process for backtesting ΔCoVaR. The components $g^\Delta_{1,t}$ and $g^\Delta_{2,t}$ correspond to those in $\boldsymbol{g}^c_t$. The last two components, $g^\Delta_{3,t}$ and $g^\Delta_{4,t}$, are required to backtest the $(\beta,\alpha_{\text{inf}},\alpha_{\text{sup}})$-CoVaR. Consequently, consistent backtesting of ΔCoVaR requires four orthogonality conditions (two for the stressed CoVaR and two for the median-band CoVaR).

Moment Condition under Correct Specification. Under correct specification, each process exhibits the martingale difference sequence (MDS) property with respect to the systemic risk measure it targets. Correct specification requires that the model-based objects coincide with their theoretical counterparts, namely $v_{m,t}(\alpha, \boldsymbol{\theta}_0) = VaR_{m,t}(\alpha)$, $\mu_{i,t}(\alpha, \boldsymbol{\theta}_0) = MES_{i,t}(\alpha)$, $c_{i,t}(\beta, \alpha, \boldsymbol{\theta}_0) = CoVaR_{i,t}(\beta, \alpha)$, $v_{m,t}(\alpha_{\text{inf}}, \boldsymbol{\theta}_0) = VaR_{m,t}(\alpha_{\text{inf}})$, $v_{m,t}(\alpha_{\text{sup}}, \boldsymbol{\theta}_0) = VaR_{m,t}(\alpha_{\text{sup}})$, $c_{i,t}(\beta, \alpha_{\text{inf}}, \alpha_{\text{sup}}, \boldsymbol{\theta}_0) = CoVaR_{i,t}(\beta, \alpha_{\text{inf}}, \alpha_{\text{sup}})$. Under these conditions, the MES, CoVaR, and ΔCoVaR processes satisfy $\mathbb{E}[\boldsymbol{g}_t^{\mu}(\alpha, \boldsymbol{\theta}_0) \mid \mathcal{F}_{t-1}] = \mathbf{0}$, $\mathbb{E}[\boldsymbol{g}_t^{c}(\beta, \alpha, \boldsymbol{\theta}_0) \mid \mathcal{F}_{t-1}] = \mathbf{0}$, $\mathbb{E}[\boldsymbol{g}_t^{\Delta}(\beta, \alpha, \alpha_{\text{inf}}, \alpha_{\text{sup}}, \boldsymbol{\theta}_0) \mid \mathcal{F}_{t-1}] = \mathbf{0}$, so that each process forms an MDS whenever its associated model-based risk measure is correctly specified. We provide a proof of the MDS property of $\boldsymbol{g}_t^{\mu}$, $\boldsymbol{g}_t^{c}$, and $\boldsymbol{g}_t^{\Delta}$ in the Appendix, where the result is stated in Lemma 1. The MDS property follows directly from the construction in Fissler and Hoga (2024) considering the identification functions for the corresponding systemic risk functionals.

3.2 General Testing Framework

We use $\boldsymbol{g}_t^{j}(\boldsymbol{\theta}_0)$ to denote, respectively, $\boldsymbol{g}_t^{\mu}$, $\boldsymbol{g}_t^{c}$, and $\boldsymbol{g}_t^{\Delta}$ in the MES, CoVaR, and ΔCoVaR cases. Under correct specification of the forecasts, the processes satisfy the MDS property with the following null hypothesis $H_0 : \mathbb{E}\Big[\boldsymbol{g}_t^{j}(\boldsymbol{\theta}_0) \mid \mathcal{F}_{t-1}\Big] = \mathbf{0} \quad \text{for all } t, \text{ and } j \in \{\mu, c, \Delta\}$.

We introduce a test matrix $\boldsymbol{H}_t(\boldsymbol{\theta}_0)$ that incorporates conditioning information and construct moment conditions based on the product $\boldsymbol{H}_t(\boldsymbol{\theta}_0)\boldsymbol{g}_t^{j}(\boldsymbol{\theta}_0)$. Throughout, we take $\boldsymbol{H}_t(\boldsymbol{\theta}_0)$ to be $\mathcal{F}_{t-1}$-measurable so that $\boldsymbol{H}_t(\boldsymbol{\theta}_0)\boldsymbol{g}_t^{j}(\boldsymbol{\theta}_0)$ inherits the MDS structure under the null. This formulation mirrors the logic of standard VaR backtesting. In the VaR setting, the testing problem builds on the sequence of exceedances, which we denote by $g_{1,t}$, with $g_{1,t} = g_{1,t}^{\mu} = g_{1,t}^{c}$. Christoffersen (1998) shows that VaR forecast validity can be assessed through two properties on exceedances: unconditional coverage, expressed as $\mathbb{E}[g_{1,t}] = 0$, and independence, meaning that $g_{1,t}$ is independent of $g_{1,t-k}$

for $k \neq 0$. In the univariate case, these properties are recovered by scalar forms of the test matrix $\boldsymbol{H}_t(\boldsymbol{\theta}_0)$. In particular, $\boldsymbol{H}_t(\boldsymbol{\theta}_0) = 1$ yields the unconditional coverage test of Kupiec (1995), while $\boldsymbol{H}_t(\boldsymbol{\theta}_0) = g_{1,t-1}$, (or more generally $\boldsymbol{H}_t(\boldsymbol{\theta}_0) = g_{1,t-k}$) leads to independence tests analogous to the test of Engle and Manganelli (2004).

We define the general null hypothesis as the unconditional moment restriction

$$H_0: \quad \mathbb{E}\Big[\boldsymbol{m}_t^j(\boldsymbol{\theta}_0)\Big] = \mathbb{E}\Big[\boldsymbol{H}_t(\boldsymbol{\theta}_0)\, \boldsymbol{g}_t^j(\boldsymbol{\theta}_0)\Big] = \boldsymbol{0}, \quad \text{for } j \in \{\mu, c, \Delta\}. \tag{9}$$

Under correct specification, $\boldsymbol{g}_t^j(\boldsymbol{\theta}_0)$ is a martingale difference sequence (MDS) and $\boldsymbol{H}_t(\boldsymbol{\theta}_0)$ is $\mathcal{F}_{t-1}$-measurable; hence $\boldsymbol{m}_t^j(\boldsymbol{\theta}_0) = \boldsymbol{H}_t(\boldsymbol{\theta}_0)\boldsymbol{g}_t^j(\boldsymbol{\theta}_0)$ is also an MDS and in particular satisfies (9). Equivalently, $\mathbb{E}[\boldsymbol{m}_t^j(\boldsymbol{\theta}_0)] = \mathbb{E}\Big[\boldsymbol{H}_t(\boldsymbol{\theta}_0)\, \mathbb{E}\Big(\boldsymbol{g}_t^j(\boldsymbol{\theta}_0) \mid \mathcal{F}_{t-1}\Big)\Big]$.

In our setting, $\boldsymbol{H}_t(\boldsymbol{\theta}_0) \equiv \boldsymbol{I}_{d_j}$ yields the unconditional test for MES, CoVaR, or ΔCoVaR, while alternative specifications of $\boldsymbol{H}_t$ embed conditional or dependence-based tests in direct analogy with the VaR framework. A related matrix-based approach is proposed by Barendse, Kole, and van Dijk (2023), who develop a similar structure for backtesting ES jointly with VaR.

The general null in (9) can be tested through its empirical counterpart computed out-of-sample. We partition the sample $\{\boldsymbol{y}_t\}_{t=1}^T$, where T denotes the full sample size, into an initial estimation period of length M and a backtesting period of length $N = T-M$. At each out-of-sample time t, the parameter vector $\boldsymbol{\theta}_0$ is estimated using a set of in-sample observations, yielding $\hat{\boldsymbol{\theta}}_t$ and the corresponding risk measure forecasts, which are then evaluated over the N backtesting observations. We consider fixed, rolling, and expanding forecasting schemes, as in West and McCracken (1998), McCracken (2000), and Escanciano and Olmo (2010), which differ in the observations used to estimate $\boldsymbol{\theta}_0$.

Let $\hat{\boldsymbol{\theta}}_t$ denote an estimator of $\boldsymbol{\theta}_0$ at time t computed using the set of in-sample time indices W_t of size $|W_t|$ (i.e., using information available up to time $t-1$). Under the fixed and rolling schemes, the estimation window has a fixed size, with $|W_t| = M$, where $W_t = \{1, \ldots, M\}$ and $W_t = \{t-M, \ldots, t-1\}$, respectively. Under the expanding scheme, $W_t = \{1, \ldots, t-1\}$, so that $|W_t| = t-1$. Define the instrumented moment vector $\boldsymbol{m}_t^j(\boldsymbol{\theta}) := \boldsymbol{H}_t(\boldsymbol{\theta})\, \boldsymbol{g}_t^j(\boldsymbol{\theta})$. Accordingly, we consider the empirical statistic

$$\boldsymbol{S}_N^j(\hat{\boldsymbol{\theta}}_t) = \frac{1}{\sqrt{N}} \sum_{t=M+1}^{T} \boldsymbol{m}_t^j(\hat{\boldsymbol{\theta}}_t) = \frac{1}{\sqrt{N}} \sum_{t=M+1}^{T} \boldsymbol{H}_t(\hat{\boldsymbol{\theta}}_t)\, \boldsymbol{g}_t^j(\hat{\boldsymbol{\theta}}_t), \tag{10}$$

which provides an out-of-sample sample analogue of the moment conditions in (9).

Let $\boldsymbol{\ell}_t(\boldsymbol{\theta})$ denote the d_θ-dimensional estimating function associated with the estimator $\hat{\boldsymbol{\theta}}_t$. We assume that $\hat{\boldsymbol{\theta}}_t$ admits the asymptotic linear representation $\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_0 = \boldsymbol{B}\, |W_t|^{-1} \sum_{s \in W_t} \boldsymbol{\ell}_s(\boldsymbol{\theta}_0) + o_p(|W_t|^{-1/2})$, where W_t denotes the estimation window used at time t, $|W_t|$ its size, and $\boldsymbol{B}$ is a non-random full-rank matrix of size $d_\theta \times d_\theta$ arising from this linearization, i.e., the bread matrix. See Assumption 6 in Appendix A for details.

Theorem 1. *Under Assumptions A.1–A.8 (listed in Appendix A) with $M \to \infty$, $N \to \infty$, and $N/M \to p$ with $0 \leq p < \infty$, for each $j \in \{\mu, c, \Delta\}$,*

$$\boldsymbol{S}_N^j \xrightarrow{d} \mathcal{N}\Big(\mathbf{0}, \boldsymbol{\Omega}_j\Big), \qquad \boldsymbol{\Omega}_j = \boldsymbol{\Sigma}_j + \kappa_{hl}(p)\Big(\boldsymbol{A}_j \boldsymbol{B} \boldsymbol{q}_j + \boldsymbol{q}_j' \boldsymbol{B}' \boldsymbol{A}_j'\Big) + \kappa_{ll}(p)\, \boldsymbol{A}_j \boldsymbol{B} \boldsymbol{V} \boldsymbol{B}' \boldsymbol{A}_j'.$$

where $\boldsymbol{m}_t^j(\boldsymbol{\theta}) := \boldsymbol{H}_t(\boldsymbol{\theta})\, \boldsymbol{g}_t^j(\boldsymbol{\theta})$ and $\boldsymbol{m}_t^j := \boldsymbol{m}_t^j(\boldsymbol{\theta}_0)$, and

$$\boldsymbol{\Sigma}_j = \sum_{k \in \mathbb{Z}} \operatorname{Cov}\Big(\boldsymbol{m}_t^j,\, \boldsymbol{m}_{t-k}^j\Big)$$

is the $d_j \times d_j$ long-run variance-covariance matrix of the instrumented moment sequence. Under H_0, since $\boldsymbol{g}_t^j(\boldsymbol{\theta}_0)$ is an MDS and $\boldsymbol{H}_t(\boldsymbol{\theta}_0)$ is $\mathcal{F}_{t-1}$-measurable, $\{\boldsymbol{m}_t^j\}$ is also an MDS and thus $\boldsymbol{\Sigma}_j = \operatorname{Var}(\boldsymbol{m}_t^j) = \operatorname{Var}\Big(\boldsymbol{H}_t(\boldsymbol{\theta}_0) \boldsymbol{g}_t^j(\boldsymbol{\theta}_0)\Big)$. Moreover, $\boldsymbol{A}_j = \partial_{\boldsymbol{\theta}}\, \mathbb{E}\Big[\boldsymbol{m}_t^j(\boldsymbol{\theta})\Big]\Big|_{\boldsymbol{\theta}_0}$ is the $d_j \times d_\theta$ matrix of derivatives of $\mathbb{E}[\boldsymbol{m}_t^j(\boldsymbol{\theta})]$ with respect to $\boldsymbol{\theta}$, $\boldsymbol{V} = \sum_{k \in \mathbb{Z}} \mathbb{E}[\boldsymbol{\ell}_t(\boldsymbol{\theta}_0)\, \boldsymbol{\ell}_{t-k}(\boldsymbol{\theta}_0)']$ is the $d_\theta \times d_\theta$ long-run covariance matrix of the estimating function $\boldsymbol{\ell}_t$, $\boldsymbol{B}$ is the bread matrix defined above, $\boldsymbol{q}_j = \sum_{k \in \mathbb{Z}} \operatorname{Cov}\Big(\boldsymbol{\ell}_{t-k}(\boldsymbol{\theta}_0),\, \boldsymbol{m}_t^j(\boldsymbol{\theta}_0)\Big)$ is the $d_\theta \times d_j$ long-run lagged cross-covariance matrix between $\boldsymbol{\ell}_t$ and $\boldsymbol{m}_t^j$, and $\kappa_{hl}(p)$ and $\kappa_{ll}(p)$ are some predetermined weights whose values depend on the forecasting scheme (fixed/rolling/expanding).

Theorem 1 provides the joint asymptotic normality of $\boldsymbol{S}_N^j$. The variance-covariance matrix of $\boldsymbol{S}_N^j$ depends on two components, namely $\boldsymbol{\Sigma}_j$ and the estimation-risk corrections. Formally, this decomposition separates the sampling variability of the instrumented moment condition $\boldsymbol{m}_t^j(\boldsymbol{\theta}_0)$ through $\boldsymbol{\Sigma}_j$, from estimation-risk components induced by parameter estimation $\hat{\boldsymbol{\theta}}_t$ and driven by both the long-run variance $\boldsymbol{V}$ and bread matrix $\boldsymbol{B}$ of the estimating function $\boldsymbol{\ell}_t$ and by the long-run cross-covariance $\boldsymbol{q}_j$ between $\boldsymbol{\ell}_t$ and $\boldsymbol{m}_t^j$. Hence, the estimation-risk corrections inherit the dependence on the test matrix $\boldsymbol{H}_t$ through $\boldsymbol{A}_j$ and $\boldsymbol{q}_j$. Note that both $\boldsymbol{A}_j = \partial_{\boldsymbol{\theta}} \mathbb{E}\Big[\boldsymbol{H}_t(\boldsymbol{\theta})\boldsymbol{g}_t^j(\boldsymbol{\theta})\Big]\Big|_{\boldsymbol{\theta}_0}$ and $\boldsymbol{q}_j = \sum_{k\in\mathbb{Z}} \operatorname{Cov}\Big(\boldsymbol{\ell}_{t-k}(\boldsymbol{\theta}_0),\, \boldsymbol{H}_t(\boldsymbol{\theta}_0)\boldsymbol{g}_t^j(\boldsymbol{\theta}_0)\Big)$ depend on the choice of $\boldsymbol{H}_t$. Furthermore, the weights $\kappa_{hl}(p)$ and $\kappa_{ll}(p)$ depend only on the forecasting scheme (fixed, rolling, or expanding) and on the ratio $p = N/M$. We provide their closed-form expressions in Proposition 2 in Appendix A. Finally, under the MDS property, the long-run variance $\boldsymbol{\Sigma}_j$ collapses to $\boldsymbol{\Sigma}_j = \operatorname{Var}\Big(\boldsymbol{H}_t(\boldsymbol{\theta}_0)\boldsymbol{g}_t^j(\boldsymbol{\theta}_0)\Big)$. Importantly, the MDS property removes serial correlation in $\{\boldsymbol{m}_t^j\}$, but it does not restrict the choice of $\boldsymbol{H}_t$; hence $\boldsymbol{\Sigma}_j$ generally depends on $\boldsymbol{H}_t$ through $\operatorname{Var}(\boldsymbol{H}_t(\boldsymbol{\theta}_0)\boldsymbol{g}_t^j(\boldsymbol{\theta}_0))$. In the unconditional case $\boldsymbol{H}_t \equiv \boldsymbol{I}_{d_j}$, this reduces to $\operatorname{Var}(\boldsymbol{g}_t^j(\boldsymbol{\theta}_0))$.

Corollary 1. *When there is no estimation risk, i.e., when $p = 0$; under Assumptions A.1–A.8 in Appendix A, we have $\boldsymbol{\Omega}_j = \boldsymbol{\Sigma}_j$.*

Corollary 1 highlights the simplified form of $\boldsymbol{\Omega}_j$ in the absence of estimation risk. As $p \to 0$, both weighting terms $\kappa_{hl}(p)$ and $\kappa_{ll}(p)$ vanish, so that the estimation-risk corrections disappear and $\boldsymbol{\Omega}_j$ converges to $\boldsymbol{\Sigma}_j$.

Corollary 2. *Consider the Wald-type test statistic*

$$W_N^j = \boldsymbol{S}_N^j(\hat{\boldsymbol{\theta}}_t)'\,\hat{\boldsymbol{\Omega}}_j^{-1}\,\boldsymbol{S}_N^j(\hat{\boldsymbol{\theta}}_t), \tag{11}$$

where $\hat{\boldsymbol{\Omega}}_j$ is a consistent estimator of the asymptotic covariance matrix $\boldsymbol{\Omega}_j$ defined in Theorem 1, and $\mathrm{d}_j = \dim(\boldsymbol{S}_N^j)$. Under the null hypothesis and as $N \to \infty$, the test statistic W_N^j converges in distribution to a chi-square random variable with d_j degrees of freedom.

Corollary 2 introduces a Wald-type test statistic indexed by j, which allows for backtesting different systemic risk measures through the corresponding moment conditions $\boldsymbol{g}_t^j$. Since the true parameter vector $\boldsymbol{\theta}_0$ is typically unknown, the functions $\boldsymbol{H}_t(\boldsymbol{\theta})$ and

$\boldsymbol{g}_t^j(\boldsymbol{\theta})$ are evaluated at the estimator $\hat{\boldsymbol{\theta}}_t$ rather than at $\boldsymbol{\theta}_0$. A feasible plug-in estimator $\hat{\boldsymbol{\Omega}}_j$ of the asymptotic covariance matrix is provided in Proposition 3 in Appendix A.

Under the null hypothesis of correctly specified forecasts, the test statistic converges in distribution to a chi-square random variable with d_j degrees of freedom. This result justifies a level-η test that rejects H_0 whenever $W_N^j > \chi^2_{d_j,1-\eta}$ where $\chi^2_{d_j,1-\eta}$ denotes the $(1-\eta)$ quantile of the chi-square distribution with d_j degrees of freedom.

In empirical work, we report both (i) a standard implementation that ignores estimation-risk corrections by using $\hat{\boldsymbol{\Omega}}_j = \hat{\boldsymbol{\Sigma}}_j$ (a consistent estimator of $\boldsymbol{\Sigma}_j$), and (ii) a robust implementation that uses $\hat{\boldsymbol{\Omega}}_j$ with the estimation-risk corrections implied by Theorem 1. We consider three joint tests for (VaR, MES), (VaR, CoVaR), and ΔCoVaR, varying in the specification of $\boldsymbol{H}_t(\boldsymbol{\theta})$. We adhere to the conventional terminology in the literature, as seen in Christoffersen (1998), distinguishing between unconditional and conditional tests. Unconditional tests aim to identify significant forecast errors on average, corresponding to constant instruments (e.g., $\boldsymbol{H}_t \equiv \boldsymbol{I}_{d_j}$). In contrast, conditional tests use $\mathcal{F}_{t-1}$-measurable instruments $\boldsymbol{H}_t$ to detect state-dependence or clustering in forecast errors. Typical choices include lagged exceedances (which recover lag-based moment conditions such as $\mathbb{E}[g_{i,t}(\theta_0)g_{k,t-1}(\theta_0)]$) or other predetermined variables (e.g., lagged returns), depending on the application.

4 Size Properties of the Joint Backtest

4.1 Simulation Design

We conduct Monte Carlo simulations to evaluate the finite-sample size properties of our joint backtesting procedure under correct model specification. The simulation design follows Banulescu-Radu, Hurlin, Leymarie, and Scaillet (2021) and considers two data generating processes (DGP): *a marginal DGP* with i.i.d. bivariate returns (no dynamics), and *a conditional DGP* where returns follow a bivariate dynamic structure.[2] For each replication, we estimate model parameters using an in-sample of size M observa-

[2] The precise DGP specifications and the corresponding risk-measure forecast formulas used in the simulations are provided in Appendix B, Sections B.1–B.2.

tions and evaluate forecast performance over an out-of-sample backtesting period of size N observations, so that the full sample size is $T = M + N$. All Monte Carlo experiments are conducted under the fixed forecasting scheme, i.e., parameters are estimated once using $\{1, \ldots, M\}$ and the resulting forecasts are evaluated for all $t = M + 1, \ldots, T$. We consider 1,000 Monte Carlo replications for each configuration.

The parameter combinations examined are: $M \in \{250, 500, 1000\}$ and $N \in \{250, 500, 1000\}$, which represent realistic sample sizes encountered in practice. For instance, $M = 250$ corresponds to approximately one year of daily returns for parameter estimation, while $N = 500$ represents two years of out-of-sample evaluation. The ratio $p = N/M$ varies from 0.5 to 2, allowing us to assess the impact of relative sample sizes on test performance.

For each specification, we implement six types of tests: unconditional and conditional versions of the (VaR, MES) test, the (VaR, CoVaR) test, and the ΔCoVaR test. *Unconditional tests* set $\boldsymbol{H}_t \equiv \boldsymbol{I}_{d_j}$, so that $\boldsymbol{m}_t^j(\theta) = \boldsymbol{g}_t^j(\theta)$. *Conditional tests* use a bounded $\mathcal{F}_{t-1}$-measurable scalar instrument z_t based on the magnitude of the lagged market return $y_{m,t-1}$. Let

$$\hat{s}_M := \operatorname{sd}(y_{m,1}, \ldots, y_{m,M})$$

be an in-sample scale estimate computed from the estimation period, and define

$$z_t \;=\; 1 + \tanh\!\left(\frac{|y_{m,t-1}|}{\hat{s}_M}\right) \in (1, 2), \qquad \boldsymbol{H}_t \equiv z_t \boldsymbol{I}_{d_j},$$

so that $\boldsymbol{m}_t^j(\theta) = z_t\, \boldsymbol{g}_t^j(\theta)$. By construction, $\boldsymbol{H}_t$ is $\mathcal{F}_{t-1}$-measurable and bounded, ensuring that the conditional tests remain within the framework of Section 3.2. Each test is computed in two versions: *a standard version* that ignores estimation risk (using $\hat{\Omega}_j = \hat{\Sigma}_j$), and *a robust version* that accounts for parameter uncertainty through the estimation-risk corrections in Theorem 1. All tests are conducted at the nominal level $\eta = 0.05$. We set $\alpha = 0.05$, $\beta = 0.5$, and $(\alpha_{\text{inf}}, \alpha_{\text{sup}}) = (0.25, 0.75)$.

4.2 Size Results

Table 1 presents the empirical rejection frequencies under the null hypothesis of correct model specification. Several important findings emerge from these results.

4.3 Discussion of Size Results

The results in Table 1 reveal several critical patterns regarding the finite-sample behavior of our joint backtesting procedures.

Severe size distortions of standard tests. First and foremost, the standard tests that ignore estimation risk exhibit substantial size distortions across all configurations. For instance, when $M = 250$ and $N = 500$ under the marginal DGP, the unconditional (VaR, MES) standard test rejects 13.0% of the time, more than 2.6 times the nominal 5% level. These distortions become more pronounced as the ratio $p = N/M$ increases. For instance, when $M = 500$ and $N = 1000$ (i.e., $p = 2$), the rejection frequency reaches 14.0% for the (VaR, MES) test. This pattern is consistent with the theoretical prediction that the standard test statistic is affected by estimation risk when M is fixed and N increases, as documented by Banulescu-Radu, Hurlin, Leymarie, and Scaillet (2021) for the MES case.

The intuition behind these size distortions is straightforward: parameter estimation introduces additional variability in the test statistic that is not accounted for in the standard asymptotic distribution. When the out-of-sample period is relatively large compared to the estimation period, the impact of estimation error on the cumulative violation process accumulates over time, leading to inflated rejection frequencies. The standard test incorrectly attributes this estimation-induced variability to model misspecification, resulting in spurious rejections.

Excellent performance of robust tests. In sharp contrast, the robust tests that explicitly account for estimation risk demonstrate empirical sizes remarkably close to the nominal 5% level across all configurations. For the same scenario with $M = 250$ and $N = 500$, the robust (VaR, MES) test exhibits a rejection frequency of only 4.8%, nearly identical to the nominal level. Even in the most challenging configuration with $M = 500$ and $N = 1000$, the robust test maintains a rejection frequency of 5.5%, well within acceptable bounds.

This excellent size control holds uniformly across all three test types (VaR-MES, VaR-CoVaR, and ΔCoVaR) and both data generating processes (marginal and conditional). The robust tests also maintain appropriate size for both unconditional and conditional versions. These findings validate our theoretical derivation of the asymptotic distribution under estimation uncertainty and demonstrate that the proposed corrections to the covariance matrix are effective in finite samples.

Sensitivity to sample size ratio. The magnitude of size distortions in standard tests increases monotonically with the ratio $p = N/M$. When M is large relative to N (e.g., $M = 1000$, $N = 500$, $p = 0.5$), the standard test performs reasonably well, with rejection frequencies around 6.7%, only moderately above the nominal level. However, as p increases beyond unity, the size distortions become severe. This pattern underscores the practical importance of accounting for estimation risk, particularly in regulatory contexts where out-of-sample evaluation periods may be extended to assess forecast performance over multiple years.

Impact of DGP and test type. The conditional DGP generally produces more severe size distortions than the marginal DGP, especially for larger values of p. For example, with $M = 250$ and $N = 500$, the standard unconditional (VaR, MES) test rejects 16.5% of the time under the conditional DGP compared to 13.0% under the marginal DGP. This difference likely reflects the additional complexity in estimating the dependence structure between firm and market returns in the conditional case, which amplifies the impact of estimation uncertainty.

Comparing across test types, the (VaR, MES), (VaR, CoVaR), and ΔCoVaR tests exhibit similar size properties, with rejection frequencies typically within 1–2 percentage points of each other for a given configuration. This similarity is expected given that all three tests are based on the same underlying cumulative violation process framework, differing primarily in their definition of the conditional quantile being tested.

Conditional versus unconditional tests. The conditional versions of the tests, which incorporate lagged information in the test matrix $\boldsymbol{H}_t(\theta)$, generally exhibit size properties similar to their unconditional counterparts. Under the marginal DGP, conditional tests occasionally show slightly better size control than unconditional tests (e.g., when $M = 250, N = 500$, the conditional robust test has size 4.7% versus 4.8% for the unconditional version). However, these differences are minor and do not suggest a systematic advantage of one approach over the other in terms of size control.

Practical implications. These simulation results have important implications for practitioners and regulators. First, accounting for estimation risk is not merely a theoretical refinement but a practical necessity for obtaining valid backtests, particularly when the evaluation period is substantial relative to the estimation period. Second, the robust tests perform well even when M is as small as 250 observations (approximately one year of daily data), suggesting that our methodology can be reliably applied in realistic regulatory settings. Third, the consistent performance across different systemic risk measures (MES, CoVaR, ΔCoVaR) indicates that our framework provides a unified approach for validating diverse specifications of systemic risk.

Recommendations. Based on these findings, we strongly recommend the use of robust test statistics that account for estimation risk in all practical applications. While standard tests may suffice when the in-sample period is very large relative to the out-of-sample period ($p \ll 1$), this condition rarely holds in regulatory practice where models are regularly evaluated over extended horizons. The modest computational cost of implementing the robust corrections is far outweighed by the benefits of correct size and valid inference.

5 Power Properties of the Joint Backtest

5.1 Simulation Design

We consider four forecast-misspecification scenarios. Scenarios A1–A3 are implemented under both the marginal and conditional DGPs and are indexed by a severity parameter

$\tau \in \{25\%, 50\%, 75\%\}$. Scenario B4 is considered only under the conditional DGP and has no severity parameter.

- **Alternative A1 (institution volatility underestimated):** The forecasting model understates the institution-side conditional variance. In our implementation, the institution variance forecasts are scaled down by a factor $(1-\tau)$, while the market-side variance and dependence forecasts are left unchanged. This mainly perturbs the institution-side SRM forecasts (MES/CoVaR/ΔCoVaR) and does not alter the market VaR forecast that defines the market distress indicator.
- **Alternative A2 (market volatility underestimated):** The forecasting model understates the market-side conditional variance. We scale down the market variance forecasts by $(1 - \tau)$, which directly affects the market VaR forecast and therefore the market distress indicator entering the joint backtests (and the SRM definitions that condition on market distress).
- **Alternative A3 (dependence underestimated):** The forecasting model understates dependence between institution and market returns. Specifically, the correlation forecast is shrunk toward zero by a factor $(1-\tau)$ (pointwise in t under the conditional DGP). This captures misspecification in systemic linkages.
- **Alternative B4 (conditional DGP only; dynamic DGP with static forecasts):** The data exhibit time-varying conditional variances and dependence, but forecasts are generated using static values fixed at in-sample averages (computed over the estimation window of length M) of the estimated conditional variances and correlation. Misspecification here stems from ignoring dynamics rather than from incorrect average levels.

5.2 Power Results

Tables 2 and 3 present the size-corrected power results of our joint backtests under various misspecification scenarios described previously.

Methodological note on size correction: Empirical size under H_0 using standard χ^2 critical values at the 5% nominal level is reported in Table 1. For power calculations under H_1, we apply size correction separately for each test (VaR–MES, VaR–CoVaR, and ΔCoVaR), each version (standard vs. robust), and each backtest type (unconditional vs. conditional). Specifically, we obtain empirical critical values as the $(1 - 0.05)$ quantiles of the simulated H_0 distributions of the corresponding test statistics, and we then compute power under H_1 by comparing the H_1 statistics to these empirical critical values. This ensures that all reported H_1 rejection frequencies are size-corrected for finite-sample distortions.

5.3 Discussion of Power Results

Tables 2 and 3 summarize size-corrected power for $M = N = 250$ at the 5% nominal level. In these settings, power increases monotonically with the severity parameter τ in A1–A3, and is highest when the market-side risk driving the distress event is misspecified.

A2 (market volatility underestimated) delivers the strongest power. Across both DGPs and all three SRM settings, power is already substantial at $\tau = 25\%$ (roughly 0.36–0.48) and becomes very high at $\tau = 50\%$ (about 0.95–0.99), reaching (near) one at $\tau = 75\%$. This pattern is consistent with A2 directly distorting the market VaR forecast and hence the market distress indicator entering the joint backtests.

A1 (institution volatility underestimated) yields moderate power. Since the market VaR (and thus the distress event) remains correctly specified, misspecification mainly affects SRM forecasts through the institution-side scale component. Power is therefore lower than in A2, especially under the conditional DGP. For example, for the (VaR, MES) test, robust power reaches about 0.60 under the marginal DGP at $\tau = 75\%$, and about 0.41–0.42 under the conditional DGP.

A3 (dependence underestimated) is detectable, with strongest power for (VaR, MES). Power increases sharply with τ and is highest for the (VaR, MES) joint test. At $\tau = 75\%$, robust power is about 0.89 under the marginal DGP and about 0.71

under the conditional DGP, whereas the CoVaR- and ΔCoVaR-based joint tests are generally less powerful in this scenario.

B4 (dynamic data with static forecasts) produces moderate but nontrivial power. Under the conditional DGP, static variance and correlation forecasts (fixed at in-sample averages) lead to rejection frequencies around 0.20–0.31 for (VaR, MES) and (VaR, CoVaR), while ΔCoVaR achieves the highest power (about 0.36–0.38).

Finally, differences between conditional and unconditional versions are typically small in our tables (often within a few percentage points). Robust variants often deliver comparable or higher size-corrected power for (VaR, MES), consistent with their improved finite-sample calibration under H_0 (see Table 1).

6 Conclusion

This paper develops a comprehensive backtesting framework for systemic risk measures that addresses fundamental challenges in validating forecasts of the Marginal Expected Shortfall, Conditional Value-at-Risk, and related systemic risk indicators.

Our first major contribution is methodological: we are among the first papers to exploit orthogonality conditions for backtesting systemic risk measures, but our approach fundamentally differs from the pioneering work of Banulescu-Radu, Hurlin, Leymarie, and Scaillet (2021). For CoVaR backtesting, while both approaches employ a closely related orthogonality condition based on joint violations of the institution return relative to its CoVaR forecast and the market return relative to its VaR forecast, our framework incorporates a crucial second orthogonality condition that yields a strictly identifying moment system for detection. This additional condition is not merely a technical refinement but addresses a fundamental identification issue: without it, the test may fail to detect certain forms of model misspecification. For MES backtesting, the contrast between our approach and Banulescu-Radu, Hurlin, Leymarie, and Scaillet (2021) is even more striking. Their test is based on integrating CoVaR violations across risk levels, providing an indirect path to MES validation through the relationship between MES and CoVaR. In contrast, our approach directly exploits identification functions that are explicit to the MES itself, derived from the strict joint identifiability of the (VaR, MES) pair. This fundamental difference means that the two routes to backtesting MES (and by extension SES and SRISK) are fundamentally different in their construction. Our direct approach provides clearer statistical interpretability and more transparent connections between test rejections and specific aspects of MES misspecification.

Our second major contribution is practical: we provide forecast-based backtests for systemic risk measures with valid inference under parameter estimation uncertainty. This represents a significant advance for practitioners and regulators. Traditional approaches, including Banulescu-Radu, Hurlin, Leymarie, and Scaillet (2021), typically require explicit knowledge of the parametric model used to generate forecasts—specifically, they need the cumulative distribution function of the joint distribution to be available in analytical or computable form. This requirement substantially limits the applicability of such tests, excluding semiparametric and nonparametric forecasting approaches, and making implementation highly model-specific. Our backtests, by contrast, operate directly on the systemic risk forecasts themselves and do not require access to the full predictive distribution. The test statistic can be computed given only the sequence of market VaR forecasts and MES or CoVaR forecasts and the realized returns, regardless of whether these forecasts come from a GARCH model, a factor model, a machine learning approach, or any other methodology. Robust inference under parameter estimation uncertainty additionally requires estimating the associated covariance correction terms under the chosen estimation scheme, but this can be done without re-implementing the full forecasting model. This forecast-based implementation has two critical advantages. First, the moment conditions and test statistics are the same across different forecasting methods, which simplifies horse races and model selection; only the estimation-risk covariance correction depends on estimator-specific inputs. Second, the forecast-based nature dramatically expands the applicability of backtesting to modern forecasting environments where semiparametric, nonparametric, or machine learning methods may not provide closed-form distributional assumptions.

To our knowledge, we are among the first to provide such forecast-based backtests with estimation-risk-robust inference specifically for systemic risk measures. This paradigm follows the forecast-based approach established for Expected Shortfall by Barendse, Kole, and van Dijk (2023) and Bayer and Dimitriadis (2022), and parallels traditional VaR backtesting which relies on exceedance-based calibration rather than full distributional assumptions.

Moreover, our paper makes several additional advances. We provide rigorous treatment of estimation risk in the backtesting context, deriving explicit asymptotic corrections that account for parameter uncertainty. Our Monte Carlo evidence demonstrates that these corrections are not merely theoretical niceties but practically essential: uncorrected tests exhibit size distortions reaching roughly 250–300% of the nominal level when the evaluation period is large relative to the estimation period. We document that robust tests maintain accurate size across all sample configurations examined while preserving competitive size-corrected power against relevant misspecification alternatives.

For regulatory practice, our findings carry clear implications. The forecast-based nature of our backtests makes them readily implementable in existing regulatory workflows without requiring access to institutions' full predictive distributions. Supervisors can apply our tests to the systemic risk forecasts computed on banks, insurers, or alternative financial services providers regardless of the underlying methodology. The substantial size distortions we document when ignoring estimation risk underscore the importance of using robust test versions in regulatory applications to avoid erroneous rejections that would impose unnecessary compliance burdens. Our decomposition of ΔCoVaR into stressed and median components provides regulators with diagnostic information about which aspects of model specification require improvement.

References

Acharya, V., R. Engle, and M. Richardson (2012): "Capital shortfall: A new approach to ranking and regulating systemic risks," *American Economic Review*, 102(3), 59–64. 7, 8

Acharya, V. V., L. H. Pedersen, T. Philippon, and M. Richardson (2017): "Measuring systemic risk," *The review of financial studies*, 30(1), 2–47. 2, 6, 7

Adrian, T., and M. K. Brunnermeier (2016): "CoVaR," *The American Economic Review*, 106(7), 1705. 2, 7

Banulescu-Radu, D., C. Hurlin, J. Leymarie, and O. Scaillet (2021): "Backtesting marginal expected shortfall and related systemic risk measures," *Management science*, 67(9), 5730–5754. 3, 4, 7, 9, 10, 11, 16, 18, 23, 24

Barendse, S., E. Kole, and D. van Dijk (2023): "Backtesting value-at-risk and expected shortfall in the presence of estimation error," *Journal of Financial Econometrics*, 21(2), 528–568. 3, 4, 13, 25

Bayer, S., and T. Dimitriadis (2022): "Regression-based expected shortfall backtesting," *Journal of Financial Econometrics*, 20(3), 437–471. 25

Brownlees, C., and R. F. Engle (2017): "SRISK: A conditional capital shortfall measure of systemic risk," *The Review of Financial Studies*, 30(1), 48–79. 6, 7, 8

Christoffersen, P. F. (1998): "Evaluating interval forecasts," *International economic review*, 39(4), 841–862. 3, 12, 16

Du, Z., and J. C. Escanciano (2017): "Backtesting expected shortfall: accounting for tail risk," *Management Science*, 63(4), 940–958. 3

Engle, R. F., and S. Manganelli (2004): "CAViaR: Conditional autoregressive value at risk by regression quantiles," *Journal of business & economic statistics*, 22(4), 367–381. 13

Fissler, T., and Y. Hoga (2024): "Backtesting systemic risk forecasts using multi-objective elicitability," *Journal of Business & Economic Statistics*, 42(2), 485–498. 9, 10, 11, 12

Fissler, T., and J. F. Ziegel (2016): "Higher order elicitability and Osband's principle," *The Annals of Statistics*, 44(4), 1680 – 1707. 3

Fissler, T., J. F. Ziegel, and T. Gneiting (2016): "Expected Shortfall is jointly elicitable with value-at-risk: Implications for backtesting," *Journal of Risk*, pp. 58–61. 3

Francq, C., and J.-M. Zakoïan (2025): "Inference on dynamic systemic risk measures," *Journal of Econometrics*, 247, 105936. 7

Girardi, G., and A. T. Ergün (2013): "Systemic risk measurement: Multivariate GARCH estimation of CoVaR," *Journal of Banking & Finance*, 37(8), 3169–3180. 7

Kupiec, P. (1995): "Techniques for Verifying the Accuracy of Risk Measurement Models," *Journal of Derivatives*, 3(2), 73–84. 13

Newey, W. K., and D. McFadden (1994): "Large sample estimation and hypothesis testing," *Handbook of econometrics*, 4, 2111–2245. 9

Osband, K. H. (1985): *Providing Incentives for Better Cost Forecasting (Prediction, Uncertainty Elicitation)*. University of California, Berkeley. 9

Scaillet, O. (2004): "Nonparametric estimation and sensitivity analysis of expected shortfall," *Mathematical Finance: An International Journal of Mathematics, Statistics and Financial Economics*, 14(1), 115–129. 6

Table 1: **Empirical Size of Joint Backtests at 5% Nominal Level**

Notes: This table reports empirical rejection frequencies at the 5% nominal significance level for various joint backtests based on 1,000 Monte Carlo replications. The data generating process (DGP) is either marginal (i.i.d. returns) or conditional (dependent returns with dynamics). Tests include unconditional versions with $\boldsymbol{H}_t \equiv \boldsymbol{I}_{d_j}$ and conditional versions with $\boldsymbol{H}_t \equiv z_t \boldsymbol{I}_{d_j}$, where $x_t := y_{m,t-1}$, $z_t = 1 + \tanh(|x_t|/\hat{s}_M)$, and $\hat{s}_M = \mathrm{sd}(y_{m,1}, \ldots, y_{m,M})$ is computed from the in-sample period. M denotes the in-sample estimation period, and N denotes the out-of-sample backtesting period (so $T = M + N$). Standard tests ignore estimation risk, while robust tests account for parameter uncertainty through the corrected asymptotic covariance matrix. Results are based on market VaR level $\alpha = 0.05$, CoVaR quantile level $\beta = 0.5$, median-band parameters $(\alpha_{\text{inf}}, \alpha_{\text{sup}}) = (0.25, 0.75)$, and nominal test level $\eta = 0.05$.

	(VaR, MES)		(VaR, CoVaR)		ΔCoVaR	
Configuration	Standard	Robust	Standard	Robust	Standard	Robust
Panel A: Marginal DGP, Unconditional Test						
$M = 250, N = 250$	0.076	0.033	0.079	0.047	0.086	0.049
$M = 250, N = 500$	0.130	0.048	0.118	0.048	0.122	0.054
$M = 500, N = 250$	0.059	0.043	0.062	0.056	0.056	0.056
$M = 500, N = 500$	0.084	0.046	0.093	0.059	0.097	0.065
$M = 500, N = 1000$	0.140	0.055	0.132	0.054	0.131	0.066
$M = 1000, N = 500$	0.067	0.041	0.068	0.034	0.060	0.037
Panel B: Conditional DGP, Unconditional Test						
$M = 250, N = 250$	0.088	0.048	0.100	0.051	0.097	0.055
$M = 250, N = 500$	0.165	0.078	0.159	0.080	0.157	0.061
$M = 500, N = 250$	0.069	0.038	0.067	0.035	0.076	0.040
$M = 500, N = 500$	0.102	0.044	0.097	0.041	0.100	0.055
$M = 500, N = 1000$	0.134	0.071	0.126	0.064	0.135	0.061
$M = 1000, N = 500$	0.067	0.056	0.066	0.048	0.075	0.050
Panel C: Marginal DGP, Conditional Test						
$M = 250, N = 250$	0.074	0.031	0.070	0.039	0.076	0.051
$M = 250, N = 500$	0.127	0.047	0.115	0.047	0.113	0.049
$M = 500, N = 250$	0.058	0.047	0.059	0.046	0.066	0.052
$M = 500, N = 500$	0.085	0.044	0.089	0.052	0.092	0.062
$M = 500, N = 1000$	0.138	0.054	0.122	0.056	0.134	0.066
$M = 1000, N = 500$	0.068	0.039	0.069	0.039	0.063	0.044
Panel D: Conditional DGP, Conditional Test						
$M = 250, N = 250$	0.079	0.043	0.088	0.050	0.087	0.050
$M = 250, N = 500$	0.149	0.074	0.141	0.067	0.144	0.068
$M = 500, N = 250$	0.062	0.038	0.056	0.037	0.073	0.043
$M = 500, N = 500$	0.094	0.048	0.093	0.038	0.101	0.049
$M = 500, N = 1000$	0.122	0.069	0.119	0.058	0.135	0.056
$M = 1000, N = 500$	0.070	0.048	0.063	0.050	0.072	0.049

Table 2: **Size-Corrected Power: Marginal DGP ($M = 250$, $N = 250$)**

Notes: This table reports size-corrected power at the 5% nominal significance level based on 1,000 Monte Carlo replications under the marginal DGP. All entries report size-corrected power computed using empirical critical values obtained from the simulated H_0 distribution of the corresponding test statistic. Misspecification scenarios: institution variance underestimated (A1), market variance underestimated (A2), and correlation shrunk toward zero (A3), with severity levels $\tau \in \{25\%, 50\%, 75\%\}$. Standard and robust refer to tests ignoring or accounting for estimation risk, respectively. Unc. and Cond. refer to unconditional and conditional test versions.

		(VaR, MES)				**(VaR, CoVaR)**				**ΔCoVaR**			
Scenario	τ	**Std.**	**Rob.**	**Std.**	**Rob.**	**Std.**	**Rob.**	**Std.**	**Rob.**	**Std.**	**Rob.**	**Std.**	**Rob.**
		Unc.	**Unc.**	**Cond.**	**Cond.**	**Unc.**	**Unc.**	**Cond.**	**Cond.**	**Unc.**	**Unc.**	**Cond.**	**Cond.**
Power: A1													
A1	25%	0.086	0.098	0.091	0.090	0.065	0.064	0.075	0.080	0.084	0.075	0.078	0.073
A1	50%	0.209	0.261	0.212	0.252	0.167	0.169	0.166	0.181	0.160	0.151	0.163	0.155
A1	75%	0.491	0.598	0.486	0.580	0.377	0.382	0.376	0.405	0.320	0.323	0.310	0.315
Power: A2													
A2	25%	0.423	0.435	0.404	0.419	0.422	0.403	0.409	0.407	0.479	0.401	0.461	0.397
A2	50%	0.968	0.972	0.967	0.970	0.964	0.963	0.964	0.964	0.989	0.981	0.987	0.976
A2	75%	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Power: A3													
A3	25%	0.146	0.171	0.142	0.160	0.109	0.107	0.106	0.113	0.116	0.107	0.109	0.101
A3	50%	0.438	0.556	0.435	0.538	0.325	0.329	0.331	0.354	0.285	0.283	0.277	0.272
A3	75%	0.799	0.889	0.797	0.880	0.582	0.584	0.563	0.600	0.503	0.501	0.485	0.488

Table 3: **Size-Corrected Power: Conditional DGP ($M = 250$, $N = 250$)**

Notes: This table reports size-corrected power at the 5% nominal significance level based on 1,000 Monte Carlo replications under the conditional DGP with dynamic conditional variances and dependence. All entries report size-corrected power computed using empirical critical values obtained from the simulated H_0 distribution of the corresponding test statistic. Misspecification scenarios: institution variance underestimated (A1), market variance underestimated (A2), and correlation shrunk toward zero (A3), with severity levels $\tau \in \{25\%, 50\%, 75\%\}$, and a static-forecast scenario (B4) in which variance and correlation forecasts are fixed at in-sample averages. Scenario B4 has no severity parameter. Standard and robust refer to tests ignoring or accounting for estimation risk, respectively. Unc. and Cond. refer to unconditional and conditional test versions.

		(VaR, MES)				**(VaR, CoVaR)**				**ΔCoVaR**			
Scenario	τ	**Std.**	**Rob.**	**Std.**	**Rob.**	**Std.**	**Rob.**	**Std.**	**Rob.**	**Std.**	**Rob.**	**Std.**	**Rob.**
		Unc.	**Unc.**	**Cond.**	**Cond.**	**Unc.**	**Unc.**	**Cond.**	**Cond.**	**Unc.**	**Unc.**	**Cond.**	**Cond.**
Power: A1													
A1	25%	0.044	0.061	0.049	0.066	0.043	0.061	0.051	0.061	0.050	0.056	0.044	0.061
A1	50%	0.099	0.175	0.110	0.168	0.109	0.142	0.119	0.145	0.090	0.105	0.088	0.114
A1	75%	0.232	0.409	0.267	0.420	0.253	0.334	0.281	0.340	0.213	0.253	0.205	0.262
Power: A2													
A2	25%	0.382	0.359	0.383	0.356	0.378	0.386	0.389	0.366	0.440	0.378	0.418	0.384
A2	50%	0.955	0.959	0.956	0.957	0.950	0.961	0.951	0.954	0.980	0.975	0.979	0.976
A2	75%	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Power: A3													
A3	25%	0.063	0.103	0.076	0.122	0.075	0.100	0.093	0.105	0.079	0.085	0.078	0.098
A3	50%	0.231	0.383	0.255	0.399	0.232	0.284	0.250	0.287	0.213	0.221	0.206	0.239
A3	75%	0.448	0.716	0.492	0.711	0.411	0.497	0.428	0.494	0.377	0.391	0.355	0.402
Power: B4													
B4	—	0.240	0.272	0.282	0.312	0.196	0.243	0.243	0.283	0.368	0.361	0.377	0.376

Appendix A

A.1 Assumptions

1 Covariance stationarity. The processes $\{\boldsymbol{g}_t^j(\theta_0)\}_{t\in\mathbb{Z}}$ and $\{\ell_t(\theta_0)\}_{t\in\mathbb{Z}}$ (and any $\mathcal{F}_{t-1}$-measurable instruments entering the tests) are covariance-stationary with $\mathbb{E}[\|\boldsymbol{g}_t^j(\theta_0)\|^2] < \infty$ and $\mathbb{E}[\|\ell_t(\theta_0)\|^2] < \infty$.

2 Strong mixing. The data and model-based processes underlying the tests (in particular $\boldsymbol{g}_t^j(\theta_0)$ and $\ell_t(\theta_0)$) are α-mixing with coefficients $\alpha(k) = O(k^{-a})$ for some $a > 2$. All $\mathcal{F}_t$-measurable transforms used in the paper inherit α-mixing, ensuring LLN/CLT for our statistics.

3 Finite moments. There exist $\delta > 0$ and $r > 1$ such that $\mathbb{E}[\|\boldsymbol{g}_t^j(\theta_0)\|^{2+\delta}] < \infty$ and $\mathbb{E}[|y_{i,t}|^{2r} + |y_{m,t}|^{2r}] < \infty$. Moreover, $\mathbb{E}[\|\ell_t(\theta_0)\|^{2+\delta}] < \infty$.

4 Model regularity and predictability. (i) The conditional CDF admits a bounded, Lipschitz-continuous density in neighborhoods of the relevant quantiles (e.g., levels α, β and the band $[\alpha_{\text{inf}}, \alpha_{\text{sup}}]$); in particular, conditional densities at all decision thresholds exist and are continuous at those points. (ii) For every t, the forecasts entering the identification vectors (e.g., $v_{m,t}(\theta_0)$, $\mu_t(\theta_0)$, $c_{i,t}(\theta_0)$ and their ΔCoVaR analogues) are $\mathcal{F}_{t-1}$-measurable. Under correct specification this implies $\mathbb{E}[\boldsymbol{g}_t^j(\theta_0) \mid \mathcal{F}_{t-1}] = \mathbf{0}$ (MDS property).

5 Estimation scheme and asymptotic regime. Let the in-sample length be $M \to \infty$, the out-of-sample length $N \to \infty$, with $T = M + N$, and define $p := N/M \to p_0 \in [0, \infty)$. Parameters are estimated by one of the following schemes: fixed (single in-sample estimate $\hat{\theta}_M$ reused for all $t > M$), rolling (window size M re-estimated each t using $\{t - M, \dots, t - 1\}$), or expanding (window $\{1, \dots, t - 1\}$). The special case $p_0 = 0$ corresponds to vanishing estimation risk.

6 Asymptotic linearity of the estimator. Let $d_\theta := \dim(\theta)$. There exist a nonrandom full-rank matrix $B \in \mathbb{R}^{d_\theta \times d_\theta}$ and a zero-mean, square-integrable influence process $\{\ell_t(\theta_0)\}_{t\in\mathbb{Z}} \subset \mathbb{R}^{d_\theta}$ such that:

- fixed: $\hat{\theta}_M - \theta_0 = B \frac{1}{M} \sum_{s=1}^{M} \ell_s(\theta_0) + o_p(M^{-1/2})$;
- rolling: $\hat{\theta}_t - \theta_0 = B \frac{1}{M} \sum_{s=t-M}^{t-1} \ell_s(\theta_0) + o_p(M^{-1/2})$ uniformly in $t \in \{M+1, \dots, T\}$;
- expanding: $\hat{\theta}_t - \theta_0 = B \frac{1}{t-1} \sum_{s=1}^{t-1} \ell_s(\theta_0) + o_p((t-1)^{-1/2})$ uniformly except finitely many initial t.

Assume $\mathbb{E}[\ell_t(\theta_0)] = 0$ and $\{\ell_t(\theta_0)\}$ is α-mixing. The longrun covariance $V := \sum_{k\in\mathbb{Z}} \mathbb{E}[\ell_t(\theta_0)\ell_{t-k}(\theta_0)']$ exists and is positive definite.

7 Differentiability in expectation. For each $j \in \{\mu, c, \Delta\}$, let $\boldsymbol{m}_t^j(\boldsymbol{\theta}) := \boldsymbol{H}_t(\boldsymbol{\theta})\, \boldsymbol{g}_t^j(\boldsymbol{\theta})$. The map $\boldsymbol{\theta} \mapsto \mathbb{E}[\boldsymbol{m}_t^j(\boldsymbol{\theta})]$ is differentiable at $\boldsymbol{\theta}_0$ with sensitivity matrix $\boldsymbol{A}_j := \partial_{\boldsymbol{\theta}} \mathbb{E}[\boldsymbol{m}_t^j(\boldsymbol{\theta})]\Big|_{\boldsymbol{\theta}_0}$ finite and time-invariant. Moreover, the linearization $\boldsymbol{m}_t^j(\hat{\boldsymbol{\theta}}_t) = \boldsymbol{m}_t^j(\boldsymbol{\theta}_0) + \boldsymbol{A}_j(\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_0) + \boldsymbol{r}_{t,N}$ holds with $N^{-1/2} \sum_{t=M+1}^{T} \boldsymbol{r}_{t,N} = o_p(1)$.

8 Longrun covariance matrices and nonsingularity. Let $\boldsymbol{m}_t^j(\boldsymbol{\theta}) := \boldsymbol{H}_t(\boldsymbol{\theta})\, \boldsymbol{g}_t^j(\boldsymbol{\theta})$ and define $S_{gg} := \sum_{k\in\mathbb{Z}} \mathbb{E}[\boldsymbol{m}_t^j(\boldsymbol{\theta}_0)\boldsymbol{m}_{t-k}^j(\boldsymbol{\theta}_0)']$, $S_{gh} := \sum_{k\in\mathbb{Z}} \mathbb{E}[\boldsymbol{m}_t^j(\boldsymbol{\theta}_0)\ell_{t-k}(\boldsymbol{\theta}_0)']$, and $S_{hh} := \sum_{k\in\mathbb{Z}} \mathbb{E}[\ell_t(\boldsymbol{\theta}_0)\ell_{t-k}(\boldsymbol{\theta}_0)'](= V)$. These sums exist (absolute summability) and are finite; V is positive definite; and the relevant submatrices of S_{gg} are nonsingular so that Wald statistics are well-defined. Under correct specification and for $\mathcal{F}_{t-1}$-measurable instruments, $\boldsymbol{g}_t^j(\boldsymbol{\theta}_0)$ is an MDS and hence $\boldsymbol{m}_t^j(\boldsymbol{\theta}_0)$ is also an MDS, so that $S_{gg} = \mathrm{Var}(\boldsymbol{m}_t^j(\boldsymbol{\theta}_0))$. (In particular, for unconditional tests with $\boldsymbol{H}_t \equiv I$, $S_{gg} = \mathrm{Var}(\boldsymbol{g}_t^j(\boldsymbol{\theta}_0))$.) Throughout, assume $0 < \alpha, \beta, \gamma < 1$.

A.2 Closed-form $\boldsymbol{\Sigma}_j$ under unconditional tests

Proposition 1 (Closed-form long-run variance under unconditional tests)**.** *Under H_0 and for unconditional tests with $\boldsymbol{H}_t \equiv \boldsymbol{I}_{d_j}$, the long-run variance reduces to $\boldsymbol{\Sigma}_j = \mathrm{Var}(\boldsymbol{g}_t^j(\boldsymbol{\theta}_0))$..*

(i) For the (VaR, MES) backtest ($j = \mu$),

$$\boldsymbol{\Sigma}_\mu = \begin{pmatrix} \alpha(1-\alpha) & 0 \\ 0 & \alpha\, \sigma^2_{i|\text{stress}} \end{pmatrix}, \qquad \sigma^2_{i|\text{stress}} := \mathbb{E}\Big[(y_{i,t} - \mu_{i,t}(\alpha, \boldsymbol{\theta}_0))^2 \,\Big|\, y_{m,t} \le v_{m,t}(\alpha, \boldsymbol{\theta}_0)\Big].$$

(ii) For the (VaR, CoVaR) backtest ($j = c$),

$$\boldsymbol{\Sigma}_c = \begin{pmatrix} \alpha(1-\alpha) & 0 \\ 0 & \alpha\, \beta(1-\beta) \end{pmatrix}.$$

(iii) For the ΔCoVaR backtest ($j = \Delta$), let $\gamma := \alpha_{\sup} - \alpha_{\inf}$ with $\alpha < \alpha_{\inf} < \alpha_{\sup}$. Then

$$\boldsymbol{\Sigma}_\Delta = \begin{pmatrix} \alpha(1-\alpha) & 0 & -\alpha\gamma & 0 \\ 0 & \alpha\beta(1-\beta) & 0 & 0 \\ -\alpha\gamma & 0 & \gamma(1-\gamma) & 0 \\ 0 & 0 & 0 & \gamma\beta(1-\beta) \end{pmatrix}.$$

A.3 Closed-form weight functions for estimation risk

Proposition 2 (Weight functions $\kappa_{hl}(p)$ and $\kappa_{ll}(p)$)**.** *Under Assumptions A.1–A.8, the weight functions $\kappa_{hl}(p)$ and $\kappa_{ll}(p)$ depend only on the estimation window scheme and on $p = N/M > 0$, with closed forms:*

- Fixed window: $\kappa_{hl}(p) = 0, \quad \kappa_{ll}(p) = p.$

- Rolling window:

$$\kappa_{hl}(p) = \begin{cases} p/2, & 0 < p \le 1, \\ 1 - \dfrac{1}{2p}, & p > 1, \end{cases} \qquad \kappa_{ll}(p) = \begin{cases} p - \dfrac{p^2}{3}, & 0 < p \le 1, \\ 1 - \dfrac{1}{3p}, & p > 1. \end{cases}$$

- Expanding window: $\kappa_{hl}(p) = 1 - \dfrac{\ln(1+p)}{p}, \quad \kappa_{ll}(p) = 2\left(1 - \dfrac{\ln(1+p)}{p}\right).$

As $p \to 0$, $\kappa_{hl}(p), \kappa_{ll}(p) \to 0$.

A.4 Plug-in estimation of $\boldsymbol{\Omega}_j$

A.4.1 Estimation of the sensitivity matrix $\boldsymbol{A}_j$. For $j \in \{\mu, c, \Delta\}$ recall that $\boldsymbol{m}_t^j(\boldsymbol{\theta}) = \boldsymbol{H}_t(\boldsymbol{\theta})\, \boldsymbol{g}_t^j(\boldsymbol{\theta})$ and $\boldsymbol{A}_j = \partial_{\boldsymbol{\theta}}\, \mathbb{E}\Big[\boldsymbol{m}_t^j(\boldsymbol{\theta})\Big]\Big|_{\boldsymbol{\theta}_0}$. A feasible estimator is obtained from the sample analogue of the Jacobian:

$$\hat{\boldsymbol{A}}_j = \frac{1}{N} \sum_{t=M+1}^{T} \partial_{\boldsymbol{\theta}} \boldsymbol{m}_t^j(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_t}.$$

When $\boldsymbol{H}_t(\boldsymbol{\theta})$ does not depend on $\boldsymbol{\theta}$, this simplifies to $\partial_{\boldsymbol{\theta}} \boldsymbol{m}_t^j(\boldsymbol{\theta}) = \boldsymbol{H}_t\, \partial_{\boldsymbol{\theta}} \boldsymbol{g}_t^j(\boldsymbol{\theta})$. In the presence of indicator functions (e.g., $\mathbf{1}\{y_{m,t} \le v_{m,t}(\alpha, \boldsymbol{\theta})\}$), $\partial_{\boldsymbol{\theta}} \boldsymbol{g}_t^j(\boldsymbol{\theta})$ can be computed either (i) by replacing indicators with smooth approximations (e.g., logistic smoothing), or (ii) by using plug-in estimates of the relevant conditional densities at the decision thresholds. Both approaches yield consistent estimators under Assumption A.4(i).

Proposition 3 (Plug-in estimator of $\boldsymbol{\Omega}_j$)**.** *Let $\hat{\Sigma}_j$ be a consistent estimator of Σ_j under the MDS null, and define $\hat{\boldsymbol{m}}_t^j := \boldsymbol{H}_t(\hat{\boldsymbol{\theta}}_t)\, \hat{\boldsymbol{g}}_t^j(\hat{\boldsymbol{\theta}}_t)$. Let $\hat{q}_j$ and $\hat{V}$ be HAC-type estimators of q_j and V based on $\{\hat{\boldsymbol{m}}_t^j\}$ and a scheme-consistent influence-function path $\{\hat{\ell}_t\}$, with bandwidths $L_q, L_V \to \infty$ such that $L_q/N \to 0$ and $L_V/M \to 0$. Let $\hat{A}_j \to_p A_j$ (e.g., computed as in Paragraph A.4.1) and $\hat{B} \to_p B$. Then a feasible plug-in estimator of Ω_j is*

$$\hat{\Omega}_j = \hat{\Sigma}_j + \kappa_{hl}(p)\Big(\hat{A}_j \hat{B} \hat{q}_j + \hat{q}_j' \hat{B}' \hat{A}_j'\Big) + \kappa_{ll}(p)\, \hat{A}_j \hat{B} \hat{V} \hat{B}' \hat{A}_j',$$

where $\kappa_{hl}(p)$ and $\kappa_{ll}(p)$ are given in Proposition 2. Under Assumptions A.1–A.8, $\hat{\Omega}_j \to_p \Omega_j$ and, if Ω_j is nonsingular, the Wald statistic $W_N^j = \boldsymbol{S}_N^{j\prime} \hat{\Omega}_j^{-1} \boldsymbol{S}_N^j$ is asymptotically χ^2 under H_0.

A.5 Martingale-difference property

Lemma 1 (Zero conditional mean)**.** *Assume $\mathbb{E}\Big[\|\boldsymbol{g}_t^j(\theta_0)\|^2\Big] < \infty$. If the model is correctly specified, then for every $j \in \{\mu, c, \Delta\}$*

$$\mathbb{E}\Big[\boldsymbol{g}_t^j(\theta_0) \mid \mathcal{F}_{t-1}\Big] = \mathbf{0} \qquad \textit{a.s. for all } t,$$

so $\{\boldsymbol{g}_t^j(\theta_0), \mathcal{F}_t\}_{t\in\mathbb{Z}}$ *is a martingale-difference sequence.*

Proof. We treat the three cases separately.

(i) VaR–MES identification vector $\Big(j = \mu\Big)$**.** Recall $I_t = \mathbf{1}\{y_{m,t} \leq v_{m,t}(\theta_0)\}$ with $\Pr(I_t = 1 \mid \mathcal{F}_{t-1}) = \alpha$, $U_t = \mu_t(\theta_0) - y_{i,t}$ and $\mu_t(\theta_0) = \mathbb{E}[y_{i,t} \mid I_t = 1, \mathcal{F}_{t-1}]$. The two components of $\boldsymbol{g}_t^\mu(\theta_0) = \Big(I_t - \alpha,\ I_t U_t\Big)'$ satisfy

$$\mathbb{E}[I_t - \alpha \mid \mathcal{F}_{t-1}] = \alpha - \alpha = 0,$$

and, using the conditional–expectation identity $\mathbb{E}[X\mathbf{1}_C \mid \mathcal{F}] = \Pr(C \mid \mathcal{F})\mathbb{E}[X \mid C, \mathcal{F}]$,

$$\mathbb{E}[I_t U_t \mid \mathcal{F}_{t-1}] = \mu_t(\theta_0)\alpha - \alpha\,\mathbb{E}[y_{i,t} \mid I_t = 1, \mathcal{F}_{t-1}] = 0.$$

Hence $\mathbb{E}[\boldsymbol{g}_t^\mu(\theta_0) \mid \mathcal{F}_{t-1}] = \mathbf{0}$.

(ii) VaR–CoVaR identification vector $\Big(j = c\Big)$**.** Let the second indicator be $J_t = \mathbf{1}\{y_{i,t} \leq c_{i,t}(\beta, \alpha, \theta_0)\}$ and define $\beta = \Pr(J_t = 1 \mid I_t = 1, \mathcal{F}_{t-1})$. Then $\boldsymbol{g}_t^c(\theta_0) = \Big(I_t - \alpha,\ I_t(J_t - \beta)\Big)'$. The first component is identical to case (i). For the second:

$$\mathbb{E}\Big[I_t(J_t - \beta) \mid \mathcal{F}_{t-1}\Big] = \alpha\,\mathbb{E}\Big[J_t - \beta \mid I_t = 1, \mathcal{F}_{t-1}\Big] = \alpha\,(0) = 0,$$

because $J_t \mid (I_t = 1, \mathcal{F}_{t-1}) \sim \text{Bernoulli}(\beta)$. Therefore $\mathbb{E}[\boldsymbol{g}_t^c(\theta_0) \mid \mathcal{F}_{t-1}] = \mathbf{0}$.

(iii) ΔCoVaR identification vector $\Big(j = \Delta\Big)$**.** Recall the four-dimensional process in Equation (8). Define the stress and median-band indicators

$$I_t^s := \mathbf{1}\{y_{m,t} \leq v_{m,t}(\alpha, \theta_0)\}, \qquad I_t^m := \mathbf{1}\{v_{m,t}(\alpha_{\inf}, \theta_0) \leq y_{m,t} \leq v_{m,t}(\alpha_{\sup}, \theta_0)\},$$

and the institution indicators

$$J_t^s := \mathbf{1}\{y_{i,t} \leq c_{i,t}(\beta, \alpha, \theta_0)\}, \qquad J_t^m := \mathbf{1}\{y_{i,t} \leq c_{i,t}(\beta, \alpha_{\inf}, \alpha_{\sup}, \theta_0)\}.$$

Under correct specification, $\Pr(I_t^s = 1 \mid \mathcal{F}_{t-1}) = \alpha$ and $\Pr(I_t^m = 1 \mid \mathcal{F}_{t-1}) = \gamma := \alpha_{\sup} - \alpha_{\inf}$, while

$$\Pr(J_t^s = 1 \mid I_t^s = 1, \mathcal{F}_{t-1}) = \beta, \qquad \Pr(J_t^m = 1 \mid I_t^m = 1, \mathcal{F}_{t-1}) = \beta.$$

Hence, componentwise,

$$\mathbb{E}[I_t^s - \alpha \mid \mathcal{F}_{t-1}] = 0, \qquad \mathbb{E}[I_t^s(J_t^s - \beta) \mid \mathcal{F}_{t-1}] = \alpha\,\mathbb{E}[J_t^s - \beta \mid I_t^s = 1, \mathcal{F}_{t-1}] = 0,$$

$$\mathbb{E}[I_t^m - \gamma \mid \mathcal{F}_{t-1}] = 0, \qquad \mathbb{E}[I_t^m(J_t^m - \beta) \mid \mathcal{F}_{t-1}] = \gamma\,\mathbb{E}[J_t^m - \beta \mid I_t^m = 1, \mathcal{F}_{t-1}] = 0.$$

Therefore $\mathbb{E}[\boldsymbol{g}_t^\Delta(\theta_0) \mid \mathcal{F}_{t-1}] = \mathbf{0}$. □

Appendix B

B.1 Data generating processes

In each replication we generate a bivariate return vector $\boldsymbol{y}_t = (y_{m,t}, y_{i,t})'$ for $t = 1, \ldots, T$ with $T = M + N$, where $y_{m,t}$ denotes the market return and $y_{i,t}$ the institution return. Unless stated otherwise, innovations are Gaussian.

Marginal DGP (i.i.d. bivariate returns). Returns are i.i.d. bivariate normal,

$$\boldsymbol{y}_t \overset{i.i.d.}{\sim} \mathcal{N}(\mathbf{0}, \Sigma), \qquad \Sigma = \begin{pmatrix} \sigma_m^2 & \rho\, \sigma_i \sigma_m \\ \rho\, \sigma_i \sigma_m & \sigma_i^2 \end{pmatrix}, \tag{B1}$$

with $(\sigma_i^2, \sigma_m^2, \rho) = (3.506,\ 0.722,\ 0.663)$.

Conditional DGP (non-i.i.d. bivariate returns based on GJR–DCC GARCH). Let $\boldsymbol{z}_t \mid \mathcal{F}_{t-1} \sim \mathcal{N}(\mathbf{0}, R_t)$ and define

$$\boldsymbol{y}_t = D_t\, \boldsymbol{z}_t, \qquad D_t = \operatorname{diag}\!\Big(\sqrt{h_{m,t}}, \sqrt{h_{i,t}}\Big), \qquad \operatorname{Cov}(\boldsymbol{y}_t \mid \mathcal{F}_{t-1}) = D_t R_t D_t. \tag{B2}$$

The conditional variances follow GJR–GARCH(1,1) recursions:

$$h_{i,t+1} = a_{10} + a_{11} y_{i,t}^2 + a_{12} y_{i,t}^2 \mathbf{1}\{y_{i,t} < 0\} + a_{13} h_{i,t}, \tag{B3}$$

$$h_{m,t+1} = a_{20} + a_{21} y_{m,t}^2 + a_{22} y_{m,t}^2 \mathbf{1}\{y_{m,t} < 0\} + a_{23} h_{m,t}, \tag{B4}$$

and the correlation matrix R_t is generated by a DCC(1,1) recursion. Let $\boldsymbol{\varepsilon}_t = D_t^{-1} \boldsymbol{y}_t = (\varepsilon_{m,t}, \varepsilon_{i,t})'$. Then

$$Q_{t+1} = (1 - a - b)\, \bar{Q} + a\, \boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}_t' + b\, Q_t, \tag{B5}$$

$$R_t = \operatorname{diag}(Q_t)^{-1/2} Q_t \operatorname{diag}(Q_t)^{-1/2}, \tag{B6}$$

with $\bar{Q} = \begin{pmatrix} 1 & \rho_0 \\ \rho_0 & 1 \end{pmatrix}$. We set $(a_{10}, a_{11}, a_{12}, a_{13}) = (0.149, 0.065, 0.075, 0.855)$, $(a_{20}, a_{21}, a_{22}, a_{23}) = (0.057, 0.001, 0.268, 0.786)$, $(a, b, \rho_0) = (0.041, 0.754, 0.663)$. Initialization uses the corresponding unconditional variances and $Q_1 = \bar{Q}$. A burn-in of 2000 observations is discarded prior to keeping the final $T = M + N$ observations.

B.2 Risk-measure forecasts under the DGPs

Let $\alpha \in (0,1)$ denote the market tail probability, $\beta \in (0,1)$ the conditional tail level, and $z_\alpha = \Phi^{-1}(\alpha)$. For ΔCoVaR we use a "median band" event $v_{m,t}(\alpha_{\inf}) \le y_{m,t} \le v_{m,t}(\alpha_{\sup})$ with $\gamma := \alpha_{\sup} - \alpha_{\inf}$.

VaR and MES. Under both DGPs, $y_{m,t} \mid \mathcal{F}_{t-1} \sim \mathcal{N}(0, h_{m,t})$ (with $h_{m,t} \equiv \sigma_m^2$ under the marginal DGP), so

$$v_{m,t}(\alpha) = \sqrt{h_{m,t}}\, z_\alpha. \tag{B7}$$

With conditional Gaussian dependence and correlation ρ_t ($\rho_t \equiv \rho$ under the marginal DGP), the MES forecast is

$$\mu_t(\alpha) := \mathbb{E}[y_{i,t} \mid y_{m,t} \leq v_{m,t}(\alpha), \mathcal{F}_{t-1}] = -\,\rho_t \sqrt{h_{i,t}}\, \frac{\phi(z_\alpha)}{\alpha}. \tag{B8}$$

CoVaR and band-CoVaR. Let $Z_{i,t} = y_{i,t}/\sqrt{h_{i,t}}$ and $Z_{m,t} = y_{m,t}/\sqrt{h_{m,t}}$. Conditional on $\mathcal{F}_{t-1}$, $(Z_{i,t}, Z_{m,t})$ is standard bivariate normal with correlation ρ_t and CDF $\Phi_2(\cdot, \cdot; \rho_t)$. The stressed CoVaR forecast $c_{s,t}$ is defined by

$$\Pr\Big(y_{i,t} \leq c_{s,t} \;\Big|\; y_{m,t} \leq v_{m,t}(\alpha), \mathcal{F}_{t-1}\Big) = \beta, \tag{B9}$$

equivalently $c_{s,t} = \sqrt{h_{i,t}}\, z_{c,s}(\rho_t)$ where $z_{c,s}(\rho)$ solves

$$\frac{\Phi_2\Big(z_{c,s}(\rho),\ z_\alpha;\ \rho\Big)}{\Phi(z_\alpha)} = \beta. \tag{B10}$$

For the median-band CoVaR, define $z_{\text{inf}} = \Phi^{-1}(\alpha_{\text{inf}})$ and $z_{\text{sup}} = \Phi^{-1}(\alpha_{\text{sup}})$. Then $c_{m,t} = \sqrt{h_{i,t}}\, z_{c,m}(\rho_t)$ where $z_{c,m}(\rho)$ solves

$$\frac{\Phi_2\Big(z_{c,m}(\rho),\ z_{\text{sup}};\ \rho\Big) - \Phi_2\Big(z_{c,m}(\rho),\ z_{\text{inf}};\ \rho\Big)}{\Phi(z_{\text{sup}}) - \Phi(z_{\text{inf}})} = \beta. \tag{B11}$$

In implementations, $z_{c,s}(\rho)$ and $z_{c,m}(\rho)$ are obtained by one-dimensional root finding (or, for conditional DGP, by interpolation from a precomputed ρ-grid).