Multilevel and Tail Risk Management*

Lynda Khalaf†, Arturo Leccadito‡, Giovanni Urga§

August 31, 2020

Abstract

We introduce backtesting methods to assess Value-at-Risk (VaR) and Expected Shortfall (ES) that require no more than desktop VaR violations as inputs. Maintaining an integrated VaR perspective, our methodology relies on multiple testing to combine evidence on the frequency and dynamic evolution of violations, and to capture more information than a single threshold can provide about the magnitude of violations. Contributions include a formal finite sample analysis of the joint distribution of multi-threshold violations, and limiting results that unify discrete and continuous definitions of cumulative violations across thresholds. Simulation studies demonstrate the power advantages of the proposed tests, particularly with small samples and when underlying models are unavailable to assessors. Results also reinforce the usefulness of CaViaR approaches not just for VaR but also as ES backtests. Empirically, we assess desktop data by Bloomberg on exchange traded funds. We find that tail risk is not adequately reflected via a wide spectrum of models and available measures. Results provide useful prescriptions for empirical practice and, more generally, reinforce the recent arguments in favor of combined tests and forecasts in tail risk management.

Keywords: Value-at-Risk, Expected Shortfall, Backtesting, CaViaR, Exchange-Traded Funds, Multiple Testing
JEL Classification: C12, C15, C58, G17, G32

*We wish to thank the Editor, Fabio Trojani, an Associate Editor and two anonymous referees for very useful comments and suggestions which have helped to develop and improve the paper further. The usual disclaimer applies. Lynda Khalaf acknowledges financial support from the Social Sciences and Humanities Research Council of Canada (SSHRC), the "Azione 2 - Progetto STaRs" at the Department of Management, Economics and Quantitative Methods of University of Bergamo (Italy), and the Centre for Econometric Analysis, Cass Business School, London (UK).

†Economics Department, Carleton University, Loeb Building 1125 Colonel By Drive, Ottawa, Ontario, K1S 5B6 Canada, and Centre interuniversitaire de recherche en économie quantitative (CIREQ). E-mail: Lynda_Khalaf@carleton.ca.

‡Department of Economics, Statistics and Finance, Università della Calabria, Ponte Bucci cubo 3 C, 87030 Rende (CS), Italy. E-mail: arturo.leccadito@unical.it

§Centre for Econometric Analysis, Faculty of Finance, Cass Business School, City University of London, 106 Bunhill Row, EC1Y 8TZ, London (UK) and Department of Management, Economics and Quantitative Methods, University of Bergamo (Italy). E-mail: g.urga@city.ac.uk

1 Introduction

Statistical risk assessment techniques have attracted increased attention, reflecting the evolution of regulatory environments. In this context, diagnostics based on Value-at-Risk (VaR) are in wide use, despite research on arguably more informative alternative measures such as expected shortfall (ES, sometimes also denoted as CVaR); see Perignon and Smith (2010) and Berkowitz et al. (2011) for a unified treatment on VaR backtests.

The original appeal of VaR-based methods was their simplicity. Their long-standing endorsement in the industry can also be tied to input requirements. Conformably, Berkowitz et al. (2011) recommend VaR backtests that do not require model assumptions and depend only on the observed profit and loss data. However, VaR is insensitive to the severity of tail losses, by its definition as a quantile. Recent regulatory accords have thus shifted emphasis to ES. Testing ES is much harder, particularly without assuming an underlying model and when returns or risk measures are not observed (by the assessor).

This paper proposes VaR and ES backtests that account for tail risk and, in line with Berkowitz et al. (2011), require no more than desktop VaR violations as inputs. Such procedures are particularly useful to bridge the gap between academic research and risk management practice. In line with this objective, we illustrate the usefulness of our proposed methods using Exchange-Traded Funds (ETFs) VaR and CVaR data. Despite remarkable growth and widespread acceptance in the market, formal backtests are surprisingly lacking on available ETF risk measures. Recent wild market swings have drawn the industry's attention, amid the blow-up of volatility ETFs. The swelling importance of ETFs is not just challenging traditional managers but starting to profoundly impact financial systems. Regulators are reacting to these challenges as reflected in e.g. Europe's Markets in Financial Instruments Directive (Mifid) II that came into effect in January 2018.

Our methodology relies on model-free induced testing, which involves: (i) combining several statistical procedures, and (ii) a non-parametric treatment of tail risk.[1] Our primary motivation for combined backtests is to exploit the incremental advantage of various statistics on, *e.g.*, the number, dynamic evolution and magnitude of violations. Another more fundamental objective is to reconcile model-free VaR and ES monitoring, building on the definition of ES as an integrated VaR. The crucial point is the appropriate definition of VaR thresholds, in conjunction with simulation-based algorithms that provide a model-free approach to assess ES.

In this context, our contributions include: (i) a formal finite sample analysis of the joint distribution of multi-threshold violations, in contrast to existing methods [Leccadito et al. (2014), Perignon and Smith (2008), Du and Escanciano (2017), Kratz et al. (2018)] that analyze cumulative violations (CVs) which in the discrete case amount to their sum (over thresholds); (ii) limiting results that unify discrete and continuous definitions of CVs; (iii) an induced test approach that allows us to channel the wide set of tools that have been shown to inform on VaRs, towards inference on ES; (iv) an adaptive combinatorial procedure that embeds non-uniform weights across the considered thresholds to capture revealing violations more effectively than available statistical approaches.

Additionally we extend the so called CaViaR methods to assess ES. CaViaR tests rely on a regression of VaR violations, on lags and other possible predictors. Significant predictors refute VaR accuracy at the considered threshold. Further restrictions on the regression intercept resulting from the VaR's frequency implications provide additional unconditional coverage checks. ES analysis in this way is another contribution of our paper.

[1] For recent references on combination-based risk analysis, see Hurlin and Tokpavi (2006), Perignon and Smith (2008), Colletaz et al. (2013), Leccadito et al. (2014), Danielsson and Zhou (2016) and Kratz et al. (2018), on combinations across various VaR thresholds, or Taylor (2020) on combining forecasts across competing models; for insights and perspectives on model-free forecasting environments broadly, refer to Diebold (2015) and companion comments.

Three primary advantages underscore the practical relevance of our approach, when it comes down to making inference with desk level data. *First*, our VaR-based ES tests may be preferred on the basis of robustness and applicability. Indeed, VaR is more commonly used than ES yet is less informative on tail risk because of its insensitivity to the latter, which controversially, makes it a more robust measure to backtest. *Second*, our CaViaR perspective allows risk managers to expand the information set they use beyond past violations, which, in addition to statistical power, may shed light on possible modeling improvements.

The *third* and most crucial advantage is our Monte Carlo [Dufour (2006)] test (MCT) based finite sample implementation. A large number of observations is rarely available for backtesting, so asymptotic irregularities cannot be ruled out; see e.g. Danielsson and Zhou (2016). Despite the important contributions of related research, exact distribution-free methods remain restricted to single-threshold backtests [Christoffersen and Pelletier (2004), Berkowitz et al. (2011)] or to CVs [Leccadito et al. (2014), Kratz et al. (2018)]. In contrast, we show that the exact joint distribution of the multi-threshold violation series is nuisance-parameter free under a correct risk model and can be simulated with no further assumptions. It follows that an exact distribution-free MCT p-value can be derived for *any* statistic that depends on the data only through these violations. The generality of this result and the wide horizon in flexibility that it entails distinguish our contribution from existing distribution-free risk management tools.

The properties of our proposed tests are illustrated via empirically relevant simulation experiments. We document concrete advantages relative to available methods, maintaining our model agnostic approach. Our empirical analysis particularly with Bloomberg ETFs further illustrates the usefulness of a mode-free analysis. While ETF markets have grown dramatically, research assessing their ex-ante VaR and CVaR forecasts was evidently slow to follow suit. Our approach is useful not only because of its simplicity, but because we require no more information than was originally disclosed by Bloomberg. This suggests that informative backtests can be customized into professional desktop tools.

The paper is organized as follows. Section 2 discusses our multi-threshold VaR and ES perspectives. Section 3 presents presents our combined backtesting procedures. In Section 4, we evaluate the performance of the proposed tests relative to several alternative testing procedures via an extensive Monte Carlo analysis. Our empirical analysis is discussed in Section 5. Section 6 concludes.

2 ES: a multi-threshold VaR perspective

VaR violations are time periods where, at a given threshold, the realized loss exceeds the risk estimate arising from the strategy or model under test. Formally, let $\mathcal{F}_t$ denote the information set up to time t. The VaR for threshold α at time $t+1$, given $\mathcal{F}_t$, satisfies $P\left(r_{t+1} \leq -\text{VaR}_{t+1|t}(\alpha)|\mathcal{F}_t\right) = \alpha$, where r_{t+1} is the realized time $t+1$ return. In this paper, we use the term "threshold" to indicate the VaR or ES coverage probability, and "level" to denote the critical rejection probability associated with a backtest. Conformably, violations are defined via the exception indicator series:

$$I_{t+1} = \begin{cases} 1 & \text{if } r_{t+1} \leq -\text{VaR}_{t+1|t}(\alpha) \\ 0 & \text{if } r_{t+1} > -\text{VaR}_{t+1|t}(\alpha) \end{cases}. \tag{1}$$

Two broad approaches are available to assess the resulting process of zeros and ones, where clearly the latter correspond to a violation. The first, which includes the test of Kupiec (1995), is the so-called unconditional coverage analysis based on comparing the observed frequency of violations to its expected counterpart, assuming forecast accuracy or a correct model specification. The second approach includes e.g. the Markov tests of Christoffersen (1998), the duration based test of Christoffersen and Pelletier

(2004) and the above cited CaViaR tests of Engle and Manganelli (2004). When combined with unconditional coverage checks, the latter can be described as conditional checks in the following sense: accurately predicted time-t violations cannot be explained using information available at time t, while clustering in the violation sequence should signal rejection provided the underlying estimate of VaR has been updated conformably.

The MCT methodology (Christoffersen and Pelletier, 2004; Berkowitz et al., 2011) which we also extend here, defines the null hypothesis as

$$I_{t+1} \overset{iid}{\sim} \text{Bernoulli } (\alpha, \alpha(1-\alpha)). \tag{2}$$

Conformably, given α, a series of violations at threshold α under the null hypothesis can be easily simulated. If a statistic depends on the data only via violations, then a bootstrap-type p-value can be obtained by computing a simulated counterpart to the observed statistic based on each draw from the Bernoulli distribution knowing α, and then computing an empirical p-value based on the rank of the observed statistic relative to the simulated one.

2.1 The cumulative violations series: distribution-free criteria

Testing expected shortfall is by far more challenging, particularly from the non-parametric and model agnostic perspective we adopt in this paper. In fact, because the measure is a population mean, the well known Bahadur and Savage (1956) critique cautions that its size-correct tests would lack power unless - typically parametric - restrictions are imposed. From a non-parametric perspective, tests that are almost surely discontinuous may circumvent this problem; see Dufour (1997), Romano (2004) and Bertanha and Moreira (2020) for recent insights. The procedures we propose here are motivated by these considerations.

Formally, ES at threshold α and time $t+1$ given $\mathcal{F}_t$ - denoted thereafter as $ES_{t+1|t}(\alpha)$ - is defined as the expected loss given that the loss exceeds $\text{VaR}_{t+1|t}(\alpha)$. Consequently, $ES_{t+1|t}(\alpha)$ can be represented as an integrated VaR:

$$ES_{t+1|t}(\alpha) = \frac{1}{\alpha}\int_0^{\alpha} VaR_{t+1|t}(u)\, du. \tag{3}$$

This suggests that averaging VaRs over a large enough and judicially chosen set of thresholds should approximate ES. In particular, the integral in (3) can be approximated with a Riemann sum with K terms, leading to

$$ES_{t+1|t}(\alpha) \simeq ES^K_{t+1|t}(\alpha) = \frac{1}{K}\sum_{j=1}^{K} \text{VaR}_{t+1|t}\left(j\frac{\alpha}{K}\right). \tag{4}$$

See for example Acerbi and Tasche (2002) who consider (4) with $K = 4$.

In contrast, Du and Escanciano (2017) show that it is possible to avoid approximations when the objective is backtesting rather than estimation. The idea is to construct an aggregated violation process building on (3) and assess the properties of this process under a correct risk model. The so-called time t CV process is given by

$$H_t(\alpha) = \frac{1}{\alpha}\left[\alpha - u_t\right]\mathbb{I}\left(u_t \leq \alpha\right) \tag{5}$$

where $u_t = G(r_t|\,\mathcal{F}_{t-1})$ is the Probability Integral (PIT) transform [(Rosenblatt, 1952)] and $G(\cdot|\,\mathcal{F}_{t-1}) = \mathbb{P}(r_t \leq \cdot|\,\mathcal{F}_{t-1})$ is the conditional cumulative distribution function (cdf) of r_t.[2]

[2]We refer the reader to the surveys of Corradi and Swanson (2006) and Tay and Wallis (2000) for more details on density forecast and its evaluation using the PIT and to contributions like Diebold et al. (1998), Berkowitz (2001), or Diks and Fang (2020) for a number of applications in finance and risk management.

The principle difficulty in developing formal procedures based on the $H_t(\alpha)$ series is that it requires the parametric estimation of $G(\cdot \,|\, \mathcal{F}_{t-1})$. This impedes applications beyond fully parametrized risk models, and even there, the asymptotic theory on which estimators of $G(\cdot \,|\, \mathcal{F}_{t-1})$ are based necessitates a large sample. Instead, Kratz et al. (2018) propose a discrete alternative to (5) by aggregating the sequence of VaRs from (4):

$$\bar{N}^K_{t+1}(\alpha) = \sum_{j=1}^{K} \bar{I}^K_{j,t+1}(\alpha), \tag{6}$$

$$\bar{I}^K_{j,t+1}(\alpha) = \begin{cases} 1 & \text{if } r_{t+1} \leq -\text{VaR}_{t+1|t}\left((K-j+1)\frac{\alpha}{K}\right) \\ 0 & \text{if } r_{t+1} > -\text{VaR}_{t+1|t}\left((K-j+1)\frac{\alpha}{K}\right) \end{cases} \quad j = 1, \ldots, K. \tag{7}$$

Building on the properties of this measure under a correct risk model, Kratz et al. (2018) propose a non-parametric test of the underlying VaRs, all together, and argue intuitively that such a test *implicitly* assesses ES.

Expanding on the multithreshold VaR idea, we first show that $\bar{N}^K_{t+1}(\alpha)/K$ converges weakly to $H_t(\alpha)$ for large K. This result which formalizes the intuitive argument of Kratz et al. (2018) and provides a useful unification of existing works, forms the basis for the alternative induced test procedure we propose in this paper.

Theorem 1. *Given an ES threshold α, consider the K equally spaced measures from (4), $\text{VaR}_{t+1|t}\left(j\frac{\alpha}{K}\right)$, $j = K, \ldots, 1$, and the associated total violations criteria $\bar{N}^K_{t+1}(\alpha)$ as defined in (6). Then as $K \to \infty$ and assuming that underlying returns are absolutely continuous*

$$\frac{\bar{N}^K_{t+1}(\alpha)}{K} \xrightarrow{D} H_t(\alpha).$$

Proof: See Appendix B.1.

Independently, Leccadito et al. (2014) and Perignon and Smith (2008) use an aggregate of this form, given a vector of thresholds $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)'$, with $\alpha_1 > \alpha_2 > \ldots > \alpha_K$:

$$N_{t+1}(\boldsymbol{\alpha}) = \sum_{j=1}^{K} I_{j,t+1}, \tag{8}$$

$$I_{j,t+1} = \begin{cases} 1 & \text{if } r_{t+1} \leq -\text{VaR}_{t+1|t}(\alpha_j) \\ 0 & \text{if } r_{t+1} > -\text{VaR}_{t+1|t}(\alpha_j) \end{cases}. \tag{9}$$

It is straightforward to see that (6) corresponds to (8) with

$$\alpha_j = (K-j+1)\frac{\alpha}{K}. \tag{10}$$

For further reference, the special case $K = 4$ leads to

$$\alpha_1 = \alpha, \quad \alpha_2 = .75\alpha, \quad \alpha_3 = .5\alpha, \quad \alpha_4 = .25\alpha. \tag{11}$$

The key main difference in the construction of (6) is that the VaR thresholds are based on (4); $\bar{N}^K_{t+1}(\alpha)/K$ thus converges to $H_t(\alpha)$ for large K yet the properties of the former can be assessed non-parametrically for fixed K. A discontinuous formulation thus enables one to: (i) avoid approximating the integral in (3) and capture the information on tail risk without imposing any further distributional

assumptions; and (ii) eschew Bahadur-Savage type critiques and ensure size-controlled power non-parametrically in finite samples.

Differently from available methods, here we use the properties of the $\bar{N}^K_{t+1}(\alpha)$ process to derive the finite sample *joint* distribution of the K series of violations $\bar{I}^K_{j,t+1}(\alpha)$. Formally, we show that this distribution is nuisance-parameter-free under a correct risk model, does not rely on any parametric dynamics for the underlying returns, and can be simulated without any further assumptions. Then any statistic that depends on the data only through these violations will inherit this property. Our results open up a wide spectrum of possibilities for ES testing, that are not restricted to CVs as has been suggested so far. From there on, we propose an induced test approach along with a novel adaptive combinatorial approach that allows us to better capture tail behavior.

Moving from the theory underlying Theorem 1 to practical reliance on many quantiles raises enduring overlap issues, particularly when the thresholds are small or are too close to each other. In principal, one can consider recognizably different thresholds $\alpha_1 > \alpha_2 > \ldots > \alpha_K$ so that, under standard continuity assumptions, associated quantiles are distinct. The fact remains that finite samples counterparts are not necessarily distinguishable as K increases and one is driven far into the tail. Standard inference procedures may lead to spurious rejections in this case. This is in contrast to our induced test approach, as tests on each of the thresholds are performed independently, and then combined. As may become clear in what follows, the simulation-based procedure we propose to control the family-wise error rate is not invalidated by finite sample overlaps.

2.2 The multi-threshold violations series: joint distributional theory

By the definition of a VaR under the correct risk model, each of the measures in Theorem 1 satisfies the following hypothesis:

$$\mathcal{H}_{0j} : P\left(r_{t+1} \leq -\text{VaR}_{t+1|t}\left(j\frac{\alpha}{K}\right)|\mathcal{F}_t\right) = \alpha, \quad j \in \{K, \ldots, 1\}, \tag{12}$$

so a correct ES corresponds to these hypotheses jointly, viewed as a continuum. Said differently, the adequacy of (3) can be mapped to $\mathcal{H}_* = \cap_{j=1}^K \mathcal{H}_{0j}$, $K \to \infty$. This forms the basis of the discretized counterpart, for fixed K:

$$\mathcal{H}_0 = \cap_{j=1}^K \mathcal{H}_{0j}, \quad j = K, \ldots, 1. \tag{13}$$

In this context, we next derive finite sample distributional results for the multivariate violation vector $\bar{I}^K_{j,t+1}(\alpha)$ from (7). To do this, we first consider the joint distribution of the $I_{j,t+1}$ series from (9) which is associated to a vector of unrestricted thresholds, that is, relaxing (10).

Theorem 2. *Given a vector of thresholds $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)'$ with $\alpha_1 > \alpha_2 > \ldots > \alpha_K$, and K series of exception indicators $I_{j,t+1}$ as defined in* (9)*. Then any set of test statistics to assess (13) that depend on the data only through the $I_{j,t+1}$ series will have a nuisance-parameter free joint null distribution that can be simulated as follows:*

$$\begin{gathered}\{I_{1,t+1} = \ldots = I_{j,t+1} = 1, \quad I_{j+1,t+1} = \ldots = I_{K,t+1} = 0\} \Leftrightarrow N_{t+1}(\boldsymbol{\alpha}) = j \\ \{I_{1,t+1} = I_{2,t+1} = \ldots = I_{K,t+1} = 0\} \Leftrightarrow N_{t+1}(\boldsymbol{\alpha}) = 0\end{gathered} \tag{14}$$

where $N_{t+1}(\boldsymbol{\alpha})$ as defined in (8) satisfies the following

$$\begin{gathered}N_{t+1}(\boldsymbol{\alpha}) = j \text{ with probability } \theta_j \text{ for } j = 0, \ldots, K, \\ N_{t+1}(\boldsymbol{\alpha}) \perp\!\!\!\perp N_{t+1-h}(\boldsymbol{\alpha}),\ \forall h \neq 0,\end{gathered} \tag{15}$$

$\perp\!\!\!\perp$ *denotes independence, $\alpha_{K+1} = 0$, $\alpha_0 = 1$, and*

$$\theta_j = \alpha_j - \alpha_{j+1},\ j = 0, 1, \ldots, K. \tag{16}$$

Proof: See Appendix B.2.
For further reference, two key ingredients underlying the proof of Theorem 2 are worth emphasizing.

1. The K critical thresholds considered are ordered such that $\text{VaR}_{t+1|t}(\alpha_1) < \text{VaR}_{t+1|t}(\alpha_2) < \ldots < \text{VaR}_{t+1|t}(\alpha_K)$ and $\alpha_j - \alpha_{j+1} > 0$ for any given finite K.

2. The probability that r_{t+1} falls in between the VaR quantiles associated to α_j and α_{j+1} is equal to θ_j under a correct risk specification in which case it is easy to see that

$$P\left(N_{t+1}(\boldsymbol{\alpha}) = x, N_{t+1-j}(\boldsymbol{\alpha}) = y\right) = \theta_x \theta_y, \quad \forall x, y. \tag{17}$$

Theorem 2 extends the simulation-based implications of (2). Indeed, given $\alpha_1 > \alpha_2 > \ldots > \alpha_K$ which uniquely define the θ_js: (i) a series of N_ts of length T can be easily simulated under a correct risk model that entails a multiple hypothesis involving the K VaRs, and (ii) it is straightforward to map this generated series into the corresponding $K \times T$ matrix of VaR violations $I_{i,t+1}$. This allows us to propose tests of (13) that utilize violations beyond their sum across thresholds, in contrast to available works.

The joint distribution of the indicators $\bar{I}^K_{j,t+1}(\alpha)$ follows from Theorem 2 by substituting $\alpha_j = (K - j + 1)\frac{\alpha}{K}$, $j = 1, \ldots, K$, so the thresholds are equally spaced, that is $\alpha_j - \alpha_{j+1} = \alpha/K$. For completion, and to reinforce the ES testing foundations, we state this result in the following corollary.

Corollary 1. *Given an ES threshold α and K series of exception indicators $\bar{I}^K_{j,t+1}(\alpha)$ as defined in (7) with reference to equally spaced VaR measures. Then under a correct risk model which entails (13), any set of test statistics that depend on the data only through the $\bar{I}^K_{j,t+1}(\alpha)$ series will have a nuisance-parameter free joint distribution that can be simulated as follows*

$$\begin{gathered}\left\{\bar{I}^K_{1,t+1}(\alpha) = \ldots = \bar{I}^K_{j,t+1}(\alpha) = 1, \quad \bar{I}^K_{j+1,t+1}(\alpha) = \ldots = \bar{I}^K_{K,t+1}(\alpha) = 0\right\} \Leftrightarrow \bar{N}^K_{t+1}(\alpha) = j \\ \left\{\bar{I}^K_{1,t+1}(\alpha) = \bar{I}^K_{2,t+1}(\alpha) = \ldots = \bar{I}^K_{K,t+1}(\alpha) = 0\right\} \Leftrightarrow \bar{N}^K_{t+1}(\alpha) = 0\end{gathered} \tag{18}$$

where $\bar{N}^K_{t+1}(\alpha)$ as defined in (6) satisfies the following

$$\begin{gathered}\bar{N}^K_{t+1}(\alpha) = j \text{ with probability } \alpha/K,\ j = 0, \ldots, K \\ \bar{N}^K_{t+1}(\alpha) \perp\!\!\!\perp \bar{N}^K_{t+1-h}(\alpha),\ \forall h \neq 0.\end{gathered} \tag{19}$$

The above results imply that the joint distribution of the K series of violations can be fully simulated under the null hypothesis (13), which is itself the intersection of adequacy at each threshold. We propose induced test procedures, by statistically combining VaR tests for each threshold. The primary statistical challenge in testing ES is thus diverted from the familiar problem of comparing a forecasted expectation with its model-based estimate, to the problem of controlling false-positive multi-threshold VaR backtests. The former problem is much more tedious than the latter and, perhaps more to the point here, often calls for additional inputs or assumptions.

Theorem 2 and Corollary 1 are not incompatible with possible finite sample quantile overlap, which would entail identical violation series. Some violation series may be identical for various other reasons, even if the thresholds are far apart. Such occurrences can actually reveal useful information, beyond what can be gleaned from a cumulative series. Conformably, and building on the distributional results we have shown, the tests we propose next require no more inputs than the violation series.

3 Induced test procedures

Regardless of the choice of K, the thresholds themselves and the tests to combine, false discoveries should be managed because the criteria in question are not independent. In fact, it is this dependence that conveys a formal ES interpretation to our multi-threshold formulation. To account for correlation of tests thus ensures the definitional relation from our multiple hypothesis to (4). For this purpose, we consider two strategies.

We first reformulate the existing multi-threshold tests of Perignon and Smith (2008) (PS henceforth) and Leccadito et al. (2014) for our purpose, by selecting the thresholds as in (4). The PS criterion is given by $LR^{PS} = 2\left(\sum_{i=0}^{K} \ln(\widehat{\pi}_i/\theta_i)^{\sum_{t=1}^{T} J_{i,t}}\right)$, where $\widehat{\pi}_i = \frac{1}{T}\sum_{t=1}^{T} J_{i,t}$ and $J_{i,t}$ is defined in Appendix B.2, eq. (33) . Its (limiting) null distribution is $\chi^2(K)$. Here we propose a MCT-based modification which will control its size in finite samples under our agnostic adequacy null. The Pearson test of Leccadito et al. (2014) assesses (8) in conjunction with (17). Denoting by $T_{x,y}^{(j)}$ the number of observations for which $N_t = x$ and $N_{t-j} = y$, the test statistic is

$$X_m = \sum_{j=1}^{m} \sum_{x,y} \frac{(T_{x,y}^{(j)} - (T-j)\theta_x\theta_y)^2}{(T-j)\theta_x\theta_y}. \tag{20}$$

Leccadito et al. (2014) propose a MCT p-value for X_m, which we also consider here.

Both statistics may be restrictive for various reasons, in view of the broad range of methods available to assess VaR. In particular, no information except lagged violations can be used, even when available and relevant. We thus next propose alternative statistical combinations methods. The following discussion uses any given set of α_is, not necessarily as in (11).

3.1 Combined criteria

To concretize concepts, we focus on combining CaViaR tests from Engle and Manganelli (2004) across thresholds, which allows one to add relevant predictors. CaViaR tests have also been reported by Berkowitz et al. (2011) to outperform various contenders. Consider K individual autoregressions for each threshold i, each of order m:

$$I_{i,t} = \nu_i + \sum_{k=1}^{m} \beta_{i,k} I_{i,t-k} + \sum_{q=1}^{Q} \bar{\beta}_{i,q} \bar{Z}_{i,t-q} + u_{i,t} \tag{21}$$

where the $\bar{Z}$ are desired predictors. Since each dependent variable is binary, we estimate K individual logit regressions and implement, in each case, a Wald test to verify, jointly, whether the $\beta_{i,k}$ and $\bar{\beta}_{i,q}$ are statistically significant and whether $\mathbb{P}\left(I_{i,t} = 1 \,\middle|\, \beta_{i,1} = \ldots = \beta_{i,m} = \bar{\beta}_{i,1} = \ldots = \bar{\beta}_{i,Q} = 0\right) = \mathrm{e}^{\nu_i}/(1+\mathrm{e}^{\nu_i}) = \alpha_i$. The (limiting) null distribution of the resulting test statistic is $\chi^2(m+1+Q)$. The individual tests are then combined, as described below.

The statistics literature has established various procedures to combine tests; the related body of works is too large to be usefully reviewed here. We focus on concepts that can be traced back to Fisher (1932) and Simes (1986). Recent modifications of Fisher's test include the truncated product method of Zaykin et al. (2002), the adaptive rank truncated product method of Yu et al. (2009) and Zhang et al. (2013). Improved versions of both procedures have also been proposed by Dong et al. (2015). These works also provide useful reviews and statistical references.

Denote by p_i the p-value associated to the single-threshold test with threshold α_i, $i = 1, \ldots, K$ and by p_{PS} the p-value of the PS procedure for the same thresholds. Let $\bar{p}_i$ be the i-th element of the $(K+1)$-vector $(p_1, \ldots, p_K, p_{PS})'$.

The multi-threshold combinations we first consider are

$$[\text{Fisher}]: \quad S_F = -2\sum_{i=1}^{K} \log p_i, \qquad [\text{Simes}]: \quad S_{Si} = -2\log\left[\min_{1\leq i\leq K}\left\{p_{(i)}\frac{K}{i}\right\}\right] \tag{22}$$

where $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(K)}$ are the ordered p-values. When applying the Simes method, we consider a monotonic transformation of the test statistic commonly considered in the statistical literature, i.e. $\min_{1\leq i\leq K}\left\{p_{(i)}\frac{K}{i}\right\}$. The transformation, which has no impact when using the MCT technique of Dufour (2006), maps the interval $[0,1]$ into $[0,+\infty]$, so that S_{Si} has the same order of magnitude of S_F. This, in turn, allows to use (23) below as a proper test statistic.

Since Simes' method performs well when the total evidence against the global null is concentrated in few of the combined p-values, whereas Fisher's method is powerful when the number of false individual hypotheses is instead large, we consider: the sum of the two

$$[\text{Fisher + Simes}]: \quad S_{FSi} = S_F + S_{Si}, \tag{23}$$

and the sum of individual single-threshold test statistics

$$[\text{Sum of CaViaR}]: \quad S_{\Sigma} = \sum_{i=1}^{K} LR_i^{CaViaR}, \tag{24}$$

where LR_i^{CaViaR} is the Wald statistic introduced after (21). We also consider the combinations of the K single-threshold tests with the PS procedure. The resulting statistics, denoted by $\bar{S}_F$, $\bar{S}_{Si}$, and $\bar{S}_{FSi}$, are obtained by using $\bar{p}_i$ instead of p_i. Using the LR^{PS} statistic of Section 3, the sum statistic is $\bar{S}_{\Sigma} = \sum_{i=1}^{K} LR_i^{CaViaR} + LR^{PS}$.

3.2 Which violations to boost? Adaptive combinations

So far, we have not discussed the choice of K. In fact, we find - via simulations - some evidence in favor of $K = 4$ as proposed by Acerbi and Tasche (2002), yet no clear optimal choice emerges. To address this concern, we propose an omnibus procedure that does not only take several VaR thresholds into account, but also flexibly adjusts to their information content. Contrary to the application of the average in (4), the K VaR measures are not directly aggregated to approximate the integral in (3). In fact, little is known about their consolidation and joint interpretation for backtesting ES, and there is no reason to expect that VaR deviations are evenly revealing across thresholds. The issue can be resolved by weighting thresholds unequally as backtests are combined. We assess the various standard combination strategies for this purpose as discussed in section (3.1); again, no clear choice emerges. We thus propose instead to combine a series of tests that are themselves combination tests across several values of K. For a clarity, and to emphasize its distinctive features relative to existing ones, we formally define the proposed combinatorial strategy as follows.

Definition 1. *Let $S(K_j)$ be the test statistic (e.g. Fisher's) that combines a given series of VaR tests with thresholds $i\frac{\alpha}{K_j}$, $i = 1, \ldots, K_j$, as in Theorem 1. The test based on the statistic*

$$S = \sum_{j\in J} S(K_j) \tag{25}$$

where J is a set of indices used to identify the number of combined thresholds, is defined as a "double combination" [DC] procedure. In particular, if $J = \{\underline{j}, \underline{j}+1, \ldots, \bar{j}\}$, then $K_{\underline{j}}$ and $K_{\bar{j}}$ are the smallest and largest K entering the combination.

The "DC" notation we introduce here aims to mimic the concept of the "double bootstrap" where in our framework, the idea of combining the combined tests mimics bootstrapping the bootstrap. As an alternative to optimizing K, which in our model-agnostic setting may be hard (if not impossible) to formalize, DC provides non-uniform weights across the considered thresholds which seem to capture revealing violations more effectively.

To see this, consider e.g. moving from the $S(2)$ to the $S(4)$ statistic. Increasing K from 2 to 4 expands the set of tested thresholds from $\{0.5\alpha, \alpha\}$ to $\{0.25\alpha, 0.5\alpha, 0.75\alpha, \alpha\}$, which enables more complete screening through the tail. Yet this breadth of scope may dilute signal strength from violations that are most at odds with the null hypothesis. Instead, using $S = S(2) + S(4)$ increases the weights attributed to VaR_α and $\text{VaR}_{0.5\alpha}$, which will leverage information at these thresholds thereby increasing power when effective detection is associated with them. Doubling is thus equivalent to a weighted combinatorial method. If, by means of that, informative signals are enhanced, substantial gains of power can result.

It is broadly known that informative violations are not uniformly spread across the tail. Furthermore, because the true spread is unknown and variable, there is no clear indication as to which violations to boost, and this is likely to vary across models and tests. Doubling - rather than just increasing K - provides a flexible omnibus "signal booster", to improve detection within a broad range of departures from accurate VaRs.

3.3 Combined p-values, finite sample validity

All of the above combined criteria have non-standard null distributions. Yet, importantly, all satisfy the conditions of Theorem 2 or Corollary 1. Indeed, the combined functions of the data depend, on the data, only through the joint series of underlying violations. It follows that their null distributions - individually and jointly - can be simulated with no further assumptions so that the MCT technique of Dufour (2006) is applicable, although tie-breaking is necessary, since the statistics are not strictly continuous. To obtain a p-value for any of the above test statistics (and right tailed tests), we apply the following.

Step 1: using (15) or (19), generate M time series replications of N_ts or $\bar{N}_t$s, each of length T. In each case, using (14) or (18), map the generated series into the corresponding $K \times T$ matrix of VaR violations.

Step 2: for each $j = 1, \ldots, M$, calculate the test statistic and denote it by S_j;

Step 3: compute the p-value as

$$\begin{aligned} \tilde{p}_M(S_0) &= \frac{M \times \tilde{G}_M(S_0) + 1}{M + 1}, \\ \tilde{G}_M(S_0) &= 1 - \frac{1}{M}\sum_{j=1}^{M} I(S_0 \geq S_j) + \frac{1}{M}\sum_{j=1}^{M} I(S_0 = S_j) \times I(U_0 \leq U_j) \end{aligned} \tag{26}$$

where $I(\cdot)$ is the indicator function, S_0 is the test statistic calculated from the original sample, and U_j, $j = 0, \ldots, M$, are independent standard uniform random variates.

Theorem 3. *Given any test statistic that satisfies the conditions of Theorem 2 or Corollary 1, let $\tilde{p}_M(S_0)$ denote its MCT p-value derived as in (26) where S_0 refers to the test statistic calculated from the original sample, and its simulated counterparts S_j, $j = 1, \ldots, M$, are derived applying (14) or (18) to independant draws from the discrete process in (15) or (19). Then for $0 < \bar{\alpha} < 1$ and when $\bar{\alpha}\,(M+1)$ is an integer:*

$$\mathbb{P}[\tilde{p}_M(S_0) \leq \bar{\alpha}] = \bar{\alpha}. \tag{27}$$

It is evident from Theorem 3 that quantile overlap that may entail identical violation series for different thresholds (in observed or simulated samples for that matter) does not invalidate our proposed method, in the sense that the resulting p-value will nevertheless satisfy (27). Quantile overlap conveys factual information on the tail of the distribution, despite the irregularities it may reflect. This reinforces the usefulness of finite sample and discrete procedures such as those we propose here, that require no regularity assumptions on the tail.

All of the above can be used with violations from parametric, non-parametric, Historical Simulation (HS) or Monte Carlo risk models as inputs. This paper does not take a stand on any risk model, as proposed procedures are model-agnostic. On balance, power remains a documented general problem for backtesting, which motivates our multiple testing approach.

From a practical perspective, VaRs for the various thresholds underlying Corollary 1 may be not be readily supplied by financial institutions. At the same time, PIT-based and parametric counterparts presume that: (i) institutions would readily part with information on underlying models, and (ii) stable models that allow for consistent estimation of intervening parameters are in fact maintained. On such trade-offs, expanding the set of quantiles to backtest may be easier to request relative to full model disclosure, as the regulatory environment evolves towards tail risk monitoring.

4 Statistical properties of proposed procedures

To illustrate the usefulness of our proposed procedure, we first report a numerical case study that documents concrete shortcomings of single-threshold methods even when ES is not the ultimate objective. We next consider various simulation experiments, all of which are calibrated for a fair comparison with existing methods.

4.1 Monte Carlo Experiments

In this section, we evaluate the proposed tests relative to several alternative procedures: the standard model-free regression method (see Christoffersen, 2012, p. 308-309), the conditional [denoted C_{ES}] and unconditional [denoted U_{ES}] tests of Du and Escanciano (2017, equations (5) and (6)), and two tests [denoted Z_1 and Z_2] from Acerbi and Székely (2014, equations (4) and (6)). Designs are calibrated to provide fair comparisons with relevant contenders. Three settings are maintained for this purpose.

First, C_{ES} and U_{ES} are applied with a large estimation window to avoid the over-rejections - or the costly corrections - as documented by the authors in small samples. Second, we calculate the p-values for the PS test using the asymptotic chi-square distribution and our MCT correction which we refer to under the heading PS_MC. Third, we implement (21) with just lagged violations. We consider two choices for equally-weighted coverage probabilities:

$$K = 3, \text{ with } \alpha_1 = 5\%, \quad \alpha_2 = 2.5\%, \quad \alpha_3 = 1\%, \tag{28}$$

and our ES-motivated prescription as in (11):

$$K = 4, \text{ with } \alpha_1 = 5\%, \quad \alpha_2 = 3.75\%, \quad \alpha_3 = 2.5\%, \quad \alpha_4 = 1.25\%, \tag{29}$$

in which case we compare power relative to ES(5%) benchmarks. For the test in (25) the values of K we combine are $K \in 2, 4, \ldots, 10$, leading to $S = S(2) + S(4) + ... + S(10)$. In all cases, the p-values for our proposed tests are applied with our MCT method. Some designs are based on a GARCH(1,1) model with intercept 0.05, ARCH parameter =0.05, and GARCH parameter =0.9. In addition, we simulate returns according to the Generalized Lambda Distribution (GLD). The GLD distribution has

four parameters: location and scale (λ_1 and λ_2), and shape parameters (λ_3 and λ_4) explicitly related to skewness and kurtosis (κ_3 and κ_4). We set $\kappa_3 = -0.8$, $\kappa_4 = 3.5$ and fix λ_1 and λ_2 leading to the GLD distribution with zero mean and unit variance.

4.2 Varying K in ES Tests

We first report our baseline study designed to document the choice of K in standard combined methods. Panel (a) of Figure 1 plots against K the rejection frequencies for $\bar{S}_{FSi}$ for the case in which returns are generated using a normal-GARCH model, VaR is estimated assuming a normal distribution and the ES level is $\alpha = 5\%$. Panel (b) is instead for the case in which returns are generated using a t-GARCH model and VaR is again estimated assuming a normal distribution.

[Figure 1 about here.]

One way to read these figures is that $K = 4$ provides the most efficient power improvement, which lends support to the thresholds proposed by Acerbi and Tasche (2002); see also Kratz et al. (2018). However, further improvements are not ruled out, particularly as the sample size grows. In this baseline design, the hypothesized VaR is the simple i.i.d. normal and, nevertheless, no further concrete recommendations emerge regarding K. This study motivated our proposed doubling strategy, since it reveals that just increasing K although useful, does not suffice. Our results with the DC methods reported below confirm their power advantage over all available ES tests that control size.

4.3 Size

When (15) is used to draw violations directly, (27) implies that all tests that depend on the data via no more than these violations[3], which includes the formula for the test statistic and the critical point, will be provably size-correct. To investigate the size of the tests when (15) is not explicitly imposed, we draw returns rather than violations using DGPs with: (i) GARCH-type volatility clustering imposing and relaxing conditional normality, and (ii) *i.i.d.* non-normal fundamentals. Yet to calculate the observed VaRs, we proceed as follows.

In experiment (i), we maintain the hypothesized DGPs yet we estimate the underlying parameters. Estimation error may intervene; for this design in particular, it matters with reference to the considered competing methods. In the (ii) case, we use HS since the unconditional quantile comes close to approximating the null VaR because of the *i.i.d.* setting. We aim to evaluate the practical worth of our tests under these empirically relevant parametric and non-parametric assumptions.

As a first experiment, we generate 5,000 time series according to the GARCH(1,1) model

$$h_{t+1} = \omega + \alpha e_t^2 + \beta h_t, \tag{30}$$

with $e_t = \sqrt{h_t} u_t$. We use the following parameters $\omega = 0.05$, $\alpha = 0.05$, $\beta = 0.9$. The results reported in Panel A of Table 1 refer to the case where $u_t \sim t(5)$, *i.e.* the distribution of the innovation is Student-t with 5 degrees of freedom, and the VaR is estimated out-of-sample using the GARCH-t VaR. Panel B of Table 1 reports instead the results for the case in which u_t has a standard normal distribution and the VaR is estimated out-of-sample using the Normal-GARCH model. In both cases we use a rolling window of length 500 for VaR estimation.

[Table 1 about here.]

[3]Tests will also depend on underlying exogenous predictors when considered, as in CaViaR tests.

From the results of Panel A and B of Table 1 it is evident that size is controlled which is noteworthy since parameters are estimated here. The parametric DGP in question is prominently considered in this literature, in both academia and practice. Escanciano and Olmo (2010) and Escanciano and Pei (2012) document problems arising from estimation and model specification risk in the asymptotic distribution of some out-of-sample backtests. Many methods require quite large estimation windows to satisfy the regularity conditions underlying usual asymptotic methods. These include the ES methods we consider below, where in contrast to the present estimation window of 500 observations, we use 2500. This point deserves notice for further reference.

We next simulate returns according to the Generalized Lambda Distribution (GLD) and estimate VaR using the HS method. Panel C of Table 1 reports the results. The GLD distribution has four parameters $(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$: λ_1 and λ_2 are the location and scale parameters and λ_3 and λ_4 are shape parameters explicitly related to the coefficients of skewness and kurtosis (κ_3 and κ_4). In our experiments we set $\kappa_3 = -0.8$ and $\kappa_4 = 3.5$ and fix the location and scale parameters so that we simulate from the GLD distribution with zero mean and variance equal to one.

From Panel C of Table 1, it is clear that the proposed tests have good size properties even when returns are drawn without explicitly imposing (15). This experiment also serves as a baseline to assess deviation from a normal VaR. Given the empirical properties of returns, any test to be useful should detect violations of this nature with high probability.

Finally, we also verify that size is controlled for the designs underlying the power study reported in Tables 6–7. These results, available upon request, pertain to an estimation window of length 2500 to respect the asymptotic theory in Du and Escanciano (2017). To sum up, our simulations - reported below - will objectively inform on the relative advantage of all competing methods, maintaining our focus on practical considerations.

4.4 Power Analysis

To study the power of the tests we consider the following cases:

Design	DGP	Table	Hypothesized VaR	Estimation window	Thresholds
I	GARCH(1,1)-t	2/3	HS/FHS	500/2500	(29)/(29) and **DC**
II	AR(1) N_t	4	–	–	(28)
III	*i.i.d.* GLD	5/6	*i.i.d.* Gaussian	250	(28)/(29) and **DC**
IV	GARCH(1,1)-t/GARCH(1,1)-Normal	7/8	*i.i.d.* Gaussian	2500	(29) and **DC**/(29) and **DC**

Design I: Conditional heteroskedasticity and HS. We generate 5,000 time series according to a GARCH(1,1) model with Student-t innovations. VaR is estimated out-of-sample using HS and a rolling window of length 500. The relevance of this power experiment stems from the fact that HS slowly adapts to the dependence of returns. Yet the latter DGP features key empirical properties of returns, on which informative backtests are required. Results are reported in Table 2 conforming with an ES(5%) test.

[Table 2 about here.]

Two key findings emerge from Table 2. First is the good relative power properties of our tests. Although less vulnerable to estimation error, HS-based tests raise other methodological issues summarized in e.g. Escanciano and Pei (2012) and Pelletier and Wei (2016). HS estimates VaR via unconditional empirical quantiles that account for return dynamics mainly through rolling samples. By contrast, VaRs are quantiles of the distribution of returns conditional on available information.

We study below methods based on filtered HS, that have more recently been proposed to circumvent this problem.

With (unfiltered) HS, unconditional quantiles by construction forgo conditional features such as time varying conditional volatility, which will result in underestimating the probability of an increase in VaR at time $t+1$ given it was correctly estimated at time t (Pritsker, 2006). Said differently, HS slowly updates risk measures as volatility rises, which may translate into power losses. This is clearly reflected in the power of the PS test, and all tests based on lagged violations with $m = 1$. As T and m increase, our combined CaViaR tests dominate although we did not make full use of their flexibility by extending the information set beyond violations. Despite ignoring this clear and evident advantage, our result underscore of superiority of our multiple test procedure.

The second result which emerges from Table 2, is the drastically poor performance of the familiar ES backtest even with $T = 2500$. To construct the test, on each window we calculate ES as the average of the 25 observations which are smaller or equal than the 5% quantile. It is worth recalling the prevalence of GARCH features empirically. Our combined VaR approach succeeds where this popular test fails, while maintaining a non-parametric perspective. This result is worth noting as we interpret the recent parametric ES tests below.

We generate 5,000 time series according to a GARCH(1,1) model with t-innovations. VaR is estimated out-of-sample using FHS and a rolling window of length 2500. Results are reported on Table 3. Filtering imposed the same GARCH process with estimated parameters. Nevertheless, since FHS is not a fully parametric method, the reported empirical rejections may be analyzed as power. We do not rule out their interpretation as mis-specification signals, yet in line with our agnostic approach and given the main motivation for filtering, comparing Table 3 to Table 2 may be seen a legitimate illustration of the power advantage of FHS. Results indeed confirm that filtering improves power. More important are our findings on our proposed DC method. Indeed, in addition to filtering, Panels B and C also compare our proposed ES test with $K = 4$ to our DC method. The power improvements of doubling are concrete and sizeable. These results are reinforced in other designs as we compare this method to more recent parametric ES tests.

[Table 3 about here.]

Design II: non-parametric scenarios, correlated violations. We simulate 5,000 time series each of length $T \in \{250, 500, 1000, 2500\}$ for the total number of violations in day t, that we now denote N'_t. The simulations are such that N'_t has still associated the vector $(\alpha_1, \alpha_2, \alpha_3)'$ but N'_t and N'_{t-1} are not independent. In particular, we draw from the following Gaussian AR(1) process: $z_t = \beta z_{t-1} + \epsilon_t$, for $t = 2, \ldots, T$, $z_1 \sim N(0, s)$ and $\epsilon_t \sim N(0, \sigma)$ and obtain N'_t as

$$N'_t = F^{\leftarrow}_{N_t}\left(\Phi(z_t/s)\right), \quad s^2 = \sigma^2/(1-\beta^2) \tag{31}$$

where $F^{\leftarrow}_{N_t}$ is the generalized inverse of the cdf of N_t, and Φ is the standard normal cdf. Results for $\sigma = 1$ and $\beta = 0.95$ are reported on Table 4. This design is intended to mimic realistic desk-level systematic unobservables leading to correlated violations. Table 4 illustrates the superiority of our combined CaViaR tests with small T. Unconditional coverage tests cannot detect such violations; the relatively weak performance of the PS is thus not surprising. However, rejections with the Pearson test are at most of 38.2%, while rejections via our tests range from 93.5% to 100% with $T = 250$. Any of the combination methods we propose is thus particularly useful as an addition to practitioners' toolkits.

[Table 4 about here.]

Design III: Benchmarking Gaussian VaR, i.i.d. directions. In Table 5, we simulate returns according to the Generalized Lambda Distribution (GLD) and estimate VaR assuming a normal model. While all tests show very good power for large T, our proposed tests are clearly superior for small samples, in particular for the case of $T = 250$. While lagged violations are not the source of power in this design, we find that the CaViaR tests detect deviations in frequency quite well, via their implications on the regression intercept. With small T, the Pearson test has no power, and despite its good relative performance, the PS test is dominated by any combination method we propose. The null and alternative models are both i.i.d. in this design, and this may drive the power (or lack thereof) with just 250 violations. The fact remains that the excess skewness and kurtosis that one realistically may miss here allow serious extreme events. Again, our multiple hypothesis approach achieves very useful improvements, when and where they are mostly needed.

[Table 5 about here.]

The design underlying Table 5 is next utilized by varying the thresholds, to compare our tests with the ES tests of Du and Escanciano (2017). Recall that we restrict our analysis here to an estimation window of 2500 observations. Results are reported in Table 6.

[Table 6 about here.]

Furthermore, recall that the Z_1 and Z_2 tests were oversized, despite the simplicity if the considered null model. While results are reported for completion, their interpretation for power comparisons lacks validity. One of the findings in Table 6 is the lack of power of $C_{ES}(1)$ and $C_{ES}(5)$. This test perhaps forgoes the frequency implications, in contrast to the CaViaR approach. The U_{ES} test dominates in this design relative to our proposed test based on $K = 4$. However, our proposed DC method outperforms U_{ES} for all sample sizes.

One broad conclusion we also draw from this design is that some conditional tests may weakly detect frequency deviations, when their key constituents focus mostly on the time dependence and clustering properties of violations. Indeed, we find that our tests which combine the PS and CaViaR tests together almost match the parametric unconditional test of Du and Escanciano (2017), although they just approximate ES. This result illustrates another advantage of our test combinations, particularly when large estimation windows are infeasible, a stable parametric model is not utilized or is unavailable to the assessor.

Design IV: Benchmarking Gaussian VaR, dependent directions. We now move to a design which should suit the conditional test of Du and Escanciano (2017). In line with their own designs, we reconsider the same GARCH-t model of the above experiment I, and assume that the model used in estimation is the normal distribution. Table 7 reports the results using a window of length 2500. Additionally, Table 8 displays the power of the different tests when returns are generated according to a GARCH model with normal innovations and VaR is estimated using the normal model.

[Table 7 about here.]

Interestingly, and despite the dependent alternative we purposely consider, U_{ES} still outperforms $C_{ES}(5)$ for $T = 250$ and $C_{ES}(1)$ for T up to 1000. Both tests are however dominated, and by a wide margin, by our modified PS and our combined CaViaR tests, whether $K = 4$ or using our DC approach.

In the context of this experiment, in addition to setting $K = 4$, we also considered combining $K = 20$ CaViaR tests. The comparison is meaningful because 20 is the number of different thresholds

that are used in the combinations of combinations tests for $K \in 2, 4, \ldots, 10$. The results (not reported here, but available upon request) show that the DC combination method systematically outperforms the tests based on $K = 20$. This reinforces the power advantage of doubling, in contrast to just increasing K.

[Table 8 about here.]

Results collected together reinforce the superiority of multi-threshold combined methods for backtesting ES. In view of the minimal inputs we need relative to clear and evident power advantages, our proposed MC and model-free combination method emerges with concrete promise.

5 Empirical Analysis: ETF VaR and ES

5.1 A case study: ETFs in financial crises

To concretize the multiple hypothesis concerns with familiar data and measures, consider assessing the HS VaR for the iShares Core S&P 500 ETF, over the 2007-2010 period. The financial crisis underscores the considered time span which, for the record, pre-dates the Basel III international regulatory framework for banks. Out of sample backtests are performed based on $T = 250$ observations and coverage probabilities $\alpha_1 = 5\%$, $\alpha_2 = 2.5\%$, and $\alpha_3 = 1\%$. Though it departs from (11), this choice reflects available multi-threshold works in line with our benchmarking objective in this section.

We first apply three single-threshold CaViaR tests, using (21) with just lagged violations and $m = 1$. The p-values we obtain are 14.106%, 0.350%, and 6.620%, respectively. A single-threshold interpretation at the 5% significance level thus rejects VaR(2.5%) and fails to reject both VaR(5%) and VaR(1%). What then, can we conclude about the underlying HS VaR? The familiar 1% threshold does not reveal enough to reject the model relative to its 2.5% counterpart. Conceivably, extreme tail events can be rare with the aggregate data at hand, which may drive such discordance across thresholds. However, the time frame we purposely consider spans the financial crisis; such disclaimers are thus weakly relevant here. Regardless, any decision on the HS VaR based on the three CaViaR tests interpreted *as is*, entails a type I error of up to $3 \times 5 = 15\%$. A Bonferroni-based equal weight decision would validate rejection at any level larger than $3 \times 0.350 = 1.050\%$. Yet, this approach side-steps rather than models the dependence between tests. In contrast, dependence should be intrinsically accounted for to accurately approximate tail risk, our ultimate objective.

So using the statistics (22)–(24) we formally combine for this purpose, we find $S_F = 20.657$, $S_{Si} = 9.113$, $S_{FSi} = 29.770$, and $S_\Sigma = 20.657$ with MCT p-values 0.464%, 0.370%, 0.390%, and 1.118%, respectively. Applying the PS test, we find $LR_{uc} = 13.003$ and a p-value equal to 0.463%. Integrating the latter into our combination strategy, we find $\bar{S}_F = 31.408$, $\bar{S}_{Si} = 9.364$, $\bar{S}_{FSi} = 40.772$, and $\bar{S}_\Sigma = 33.660$ with MCT p-values of 0.292%, 0.402%, 0.306%, and 0.498%, respectively. All of these confirm rejecting the underlying VaRs, controlling type I error with each statistic, this time, to the desired 5%. The Pearson test, eq (20), is not significant at the 5% level, given that the resulting p-values are 39.573% for the test with $m = 1$, which is noteworthy since our CaViaR tests use no more than lagged violations as predictors.

For completeness, we replicate the above experiment in the case in which the proposed tests (again with $m = 1$) are used to backtest 5%-ES assuming the HS method to predict VaRs. When we use the ES approximation based on 4 thresholds, see eq. (4) and eq. (11), all the testing procedure we consider give p-values smaller than 1%, with the only exception of the 'Sum of CaViaR' test which gives an observed test statistic $S_\Sigma = 27.964$ corresponding to a p-value of 1.458%. When using instead the DC combination scheme, in particular, a combination of tests with $K \in 2, 4, \ldots, 10$, the smallest

p-values we obtain is again the one associated to S_Σ (p-value of 0.04%). Finally we consider the filtered historical simulation method (based on the residuals from a GARCH(1,1) model with t-distributed innovations) instead of HS. Also in this case we strongly reject the null (all p-values are smaller or equal than 0.03%) both in the case $K = 4$ and in the case in which we combine tests across K.

5.2 Bloomberg ETFs

We consider the iShares Core S&P 500 ETF (Ticker IVV), the iShares Core FTSE 100 UCITS ETF (Ticker BCYIF), and the iShares MSCI Hong Kong ETF (Ticker EWH). We solely rely on the published VaR and CVaR of Bloomberg according to three methods: 1 year Historical (HS based on 1 year of past data), Monte Carlo[4], and parametric (assuming that returns are normally distributed). We consider the coverage probabilities as in (29) to assess ES(5%). Since VaRs and CVaRs are only available from 2009, we consider the 2009-2016 period. We also analyze the 2010-2014 sub-sample, to reflect the fundamentals leading to and underlying the Basel III accords. Specifically, the G20 had advocated that "national authorities should develop and agree by 2010 a global framework for promoting stronger liquidity buffers at financial institutions." Conformably, the Basel Committee on Banking Supervision proposed reforms on the Liquidity Coverage Ratio to be implemented as of January 2015. The Basel framework is relevant for the financial system beyond its direct targets (namely banks), as surrounding consultations and resulting regulations reflect the risk environment at large, and shapes monitoring tools and measures. Tables 9, 10 and 11 report results on IVV, BCYIF, and EWH ETF, respectively.

[Table 9 about here.]

[Table 10 about here.]

[Table 11 about here.]

In the following discussion, *rejection* refers to the 5% statistical significance level, and the *existing* tests/methods refer to our modified ES-targeted implementation of the PS and Pearson tests as well as the standard regression-based model-free ES test, as we analyze the value added of our combined CaViaR tests.

Regarding the IVV ETF, Table 9 reveals clear differences between the sub-period and full-sample results. Specifically, when focus is restricted to 2010-2014: (i) all tests fail to reject accuracy for all considered models if $m = 1$ (when relevant), and (ii) all combined tests concur in rejecting all models if $m = 5$ while the existing tests are insignificant, with the exception of the Pearson test which rejects the parametric VaR. Evidence on the Monte Carlo measure is similar over the full sample, in that no test rejects accuracy with $m = 1$ while when $m = 5$ all combined tests are significant, in contrast to the PS and conformable Pearson test. Evidence on the full sample HS measure stands in sharp contrast: existing tests are insignificant whereas combined tests are significant. The full sample normal model is rejected throughout except with the Pearson test and $m = 1$.

We find more noteworthy conflicts regarding the other two ETFs, with which combined tests are significant throughout for all three VaR estimation methods over the full and sub-sample, and for both $m = 1$ and $m = 5$. In contrast, in each case except the Normal BCYIF, there is at least one of the existing tests that is insignificant.

[4]Bloomberg uses linear factor models for VaR estimation. In the case of the Monte Carlo approach, the joint distribution of future factor is estimated assuming Student's t marginal distributions and a normal or a Student's t copula to model the dependence structure between factors. VaR and ES estimates are based on 10,000 random simulations of factor returns.

Conflict in decisions with different values of m are a familiar problem. Indeed, with daily data, a 5 days lag can be more informative. One result stands out in this regard: with the IVV ETF, many of our tests are insignificant with $m = 1$ whereas one lag suffices to reject accuracy for the other markets. This may reflect structural specificities of the aggregate S&P 500 index. More to the point are the disagreements we find between existing and combined tests, which illustrate the superiority of the latter for both lag choices.

These results underscore the usefulness of combined tests. Aside from methodological improvements, our findings reveal that tail risk is not adequately reflected via the considered publicized measures, a result which may have escaped formal notice to date. Admittedly, the profession still struggles to define useful measures so our findings do not single Bloomberg out. As model-agnostic checks gain credibility, backtests that share the flexibility of our methodology may easily be built interactively and perhaps in real-time, into portfolio managers' and on-line toolkits.

6 Conclusions

In this paper we presented backtesting methods to assess VaR and ES that require no more than VaR violations as inputs. Contrary to existing procedures, our tests do not require modelling assumptions. Our methodology relies on multiple testing for two purposes: (i) to process evidence on the frequency and dynamic evolution of violations jointly; and (ii) to capture more information about the magnitude of violations.

Simulation results illustrate the usefulness of our combined approach, particularly with small samples and when underlying models are unavailable to assessors. Results also reinforce the superiority of CaViaR approaches beyond their original focus on VaR, as ES backtests. Empirically, we assessed desktop data by Bloomberg on ETFs, and provided evidence of statistical mis-measurement in their risk exposure. Taken collectively, our simulation and empirical results provide useful practical prescriptions and underscore the usefulness of combined tests in finance.

Appendix

A Existing ES Testing Procedures

We first consider the familiar model-free ES test based on the regression

$$r_{t+1} - \mathrm{ES}_{t+1} = b_0 + b_1 \mathrm{VaR}_t + \epsilon_{t+1} \quad \text{for } t+1 \text{ where } r_{t+1} \leq -\mathrm{VaR}_{t+1}$$

The testing procedure checks, by the means of a Wald test, the null

$$b_0 = b_1 = 0. \tag{32}$$

In addition, when the design allows us to do so, we study the test of Du and Escanciano (2017) who propose using the CVs as defined as in eq. (5), for conditional and unconditional ES tests. Basically, the demeaned CVs should be a martingale difference sequence if the risk model is appropriate. To compute these CVs, the assessor needs to impose a distribution for the underlying data[5]; the tests are also asymptotic and may require costly size corrections in finite samples. Consider a probability level

[5]Du and Escanciano (2017) suggest that distribution free conditional ES tests may "require more than the ex-ante ES and the ex-post returns" (endnote 7, p. 957). This also applies to the ES backtests of Acerbi and Székely (2014).

α, for example $\alpha = 5\%$. In the notation of Du and Escanciano (2017), the α-violation at time t is defined as $h_t(\alpha) = I\left(r_t \leq G^{-1}(\alpha|\,\mathcal{F}_{t-1})\right)$. Here $G^{-1}(\alpha|\,\mathcal{F}_{t-1}) = \inf\{u : G_i(u|\,\mathcal{F}_{t-1}) \geq \alpha\}$ is the αth percentile of the distribution G. The CV at time t is defined as the integral of the hit sequence over the coverage threshold in the left tail:

$$H_t(\alpha) = \frac{1}{\alpha}\int_0^{\alpha} h_t(u)\mathrm{d}u$$

and it is shown to be equivalent to the expression in eq. (5). Since $\{u_t\}_t$ is a sample of independent and identically distributed (i.i.d.) standard uniform variables, both violations and CVs are distribution free. It is easy to show that $H_t(\alpha)$ has mean $\alpha/2$ and variance $\alpha(1/3 - \alpha/4)$. Since the cdf G generally is unknown, Du and Escanciano (2017) suggest specifying a parametric conditional distribution $G(\cdot, \mathcal{F}_{t-1}, \theta_0)$ for some unknown parameter vector θ_0 and using in (5) $\hat{u}_t = G(r_t, \mathcal{F}_{t-1}, \hat{\theta}_n)$ where $\hat{\theta}_n$ is a consistent estimator θ_0 based on n observations. The unconditional test they propose based on T CVs is given by $U_{ES} = \frac{\sqrt{T}\left(\bar{H}(\alpha)-\alpha/2\right)}{\sqrt{\alpha(1/3-\alpha/4)}}$, where $\bar{H}(\alpha) = \sum_{t=1}^{T}\hat{H}_t(\alpha)/T$ and $\hat{H}_t(\alpha) = \frac{1}{\alpha}\left[\alpha - \hat{u}_t\right]$. Assuming that both $n \to \infty$ and $T \to \infty$, such that $T/n \to 0$, U_{ES} converges in distribution to the standard normal.

Conditional procedures test the null that autocorrelations of $H_t(\alpha) - \alpha/2$ are zero up to lag m. In our simulation studies we use the Ljung-Box test and hence the test statistic is $C_{ES}(m) = T(T+2)\sum_{j=1}^{m}\hat{\rho}_{Tj}^2$, where $\hat{\rho}_{Tj}$ is the estimated lag-j autocorrelation of $H_t(\alpha) - \alpha/2$, based on a sample of T violations. If both $n \to \infty$ and $T \to \infty$, such that $T/n \to 0$, $C_{ES}(m)$ has a chi-square distribution with m degrees of freedom.

Finally, wherever possible, we make some comparisons with two testing procedures proposed in Acerbi and Székely (2014). The first testing procedure used to evaluate the accuracy of ES with level α, is based on the results that, under the correct model $\mathbb{E}\left[\frac{r_{t+1}}{\mathrm{ES}_{t+1}} + 1 \middle| r_{t+1} + \mathrm{VaR}_t < 0\right] = 0$. Based on a sequence of T violations, the test statistic is then calculated as $Z_1 = \frac{\sum_{t=1}^{T}\frac{I_t r_t}{\mathrm{ES}_t}}{\sum_{t=1}^{T} I_t} + 1$ where I_t is the exception indicator associated to the VaR with level α, see eq. (1). The second test we consider is based on the unconditional expectation $\mathrm{ES}_{t+1} = -\mathbb{E}\left[\frac{I_{t+1}r_{t+1}}{\alpha}\right] = 0$. The resulting test is given by $Z_2 = \sum_{t=1}^{T}\frac{I_t r_t}{T\alpha\mathrm{ES}_t} + 1$. To derive the significance of the tests, the authors suggest computing the test statistic, Z^0, for the observed sample, simulating returns under the null and computing the p-value as $\hat{p}_M(Z^0) = \sum_{i=1}^{M}\frac{I(Z^i<Z^0)}{M}$, where M is the number of simulations under the null and Z^i the value of the test statistic in the i-th scenario. As in the tests of Du and Escanciano (2017), implementing the above tests requires specifying a parametric conditional distribution for returns.

B Proofs

B.1 Proof of Theorem 1

Since the result does not depend on t, for simplicity we drop the time index. Therefore we use the notation $\bar{N}_K(\alpha)$ and $H(\alpha)$.

By assumption, the K levels are:

$$\alpha_j = \alpha - (j-1)\frac{\alpha}{K} = (K-j+1)\frac{\alpha}{K} \quad j = 1, 2, \ldots, K$$

Since

$$H(\alpha) = \frac{1}{\alpha}[\alpha - u]I(u \leq \alpha)$$

with

$$u = G(r),$$

when r is a continuous r.v. its cdf is (for $0 < h \leq 1$)

$$\begin{aligned} F_{H(\alpha)}(h) &= P(H(\alpha) \leq h) = P\left(1 - \frac{u}{\alpha} \leq h\right) \\ &= P\left(u \geq \alpha(1-h)\right) = P\left(r \geq G^{-1}(\alpha(1-h))\right) \\ &= P\left(r \geq q_{\alpha(1-h)}\right) = 1 - \alpha(1-h) = 1 - \alpha + h\alpha, \end{aligned}$$

where $q_s = -\text{VaR}(s)$ denotes the s-quantile of r. Note that in the above results we have used the fact that $G(\cdot)$ is the cdf of r. Note also that

$$P(H(\alpha) = 0) = P(u > \alpha) = P(r > q_\alpha) = 1 - \alpha.$$

Consider now the r.v. $\bar{M}_K(\alpha) = \bar{N}_K(\alpha)/K$. Then

$$\begin{aligned} P\left(\bar{M}_K(\alpha) = 0\right) &= P\left(r > q_{\alpha_1}\right) = 1 - \alpha_1 = 1 - \alpha \\ P\left(\bar{M}_K(\alpha) = 1/K\right) &= P\left(q_{\alpha_2} < r < q_{\alpha_1}\right) = \alpha_1 - \alpha_2 = \alpha/K \\ P\left(\bar{M}_K(\alpha) = 2/K\right) &= P\left(q_{\alpha_3} < r < q_{\alpha_2}\right) = \alpha_2 - \alpha_3 = \alpha/K \\ &\vdots \\ P\left(\bar{M}_K(\alpha) = (K-1)/K\right) &= P\left(q_{\alpha_K} < r < q_{\alpha_{K-1}}\right) = \alpha_{K-1} - \alpha_K = \alpha/K \\ P\left(\bar{M}_K(\alpha) = 1\right) &= P\left(r < q_{\alpha_K}\right) = \alpha_K = \alpha/K \end{aligned}$$

Therefore its cdf is

$$F_{\bar{M}_K(\alpha)}(m) = \begin{cases} 0 & \text{if } m < 0 \\ 1 - \alpha & \text{if } 0 \leq m < 1/K \\ 1 - \alpha + \alpha/K & \text{if } 1/K \leq m < 2/K \\ 1 - \alpha + 2\alpha/K & \text{if } 2/K \leq m < 3/K \\ & \vdots \\ 1 - \alpha/K & \text{if } (K-1)/K \leq m < 1 \\ 1 & \text{if } m \geq 1 \end{cases}$$

More synthetically,

$$F_{M_K}(m) = \sum_{j=0}^{K-1} (1 - \alpha + j\alpha/K) I(j/K \leq m < (j+1)/K) + I(m \geq 1).$$

It is easy to see that

$$\lim_{K \to \infty} F_{\bar{M}_K(\alpha)}(m) = F_{H(\alpha)}(m) \quad \forall m \in \mathbb{R}$$

and therefore

$$\bar{M}_K(\alpha) \xrightarrow{D} H(\alpha).$$

■

B.2 Proof of Theorem 2

Using the notation of Perignon and Smith (2008), let

$$J_{j,t+1} = \begin{cases} 1 & \text{if } -\text{VaR}_{t+1|t}(\alpha_{j+1}) < r_{t+1} \leq -\text{VaR}_{t+1|t}(\alpha_j) \\ 0 & \text{otherwise} \end{cases} . \tag{33}$$

By convention, $\alpha_{K+1} = 0$, $\alpha_0 = 1$, $\text{VaR}_{t+1|t}(\alpha_{K+1}) = +\infty$, and $J_{0,t+1} = \prod_{i=1}^{K}(1 - J_{i,t+1})$. Because the VaRs are ordered, these random variables are not independent. Furthermore, each indicator can be expressed as

$$J_{j,t+1} = I_{j,t+1} - I_{j+1,t+1}, \quad j = 1, \ldots, K, \tag{34}$$

where $I_{j,t}$ represents the usual exception indicator (1), except for the multi-threshold indexing (α_j instead of α). Thus defined: (i) for any t, only one of $J_{i,t+1}$ can be equal to one, and (ii) for $i = 0, \ldots, K$, N_{t+1} takes on value i when $J_{i,t+1} = 1$. This implies

$$\begin{gathered} N_{t+1}(\boldsymbol{\alpha}) = j \text{ with probability } \theta_j, \text{ for } j = 0, \ldots, K, \\ P\left(N_{t+1}(\boldsymbol{\alpha}) = x, N_{t+1-j}(\boldsymbol{\alpha}) = y\right) = \theta_x \theta_y, \quad \forall x, y \end{gathered}$$

which proves (15); see also Perignon and Smith (2008) and Leccadito et al. (2014). (34) and the defintion of $N_{t+1}(\boldsymbol{\alpha})$ also yield (6) which completes the proof.

■

B.3 Proof of Theorem 3

When the null hypothesis from Theorem 2 or Corollary 1 holds, S_j, $j = 0, \ldots, M$, are exchangeable because the distribution underlying (26) is nuisance parameter free, so all parameters required to draw S_j, $j = 1, \ldots, M$, are known (i.e. are set by the tested null). Applying Proposition 2.4 of Dufour (2006), we then get, under the null hypothesis, the desired result. ■

References

Acerbi, C. and B. Székely (2014). Backtesting expected shortfall. *Risk 12*, 76–81.

Acerbi, C. and D. Tasche (2002). On the coherence of expected shortfall. *Journal of Banking & Finance 26*(7), 1487–1503.

Bahadur, R. R. and L. J. Savage (1956). The nonexistence of certain statistical procedures in nonparametric problems. *The Annals of Mathematical Statistics 27*(4), 1115–1122.

Berkowitz, J. (2001). Testing density forecasts, with applications to risk management. *Journal of Business & Economic Statistics 19*(4), 465–74.

Berkowitz, J., P. Christoffersen, and D. Pelletier (2011). Evaluating Value-at-Risk models with desk-level data. *Management Science 57*(12), 2213–2227.

Bertanha, M. and M. J. Moreira (2020). Impossible inference in econometrics: Theory and applications. *Journal of Econometrics*. Forthcoming.

Christoffersen, P. (2012). *Elements of Financial Risk Management* (2nd ed.). San Diego, CA: Academic Press.

Christoffersen, P. and D. Pelletier (2004). Backtesting Value-at-Risk: A duration-based approach. *Journal of Financial Econometrics 2*(1), 84–108.

Christoffersen, P. F. (1998). Evaluating interval forecasts. *International Economic Review 39*(4), 841–862.

Colletaz, G., C. Hurlin, and C. Perignon (2013). The Risk Map: A new tool for validating risk models. *Journal of Banking & Finance 37*(10), 3843–3854.

Corradi, V. and N. Swanson (2006). Predictive density evaluation. In G. Elliott, C. Granger, and A. Timmermann (Eds.), *Handbook of Economic Forecasting* (1 ed.), Volume 1, Chapter 05, pp. 197–284. Elsevier.

Danielsson, J. and C. Zhou (2016). Why Risk Is So Hard to Measure. Working paper no. 494, De Nederlandsche Bank. Available at SSRN: https://ssrn.com/abstract=2597563.

Diebold, F. X. (2015). Comparing predictive accuracy, twenty years later: A personal perspective on the use and abuse of diebold–mariano tests. *Journal of Business & Economic Statistics 33*(1), 1–9.

Diebold, F. X., T. A. Gunther, and A. S. Tay (1998). Evaluating density forecasts with applications to financial risk management. *International Economic Review 39*(4), 863–883.

Diks, C. and H. Fang (2020). Comparing density forecasts in a risk management context. *International Journal of Forecasting 36*(2), 531–551.

Dong, Z., W. Yu, and W. Xu (2015). A modified combined p-value multiple test. *Journal of Statistical Computation and Simulation 85*(12), 2479–2490.

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

Dufour, J.-M. (1997). Some impossibility theorems in econometrics with applications to structural and dynamic models. *Econometrica 65*(6), 1365–1387.

Dufour, J.-M. (2006). Monte Carlo tests with nuisance parameters: A general approach to finite-sample inference and nonstandard asymptotics. *Journal of Econometrics 133*(2), 443–477.

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.

Escanciano, J. C. and J. Olmo (2010). Backtesting parametric Value-at-Risk with estimation risk. *Journal of Business & Economic Statistics 28*(1), 36–51.

Escanciano, J. C. and P. Pei (2012). Pitfalls in backtesting historical simulation VaR models. *Journal of Banking & Finance 36*(8), 2233–2244.

Fisher, R. A. (1932). *Statistical Methods for Research Workers* (4th ed.). London: Oliver and Boyd.

Hurlin, C. and S. Tokpavi (2006). Backtesting Value-at-Risk accuracy: a simple new test. *Journal of Risk 9*(2), 19–37.

Kratz, M., Y. H. Lok, and A. J. McNeil (2018). Multinomial VaR backtests: A simple implicit approach to backtesting expected shortfall. *Journal of Banking & Finance 88*, 393–407.

Kupiec, P. H. (1995). Techniques for verifying the accuracy of risk measurement models. *Journal of Derivatives 3*(2), 73–84.

Leccadito, A., S. Boffelli, and G. Urga (2014). Evaluating the accuracy of Value-at-Risk forecasts: New multilevel tests. *International Journal of Forecasting 30*(2), 206–216.

Pelletier, D. and W. Wei (2016). The geometric-VaR backtesting method. *Journal of Financial Econometrics 14*(4), 725–745.

Perignon, C. and D. Smith (2008). A new approach to comparing VaR estimation methods. *Journal of Derivatives 16*(2), 54–66.

Perignon, C. and D. Smith (2010). The level and quality of Value-at-Risk disclosure by commercial banks. *Journal of Banking & Finance 34*(2), 362–377.

Pritsker, M. (2006). The hidden dangers of historical simulation. *Journal of Banking & Finance 30*(2), 561–582.

Romano, J. P. (2004). On non-parametric testing, the uniform behaviour of the t-test, and related problems. *Scandinavian Journal of Statistics 31*(4), 567–584.

Rosenblatt, M. (1952). Remarks on a multivariate transformation. *The Annals of Mathematical Statistics 23*(3), 470–472.

Simes, R. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika 73*, 751–754.

Tay, A. S. and K. F. Wallis (2000). Density forecasting: a survey. *Journal of Forecasting 19*(4), 235–254.

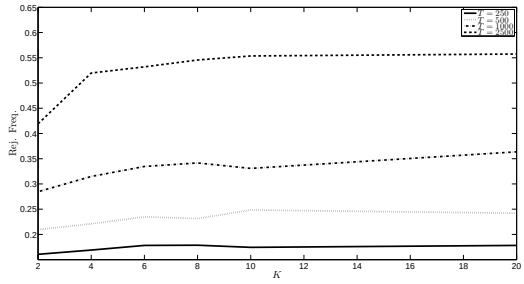
Taylor, J. W. (2020). Forecast combinations for value at risk and expected shortfall. *International Journal of Forecasting 36*(2), 428–441.

Yu, K., Q. Li, A. W. Bergen, R. M. Pfeiffer, P. S. Rosenberg, N. Caporaso, P. Kraft, and N. Chatterjee (2009). Pathway analysis by adaptive combination of p-values. *Genetic Epidemiology 33*(8), 700–709.

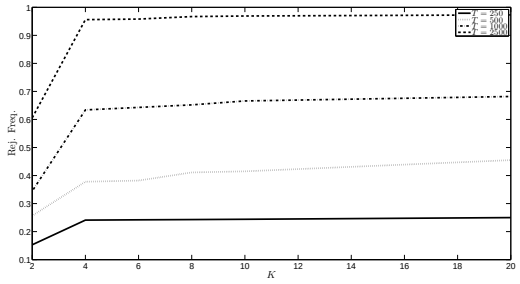
Zaykin, D. V., L. A. Zhivotovsky, P. H. Westfall, and B. S. Weir (2002). Truncated product method for combining p-values. *Genetic Epidemiology 22*(2), 170–185.

Zhang, S., H.-S. Chen, and R. M. Pfeiffer (2013). A combined p-value test for multiple hypothesis testing. *Journal of Statistical Planning and Inference 143*(4), 764–770.

Figure 1: Rejection frequencies for the $\bar{S}_{FSi}$ test with $m = 5$. Returns are generated using the normal-GARCH model (Panel a) or a t-GARCH model (Panel b) and VaR is estimated assuming a normal distribution.



(a) Normal GARCH



(b) t-GARCH

Table 1: **SIZE** The table reports rejection frequencies of multilevel tests at the 5% nominal level and vector of critical levels $(1\%, 2.5\%, 5\%)$. For Panel A, 5,000 time series of length $T \in \{250, 500, 1000, 2500\}$ have been generated according to a GARCH(1,1) model with Student-t innovations, see (30), and VaR is estimated out-of-sample with a rolling window of length 500 using the GARCH-t Model. For Panel B, 5,000 time series of length $T \in \{250, 500, 1000, 2500\}$ have been generated according to a GARCH(1,1) model with Normal innovations and VaR is estimated out-of-sample with a rolling window of length 500 using the Normal Model. For Panel C, 5,000 time series of length $T \in \{250, 500, 1000, 2500\}$ have been simulated from the GLD distribution and VaR is estimated out-of-sample using the Historical Simulation method with a rolling window of length 250. The test statistics we use are combinations of CaViaR tests based on (21).

Panel A: GARCH-t Model and GARCH-t Model out-of-sample

	Existing Tests				$m=1$				$m=5$			
T	PS	PS_MC	Pearson $m=1$	Pearson $m=5$	$\bar{S}_F$	$\bar{S}_{Si}$	$\bar{S}_{FSi}$	$\bar{S}_\Sigma$	$\bar{S}_F$	$\bar{S}_{Si}$	$\bar{S}_{FSi}$	$\bar{S}_\Sigma$
250	0.065	0.042	0.040	0.039	0.056	0.050	0.059	0.062	0.058	0.057	0.055	0.059
500	0.040	0.032	0.058	0.056	0.045	0.049	0.045	0.043	0.049	0.054	0.055	0.061
1000	0.026	0.025	0.053	0.056	0.044	0.052	0.042	0.044	0.050	0.049	0.048	0.051
2500	0.038	0.038	0.040	0.057	0.029	0.033	0.029	0.029	0.055	0.045	0.050	0.056

Panel B: GARCH Normal Model and GARCH Normal Model out-of-sample

	Existing Tests				$m=1$				$m=5$			
T	PS	PS_MC	Pearson $m=1$	Pearson $m=5$	$\bar{S}_F$	$\bar{S}_{Si}$	$\bar{S}_{FSi}$	$\bar{S}_\Sigma$	$\bar{S}_F$	$\bar{S}_{Si}$	$\bar{S}_{FSi}$	$\bar{S}_\Sigma$
250	0.069	0.055	0.034	0.034	0.057	0.050	0.053	0.060	0.057	0.057	0.055	0.065
500	0.041	0.038	0.035	0.048	0.041	0.031	0.039	0.041	0.059	0.053	0.062	0.057
1000	0.037	0.037	0.046	0.044	0.040	0.047	0.039	0.038	0.047	0.052	0.050	0.052
2500	0.046	0.045	0.052	0.057	0.041	0.052	0.043	0.041	0.061	0.059	0.065	0.062

Panel C: GLD Model ($\kappa_3 = -0.8$, $\kappa_4 = 3.5$) and HSVaR out-of-sample

	Existing Tests				$m=1$				$m=5$			
T	PS	PS_MC	Pearson $m=1$	Pearson $m=5$	$\bar{S}_F$	$\bar{S}_{Si}$	$\bar{S}_{FSi}$	$\bar{S}_\Sigma$	$\bar{S}_F$	$\bar{S}_{Si}$	$\bar{S}_{FSi}$	$\bar{S}_\Sigma$
250	0.039	0.027	0.028	0.027	0.035	0.047	0.037	0.036	0.041	0.040	0.039	0.052
500	0.008	0.007	0.037	0.021	0.020	0.021	0.020	0.018	0.032	0.038	0.035	0.041
1000	0.008	0.007	0.042	0.019	0.021	0.025	0.021	0.018	0.035	0.039	0.038	0.039
2500	0.011	0.010	0.041	0.026	0.020	0.027	0.022	0.020	0.038	0.046	0.042	0.041

Table 2: **POWER (ES Test): GARCH-t Model and HSVaR out-of-sample**.

Panel A: Existing Testing Procedures

T	ES	PS	PS_MC	Pearson $m=1$	Pearson $m=5$
250	0.105	0.069	0.059	0.043	0.047
500	0.061	0.021	0.015	0.110	0.126
1000	0.032	0.014	0.016	0.110	0.127
2500	0.051	0.014	0.013	0.183	0.186

Panel B: $m = 1$

T	S_F	S_{Si}	S_{FSi}	S_Σ	$\bar{S}_F$	$\bar{S}_{Si}$	$\bar{S}_{FSi}$	$\bar{S}_\Sigma$
250	0.113	0.100	0.115	0.116	0.103	0.093	0.105	0.096
500	0.126	0.115	0.123	0.127	0.105	0.101	0.105	0.097
1000	0.120	0.137	0.129	0.120	0.104	0.125	0.113	0.102
2500	0.195	0.215	0.202	0.195	0.159	0.189	0.176	0.148

Panel C: $m = 5$

T	S_F	S_{Si}	S_{FSi}	S_Σ	$\bar{S}_F$	$\bar{S}_{Si}$	$\bar{S}_{FSi}$	$\bar{S}_\Sigma$
250	0.142	0.116	0.143	0.147	0.123	0.114	0.127	0.136
500	0.249	0.228	0.249	0.251	0.223	0.196	0.227	0.231
1000	0.293	0.259	0.290	0.285	0.251	0.227	0.252	0.257
2500	0.495	0.449	0.487	0.497	0.455	0.413	0.449	0.460

Note: The table reports rejection frequencies of multilevel tests at the 5% nominal level and vector of critical levels $(1.25\%, 2.5\%, 3.75\%, 5\%)$. ES denotes the regression-based ES test (see for instance Christoffersen, 2012, p. 308-309). 5,000 time series of length $T \in \{250, 500, 1000, 2500\}$ have been generated according to a GARCH(1,1) model with Student-t innovations, and VaR is estimated out-of-sample using the Historical Simulation method with a rolling window of length 500. The test statistics of Panels B and C are combinations of CaViaR tests based on (21).

Table 3: **POWER (ES Test): GARCH-t Model and FHSVaR out-of-sample**.

Panel A: Existing Testing Procedures

	$K=4$			DC		
T	PS_MC	Pearson $m=1$	Pearson $m=5$	PS_MC	Pearson $m=1$	Pearson $m=5$
250	0.121	0.056	0.055	0.270	0.138	0.128
500	0.133	0.140	0.166	0.307	0.106	0.102
1000	0.131	0.166	0.185	0.324	0.060	0.061
2500	0.146	0.250	0.296	0.332	0.444	0.566

Panel B: $m=1$

	$K=4$				DC			
T	$\bar{S}_F$	$\bar{S}_{Si}$	$\bar{S}_{FSi}$	$\bar{S}_\Sigma$	$\bar{S}_F$	$\bar{S}_{Si}$	$\bar{S}_{FSi}$	$\bar{S}_\Sigma$
250	0.156	0.139	0.154	0.154	0.279	0.264	0.281	0.284
500	0.192	0.175	0.191	0.194	0.316	0.304	0.316	0.318
1000	0.213	0.205	0.215	0.205	0.408	0.394	0.412	0.423
2500	0.311	0.321	0.315	0.313	0.471	0.464	0.481	0.478

Panel C: $m=5$

	$K=4$				DC			
T	$\bar{S}_F$	$\bar{S}_{Si}$	$\bar{S}_{FSi}$	$\bar{S}_\Sigma$	$\bar{S}_F$	$\bar{S}_{Si}$	$\bar{S}_{FSi}$	$\bar{S}_\Sigma$
250	0.179	0.166	0.184	0.179	0.274	0.288	0.277	0.295
500	0.235	0.214	0.229	0.235	0.389	0.375	0.384	0.426
1000	0.312	0.285	0.304	0.322	0.515	0.488	0.515	0.566
2500	0.525	0.493	0.505	0.525	0.694	0.657	0.690	0.726

Note: The table reports rejection frequencies of multilevel tests at the 5% nominal level for the vector of critical levels $(1.25\%, 2.5\%, 3.75\%, 5\%)$ and for the double combination (DC) test (25). 5,000 time series of length $T \in \{250, 500, 1000, 2500\}$ have been generated according to a GARCH(1,1) model with normal innovations, and VaR is estimated out-of-sample with a rolling window of length 2500 with the FHS method. The test statistics of Panels B and C are combinations of CaViaR tests based on (21).

Table 4: **POWER: 'serial dependence' in violations**.

Panel A: Existing Testing Procedures

T	PS	PS_MC	Pearson $m=1$	Pearson $m=5$
250	0.776	0.724	0.382	0.379
500	0.751	0.745	0.944	0.990
1000	0.738	0.736	1.000	1.000
2500	0.728	0.727	1.000	1.000

Panel B: $m = 1$

T	S_F	S_{Si}	S_{FSi}	S_Σ	$\bar{S}_F$	$\bar{S}_{Si}$	$\bar{S}_{FSi}$	$\bar{S}_\Sigma$
250	0.999	0.990	0.998	0.998	1.000	0.994	1.000	1.000
500	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
1000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
2500	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Panel C: $m = 5$

T	S_F	S_{Si}	S_{FSi}	S_Σ	$\bar{S}_F$	$\bar{S}_{Si}$	$\bar{S}_{FSi}$	$\bar{S}_\Sigma$
250	0.954	0.962	0.957	0.820	0.968	0.975	0.972	0.935
500	0.994	0.999	0.994	0.992	0.994	1.000	0.994	0.994
1000	0.999	1.000	0.999	0.999	0.999	1.000	0.999	0.999
2500	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Note: The table reports rejection frequencies of multilevel tests at the 5% nominal level and vector of critical levels $(1\%, 2.5\%, 5\%)$. We simulate 5,000 time series of length $T \in \{250, 500, 1000, 2500\}$ for the total number of violations in day t, N_t', according to equations (31). The simulations are such that N_t' has still associated the vector $(\alpha_1, \alpha_2, \alpha_3)'$ but N_t' and N_{t-1}' are not independent. Coverage probabilities are $\alpha_1 = 5\%$, $\alpha_2 = 2.5\%$, and $\alpha_3 = 1\%$. The test statistics of Panels B and C are combinations of CaViaR tests based on (21).

Table 5: **POWER: GLD Model ($\kappa_3 = -0.8$, $\kappa_4 = 3.5$) and Normal VaR out-of-sample**.

Panel A: Existing Testing Procedures

T	PS	PS_MC	Pearson $m = 1$	Pearson $m = 5$
250	0.482	0.419	0.052	0.038
500	0.825	0.809	0.783	0.903
1000	0.998	0.998	0.996	0.999
2500	1.000	1.000	1.000	1.000

Panel B: $m = 1$

T	S_F	S_{Si}	S_{FSi}	S_Σ	$\bar{S}_F$	$\bar{S}_{Si}$	$\bar{S}_{FSi}$	$\bar{S}_\Sigma$
250	0.613	0.553	0.600	0.628	0.603	0.538	0.592	0.615
500	0.921	0.885	0.909	0.926	0.925	0.867	0.914	0.925
1000	1.000	0.999	1.000	1.000	1.000	0.999	1.000	1.000
2500	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Panel C: $m = 5$

T	S_F	S_{Si}	S_{FSi}	S_Σ	$\bar{S}_F$	$\bar{S}_{Si}$	$\bar{S}_{FSi}$	$\bar{S}_\Sigma$
250	0.563	0.528	0.559	0.612	0.576	0.494	0.562	0.627
500	0.852	0.801	0.840	0.878	0.892	0.792	0.873	0.906
1000	0.997	0.991	0.997	0.997	0.999	0.993	0.998	0.999
2500	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Note: The table reports rejection frequencies of multilevel tests at the 5% nominal level and vector of critical levels $(1\%, 2.5\%, 5\%)$. 5,000 time series of length $T \in \{250, 500, 1000, 2500\}$ have been simulated from the GLD distribution and VaR is estimated out-of-sample with a rolling window of length 250 assuming a Normal distribution. Coverage probabilities are $\alpha_1 = 5\%$, $\alpha_2 = 2.5\%$, and $\alpha_3 = 1\%$. The test statistics of Panels B and C are combinations of CaViaR tests based on (21).

Table 6: **POWER (ES Test): GLD Model ($\kappa_3 = -0.8$, $\kappa_4 = 3.5$) and Normal VaR out-of-sample**.

Panel A: Existing Testing Procedures

T	PS	PS_MC	Pearson $m=1$	Pearson $m=5$	U_{ES}	$C_{ES}(1)$	$C_{ES}(5)$	Z_1	Z_2
250	0.453	0.392	0.055	0.034	0.618	0.057	0.069	0.527	0.621
500	0.718	0.702	0.696	0.818	0.863	0.035	0.064	0.776	0.845
1000	0.962	0.960	0.960	0.979	0.990	0.039	0.036	0.960	0.983
2500	1.000	1.000	1.000	1.000	1.000	0.047	0.053	1.000	1.000

Panel B: $m = 1$

	$K=4$				DC			
T	$\bar{S}_F$	$\bar{S}_{Si}$	$\bar{S}_{FSi}$	$\bar{S}_\Sigma$	$\bar{S}_F$	$\bar{S}_{Si}$	$\bar{S}_{FSi}$	$\bar{S}_\Sigma$
250	0.577	0.516	0.566	0.583	0.713	0.670	0.715	0.724
500	0.834	0.770	0.813	0.834	0.908	0.890	0.906	0.915
1000	0.987	0.973	0.985	0.986	0.997	0.994	0.997	0.997
2500	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Panel C: $m = 5$

	$K=4$				DC			
T	$\bar{S}_F$	$\bar{S}_{Si}$	$\bar{S}_{FSi}$	$\bar{S}_\Sigma$	$\bar{S}_F$	$\bar{S}_{Si}$	$\bar{S}_{FSi}$	$\bar{S}_\Sigma$
250	0.541	0.453	0.523	0.581	0.672	0.624	0.668	0.729
500	0.797	0.704	0.778	0.815	0.891	0.854	0.888	0.913
1000	0.977	0.941	0.972	0.980	0.992	0.983	0.993	0.993
2500	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Note: The table reports rejection frequencies of multilevel tests at the 5% nominal level for the vector of critical levels $(1.25\%, 2.5\%, 3.75\%, 5\%)$ and for the double combination (DC) test (25). 5,000 time series of length $T \in \{250, 500, 1000, 2500\}$ have been simulated from the GLD distribution and VaR is estimated out-of-sample with a rolling window of length 2500 assuming a Normal distribution. The tests in the first four columns of Panel A are for $K = 4$. The test statistics of Panels B and C are combinations of CaViaR tests based on (21). U_{ES} and $C_{ES}(m)$ are the unconditional and conditional test of Du and Escanciano (2017) for 5%-ES, respectively. Z_1 and Z_2 denote the tests of Acerbi and Székely (2014).

Table 7: **POWER (ES Test): GARCH-t Model and NormalVaR out-of-sample**.

Panel A: Existing Testing Procedures

T	PS	PS_MC	Pearson $m=1$	Pearson $m=5$	U_{ES}	$C_{ES}(1)$	$C_{ES}(5)$	Z_1	Z_2
250	0.298	0.248	0.060	0.056	0.178	0.098	0.175	0.578	0.136
500	0.419	0.400	0.232	0.301	0.218	0.133	0.243	0.773	0.149
1000	0.670	0.668	0.453	0.611	0.209	0.188	0.354	0.943	0.145
2500	0.965	0.964	0.926	0.969	0.182	0.283	0.573	0.999	0.121

Panel B: $m=1$

	$K=4$				DC			
T	$\bar{S}_F$	$\bar{S}_{Si}$	$\bar{S}_{FSi}$	$\bar{S}_\Sigma$	$\bar{S}_F$	$\bar{S}_{Si}$	$\bar{S}_{FSi}$	$\bar{S}_\Sigma$
250	0.231	0.239	0.240	0.249	0.363	0.369	0.367	0.383
500	0.333	0.330	0.341	0.334	0.497	0.529	0.513	0.529
1000	0.554	0.594	0.575	0.565	0.695	0.741	0.717	0.733
2500	0.930	0.942	0.936	0.932	0.963	0.974	0.968	0.972

Panel C: $m=5$

	$K=4$				DC			
T	$\bar{S}_F$	$\bar{S}_{Si}$	$\bar{S}_{FSi}$	$\bar{S}_\Sigma$	$\bar{S}_F$	$\bar{S}_{Si}$	$\bar{S}_{FSi}$	$\bar{S}_\Sigma$
250	0.229	0.258	0.241	0.235	0.350	0.367	0.355	0.385
500	0.372	0.371	0.378	0.366	0.513	0.537	0.523	0.547
1000	0.639	0.609	0.634	0.615	0.730	0.754	0.748	0.746
2500	0.954	0.951	0.956	0.939	0.973	0.982	0.977	0.973

Note: The table reports rejection frequencies of multilevel tests at the 5% nominal level for the vector of critical levels $(1.25\%, 2.5\%, 3.75\%, 5\%)$ and for the double combination (DC) test (25). 5,000 time series of length $T \in \{250, 500, 1000, 2500\}$ have been generated according to a GARCH(1,1) model with Student-t innovations, and VaR is estimated out-of-sample with a rolling window of length 2500 assuming a Normal distribution. The tests in the first four columns of Panel A are for $K=4$. The test statistics of Panels B and C are combinations of CaViaR tests based on (21). U_{ES} and $C_{ES}(m)$ are the unconditional and conditional test of Du and Escanciano (2017) for 5%-ES, respectively. Z_1 and Z_2 denote the tests of Acerbi and Székely (2014).

Table 8: **POWER (ES Test) : GARCH-normal Model and NormalVaR out-of-sample**.

Panel A: Existing Testing Procedures

T	PS	PS_MC	Pearson $m=1$	Pearson $m=5$	U_{ES}	$C_{ES}(1)$	$C_{ES}(5)$	Z_1	Z_2
250	0.126	0.102	0.046	0.047	0.136	0.110	0.171	0.087	0.108
500	0.113	0.107	0.133	0.157	0.150	0.144	0.253	0.107	0.111
1000	0.107	0.105	0.155	0.192	0.170	0.160	0.330	0.135	0.126
2500	0.097	0.093	0.204	0.245	0.109	0.295	0.553	0.208	0.089

Panel B: $m=1$

	$K=4$				DC			
T	$\bar{S}_F$	$\bar{S}_{Si}$	$\bar{S}_{FSi}$	$\bar{S}_\Sigma$	$\bar{S}_F$	$\bar{S}_{Si}$	$\bar{S}_{FSi}$	$\bar{S}_\Sigma$
250	0.156	0.134	0.154	0.155	0.303	0.284	0.303	0.303
500	0.176	0.163	0.176	0.175	0.336	0.317	0.337	0.335
1000	0.214	0.195	0.211	0.210	0.381	0.355	0.383	0.382
2500	0.266	0.257	0.267	0.259	0.458	0.435	0.461	0.452

Panel C: $m=5$

	$K=4$				DC			
T	$\bar{S}_F$	$\bar{S}_{Si}$	$\bar{S}_{FSi}$	$\bar{S}_\Sigma$	$\bar{S}_F$	$\bar{S}_{Si}$	$\bar{S}_{FSi}$	$\bar{S}_\Sigma$
250	0.169	0.158	0.169	0.166	0.315	0.307	0.313	0.337
500	0.221	0.204	0.221	0.221	0.387	0.366	0.388	0.426
1000	0.319	0.286	0.315	0.324	0.494	0.460	0.493	0.535
2500	0.521	0.477	0.520	0.529	0.709	0.666	0.708	0.739

Note: The table reports rejection frequencies of multilevel tests at the 5% nominal level for the vector of critical levels $(1.25\%, 2.5\%, 3.75\%, 5\%)$ and for the double combination (DC) test (25). 5,000 time series of length $T \in \{250, 500, 1000, 2500\}$ have been generated according to a GARCH(1,1) model with Normal innovations and VaR is estimated out-of-sample with a rolling window of length 2500 assuming a Normal distribution. The tests in the first four columns of Panel A are for $K=4$. The test statistics of Panels B and C are combinations of CaViaR tests based on (21). U_{ES} and $C_{ES}(m)$ are the unconditional and conditional test of Du and Escanciano (2017) for 5%-ES, respectively. Z_1 and Z_2 denote the tests of Acerbi and Székely (2014).

Table 9: **P-Values (×100) of combination and benchmark tests for the IVV ETF for the periods 2009-2016 and 2010-2014 (daily data).**

VaR: Hist 1Y

	Existing Tests	ES Test	PS	PS_MC	Pearson $m=1$	Pearson $m=5$		
	2009-2016	23.698	33.745	33.700	20.500	5.100		
	2010-2014	36.888	80.851	80.800	50.700	21.200		
Multilevel Tests	S_F	S_{Si}	S_{FSi}	S_Σ	$\bar{S}_F$	$\bar{S}_{Si}$	$\bar{S}_{FSi}$	$\bar{S}_\Sigma$
2009-2016, $m=1$	1.370	3.668	1.756	2.050	4.492	2.416	1.370	2.290
2009-2016, $m=5$	0.002	0.008	0.002	0.002	0.008	0.002	0.002	0.002
2010-2014, $m=1$	41.175	50.745	43.141	47.465	64.361	49.889	41.175	49.039
2010-2014, $m=5$	0.010	0.132	0.020	0.020	0.140	0.026	0.010	0.018

VaR: MC

	Existing Tests	ES Test	PS	PS_MC	Pearson $m=1$	Pearson $m=5$		
	2009-2016	10.100	41.418	40.700	23.700	6.200		
	2010-2014	30.44	69.924	70.500	62.900	19.200		
Multilevel Tests	S_F	S_{Si}	S_{FSi}	S_Σ	$\bar{S}_F$	$\bar{S}_{Si}$	$\bar{S}_{FSi}$	$\bar{S}_\Sigma$
2009-2016, $m=1$	6.378	8.316	6.736	8.620	10.450	8.988	6.378	9.448
2009-2016, $m=5$	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002
2010-2014, $m=1$	50.141	49.915	49.859	57.607	64.361	58.223	50.141	59.385
2010-2014, $m=5$	0.010	0.078	0.012	0.012	0.084	0.016	0.006	0.010

VaR: Parametric

	Existing Tests	ES Test	PS	PS_MC	Pearson $m=1$	Pearson $m=5$		
	2009-2016	3.355	1.607	2.000	5.800	0.200		
	2010-2014	17.945	5.682	5.700	24.700	0.800		
Multilevel Tests	S_F	S_{Si}	S_{FSi}	S_Σ	$\bar{S}_F$	$\bar{S}_{Si}$	$\bar{S}_{FSi}$	$\bar{S}_\Sigma$
2009-2016, $m=1$	2.310	4.906	2.788	1.450	4.752	1.884	2.310	1.338
2009-2016, $m=5$	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002
2010-2014, $m=1$	16.534	13.414	15.740	10.402	12.522	11.026	16.534	9.422
2010-2014, $m=5$	0.002	0.012	0.002	0.002	0.012	0.002	0.002	0.002

Note: We backtest VaRs provided by Bloomberg and calculated using the Historical Simulation with a window of 1 year (Hist 1Y), the Monte Carlo method (MC) and the parametric method (Parametric). '$m=1$' and '$m=5$' denote the combinations of CaViaR tests based on (21). Coverage probabilities are $\alpha_1 = 5\%$, $\alpha_2 = 3.75\%$, $\alpha_3 = 2.5\%$, and $\alpha_4 = 1.25\%$.

Table 10: **P-Values (×100) of combination and benchmark tests for the BCYIF ETF for the periods 2009-2016 and 2010-2014 (daily data).**

VaR: Hist 1Y

Existing Tests	ES Test	PS	PS_MC	Pearson $m=1$	Pearson $m=5$		
2009-2016	0.002	13.987	13.800	0.300	0.100		
2010-2014	0.006	9.349	11.100	1.400	0.600		
Multilevel Tests	S_F	S_{Si}	S_{FSi}	S_Σ	$\bar{S}_F$	$\bar{S}_{Si}$	$\bar{S}_{FSi}$	$\bar{S}_\Sigma$
2009-2016, $m=1$	0.022	0.022	0.024	0.032	0.022	0.030	0.022	0.034
2009-2016, $m=5$	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002
2010-2014, $m=1$	0.036	0.068	0.034	0.046	0.072	0.042	0.036	0.050
2010-2014, $m=5$	0.002	0.012	0.002	0.002	0.012	0.002	0.002	0.002

VaR: MC

Existing Tests	ES Test	PS	PS_MC	Pearson $m=1$	Pearson $m=5$		
2009-2016	6.085	86.214	86.100	5.500	0.700		
2010-2014	7.426	62.078	61.600	8.900	5.400		
Multilevel Tests	S_F	S_{Si}	S_{FSi}	S_Σ	$\bar{S}_F$	$\bar{S}_{Si}$	$\bar{S}_{FSi}$	$\bar{S}_\Sigma$
2009-2016, $m=1$	0.042	0.098	0.042	0.098	0.102	0.088	0.042	0.134
2009-2016, $m=5$	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002
2010-2014, $m=1$	0.024	0.020	0.024	0.036	0.020	0.028	0.024	0.036
2010-2014, $m=5$	0.006	0.078	0.008	0.006	0.084	0.010	0.006	0.006

VaR: Parametric

Existing Tests	ES Test	PS	PS_MC	Pearson $m=1$	Pearson $m=5$		
2009-2016	0.291	3.961	4.100	2.100	0.600		
2010-2014	0.052	15.488	16.800	8.100	6.900		
Multilevel Tests	S_F	S_{Si}	S_{FSi}	S_Σ	$\bar{S}_F$	$\bar{S}_{Si}$	$\bar{S}_{FSi}$	$\bar{S}_\Sigma$
2009-2016, $m=1$	0.048	0.190	0.062	0.070	0.202	0.082	0.048	0.078
2009-2016, $m=5$	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002
2010-2014, $m=1$	0.028	0.030	0.026	0.036	0.030	0.032	0.028	0.038
2010-2014, $m=5$	0.010	0.100	0.016	0.012	0.108	0.016	0.010	0.010

Note: We backtest VaRs provided by Bloomberg and calculated using the Historical Simulation with a window of 1 year (Hist 1Y), the Monte Carlo method (MC) and the parametric method (Parametric). '$m=1$' and '$m=5$' denote the combinations of CaViaR tests based on (21). Coverage probabilities are $\alpha_1 = 5\%$, $\alpha_2 = 3.75\%$, $\alpha_3 = 2.5\%$, and $\alpha_4 = 1.25\%$.

Table 11: **P-Values (×100) of combination and benchmark tests for the EWH ETF for the periods 2009-2016 and 2010-2014 (daily data).**

VaR: Hist 1Y

Existing Tests	ES Test	PS	PS_MC	Pearson $m=1$	Pearson $m=5$		
2009-2016	0.000	95.982	95.600	4.600	2.400		
2010-2014	0.000	81.379	81.100	1.800	1.900		
Multilevel Tests	S_F	S_{Si}	S_{FSi}	S_Σ	$\bar{S}_F$	$\bar{S}_{Si}$	$\bar{S}_{FSi}$	$\bar{S}_\Sigma$
2009-2016, $m=1$	0.136	0.234	0.144	0.302	0.252	0.268	0.136	0.438
2009-2016, $m=5$	0.004	0.006	0.002	0.006	0.006	0.006	0.004	0.008
2010-2014, $m=1$	0.062	0.106	0.062	0.162	0.112	0.124	0.062	0.214
2010-2014, $m=5$	0.014	0.034	0.012	0.020	0.034	0.016	0.012	0.018

VaR: MC

Existing Tests	ES Test	PS	PS_MC	Pearson $m=1$	Pearson $m=5$		
2009-2016	56.586	0.631	0.600	0.500	0.300		
2010-2014	48.290	4.943	4.800	0.500	0.800		
Multilevel Tests	S_F	S_{Si}	S_{FSi}	S_Σ	$\bar{S}_F$	$\bar{S}_{Si}$	$\bar{S}_{FSi}$	$\bar{S}_\Sigma$
2009-2016, $m=1$	0.032	0.098	0.040	0.026	0.102	0.032	0.032	0.026
2009-2016, $m=5$	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002
2010-2014, $m=1$	0.022	0.030	0.026	0.020	0.030	0.026	0.022	0.020
2010-2014, $m=5$	0.006	0.028	0.008	0.006	0.028	0.006	0.006	0.004

VaR: Parametric

Existing Tests	ES Test	PS	PS_MC	Pearson $m=1$	Pearson $m=5$		
2009-2016	0.000	14.957	13.8	0.500	0.200		
2010-2014	0.000	69.190	69.600	0.300	1.200		
Multilevel Tests	S_F	S_{Si}	S_{FSi}	S_Σ	$\bar{S}_F$	$\bar{S}_{Si}$	$\bar{S}_{FSi}$	$\bar{S}_\Sigma$
2009-2016, $m=1$	0.016	0.020	0.016	0.020	0.020	0.022	0.016	0.024
2009-2016, $m=5$	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002
2010-2014, $m=1$	0.010	0.010	0.004	0.010	0.010	0.010	0.010	0.010
2010-2014, $m=5$	0.002	0.012	0.002	0.002	0.012	0.002	0.002	0.002

Note: We backtest VaRs provided by Bloomberg and calculated using the Historical Simulation with a window of 1 year (Hist 1Y), the Monte Carlo method (MC) and the parametric method (Parametric). '$m=1$' and '$m=5$' denote the combinations of CaViaR tests based on (21). Coverage probabilities are $\alpha_1 = 5\%$, $\alpha_2 = 3.75\%$, $\alpha_3 = 2.5\%$, and $\alpha_4 = 1.25\%$.