Predictive identification robust confidence sets with application to tail risk measures[*]

Lynda Khalaf
Department of Economics, Carleton University
Centre for Monetary Financial Economics (co-director), Carleton University

Arturo Leccadito
Department of Economics, Statistics and Finance, University of Calabria
LFIN/LIDAM, UCLouvain

Debora Loccisano
Department of Economics, Carleton University

October 29, 2024

Abstract

This paper proposes a novel method to construct confidence sets for predictive measures, that do not require identification and can be finite-sample exact. First, a joint robust confidence region for parameters that are hard to identify is constructed through the inversion of an out-of-sample specification adequacy test. This set is then projected to construct simultaneous confidence sets for any collection of measures over multiple paths. These sets provide a unified solution to confidence estimation and out-of-sample validation, without compounding type I error. We focus on tail risk metrics including Value at Risk, Expected Shortfall, and Expectile-VaR, allowing for multiple thresholds. Simultaneity further addresses joint elicitability concerns. An illustrative analysis on GARCH models is conducted, through Monte Carlo simulations and an empirical evaluation of an exchange-traded fund that tracks the technology sector.

Keywords: Identification-robust inference, Value at Risk, Expected Shortfall, Expectile-VaR, back-testing

[*] This work was supported by the Social Sciences and Humanities Research Council of Canada and the Natural Sciences and Engineering Research Council of Canada.

1 Introduction

The problem of constructing confidence sets for predictive quantities arises in a variety of econometric contexts. Important special cases include forecasts, yet as emphasized by Dufour et al. (1994), predictive quantities can be defined more generally, with an out-of-sample perspective. This paper concerns confidence set (CS) estimation of such quantities, with focus on tail risk.

Broadly defined, tail risk measures are metrics that quantify predicted exposure to extreme events beyond given thresholds. Conventional choices for such purposes, guided by the Basel agreements since 2013, include: *value-at-Risk* (VaR), defined as the highest possible loss at a given probability threshold; *expected shortfall* (ES), defined as the average loss given that loss exceeds VaR; and, more recently, *expectile-VaR* (EVaR), where expectiles are defined as least-squares analogues of quantiles.[1] Financial regulations often require large institutions to *back-test* such measures, which refers to formal comparisons of predicted against realized values to evaluate the performance of a risk forecasting model or procedure.[2] Importantly, regulations permit reliance on several measures.

The predictive quantities under consideration are commonly derived given some time series specification, although non-parametric approaches are sometimes considered. Our emphasis is on three issues: (i) identifiability of underlying model parameters; (ii) the multiplicity of available measures that can convey different and possibly conflicting assessments of risk; and (iii) the scarcity of confidence inference relative to back-testing. For example, in parametric GARCH models, popular measures (reviewed in Section 2) can be expressed

[1] See Tsay (2010) or Christoffersen (2016) for the definition of VaR and ES, and on expectiles, see Newey and Powell (1987) and Bellini and Di Bernardino (2017). A survey with formal definitions of a large spectrum of available measures is available in He et al. (2022). The Basel accords and technical requirements are publicly available from the web-site of the Bank of International Settlements.

[2] The literature on back-testing VaR is well established [see *e.g.* Kupiec (1995), Christoffersen (1998), Christoffersen and Pelletier (2004), Berkowitz et al. (2011), Leccadito et al. (2014)]; work on back-testing ES is more recent [see *e.g.* Du and Escanciano (2017), Argyropoulos and Panopoulou (2019), Banulescu-Radu et al. (2021), Hoga and Demetrescu (2023)]; very few procedures are available to back-test expectiles; see *e.g.* Bellini et al. (2019). Further work on back-testing is reviewed below.

as a function of model parameters and the cut-off point from the hypothesized error distribution. A semi-parametric alternative like the Filtered Historical Simulation replaces that cut-off point by an empirical conditional quantile. While concerns associated with (i) have escaped attention, work that addresses (ii) appears in mainly two instances. First, considering several thresholds has led to important efficiency gains for back-tests of a single metric; see *e.g.* Perignon and Smith (2008), Leccadito et al. (2014), Kratz et al. (2018), Wang and Zhao (2016), Khalaf et al. (2021), Du et al. (2023). Second, the recognition that ES and VaR are jointly elicitable[3] whereas ES - individually - is not [Fissler et al. (2016), Fissler and Ziegel (2016), Nolde and Ziegel (2017)] has spurred work on joint models and back-tests for several metrics; see *e.g.* Patton et al. (2019), Bayer and Dimitriadis (2020), Taylor (2020), Taylor (2022), Dimitriadis and Halbleib (2022), Fissler and Hoga (2023), and Dimitriadis et al. (2023). The methods we propose are related to Patton et al. (2019), in that we directly fit the measures in themselves. However, we depart from minimum distance to avoid imposing identification.

There is a shortage of works on confidence set inference, relative to back-testing. The few available exceptions are for the most part restricted to a single measure, and motivated asymptotically, imposing identification; see *e.g.* the survey by Nieto and Ruiz (2016) and Christoffersen and Goncalves (2005), Chan et al. (2007), Lan et al. (2010), Wang and Zhao (2016), Hoga (2019a), Hoga (2019b), He et al. (2022), Davison et al. (2023). Out-of-sample assessments are often conducted on (relatively) short periods; for some financial assets (*e.g.* new instruments), a long series is not even available. Yet the asymptotic distributions of relevant statistics require much larger samples. Finite sample distortions are actually documented with back-tests; see *e.g.* Barendse et al. (2021), Hurlin et al. (2017), and Dimitriadis et al. (2023). Apart from very few distribution-free tests [Christoffersen and Pelletier (2004), Berkowitz et al. (2011), Leccadito et al. (2014), Kratz et al. (2018), Khalaf

[3] A vector of functionals is elicitable if there exists a loss function for which the vector provides the unique minimizer of its expectation.

et al. (2021)], most back-tests require consistent estimation - and thus identification - of underlying model parameters. However, predictive quantities, including risk measures, can inherit the identification properties of the underlying models, which is not granted in many cases including GARCH models.

This paper proposes novel simultaneous CS procedures to address the above concerns. In the forecasting context, the simultaneous approach may be traced to Jordà and Marcellino (2010). Wang and Zhao (2016) propose simultaneous bootstrap-based CSs for VaR at multiple thresholds; Xu (2016) proposes joint non-parametric procedures for VaR and ES. Our methodology is distinct along three important dimensions. *First*, our targets are several different paths of forecasts over several time periods, for a single threshold or for multiple ones; any collection of metrics can be considered. *Second*, we derive CSs that embed an out-of sample goodness-of-fit check, which eschews the major critique associated with parametrization. *Third*, proposed CSs are valid without imposing identification. Our approach thus suggests an avenue towards an identification-robust perspective on elicitability. To do this, we proceed by projecting a joint confidence region for the parameters of the underlying econometric model that can be: (i) unbounded when identification is weak, and (ii) empty, when the specification in question lacks fits. The rationale, in line with the identification-robust literature[4], can be summarized as follows. When a model is hard to identify, a joint confidence region can still be obtained, often via test inversion, for some parameters. For *any* collection of given functions of these parameters, the projection of this region, which is its image by each function in question, will yield reliable simultaneous CSs. The predictive quantities under consideration can be defined this way, for *any* collection of measures that depend on the same history and deep parameters.

As the main contribution of this paper, we introduce joint predictive confidence regions

[4] This literature, which is now considerable, is mostly in-sample based and can be traced back to Dufour (1997) and Stock and Wright (2000); for recent references in financial econometrics, see Beaulieu et al. (2013), Beaulieu et al. (2014), Beaulieu et al. (2023) and references therein.

for this purpose, which we propose to derive by inverting back-tests, or more generally, any model assessment test that has a predictive out-of-sample basis. This will make fruitful use of the vast literature on back-tests and will yield assessments of estimation uncertainty that fulfill regulatory requirements. Modifications to the latter by the Basel Committee on Banking since 2019 emphasize "forward-looking assessment" of vulnerabilities [Basel Committee on Banking Supervision (2019)]. Our approach builds on this fact.

Inverting a test entails collecting the parameter values that are not rejected at a given level. It is thus important to be clear about what we mean by inverting a back-test, since back-tests typically rely on estimated model parameters. Instead, we propose to compute the back-test for each model parameter value that is under test, which will assess the out-of-sample fit given this specific parameter. This will check the adequacy of the risk specification jointly with each hypothesized parameter value. Inversion will thus first inform on the model parameters, and from thereon, the projections will inform on any desired collection of risk measures. Hence, even if a VaR back-test is inverted, simultaneous sets for any desired collection of measures can be obtained. The choice of back-test to be inverted is thus not limited to the specific risk measures of interest.

Given the vast literature on back-tests, it is not possible to single out uniformly the most suitable tests to invert. There is no consensus in this literature on which test is preferable for their standard application. Available simulation studies are conducted for the latter purpose, rather than for inference on model parameters. It is thus important to document the size and power properties of our proposed application of back-testing, which we do, as described below. While any back-test with reasonable finite sample properties can be inverted, our approach provides a novel avenue for synthesizing the information in available exact tests, leading to exact confidence inference for measures beyond VaR and ES, based on tractable criteria. This covers many measures on which inferential theory is scarce, complicated or seems to necessitate large assessment samples, including expectiles.[5]

Our second contribution is a detailed analysis of the GARCH case. Despite the popularity and simplicity of this model, available CSs are justified imposing regularity assumptions that typically rule out boundary values, for example at the unit boundary, although integrated GARCH models are broadly used in risk analysis [Longerstaey and Spencer (1996), Tsay (2010)]. Furthermore, the root cancellation problem intervenes pervasively in the GARCH case [Andrews (2001)]. For inversion purposes, we consider the multi-level Pearson test for VaR of Leccadito et al. (2014) which is an implicit test for ES, and the back-tests for ES of Du and Escanciano (2017). Both tests are based on some form of cumulative violations [Khalaf et al. (2021)], yet the former is an exact test. Following Patton et al. (2019), our GARCH estimates thus aim for the best-fitting tail risk (rather than volatility) forecasts, yet we view the best-fitting set as the least-rejected one.

A Monte Carlo simulation study is conducted to document the usefulness of our proposed test inversion procedure. Importantly, we study (i) the implications of weak identification, (ii) the information content of parametric versus non-parametric violation aggregators, and (iii) the usefulness of variance targeting [Engle and Mezrich (1996), Noureldin et al. (2014)] to manage identification problems when persistence in the conditional variance is high.

Our third contribution is an empirical application on the Technology Select Sector SPDR Fund ETF (Bloomberg ticker: XLK). We consider a daily prediction sample from June 22 to October 23, 2020 to track the effects of the COVID shock. The considered sector is particularly interesting given the unexpected demands on technology that resulted from social distancing restrictions. We produce simultaneous confidence bands for the VaR, ES and EVaR forecasts, within a parametric GARCH model and its filtered historical simulation [FHS] semi-parametric counterpart [(Barone-Adesi et al., 1999)]. Results reveal

[5] In addition the above cited references on expectiles, see Nolde and Ziegel (2017) for a discussion of related complications.

root cancellation and boundary problems with GARCH parameters. Nevertheless, confidence bands remain relatively informative on the risk measures themselves, which is partly driven by variance-targeting. Overall, findings quantify the vulnerability of the considered fund relative to "flash events" [see Lettau and Madhavan (2018)] including vaccine breakthroughs and major market-shaking events in the artificial intelligence sector.

The paper is organized as follows. Section 2 provides our proposed statistical framework and general inference strategy. Section 3 discusses the GARCH case with a Monte Carlo and empirical analysis. Section 4 concludes.

2 Framework and inference strategy

In this section, we present the general set-up and inference targets. We consider the problem of estimating CSs for tail measures that are associated with a time series $\{R_t\}_{t=1,\ldots,T}$. For risk analysis, $\{R_t\}$ is a sample of portfolio returns, and tail measures are corresponding end-of-sample based parameters. We thus define the prediction sample with reference to the index set $\{n+1,\ldots,T\}$ where n is predetermined. Let $\mathcal{F}_t$ represent the information set up to time t, which may be restricted to the history of returns $\{R_\tau : \tau \leq t\}$ or may include additional predictors. We make the following assumptions.

Assumption 1 - Inference targets. *The underlying data generating process (**DGP**) is sufficiently specified, up to a parameter $\theta \in \Omega$, so that the collection of the J paths of interest, which we denote by $\Lambda^j_{t,q}$, can be formulated as given functions of θ and $\mathcal{F}_{t-1}$:*

$$\Lambda^j_{t,q} := g^j_{q,t}(\theta, \mathcal{F}_{t-1}), \quad t = n+1,\ldots,T, \qquad j = 1,\ldots,J, \tag{1}$$

where $g^j_{q,t}(\cdot)$ is a known (scalar) function.

Assumption 2 - Specification adequacy. *The adequacy of (1) is formulated as*

$$f_t(R_t, \theta, \mathcal{F}_{t-1}) = u_t, \quad t = n+1,\ldots,T, \tag{2}$$

where $f_t(\cdot)$ is a known function and u_t is a random vector with zero expectation conditionally or unconditionally, that is, $E(u_t|\mathcal{F}_{t-1}) = 0$ or $E(u_t) = 0$.

The above framework does not require the identification of θ, nor it is fully parametric. Assumption 1 defines $\Lambda^j_{q,t}$ as a *predictive* quantity through its reference to the prediction sample. Assumption 2 defines end-of-sample stability through the function $f_t(\cdot)$, that generates *predictive* residuals in line with a wide spectrum of econometric models and methods [see *e.g.* Dufour et al. (1994)]. We thus denote $f_t(\cdot)$ as the *residual maker* function. Standard forms of $f_t(\cdot)$ (with formal definitions in Section 2) capture the frequency or time series behavior of violations, where the latter is defined as a binary variable, which takes value 1 if an observed data point falls beyond the considered tail threshold. The disturbances u_t are not assumed to be *i.i.d.* although such an assumption is appropriate in some contexts. These include a popular distribution-free definition of adequacy, defined below in 2, that has led to the above cited broad spectrum of exact risk measure back-tests. Adequacy is usually defined as a conditional or unconditional zero-expectation assumption on the residual maker function.

Broadly, tail risk measures which we formalize through Assumption 1 are functionals that describe the predicted tail of the conditional (or unconditional) distribution of R_t, beyond a given threshold $q \in [0, 1]$. We will refer to q as the tail *threshold* (rather than tail level) and reserve the terms *level* and *coverage* with reference to statistical tests and CSs. Popular tail measures which we consider include the following.

1. The one step ahead VaR, denoted $VaR_{t,q}$, defined as the negative return at time t such that the conditional probability to observe a more extreme return is q:

$$\mathrm{P}\left[\left(R_t < -VaR_{t,q}\right)|\mathcal{F}_{t-1}; \theta\right] = q, \quad t = n+1, \ldots, T; \tag{3}$$

2. The one step ahead ES, denoted $ES_{t,q}$, defined as the conditional tail expectation at

time t that assigns equal weight to all quantiles below q:

$$ES_{t,q} = -E\left[R_t|\left(R_t < -VaR_{t,q}\right), \mathcal{F}_{t-1}; \theta\right] = \frac{1}{q}\int_0^q VaR_{t,u}\, du, \quad t = n+1, \ldots, T; \tag{4}$$

3. The expectile-VaR, denoted $EVaR_{t,q}$, which is similar to a conditional quantile but is determined by tail expectations rather than tail probabilities:

$$EVaR_{t,q} = -\underset{x\in\mathbb{R}}{\operatorname{argmin}} E\left[\eta_q(R_t - x) - \eta_q(R_t)|\mathcal{F}_{t-1}; \theta\right], \quad t = n+1, \ldots, T, \tag{5}$$

where for any real scalar y, $\eta_q(y) = |q - \mathbb{1}(y \leq 0)|y^2$ and $\mathbb{1}(A)$ denotes the indicator function of any event A [Bellini et al. (2014), Daouia et al. (2020)].

Parametric distributions are broadly adopted that allow one to express these measures as in (1). Our approach is also compatible with semi-parametric modeling, including the FHS approach, the extreme-value theory based semi-parametric model of D'Innocenzo et al. (2023) or the dynamic specifications of Patton et al. (2019). The properties of (3), (4) and (5) are discussed at length in the literature that we have reviewed above. In particular, $EVaR_{t,q}$ is elicitable if $q \in (0, 1/2]$, and $ES_{t,q}$ and $VaR_{t,q}$ are jointly elicitable although $ES_{t,q}$, viewed on its own, is not. Such concerns have attracted concrete interest in joint testing, which motivates our simultaneous inference approach. In this context, the procedure that we propose for inference on the collection of $\Lambda^j_{t,q}$, $j = 1, ..., J$, $t = n+1, \ldots, T$, is summarized in Algorithm 1.

The least rejected values apply the Hodges-Lehman (Hodges and Lehmann, 1963) estimation method. The estimator of θ may not be unique as multiple values of θ can correspond to the largest p-value. The above algorithm abstracts from the choice of which test to invert, which is user defined, as is typically the case with statistical objective functions. Furthermore, our methodology easily accommodates a combined criterion [see *e.g.* Khalaf et al. (2021)] based on several back-tests. We next formalize the above steps, beginning with a statement of the null hypotheses underlying the tests we invert, conforming with (2).

Algorithm 1

```
1: Test inversion. A back-test of (2) is inverted over $\theta \in \Omega$, using some conventional form for $f_t(\cdot)$. This step, implemented through a numerical search within $\Omega$, will recover the set of $\theta$ values, denoted thereafter as $\mathrm{CS}(\theta; \tilde{\alpha})$, that are not rejected at level $\tilde{\alpha}$.
2: Projections. Obtain the image of $\mathrm{CS}(\theta; \tilde{\alpha})$ by each of the $g^j_{q,t}(\cdot)$, as defined in (1), which will yield the desired collection of confidence set estimates, denoted as $\mathrm{CS}^j_{t,q}(\Lambda^j_{t,q}; \tilde{\alpha})$. The infimum and supremum of each of $\mathrm{CS}^j_{t,q}(\Lambda^j_{t,q}; \tilde{\alpha})$, denoted as $\mathrm{L}^j_{t,q}(\Lambda^j_{t,q}; \tilde{\alpha})$ and $\mathrm{U}^j_{t,q}(\Lambda^j_{t,q}; \tilde{\alpha})$, further yield a collection of simultaneous confidence bands.
3: Least rejected values. Within $\mathrm{CS}(\theta; \tilde{\alpha})$, the value(s) of $\theta$ that are associated with the largest back-test $p$-value yield an estimator of $\theta$. The associated values of the measures through each of the $g^j_{q,t}(\cdot)$ yield estimators for the $\Lambda^j_{t,q}$s.
```

2.1 Hypotheses and residual making functions

Usual back-tests assess the following conditional or unconditional null hypotheses

$$\mathcal{H}^*_c : E\left[f_t(R_t, \theta, \mathcal{F}_{t-1})|\mathcal{F}_{t-1}\right] = 0, \text{ for some } \theta \in \Omega, \; t = n+1, \ldots, T, \tag{6}$$

$$\mathcal{H}^*_u : E\left[f_t(R_t, \theta, \mathcal{F}_{t-1})\right] = 0, \text{ for some } \theta \in \Omega, \; t = n+1, \ldots, T. \tag{7}$$

In contrast, our proposed *inversion step* is associated with

$$\mathcal{H}_c\left(\theta_0\right) : E\left[f_t(R_t, \theta_0, \mathcal{F}_{t-1})|\mathcal{F}_{t-1}\right] = 0, \text{ for } \theta_0 \text{ known}, \; t = n+1, \ldots, T, \tag{8}$$

$$\mathcal{H}_u\left(\theta_0\right) : E\left[f_t(R_t, \theta_0, \mathcal{F}_{t-1})\right] = 0, \text{ for } \theta_0 \text{ known}, \; t = n+1, \ldots, T. \tag{9}$$

This distinction is important. However, as in Beaulieu et al. (2014) and in line with the literature on weak identification, $\mathcal{H}^*_c$ or $\mathcal{H}^*_u$ can be rejected at the $\tilde{\alpha}$ level without imposing identification if the outcome of our inversion exercise yields $\mathrm{CS}(\theta; \tilde{\alpha}) = \varnothing$. In our simulations below, we show that our approach is highly informative on θ within the considered DGP, in addition to the traditional application of back-tests.

We also consider the subset inference problem when $\theta = (\delta', \nu')'$, where δ is the subvector that contains hard to identify parameters, and ν can be partialled-out through an estimator given δ, denoted by $\hat{\nu}\left(\delta\right)$, obtained from some pre-prediction or training sample, *i.e.* using data up to time $t-1$ or n. For example, in a Student-t GARCH model, the degrees-of-freedom parameter can be identified when the GARCH parameters are fixed. The above procedure is then modified as follows. The inversion step is conducted over δ, whereby ν is re-estimated imposing, in turn, each value of δ under the null. The associated null hypotheses correspond to the following, where the superscript s refers to subset testing: there exists $\nu_0 \equiv \nu(\delta_0)$ such that with $\hat{\nu}\,(\delta_0) \xrightarrow{P} \nu_0$ we have:

$$\mathcal{H}_c^s\,(\delta_0) : E\left[f_t(R_t, (\delta_0', {\nu_0}')', \mathcal{F}_{t-1})|\mathcal{F}_{t-1}\right] = 0, \text{ for } \delta_0 \text{ known}, t = n+1, \ldots, T, \tag{10}$$

$$\mathcal{H}_u^s\,(\delta_0) : E\left[f_t(R_t, (\delta_0', {\nu_0}')', \mathcal{F}_{t-1})\right] = 0, \text{ for } \delta_0 \text{ known}, t = n+1, \ldots, T. \tag{11}$$

Thus defined, ν_0 will likely depend on δ_0. An empty CS outcome in this case can also be interpreted as evidence against $\mathcal{H}_c^*$ or $\mathcal{H}_u^*$.

We now turn to the discussion of the considered forms for $f_t(\cdot)$ in (2) that are conformable with tail risk measures. Our objective is to provide a useful unification to this broad and diverse literature. Typically, popular back-tests rely on the following:

1. centered *scalar violation* (SV) *indicators* at the $100 \times q\%$ threshold, defined as:

$$f_{t,q}^{\mathrm{SV}}(R_t, \theta, \mathcal{F}_{t-1}) := \mathbb{1}(R_t \leq -VaR_{t,q}) - q := I_{t,q}(R_t, \theta, \mathcal{F}_{t-1}) - q; \tag{12}$$

2. centered *integrated violation* (IV) *aggregators* at the $100 \times q\%$ threshold, defined as:

$$f_{t,q}^{\mathrm{IV}}(R_t, \theta, \mathcal{F}_{t-1}) := \frac{1}{q}\int_0^q I_{t,s}(R_t, \theta, \mathcal{F}_{t-1})ds - \frac{q}{2} = \frac{1}{q}(q - w_t)\mathbb{1}(w_t \leq q) - \frac{q}{2}; \tag{13}$$

where $w_t = F_t(R_t|\mathcal{F}_{t-1})$ and $F_t(\cdot|\mathcal{F}_{t-1})$ is the conditional cumulative distribution function of R_t; see Du and Escanciano (2017) for further reference;

3. centered *vector violation* (VV) *indicators*, defined as the vector-valued functions

$$f_{t,\mathbf{q}}^{\mathrm{VV}}(R_t, \theta, \mathcal{F}_{t-1}, K) := \left(f_{t,q_1}^{\mathrm{SV}}(R_t, \theta, \mathcal{F}_{t-1}), ..., f_{t,q_K}^{\mathrm{SV}}(R_t, \theta, \mathcal{F}_{t-1})\right)';$$

where $f_{t,q_i}^{\mathrm{SV}}(R_t, \theta, \mathcal{F}_{t-1})$ are the scalar indicator functions defined in (12) and $\mathbf{q} = (q_1, \cdots, q_K)'$. The K given thresholds $q_1 > \cdots > q_K$ are ordered conforming with $VaR_{t,q_1} < \cdots <$

VaR_{t,q_K}; the special case

$$\bar{\mathbf{q}}(K,q) := (\bar{q}_1(K,q), \cdots, \bar{q}_K(K,q))', \quad \bar{q}_i(K,q) = \frac{(K-i+1)\,q}{K}, \quad \text{for } q \text{ given}, \tag{14}$$

leads to equally spaced thresholds, that is $|\bar{q}_i(K,q) - \bar{q}_{i+1}(K,q)| = \frac{q}{K}$;

4. centered *realized violation* (RV) *aggregators*:

$$f^{\mathrm{RV}}_{t,\mathbf{q}}(R_t, \theta, \mathcal{F}_{t-1}, K) := \frac{1}{K}\sum_{i=1}^{K}\left(I_{t,q_i}(R_t,\theta,\mathcal{F}_{t-1}) - q_i\right), \tag{15}$$

$$\hat{f}^{\mathrm{IV}}_{t,q}(R_t, \theta, \mathcal{F}_{t-1}, K) := \frac{1}{K}\sum_{i=1}^{K} I_{t,\bar{q}_i(K,q)}(R_t,\theta,\mathcal{F}_{t-1}) - \frac{(K+1)q}{2K}; \tag{16}$$

5. residuals based on alternative *scorings* of violations and so-called *test functions* [Nolde and Ziegel (2017)], including scores that target expected shortfall, expectiles, or joint measures; these cover the generalized residuals of Patton et al. (2019).

All back-tests considered will be based on some test statistic, which we denote by $S(\theta_0)$ or $S(\delta_0)$, that depends on the data only through the series $f_t(R_t, \theta_0, \mathcal{F}_{t-1})$ or $f_t(R_t, (\delta_0', \hat{\nu}\,(\delta_0)\,')', \mathcal{F}_{t-1})$, $t = n+1, \ldots, T$. While the above choices for $f_t(\cdot)$ are the most prominent in the literature on back-testing risk measures, in the sense that some distributional theory is available to assess $\mathcal{H}^*_c$ or $\mathcal{H}^*_u$ or their independence counterparts, our presentation in what follows will intentionally remain general. Other predictive tests can also be inverted more broadly; our focus on tail risk back-tests is motivated by our inference targets.

2.2 Simultaneous confidence sets

It is useful at this stage to contrast our perspective with standard semi-parametric approaches. The latter usually begin by optimizing a statistical objective function to estimate θ, using the training sample, that is data up to time n, or some rolling window of data up to time $t-1$. The risk measures are then computed by plugging the estimated value of θ into the dynamic structures that define them, that is using (1). Point-wise standard errors may be provided for each measure, although related works are rather scarce. Interest rather centers on validating back-tests for the measures under consideration. Estimation consistency and back-test validity often stem from regularity conditions that impose identification. Our objective is to relax this requirement, which may fail in heavy-tailed contexts. Instead, we overcome estimation problems through back-test inversion. When underlying parameters do not require identification, typical back-tests are more likely to control size which leads to CSs for θ with reliable coverage. The fact that we invert adequacy back-tests also eschews further goodness-of-fit assessments. This is because an empty CS formally corresponds to a model rejection by the considered goodness-of-fit back-test. Thereafter, projections yield simultaneous CSs for the desired path and any collection of measures without further assumptions, which also addresses elicitability concerns. The following Theorem formally presents our proposed simultaneous procedures, where again, our exposition remains intentionally general.

Theorem 2.1. *Under Assumptions 1 and 2, for $t = n+1, \ldots, T$, which indexes a predetermined prediction sample, $j = 1, ..., J$, which indexes a collection of predictive parameters for any collection of thresholds $0 \leq q \leq 1$, consider the collection of confidence regions*

$$\mathrm{CS}^{j}_{t,q}(\Lambda^{j}_{t,q}; \tilde{\alpha}) = \{\mathring{\Lambda}^{j}_{t,q} \in \mathbb{R} : \mathring{\Lambda}^{j}_{t,q} = g^{j}_{q,t}(\theta_0, \mathcal{F}_{t-1}) \textit{ for some } \theta_0 \in \mathrm{CS}(\theta; \tilde{\alpha})\}, \tag{17}$$

$$\mathrm{CS}(\theta; \tilde{\alpha}) = \{\theta_0 \in \Omega : \hat{p}(S(\theta_0)) > \tilde{\alpha}\}, \tag{18}$$

and the collection of confidence intervals

$$\mathrm{CI}^{j}_{t,q}(\Lambda^{j}_{t,q}; \tilde{\alpha}) = \left]\mathrm{L}^{j}_{t,q}(\Lambda^{j}_{t,q}; \tilde{\alpha}), \mathrm{U}^{j}_{t,q}(\Lambda^{j}_{t,q}; \tilde{\alpha})\right[,$$

where $\mathrm{L}^{j}_{t,q}(\Lambda^{j}_{t,q}; \tilde{\alpha}) = \inf[g^{j}_{q,t}(\theta_0, \mathcal{F}_{t-1}) : \theta_0 \in \mathrm{CS}(\theta; \tilde{\alpha})]$, $\mathrm{U}^{j}_{t,q}(\Lambda^{j}_{t,q}; \tilde{\alpha}) = \sup[g^{j}_{q,t}(\theta_0, \mathcal{F}_{t-1}) : \theta_0 \in \mathrm{CS}(\theta; \tilde{\alpha})]$, $0 \leq \tilde{\alpha} \leq 1$, $\hat{p}\left(S(\theta_0)\right)$ refers to the p-value of an out-of-sample test of $\mathcal{H}_c\left(\theta_0\right)$ or $\mathcal{H}_u\left(\theta_0\right)$ as defined in (8) or (9), and $S(\theta_0)$ denotes the underlying test statistic that depends on the data only through $f_t(R_t, \theta_0, \mathcal{F}_{t-1})$, $t = n+1, \ldots, T$. If the considered test is *size-correct in the sense that*

$$\mathrm{P}\left[\hat{p}\left(S(\theta_0)\right) \leq \tilde{\alpha}\right] = \tilde{\alpha} \tag{19}$$

under $\mathcal{H}_c\left(\theta_0\right)$ *or* $\mathcal{H}_u\left(\theta_0\right)$*, then* $\mathrm{CS}^j_{t,q}(\Lambda^j_{t,q};\tilde{\alpha})$ *and* $\mathrm{CI}^j_{t,q}(\Lambda^j_{t,q};\tilde{\alpha})$ *achieve coverage control, that is:*

$$\begin{aligned}\mathrm{P}[\Lambda^j_{t,q} &\in \mathrm{CS}^j_{t,q}(\Lambda^j_{t,q};\tilde{\alpha})] \geq 1-\tilde{\alpha}, &(20)\\ \mathrm{P}[\mathrm{L}^j_{t,q}(\Lambda^j_{t,q};\tilde{\alpha}) &\leq \Lambda^j_{t,q} \leq \mathrm{U}^j_{t,q}(\Lambda^j_{t,q};\tilde{\alpha})] \geq 1-\tilde{\alpha}. &(21)\end{aligned}$$

(20) and (21) still follow when under the null hypothesis $\mathrm{P}\left[\hat{p}\left(S(\theta_0)\right) \leq \tilde{\alpha}\right] \leq \tilde{\alpha}$*.*

See the proof in the Supplementary Material. When the inverted test controls size only asymptotically (under some regularity conditions), then (20) and (21) will also hold asymptotically. If ν is estimated, the above leads to the collection of confidence regions

$$\mathrm{CS}^j_{t,q}(\Lambda^j_{t,q};\tilde{\alpha}) = \{\mathring{\Lambda}^j_{t,q} \in \mathbb{R} : \mathring{\Lambda}^j_{t,q} = g^j_{q,t}(\delta_0, \mathcal{F}_{t-1}) \text{ for some } \delta_0 \in \mathrm{CS}(\delta;\tilde{\alpha})\}, \tag{22}$$

$$\mathrm{CS}(\delta;\tilde{\alpha}) = \{\delta_0 \in \Omega : \hat{p}(S(\delta_0)) > \tilde{\alpha}\}, \tag{23}$$

and the collection of confidence intervals

$$\mathrm{CI}^j_{t,q}(\Lambda^j_{t,q};\tilde{\alpha}) = \left]\mathrm{L}^j_{t,q}(\Lambda^j_{t,q};\tilde{\alpha}), \mathrm{U}^j_{t,q}(\Lambda^j_{t,q};\tilde{\alpha})\right[,$$

where $\mathrm{L}^j_{t,q}(\Lambda^j_{t,q};\tilde{\alpha}) = \inf[g^j_{q,t}(\delta_0, \mathcal{F}_{t-1}) : \delta_0 \in \mathrm{CS}(\delta;\tilde{\alpha})]$, $\mathrm{U}^j_{t,q}(\Lambda^j_{t,q};\tilde{\alpha}) = \sup[g^j_{q,t}(\delta_0, \mathcal{F}_{t-1}) : \delta_0 \in \mathrm{CS}(\theta;\tilde{\alpha})]$, $\hat{p}\left(S(\delta_0)\right)$ refers to the p-value of an out-of-sample test of $\mathcal{H}^s_c\left(\delta_0\right)$ or $\mathcal{H}^s_u\left(\delta_0\right)$ as defined in (10) or (11), and $S(\delta_0)$ denotes the underlying test statistic that depends on the data only through $f_t(R_t, (\delta_0', \hat{\nu}\left(\delta_0\right)')', \mathcal{F}_{t-1})$, $t = n+1, \ldots, T$. The estimation step does not necessarily entail that the inverted test will not be exact. Thus, we next characterize the cases leading to exact procedures, which require some restrictions on $f_t(\cdot)$ or the DGP.

The implications of (3) and (4) are that: (i) $I_{t,q}(.)$ as defined in (12) is Bernoulli with mean q; (ii) $f^{\mathrm{IV}}_{t,q}(.)$ as defined in (13) is a martingale difference sequence, and (iii) $\left\{I_{t,q_i}(.) - I_{t,q_{i+1}}(.)\right\}_{i=0,\ldots,K}$ is Bernoulli with mean $\{q_i - q_{i+1}\}_{i=0,\ldots,K}$. Under the independence assumption, the joint null distribution of the vector

$$\mathbf{I}_{t,\mathbf{q}}(R_t, \theta, \mathcal{F}_{t-1}, K) = (I_{t,q_1}(R_t, \theta, \mathcal{F}_{t-1}), ..., I_{t,q_K}(R_t, \theta, \mathcal{F}_{t-1}))' \tag{24}$$

is thus exactly nuisance parameter free and can be simulated, *e.g.* using draws from the binomial or multinomial distribution, without taking a stand on the DGP. This has motivated useful applications of the Monte Carlo test (**MCT**) technique [Dufour (2006)] to derive exact tests based on such residuals, see Theorem 2.2. Since our implementation differs from standard back-tests, we provide Algorithm 2 for completion.

Algorithm 2

```
To back-test a given $\theta_0$, using a back-test statistic denoted by $S(\theta)$ which depends on the
data only through $f_t(R_t, \theta, \mathcal{F}_{t-1}),\quad t = n+1, \ldots, T$,
1: Using the sample of $t = n+1, \ldots, T$ observations, compute the residuals $f_t(R_t, \theta_0, \mathcal{F}_{t-1})$,
   and the considered test statistic, denoted $S(\theta_0)$;
2: For $b = 1, \ldots, B$, generate independent simulated realizations of the residuals
   $f_t(R_t, \theta_0, \mathcal{F}_{t-1})$, under Assumptions 1 and 2, and the null hypothesis, leading to $B$
   independent realizations of the test statistic denoted $S_b(\theta_0)$;
3: Compute the empirical $p$-value $\hat{p}\,(S(\theta_0)) = p_B\,(S(\theta_0))$with tie-breakers, as
   $$p_B(x) = \frac{B \times G_B(x) + 1}{B + 1} \tag{25}$$
   $$G_B(x) = 1 - \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}(S_b(\theta_0) \geq x) + \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}(S_b(\theta_0) \geq x) \times \mathbb{1}(W_0 \leq W_b)$$
   and $W_b$, $b = 0, 1, \ldots, B$, are independent standard uniform random variates.
```

Theorem 2.2. *In the context of Assumptions 1 and 2, consider a test statistic $S(\theta_0)$ to assess specification adequacy for a given θ_0 when the following supplementary condition holds: under the adequacy assumption, the sequence of $f_t(R_t, \theta, \mathcal{F}_{t-1})$, $t = n+1, \ldots, T$, can be simulated given θ. If $S(\theta_0)$ depends on the data only through the residuals $f_t(R_t, \theta_0, \mathcal{F}_{t-1})$, then the test with critical region $\hat{p}\,(S(\theta_0)) \leq \tilde{\alpha}$, $0 < \tilde{\alpha} < 1$, where $\hat{p}(\theta_0)$ is obtained as in (25) and $\tilde{\alpha}\,(B+1)$ is an integer, is exact in the following sense:*

$$\mathrm{P}[\hat{p}\,(S(\theta_0)) \leq \tilde{\alpha}] = \tilde{\alpha}. \tag{26}$$

See the proof in the Supplementary Material. The key point we aim to emphasize via this Theorem, is the following: the only requirement is the ability to make draws from the null distribution of $f_t(R_t, \theta_0, \mathcal{F}_{t-1})$ - rather than from the null distribution of the data - since model adequacy is defined through $f_t(R_t, \theta_0, \mathcal{F}_{t-1})$, as in Assumption 2. The issue of course is how to implement step 2. Several cases are worth emphasizing.

Case **A**. If (24) is considered for $f_t(\cdot)$, then the latter can easily be simulated under the independence null with any parametric or semi-parametric DGP. Consequently, exact p-values in the sense of (26) can be obtained for any statistic, not just those suggested or reviewed by Christoffersen and Pelletier (2004), Berkowitz et al. (2011), Leccadito et al. (2014), Kratz et al. (2018), Khalaf et al. (2021), that depends on the data only through a known function of $\mathbf{I}_{t,\mathbf{q}}(R_t, \theta, \mathcal{F}_{t-1}, K)$. This includes, in particular, the scoring-based criteria of Nolde and Ziegel (2017), which broadly widens their applicability beyond the considered Wald form. In this case, the simulated residuals will not depend on θ_0 and thus can be drawn only once through the inversion search.

Case **B**. If the DGP is completely specified given θ so that R_t can be simulated knowing θ, any $f_t(\cdot)$ sequence can be simulated regardless of its form, for every θ_0 value under the null. If the DGP can be simulated yet it just models tail outcomes given θ, then any $f_t(\cdot)$ sequence can be simulated if it is only based on tail data. Exact p-values in the sense of (26) will also follow in this case, yet in contrast to case **A**, the residuals need to be drawn for every θ_0 value under test through the inversion search.

Case **C**. In the context of case **B** and the subset null hypotheses $\mathcal{H}_c^s(\delta_0)$ or $\mathcal{H}_u^s(\delta_0)$ as defined in (10) or (11), Algorithm 2 can be implemented using the plug-in estimator $\hat{\theta}_0 = (\delta_0', \hat{\nu}(\delta_0)')'$. The resulting MCT p-value $\hat{p}(\cdot)$ will be valid asymptotically in the sense that $\mathrm{P}[\hat{p}(S(\theta_0)) \leq \tilde{\alpha}]$ converges to $\tilde{\alpha}$, if null distribution of the test statistic is nuisance parameter free.[6] The aforementioned literature provides a wide spectrum of asymptotically

[6] The plug-in estimator can be validated under less stringent high-level assumptions on the null distri-

pivotal tests that are based on some specific form for $f_t(\cdot)$, and these typically require conditions on n and T. Underlying regularity conditions may hold only weakly if the estimated parameters need to be identified. We propose to implement these tests when the sub-parameter that is hard to identify is fixed for inversion purposes, which places the estimation and limiting distributional theory under more favorable conditions.

Otherwise, any of the available tests can be inverted without reliance on the MCT approach. Again, their limiting null distributions or alternative re-sampling schemes will perform better when the parameter that is hard to identify is not estimated. To illustrate the applicability of our proposed procedure, we next consider the GARCH baseline case.

3 Application to a baseline GARCH risk model

Consider the following GARCH process for $R_t = \sigma_t Z_t$ where Z_t is either standard normal, or follows a standardized symmetric ($\xi = 1$) or asymmetric Student-t distribution with ν degrees-of-freedom and asymmetry parameter ξ:

$$\sigma_t^2 | \mathcal{F}_{t-1} = \bar{\sigma}^2(1 - \alpha - \beta) + \alpha R_{t-1}^2 + \beta \sigma_{t-1}^2. \tag{27}$$

Var and ES admit closed forms which we supply in the Supplementary Material for completion. Expectiles do not admit closed-form expressions so $g_{q,t}^j(\theta, \mathcal{F}_{t-1})$ is computed numerically. We set $\delta = (\alpha, \beta)'$ as the sub-vector of parameters that are hard to identify, where $\alpha \geq 0$, $\beta \geq 0$, and $\alpha + \beta < 1$. GARCH parameters are hard to identify because of root-cancellation type problems [Andrews (2001)], which motivates our application.

We focus on two representative back-tests: the multi-level Pearson test for VaR [Leccadito et al. (2014)], denoted by $\mathcal{X}(m, \mathbf{q})$, and the tests for ES by Du and Escanciano (2017), denoted by $\mathcal{U}(q)$ in the unconditional case, and $\mathcal{C}(m, q)$ in the conditional one, where m is the number of lags. The first one is a distribution-free exact test based on

bution of the considered statistics as shown in Dufour (2006).

realized violations, whereas the last two rely on fully parametric integrated violations, and is justified asymptotically. We aim: (i) to explore linkages between integrated and realized violation, and (ii) to study within a unified environment a parametric versus model-agnostic approach. We provide the formula for the three tests statistics in the Supplementary Material. Both tests are implemented for a given vector $(\alpha, \beta)'$, estimated $\bar{\sigma}^2$ and when relevant, estimated υ and ξ, for each value of the tested $(\alpha, \beta)'$ vector.

3.1 Simulation study

We conduct a simulation study to assess size and power of the considered back-tests for hypotheses that fix α and β. This documents the coverage properties of associated inversion procedures, in line with the literature on weak identification. The experiment is conducted over 2000 Monte Carlo simulations, where the training sample size is $n = 2500$ and the prediction sample size $T - n \in \{90, 250, 500, 1000, 2000\}$ respectively. Samples are drawn from (27), assuming in turn standard normal, symmetric and asymmetric innovations to reflect typical features of return distributions such as fat tails and asymmetry. More specifically, we explore the impact of these common stylized facts by considering three different DGPs: Normal-GARCH(1,1), Student t-GARCH(1,1) with degrees of freedom set to 6.77, and skewed Student t-GARCH(1,1) where the shape parameter is set to 6.93 and the asymmetry parameter is set to 0.86. These values were estimated from daily S&P500 log returns using the considered GARCH model, over the 2008-2012 period, which will provide an empirically relevant design. The long-run variance $\bar{\sigma}^2$ is calibrated to 1, which also coincides with the starting value of the conditional variance process (27).

Under the null hypothesis, distributional parameters are estimated with maximum likelihood in sample using a rolling window design. For power purposes, α and β are modified relative to the null hypothesis, leaving the remaining parameters unchanged, including the long-run variance and starting value of the conditional variance process (which are calibrated). VaR thresholds considered for the $\mathcal{X}(m, \mathbf{q})$ test are $q_1 = 1.25\%$, $q_2 = 2.5\%$, $q_3 = 3.75\%$, $q_4 = 5\%$ and MCT p-values are computed using 999 replications. For the $\mathcal{U}(q)$ and $\mathcal{C}(m, q)$ test, $q = 5\%$. All tests are implemented at the nominal 5% level, and $m = 5$ lags are used. Selected results are shown; see the Supplementary Material for additional results.

Size results are reported in Tables 1 and 2. Empirical rejections with the $\mathcal{C}(m, q)$ test are around 10% to 12% when $T - n$ is 250 or below; concretely, a prediction sample size above 1000 corrects over-rejection probabilities. This rejection pattern is not affected by the distribution of innovations, despite the underlying nuisance parameter complications. Although asymptotically justified, the $\mathcal{U}(q)$ test performs well in our design even when $T - n = 90$. Under the assumptions of Du and Escanciano (2017), the associated limiting theory holds when the estimation sample is much larger than the prediction one; this may justify our findings. That said, recall that in contrast to their original formulation of Du and Escanciano (2017), the tests that we analyze fix α and β to their hypothesized values, which seems to contribute crucially to our size results. The $\mathcal{X}(m, \mathbf{q})$ test is exact which is reflected in our simulations.

Power results are reported in Table 3. Because of our emphasis on test inversion, we fix the DGP to two empirically relevant cases: case (i) $\alpha_{DGP} = 0.05$ and $\beta_{DGP} = 0.90$, and case (ii) $\alpha_{DGP} = 0.15$ and $\beta_{DGP} = 0.75$. In both cases, persistence is high and is driven by a predominantly high β, which is a pattern that is often observed in financial applications. We study power associated with several (α_0, β_0) pairs. Key results can be summarized as follows.

(1) All three tests have power for inference on $\mathcal{H}_c$ or $\mathcal{H}_u$. By contrast, existing studies have assessed their properties in the case of $\mathcal{H}_c^*$ or $\mathcal{H}_u^*$. On balance, power improves as $T - n$ increases. However, when $\alpha_0 = 0.05$, rejections remain low even when $T - n = 2000$. In this case, the $\mathcal{C}(m, q)$ test performs best, yet such an apparent dominance should be interpreted with caution since this test is over-sized. In fact, the $\mathcal{X}(m, \mathbf{q})$ and $\mathcal{U}(q)$ tests

dominate when $\alpha_0 \neq 0.05$, in which case rejection patterns evolve more regularly with $T - n$.

(2) Power clearly reacts to discrepancies between α_0 and α_{DGP}. However, the $\alpha_0 = 0.05$ is a case of interest because, as it is well known, β is hard to identify as $\alpha \simeq 0$. Our results reveal related difficulties, namely that correctly sized test have low (or almost no) power. For instance, let us focus first on case (i) and consider the evolution of rejections over the three choices for β_0 that are associated with $\alpha_0 = 0.05 = \alpha_{DGP}$. In this case, the $\mathcal{X}(m, \mathbf{q})$ and $\mathcal{U}(q)$ tests barely react as β_0 departs from β_{DGP}, and the only test with some power is based on $\mathcal{C}(m, q)$, which is oversized. Second, let us focus on the $\mathcal{X}(m, \mathbf{q})$ or the $\mathcal{U}(q)$ test for $\alpha_0 = 0.05$ and compare rejections between case (i) [where $\alpha_{DGP} = 0.05$] and case (ii) [where $\alpha_{DGP} = 0.15$], as $|\beta_{DGP} - \beta_0|$ evolves. While the former case involves larger discrepancies [compare 0.85, 0.45 and 0.15 to 0.70, 0.30 and 0.15], the tests perform visibly better in the latter case, which is characterized by $\alpha_{DGP} = 0.15$. This value may not seem too far apart from the hypothesized $\alpha_0 = 0.05$, yet the discrepancy suffices to produce a sizable power differential. Taken collectively, these results imply that the tests are more responsive to departures of α_0 from α_{DGP} than to departures of β_0 from β_{DGP}. This illustrates the expected implications of weak identification, which motivate our work.

(3) While a power ranking is not our direct objective, we find that no single test uniformly dominates with respect to power. This suggests that the realized violation aggregators are as informative as their integrated counterpart, at least in our design. Recall that: (a) the former converges to latter as the number of thresholds grows; (b) the integral underlying the latter refers to the conditional distribution of the data which should be specified, whereas realized aggregators are distribution-free; and (d) tests based on the former are exact under the independence assumption, in contrast to the latter which rely on the properties of martingale difference sequences. No test, in regards to assumptions, can be formally considered more or less restrictive than the other. We find that both ap-

proaches can work well from a test inversion perspective, when the null hypothesis fixes the parameters that are hard to identify. These results add interesting insights to the literature.

(4) The considered tests as implemented (that is to assess $\mathcal{H}_c$ or $\mathcal{H}_u$) have excellent power to reject the (α_0, β_0) pairs where the relative contributions of α and β is "flipped" with reference to the (α_{DGP},β_{DGP}) pair. Specifically, power is maximized for the null hypothesis $\alpha = .90$ and $\beta = .05$ when confronted with (i) or (ii). To appreciate the importance of this finding from a joint test perspective, re-express (27) in the following form:

$$\sigma_t^2|\mathcal{F}_{t-1} = \bar{\sigma}^2 + \alpha(R_{t-1}^2 - \bar{\sigma}^2) + \beta(\sigma_{t-1}^2 - \bar{\sigma}^2). \tag{28}$$

As emphasized by Christoffersen (2016) (page 71), α captures the correction to the long-run variance driven by deviations of past squared returns; instead, β picks up the effects driven by deviations of past variance. Our design calibrates $\bar{\sigma}^2$ to allow us to disentangle the relative contribution of these deviations. Power in this direction, which we illustrate in Table 3, is a highly desirable feature in GARCH modeling.

(5) The considered tests are highly informative about the distribution of innovations (here Z_t) when $T-n$ is large. The experiment in Table 4 is designed to explore the behavior of the considered tests when the distribution of Z_t assumed under the null hypothesis differs from its DGP counterpart, while the parameter values that are tested match the DGP. Specifically, samples are generated assuming Z_t follows a Student-t distribution with 3 degrees of freedom, while under the null hypothesis Z_t is standard Gaussian. We find that all tests are not responsive unless $T-n$ exceeds 1000, yet rejections with the $\mathcal{X}(m, \mathbf{q})$ test exceeds 90% when $T-n = 2000$. We may thus expect empty CSs when the size of the prediction sample is large or when the distribution of returns is misspecified. Since in this design $\alpha_0 = \alpha_{DGP}$ and $\beta_0 = \beta_{DGP}$, non-rejections with smaller prediction samples may be viewed as a robustness finding from a semi-parametric perspective. It is worth pointing out here the irregular behavior of the $\mathcal{C}(m, q)$ tests, whose power seems to decrease with

$T - n$. Clearly, this results from its over-size as pointed out above.

We conclude by reemphasizing that the tests analyzed rely on the hypothesized distribution of Z_t to compute the risk measures. In addition, the $\mathcal{C}(m, q)$ and $\mathcal{U}(q)$ tests further rely on this distribution to aggregate violations. The above remark on robustness is thus not meant as a formal semi-parametric directive. Distribution-free confidence bands are considered in our empirical section next.

3.2 Empirical Analysis

We study the Technology Select Sector SPDR Fund ETF (Bloomberg ticker: XLK), listed on the New York Stock Exchange Arca. Our sample ranges from August 2002 to October 2020, and our prediction sample focuses on the last 90 observations [June 22 to October 23, 2020], which leaves us with $n = 4378$ observations. We choose to study the market risk of this particular sector as it was one of the most impacted by the unprecedented and unanticipated effects of the COVID lockdowns. We apply the methods introduced in Section 2, using the $\mathcal{X}(m, \mathbf{q})$, $\mathcal{U}(q)$ and $\mathcal{C}(m, q)$ statistics within the GARCH setting. Furthermore, we consider the FHS setting, which is a semi-parametric pseudo-likelihood GARCH based tail risk specification. In this case, the distribution function of Z_t is estimated using the empirical distribution function (with data up to time t) for each tested (α_0, β_0) pair. The only feasible test in this case is $\mathcal{X}(m, \mathbf{q})$.

The static parameter $\bar{\sigma}^2$ can be estimated by via maximum likelihood imposing $\alpha = \alpha_0$ and $\beta = \beta_0$, along with the parameters of the distribution of Z_t (when present). The added complication is that $\bar{\sigma}^2$ is hard to identify as $\alpha + \beta \simeq 1$, as is evident from (28). However, as discussed by Engle and Mezrich (1996) [see also Noureldin et al. (2014)] GARCH models can instead be viewed flexibly subject to the so-called *variance targeting restriction*, whereby the static parameters are defined through a moment condition that does not involve the remaining dynamic ones. This view differs from its likelihood counterpart, yet remains compatible with useful and popular interpretations of the model. Conformably, $\bar{\sigma}^2$ can be estimated consistently using the empirical variance of the data (up to time t). We follow this approach since it eschews the boundary problem. We view the combination of *variance targeting* with our test inversion approach as a useful subset inference strategy to address identification problems that may stem from multiple sources.

For test inversion, we use a grid search over the range $0.01 < \alpha < 0.99$ and $0.01 < \beta < 0.99$, with a step of 0.01, imposing $\alpha + \beta < 1$, for a total of 4851 combinations of parameters. We consider the $\mathcal{X}(5, (1\%, 2.5\%, 5\%)')$ test with 99 MC replications for the MCT p-values, and the $\mathcal{U}(5\%)$ and $\mathcal{C}(5, 5\%)$ tests. Selected figures are reported below for $\tilde{\alpha} = 10\%$. The gray lines are the bounds of the projected confidence band for the considered risk measure. The projection is obtained by taking the minimum and maximum of the risk measure, which is itself a function of α and β and conformable estimated parameters, over all non-rejected (α, β) pairs. The least rejected pair is retained as the (possibly non-unique) point estimate that corresponds to the maximum p-value associated with the considered test. The risk measure corresponding to that point estimate is reported in blue, and its counterpart based on maximum likelihood is reported in red. All the above is computed for each prediction day, based on a rolling window estimation.

We produce simultaneous confidence bands for the VaR, ES and Expectile-VaR forecasts. All these bands are jointly interpretable over their full path and across measures. Because our data consists of log returns, the risk measures can be interpreted as percentages. For instance, $VaR_{n+1,5\%} = 0.02$ reads as follow: there is a 5% chance that the asset will lose at most 2% of its value on day n, by the next trading day.

Overall, despite a clear evidence of root cancellation and boundary problems when it comes to the GARCH parameters, results are informative on the parametric risk measures themselves, through tight enough bands. Consider for example Figure 1 which reports CSs for the parametric VaR, using the $\mathcal{X}(m, \mathbf{q})$ test. Since close to zero values of α cannot be ruled out, it is not surprising that information on β is limited. Interestingly, the projected bands on VaR are rather stable across the various distributions. Figure 2 shows the CS for parametric ES obtained from the test inversion of the $\mathcal{U}(q)$ test. While the $\mathcal{C}(m, q)$ test holds practically no information (see Supplementary Material), results associated with the $\mathcal{U}(q)$ test refute large α yet maintain close to zero values, with again the same implication about estimating β. Again, we find that the bands do not differ much across the distributions. The semi-parametric procedure, reported in Figure 3 reveal no information on the GARCH parameters. Given the size of our prediction sample, and the fact that the COVID shock is an end of sample disruption, this is not surprising. We report the conformable bands for VaR, for comparison purposes. While these bands are wider than their parametric counterparts in the absence on any information about α and β, recall that information on the tail still flows from the long-run variance estimates because of variance targeting, and from the empirical quantile. Lastly, Figure 4 reports the confidence bands for the EVaR, projected from the CS for (α, β) derived from the considered tests. Here the bands are broader for peak days when skewness is factored in, yet remain otherwise rather stable across distributions.

The well documented September effect is clearly visible through all bands. September 2020 is marked by important ground breaking announcements in particular about the COVID vaccine and a major acquisition in the artificial intelligence sector, which can explain the risk peaks that we observe. ETFs raise well documented financial fragility concerns; see the survey by Lettau and Madhavan (2018), and perspectives on liquidity and activeness risks from Bae and Kim (2020) and Easley et al. (2021). Our results quantify ETF tail risk effects in a historically unique episode for the technology sector.

4 Conclusion

A general method to construct simultaneous CSs for predictive quantities is proposed that can signal identification or misspecification when outcomes are unbounded or empty. Importantly, our results are not restricted to VaR and ES; in particular, we also study EVaR, on which results are scarce. An illustrative simulation and empirical analysis is considered within GARCH-based parametric and semi-parametric settings. Results underscore the usefulness of exact approaches, and the information content of cumulative multi-threshold violations along with variance targeting. We emphasize the importance of simultaneity to jointly interpret CSs across the prediction paths and all considered measures. Our empirical analysis for a technology ETF reveals root cancellation issues, yet remains relatively informative on the measures themselves. We find that variance targeting, or more generally, the practice of partialling-out static or steady state parameters can aid identification. Overall, we view our results as a stepping stone towards a unified perspective on elicitability and identification.

References

Andrews, D. W. K. (2001). Testing when a parameter is on the boundary of the maintained hypothesis. *Econometrica 69*(3), 683–734.

Argyropoulos, C. and E. Panopoulou (2019). Backtesting VaR and ES under the magnifying glass. *International Review of Financial Analysis 64*, 22–37.

Bae, K. and D. Kim (2020). Liquidity risk and exchange-traded fund returns, variances, and tracking errors. *Journal of Financial Economics 138*(1), 222–253.

Banulescu-Radu, D., C. Hurlin, J. Leymarie, and O. Scaillet (2021). Backtesting marginal expected shortfall and related systemic risk measures. *Management Science 67*(9), 5730–5754.

Barendse, S., E. Kole, and D. van Dijk (2021, 05). Backtesting Value-at-Risk and Expected Shortfall in the Presence of Estimation Error. *Journal of Financial Econometrics 21*(2), 528–568.

Barone-Adesi, G., K. Giannopoulos, and L. Vosper (1999). VaR without correlations for portfolios of derivative securities. *Journal of Futures Markets 19*(5), 583–602.

Basel Committee on Banking Supervision (2019, January). Explanatory note on the minimum capital requirements for market risk. Consultation paper, Bank for International Settlements.

Bayer, S. and T. Dimitriadis (2020, 09). Regression-Based Expected Shortfall Backtesting. *Journal of Financial Econometrics 20*(3), 437–471.

Beaulieu, M.-C., J.-M. Dufour, and L. Khalaf (2013). Identification-Robust Estimation and Testing of the Zero-Beta CAPM. *Review of Economic Studies 80*(3), 892–924.

Beaulieu, M.-C., J.-M. Dufour, and L. Khalaf (2014). Exact confidence sets and goodness-of-fit methods for stable distributions. *Journal of Econometrics 181*(1), 3–14.

Beaulieu, M.-C., J.-M. Dufour, L. Khalaf, and O. Melin (2023). Identification-robust beta pricing, spanning, mimicking portfolios, and the benchmark neutrality of catastrophe bonds. *Journal of Econometrics 236*(1), 105464.

Bellini, F. and E. Di Bernardino (2017). Risk management with expectiles. *The European Journal of Finance 23*(6), 487–506.

Bellini, F., B. Klar, A. Müller, and E. Rosazza Gianin (2014). Generalized quantiles as risk measures. *Insurance: Mathematics and Economics 54*(C), 41–48.

Bellini, F., I. Negri, and M. Pyatkova (2019). Backtesting VaR and expectiles with realized scores. *Statistical Methods & Applications 28*, 119–142.

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

Chan, N. H., S.-J. Deng, L. Peng, and Z. Xia (2007). Interval estimation of value-at-risk based on GARCH models with heavy-tailed innovations. *Journal of Econometrics 137*(2), 556 – 576.

Christoffersen, P. (2016). *Elements of Financial Risk Management*. Elsevier Science.

Christoffersen, P. and S. Goncalves (2005). Estimation risk in financial risk management. *The Journal of Risk 7*, 1–28.

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.

Daouia, A., S. Girard, and G. Stupfler (2020). Tail expectile process and risk assessment. *Bernoulli 26*(1), 531 – 556.

Davison, A. C., S. A. Padoan, and G. Stupfler (2023). Tail risk inference via expectiles in heavy-tailed time series. *Journal of Business & Economic Statistics 41*(3), 876–889.

Dimitriadis, T. and R. Halbleib (2022). Realized quantiles. *Journal of Business & Economic Statistics 40*(3), 1346–1361.

Dimitriadis, T., X. Liu, and J. Schnaitmann (2023). Encompassing tests for value at risk and expected shortfall multistep forecasts based on inference on the boundary. *Journal of Financial Econometrics 21*(2), 412–444.

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

Du, Z., P. Pei, X. Wang, and T. Yang (2023). Powerful backtests for historical simulation expected shortfall models. *Journal of Business & Economic Statistics 0*(0), 1–11.

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.

Dufour, J.-M., E. Ghysels, and A. Hall (1994). Generalized predictive tests and structural change analysis in econometrics. *International Economic Review 35*(1), 199–229.

D'Innocenzo, E., A. Lucas, B. Schwaab, and X. Zhang (2023). Modeling extreme events: Time-varying extreme tail shape. *Journal of Business & Economic Statistics 0*(0), 1–15.

Easley, D., D. Michayluk, M. O'Hara, and T. J. Putniņš (2021). The active world of passive investing. *Review of Finance 25*(5), 1433–1471.

Engle, R. and J. Mezrich (1996). GARCH for groups: A round-up of recent developments in GARCH techniques for estimating correlation. *Risk Magazine 9*, 36–40.

Fernández, C. and M. F. J. Steel (1998). On bayesian modeling of fat tails and skewness. *Journal of the American Statistical Association 93*(441), 359–371.

Fissler, T. and Y. Hoga (2023). Backtesting systemic risk forecasts using multi-objective elicitability. *Journal of Business & Economic Statistics 0*(0), 1–14.

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

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

He, X. D., S. Kou, and X. Peng (2022). Risk measures: Robustness, elicitability, and backtesting. *Annual Review of Statistics and Its Application 9*(1), 141–166.

Hodges, J. L. and E. L. Lehmann (1963). Estimates of location based on rank tests. *The Annals of Mathematical Statistics 34*(2), 598–611.

Hoga, Y. (2019a). Confidence intervals for conditional tail risk measures in ARMA–GARCH models. *Journal of Business & Economic Statistics 37*(4), 613–624.

Hoga, Y. (2019b). Limit Theory for Forecasts of Extreme Distortion Risk Measures and Expectiles. *Journal of Financial Econometrics 20*(1), 18–44.

Hoga, Y. and M. Demetrescu (2023). Monitoring value-at-risk and expected shortfall forecasts. *Management Science 69*(5), 2954–2971.

Hurlin, C., S. Laurent, R. Quaedvlieg, and S. Smeekes (2017). Risk measure inference. *Journal of Business & Economic Statistics 35*(4), 499–512.

Jordà, Ò. and M. Marcellino (2010). Path forecast evaluation. *Journal of Applied Econometrics 25*(4), 635–662.

Khalaf, L., A. Leccadito, and G. Urga (2021). Multilevel and tail risk management. *Journal of Financial Econometrics 20*(5), 839–874.

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. *The Journal of Derivatives 3*(2), 73–84.

Lan, H., B. L. Nelson, and J. Staum (2010). A confidence interval procedure for expected shortfall risk measurement via two-level simulation. *Operations Research 58*(5), 1481–1490.

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.

Lettau, M. and A. Madhavan (2018). Exchange-traded funds 101 for economists. *Journal of Economic Perspectives 32*(1), 135–154.

Longerstaey, J. and M. Spencer (1996). Riskmetrics — technical document. *Morgan Guaranty Trust Company of New York: New York 51*, 54.

Newey, W. K. and J. L. Powell (1987). Asymmetric least squares estimation and testing. *Econometrica 55*(4), 819–847.

Nieto, M. R. and E. Ruiz (2016). Frontiers in var forecasting and backtesting. *International Journal of Forecasting 32*(2), 475–501.

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

Noureldin, D., N. Shephard, and K. Sheppard (2014). Multivariate rotated ARCH models. *Journal of Econometrics 179*(1), 16–30.

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

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

Stock, J. H. and J. H. Wright (2000). GMM with weak identification. *Econometrica 68*(5), 1055–1096.

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

Taylor, J. W. (2022). Forecasting value at risk and expected shortfall using a model with a dynamic omega ratio. *Journal of Banking & Finance 140*, 106519.

Tsay, R. S. (2010). *Analysis of financial time series.* John Wiley & sons.

Wang, C.-S. and Z. Zhao (2016). Conditional value-at-risk: Semiparametric estimation and inference. *Journal of Econometrics 195*(1), 86–103.

Xu, K.-L. (2016). Model-free inference for tail risk measures. *Econometric Theory 32*(1), 122–153.

Table 1: Out-of-sample size at 5% nominal level, standard normal innovations.

	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.15$	$\alpha = 0.25$	$\alpha = 0.4$	$\alpha = 0.65$	$\alpha = 0.90$
$T-n$	$\beta = 0.90$	$\beta = 0.8$	$\beta = 0.75$	$\beta = 0.6$	$\beta = 0.4$	$\beta = 0.15$	$\beta = 0.05$
Panel A: $\mathcal{X}(5,(1.25\%,2.5\%,3.75\%,5\%)')$							
90	0.0450	0.0510	0.0450	0.0425	0.0510	0.0455	0.0480
250	0.0510	0.0510	0.0525	0.0485	0.0450	0.0455	0.0500
500	0.0490	0.0535	0.0410	0.0475	0.0540	0.0430	0.0420
1000	0.0560	0.0505	0.0465	0.0490	0.0465	0.0440	0.0530
2000	0.0525	0.0535	0.0480	0.0625	0.0500	0.0520	0.0560
Panel B: $\mathcal{U}(5\%)$							
90	0.0390	0.0395	0.0420	0.0355	0.0420	0.0355	0.0385
250	0.0480	0.0510	0.0435	0.0485	0.0515	0.0470	0.0410
500	0.0530	0.0490	0.0440	0.0520	0.0520	0.0465	0.0450
1000	0.0540	0.0410	0.0480	0.0395	0.0405	0.0560	0.0525
2000	0.0535	0.0480	0.0500	0.0490	0.0500	0.0545	0.0500
Panel C: $\mathcal{C}(5,5\%)$							
90	0.1175	0.1250	0.1205	0.1190	0.1195	0.1205	0.1205
250	0.0925	0.0785	0.0880	0.0750	0.0980	0.0925	0.0835
500	0.0745	0.0620	0.0710	0.0710	0.0655	0.0720	0.0830
1000	0.0720	0.0690	0.0700	0.0755	0.0605	0.0605	0.0625
2000	0.0620	0.0600	0.0540	0.0620	0.0595	0.0555	0.0520

For each panel, 2000 time series of in sample length $n = 2500$ and out of sample length $T-n \in \{250, 500, 1000, 2000\}$ have been generated according to a GARCH(1,1) model with Standard Normal innovations. The null hypothesis assumes the same model specification and (α, β) parameters of the DGP.

Table 2: Out-of-sample size at 5% nominal level, standardized skewed Student-t innovations.

	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.15$	$\alpha = 0.25$	$\alpha = 0.4$	$\alpha = 0.65$	$\alpha = 0.90$
$T - n$	$\beta = 0.90$	$\beta = 0.8$	$\beta = 0.75$	$\beta = 0.6$	$\beta = 0.4$	$\beta = 0.15$	$\beta = 0.05$
Panel A: $\mathcal{X}(5, (1.25\%, 2.5\%, 3.75\%, 5\%)')$							
90	0.0490	0.0405	0.0480	0.0380	0.0475	0.0470	0.0445
250	0.0470	0.0485	0.0460	0.0500	0.0440	0.0420	0.0505
500	0.0565	0.0470	0.0485	0.0430	0.0540	0.0490	0.0445
1000	0.0460	0.0430	0.0505	0.0545	0.0400	0.0540	0.0515
2000	0.0560	0.0490	0.0555	0.0505	0.0465	0.0415	0.0485
Panel B: $\mathcal{U}(5\%)$							
90	0.0410	0.0380	0.0335	0.0350	0.0445	0.0345	0.0380
250	0.0470	0.0515	0.0470	0.0460	0.0570	0.0540	0.0500
500	0.0560	0.0460	0.0525	0.0435	0.0530	0.0535	0.0525
1000	0.0465	0.0470	0.0485	0.0560	0.0455	0.0510	0.0500
2000	0.0520	0.0515	0.0420	0.0475	0.0535	0.0525	0.0495
Panel C: $\mathcal{C}(5, 5\%)$							
90	0.1185	0.1165	0.1220	0.1170	0.1075	0.1260	0.1265
250	0.0835	0.0815	0.0895	0.0795	0.0925	0.0860	0.0805
500	0.0805	0.0810	0.0855	0.0740	0.0815	0.0685	0.0690
1000	0.0725	0.0660	0.0575	0.0730	0.0595	0.0590	0.0660
2000	0.0480	0.0610	0.0540	0.0635	0.0670	0.0550	0.0660

For each panel, 2000 time series of in sample length $n = 2500$ and out of sample length $T - n \in \{250, 500, 1000, 2000\}$ have been generated according to a GARCH(1,1) model with standardized skewed Student-t innovations. The null hypothesis assumes the same model specification and (α, β) parameters of the DGP.

Table 3: Out-of-sample power 5% nominal level, standardized skewed Student-t innovations.

	$\alpha_{DGP} = 0.05, \beta_{DGP} = 0.9$					$\alpha_{DGP} = 0.15, \beta_{DGP} = 0.75$				
	$\alpha = 0.05$	$\alpha = 0.45$	$\alpha = 0.9$	$\alpha = 0.05$	$\alpha = 0.05$	$\alpha = 0.05$	$\alpha = 0.45$	$\alpha = 0.9$	$\alpha = 0.05$	$\alpha = 0.05$
$T - n$	$\beta = 0.05$	$\beta = 0.05$	$\beta = 0.05$	$\beta = 0.45$	$\beta = 0.75$	$\beta = 0.05$	$\beta = 0.05$	$\beta = 0.05$	$\beta = 0.45$	$\beta = 0.90$
Panel A: $\mathcal{X}(5, (1.25\%, 2.5\%, 3.75\%, 5\%)')$										
90	0.0965	0.1565	0.8965	0.0870	0.0635	0.1410	0.1275	0.8560	0.1255	0.0765
250	0.1270	0.2420	0.9990	0.1125	0.0815	0.2120	0.1930	0.9990	0.1715	0.0850
500	0.1570	0.3640	1.0000	0.1315	0.0845	0.2560	0.2420	1.0000	0.2105	0.0805
1000	0.1620	0.5600	1.0000	0.1335	0.0820	0.3780	0.3650	1.0000	0.2690	0.1075
2000	0.1960	0.8260	1.0000	0.1620	0.0865	0.5110	0.5510	1.0000	0.4150	0.1280
Panel B: $\mathcal{U}(5\%)$										
90	0.0830	0.1905	0.9690	0.0770	0.0595	0.0915	0.1320	0.9485	0.0780	0.0450
250	0.1170	0.2950	1.0000	0.1250	0.0940	0.1660	0.1940	1.0000	0.1470	0.0425
500	0.1590	0.4360	1.0000	0.1400	0.1035	0.1840	0.2410	1.0000	0.1700	0.0540
1000	0.1550	0.6570	1.0000	0.1410	0.1150	0.1940	0.3460	1.0000	0.1845	0.0610
2000	0.1580	0.8850	1.0000	0.1595	0.1180	0.1900	0.5120	1.0000	0.2065	0.0830
Panel C: $\mathcal{C}(5, 5\%)$										
90	0.1810	0.0860	0.0670	0.1705	0.1505	0.2915	0.1445	0.0755	0.2680	0.2085
250	0.1760	0.0900	0.2180	0.1470	0.1235	0.3700	0.176	0.1720	0.3355	0.2160
500	0.2270	0.1160	0.5090	0.1770	0.1215	0.5030	0.2240	0.4330	0.4480	0.2580
1000	0.2980	0.1660	0.9030	0.2110	0.1140	0.7150	0.3350	0.8230	0.6300	0.3500
2000	0.3940	0.2900	0.9990	0.2750	0.1255	0.9070	0.5410	0.9970	0.8310	0.4670

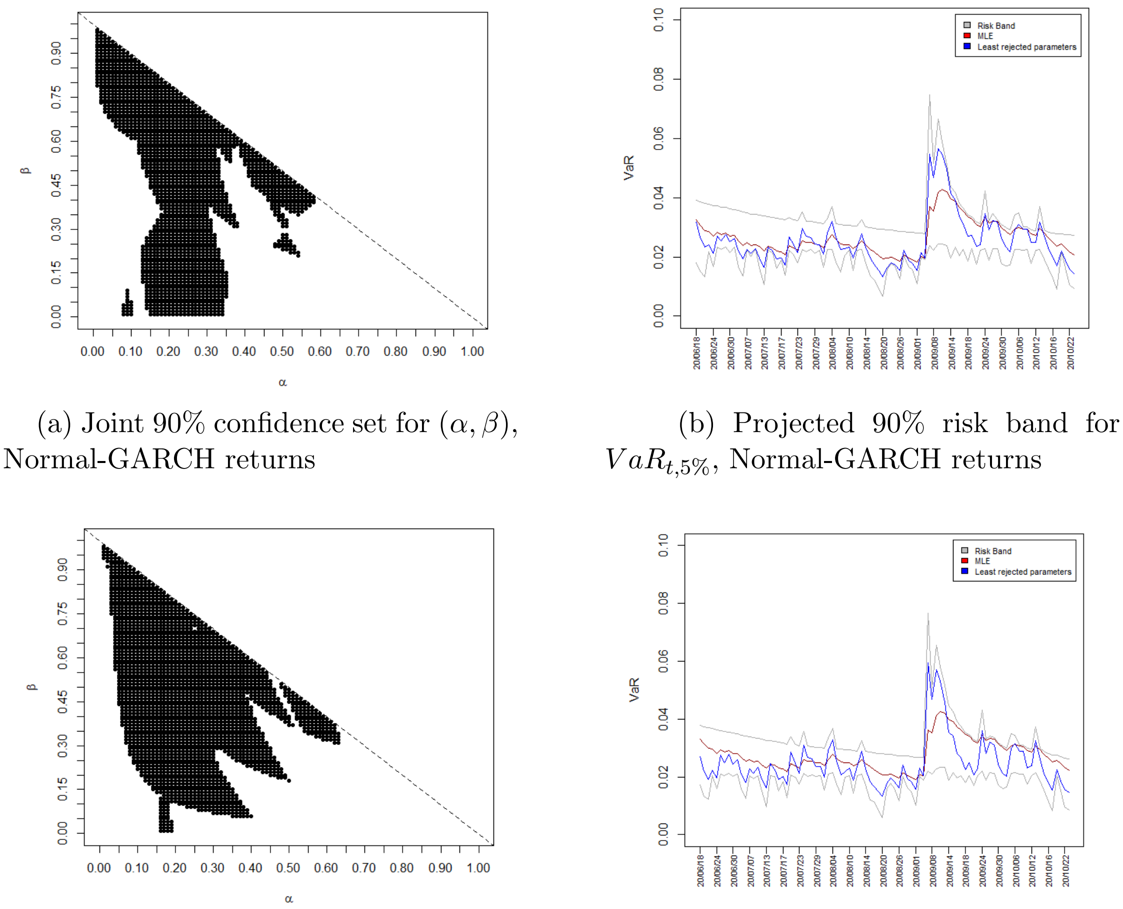
For each panel, 2000 time series of in sample length $n = 2500$ and out of sample length $T - n \in \{90, 250, 500, 1000, 2000\}$ have been generated according to a GARCH(1,1) model with standardized skewed Student-t innovations with parameters $(\alpha_{DGP}, \beta_{DGP})$. Under the null hypothesis, the same model specification of the DGP is assumed, but using the specified (α, β) parameters.

Table 4: Out-of-sample power outside the model 5% nominal level; GARCH-Student-t DGP, GARCH-Normal under the null.

	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.15$	$\alpha = 0.25$	$\alpha = 0.4$	$\alpha = 0.65$	$\alpha = 0.90$
$T - n$	$\beta = 0.90$	$\beta = 0.8$	$\beta = 0.75$	$\beta = 0.6$	$\beta = 0.4$	$\beta = 0.15$	$\beta = 0.05$
Panel A: $\mathcal{X}(5, (1.25\%, 2.5\%, 3.75\%, 5\%)')$							
90	0.0075	0.0070	0.0080	0.0060	0.0145	0.0065	0.0100
250	0.0080	0.0075	0.0120	0.0120	0.0065	0.0085	0.0120
500	0.0240	0.0165	0.0170	0.0210	0.0245	0.0245	0.0230
1000	0.2230	0.2375	0.2190	0.2265	0.2170	0.2175	0.2110
2000	0.9385	0.9345	0.9375	0.9465	0.9375	0.9310	0.9335
Panel B: $\mathcal{U}(5\%)$							
90	0.0210	0.0200	0.023	0.0230	0.0325	0.0260	0.0245
250	0.0730	0.0745	0.077	0.0830	0.0780	0.0795	0.0790
500	0.1190	0.1235	0.123	0.1175	0.1250	0.1195	0.1175
1000	0.1875	0.1950	0.187	0.1945	0.1850	0.1855	0.1885
2000	0.3165	0.3015	0.336	0.3125	0.3285	0.3090	0.3055
Panel C: $\mathcal{C}(5, 5\%)$							
90	0.1780	0.1625	0.1690	0.1815	0.1685	0.1720	0.1830
250	0.1030	0.1245	0.1115	0.1135	0.1035	0.1085	0.0995
500	0.0940	0.1010	0.0805	0.0845	0.0930	0.1025	0.0985
1000	0.0815	0.0805	0.0795	0.0850	0.0885	0.0740	0.0865
2000	0.0700	0.0690	0.0660	0.0720	0.0805	0.0680	0.0775

For each panel, 2000 time series of in sample length $n = 2500$ and out of sample length $T - n \in \{250, 500, 1000, 2000\}$ have been generated according to a GARCH(1,1) model with standardized Student-t innovations. The null hypothesis assumes a GARCH-Normal specification with same (α, β) parameters of the DGP.

Figure 1: $\mathcal{X}$ test inversion on the Technology Select Sector SPDR Fund (XLK)



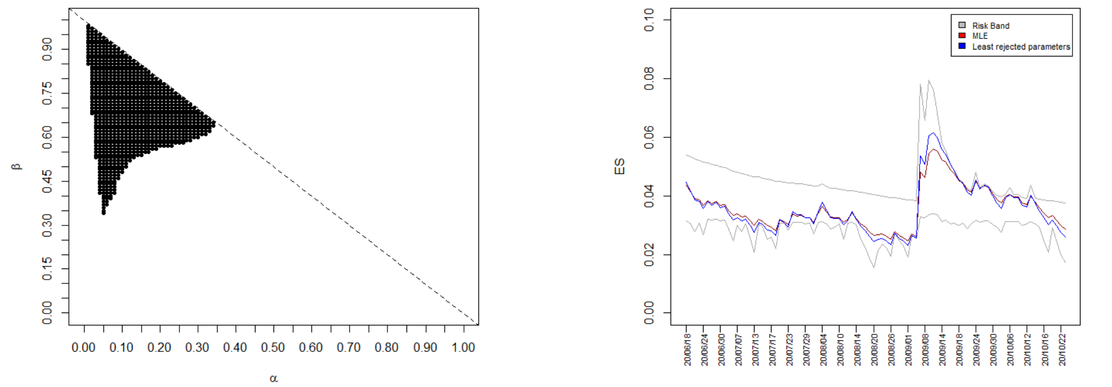
(a) Joint 90% confidence set for (α, β), Normal-GARCH returns

(b) Projected 90% risk band for $VaR_{t,5\%}$, Normal-GARCH returns

(c) Joint 90% confidence set for (α, β), skewed Student t-GARCH returns

(d) Projected 90% risk band for $VaR_{t,5\%}$, skewed Student t-GARCH returns

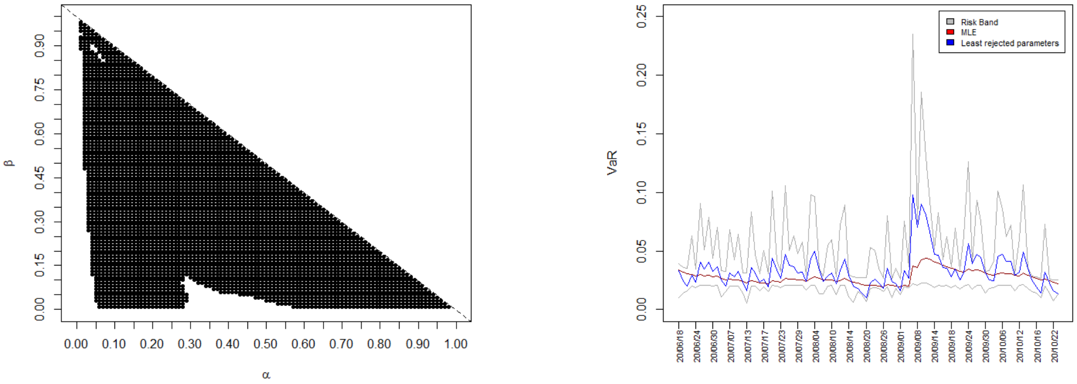
Figure 2: $\mathcal{U}$ test inversion on the Technology Select Sector SPDR Fund (XLK)



(a) Joint 90% confidence set for (α, β), Student t-GARCH returns

(b) Projected 90% risk band for $ES_{t,5\%}$, Student t-GARCH returns

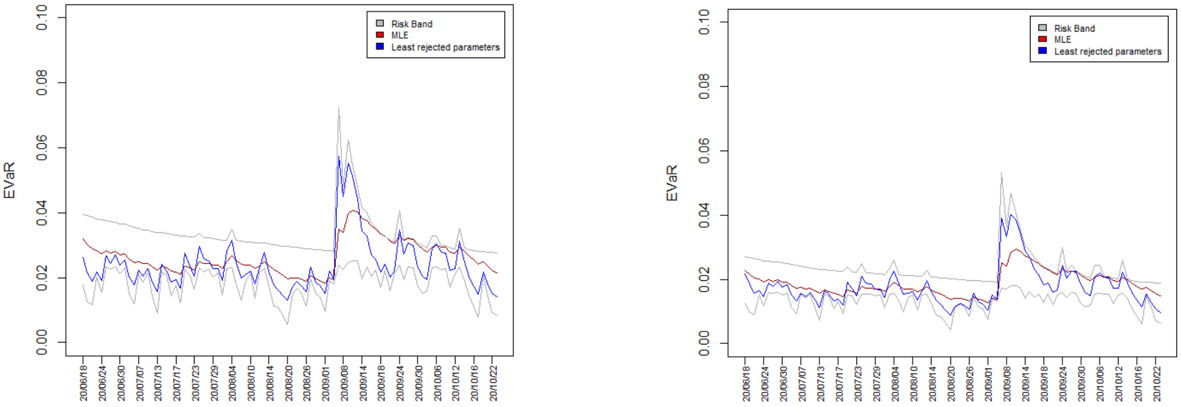
Figure 3: $\mathcal{X}$ test inversion using Filtered Historical Simulation VaR, on the Technology Select Sector SPDR Fund (XLK)



(a) Joint 90% confidence set for (α, β), GARCH returns

(b) Projected 90% risk band for $VaR_{t,5\%}$, GARCH returns

Figure 4: EVaR confidence intervals based on $\mathcal{X}$ test inversion for the Technology Select Sector SPDR Fund (XLK)



(a) Projected 90% risk band for $EVaR_{t,5\%}$, skewed Student t-GARCH returns

(b) Projected 90% risk band for $EVaR_{t,5\%}$, Student t-GARCH returns

Supplementary Material

A Proofs

Proof of Theorem 2.1 As defined, each of the regions $\mathrm{CS}^j_{t,q}(\Lambda^j_{t,q};\tilde{\alpha})$ is the image of the set $\mathrm{CS}(\theta;\tilde{\alpha})$ by the function $g^j_{q,t}(\cdot)$, that is $\theta \in \mathrm{CS}(\theta;\tilde{\alpha}) \implies g^j_{q,t}(\theta,\mathcal{F}_{t-1}) \in \mathrm{CS}^j_{t,q}(\Lambda^j_{t,q};\tilde{\alpha})$ so

$$\mathrm{P}[g^j_{q,t}(\theta,\mathcal{F}_{t-1}) \in \mathrm{CS}^j_{t,q}(\Lambda^j_{t,q};\tilde{\alpha})] \geq \mathrm{P}[\theta \in \mathrm{CS}(\theta;\tilde{\alpha})] \geq 1-\tilde{\alpha}, \text{ for all } \theta \in \Omega, \tag{29}$$

which proves (24).

Furthermore, $\theta \in \mathrm{CS}(\theta;\tilde{\alpha}) \implies \mathrm{L}^j_{t,q}(\Lambda^j_{t,q};\tilde{\alpha}) \leq g^j_{q,t}(\theta,\mathcal{F}_{t-1}) \leq \mathrm{U}^j_{t,q}(\Lambda^j_{t,q};\tilde{\alpha})$ so

$$\mathrm{P}[\mathrm{L}^j_{t,q}(\Lambda^j_{t,q};\tilde{\alpha}) \leq g^j_{q,t}(\theta,\mathcal{F}_{t-1}) \leq \mathrm{U}^j_{t,q}(\Lambda^j_{t,q};\tilde{\alpha})] \geq \mathrm{P}[g^j_{q,t}(\theta,\mathcal{F}_{t-1}) \in \mathrm{CS}^j_{t,q}(\Lambda^j_{t,q};\tilde{\alpha})] \geq 1-\tilde{\alpha} \tag{30}$$

which proves (25). ■

Proof of Theorem 2.2 Under Assumption 1 and Assumption 2, S_b, $b = 1,\ldots,B$, are exchangeable because the distribution underlying (29) is nuisance parameter free, so all parameters required to draw S_b, $b = 1,\ldots,B$, are known. Under the null hypothesis, Proposition 2.4 in Dufour (2006) completes the proof. ■

B VaR and ES formulae

Let $\Phi^{-1}(x)$ refer to the quantile at threshold x of the standard normal distribution and $\phi(x)$ denotes its density function. Then for the Gaussian case

$$VaR_{t,q} = -\sigma_t\Phi^{-1}(q), \quad ES_{t,q} = -\frac{VaR_{t,q}}{\Phi^{-1}(q)}\frac{\phi(\Phi^{-1}(q))}{q}. \tag{31}$$

For the asymmetric Student-t distribution, we follow Fernández and Steel (1998) and consider the following density function

$$\tilde{h}_{\nu,\xi}(x) = \frac{2}{\xi+\xi^{-1}}\left[h_\nu(x\xi)\mathbb{1}(x\leq 0) + h_\nu(x\xi^{-1})\mathbb{1}(x>0)\right] \tag{32}$$

where $h_\nu(x)$ is the density of the conventional Student-t distribution. Let $\mathrm{t}_\nu^{-1}(x)$ refer to the quantile at threshold x of the latter distribution and

$$\epsilon_\nu(x) = \frac{\nu}{1-\nu}\left(1+\frac{x^2}{\nu}\right)h_\nu(x),$$

denote is its lower tail expectation. Then setting $\mu = 2\frac{\nu}{\nu-1}h_\nu(0)$, we have

$$VaR_{t,q} = -\frac{\sigma_{t+1}\left[H_{\nu,\xi}^{-1}(q) - \mu\left(\xi-\xi^{-1}\right)\right]}{\sqrt{\frac{\nu}{\nu-2}\left(\xi^2+\xi^{-2}-1\right) - \mu^2\left(\xi^2+\xi^{-2}-2\right)}}, \quad ES_{t,q} = \frac{\tilde{\epsilon}_{\nu,\xi}(q) - \mu\left(\xi-\xi^{-1}\right)}{H_{\nu,\xi}^{-1}(q) - \mu\left(\xi-\xi^{-1}\right)}VaR_{t,q}$$

where

$$H_{\nu,\xi}^{-1}(x) = \begin{cases} \xi^{-1}\mathrm{t}_\nu^{-1}\left(x(\xi^2+1)/2\right) & \text{if } x \leq \frac{1}{\xi^2+1} \\ \xi\mathrm{t}_\nu^{-1}\left(\xi^{-2}(x(\xi^2+1)+\xi^2-1)/2\right) & \text{if } x > \frac{1}{\xi^2+1} \end{cases},$$

$$\tilde{\epsilon}_{\nu,\xi}(x) = \frac{2}{x(\xi+\xi^{-1})}\begin{cases} \frac{1}{\xi^2}\epsilon_\nu(H_{\nu,\xi}^{-1}(x)\xi) & \text{if } x \leq \frac{1}{\xi^2+1} \\ \frac{1}{\xi^2}\epsilon_\nu(0) + \xi^2\left(\epsilon_\nu(H_{\nu,\xi}^{-1}(x)/\xi) - \epsilon_\nu(0)\right) & \text{if } x > \frac{1}{\xi^2+1} \end{cases}.$$

C Back-tests

The Pearson test statistic from Leccadito et al. (2014), based on K thresholds $\mathbf{q} = (q_1, \cdots, q_K)'$ and m lags, takes the following form, given θ. For $t = n+1, \ldots, T$, let $N_t = \sum_{i=1}^K I_{t,q_i}(R_t, \theta, \mathcal{F}_{t-1})$ where $I_{t,q_i}(R_t, \theta, \mathcal{F}_{t-1})$ is as defined in (12), and let $\mathcal{T}_{x,y}^{(j)}$ denote a $K \times K$ matrix in which the number of time for which $N_t = x$ and $N_{t-j} = y$ is reported then the test statistic is

$$\mathcal{X}(m,\mathbf{q}) = \sum_{j=1}^m \mathcal{X}^{(j)}, \quad \mathcal{X}^{(j)} = \sum_{x,y}\frac{\left(\mathcal{T}_{x,y}^{(j)} - (T-n-j)\left(q_x - q_{x+1}\right)\left(q_y - q_{y+1}\right)\right)^2}{(T-n-j)\left(q_x-q_{x+1}\right)\left(q_y-q_{y+1}\right)}. \tag{33}$$

For a given θ, the unconditional back-test statistic of Du and Escanciano (2017) is:

$$\mathcal{U}(q) = \frac{\sqrt{T-n}\left(\frac{1}{T-n}\sum_{t=n+1}^{T} f_{t,q}^{\mathrm{IV}}(R_t, \theta, \mathcal{F}_{t-1})\right)}{\sqrt{q(\frac{1}{3}-\frac{q}{4})}} \tag{34}$$

which is asymptotically standard normal under the null hypothesis. The conditional back-test statistic with m lags is the Box and Pierce type criterion:

$$\mathcal{C}(m,q) = (T-n)\sum_{j=1}^{m} \hat{\rho}_j^2, \tag{35}$$

$$\hat{\rho}_j = \frac{(T-n)\sum_{t=1+j}^{T-n}\left(f_{t+n,q}^{\mathrm{IV}}(R_{t+n}, \theta, \mathcal{F}_{t+n-1})\right)\left(f_{t+n-j,q}^{\mathrm{IV}}(R_{t+n-j}, \theta, \mathcal{F}_{t+n-j-1})\right)}{(T-n-j)\sum_{t=1}^{T-n}\left(f_{t+n,q}^{\mathrm{IV}}(R_{t+n}, \theta, \mathcal{F}_{t+n-1})\right)^2},$$

which is asymptotically $\chi^2(m)$ under the null hypothesis.

D Monte Carlo size and power of back-tests

Table 5: Out-of-sample size at 5% nominal level, standardized Student-t innovations.

	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.15$	$\alpha = 0.25$	$\alpha = 0.4$	$\alpha = 0.65$	$\alpha = 0.90$
$T-n$	$\beta = 0.90$	$\beta = 0.8$	$\beta = 0.75$	$\beta = 0.6$	$\beta = 0.4$	$\beta = 0.15$	$\beta = 0.05$
Panel A: $\mathcal{X}(5,q)$							
90	0.0445	0.0515	0.0530	0.0465	0.0400	0.0480	0.0490
250	0.0505	0.0405	0.0500	0.0510	0.0470	0.0485	0.0550
500	0.0520	0.0480	0.0460	0.0445	0.0535	0.0445	0.0470
1000	0.0490	0.0435	0.0505	0.0515	0.0510	0.0435	0.0460
2000	0.0435	0.0485	0.0495	0.0480	0.0425	0.0480	0.0435
Panel B: $\mathcal{U}(5\%)$							
90	0.0350	0.0355	0.0320	0.0395	0.0340	0.0400	0.0400
250	0.0445	0.0540	0.0505	0.0465	0.0480	0.0425	0.0510
500	0.0400	0.0580	0.0385	0.0510	0.0565	0.0475	0.0490
1000	0.0570	0.0540	0.0550	0.0475	0.0545	0.0445	0.0485
2000	0.0455	0.0475	0.0480	0.0450	0.0465	0.0460	0.0500
Panel C: $\mathcal{C}(5,5\%)$							
90	0.1325	0.1220	0.1305	0.1265	0.1275	0.1295	0.1150
250	0.0900	0.0840	0.0910	0.0745	0.0875	0.0775	0.0805
500	0.0725	0.0770	0.0740	0.0780	0.0800	0.0665	0.0770
1000	0.0645	0.0615	0.0710	0.0615	0.0655	0.0605	0.0625
2000	0.0580	0.0555	0.0560	0.0570	0.0580	0.0585	0.0555

For each panel, 2000 time series of in sample length $n = 2500$ and out of sample length $T-n \in \{90, 250, 500, 1000, 2000\}$ have been generated according to a GARCH(1,1) model with standardized Student-t innovations. The null hypothesis assumes the same model specification and (α, β) parameters of the DGP.

Table 6: Out-of-sample power 5% nominal level, standard normal innovations.

	$\alpha_{DGP} = 0.05, \beta_{DGP} = 0.9$					$\alpha_{DGP} = 0.15, \beta_{DGP} = 0.75$				
	$\alpha = 0.05$	$\alpha = 0.45$	$\alpha = 0.9$	$\alpha = 0.05$	$\alpha = 0.05$	$\alpha = 0.05$	$\alpha = 0.45$	$\alpha = 0.9$	$\alpha = 0.05$	$\alpha = 0.05$
$T - n$	$\beta = 0.05$	$\beta = 0.05$	$\beta = 0.05$	$\beta = 0.45$	$\beta = 0.75$	$\beta = 0.05$	$\beta = 0.05$	$\beta = 0.05$	$\beta = 0.45$	$\beta = 0.90$
Panel A: $\mathcal{X}(5, q)$										
90	0.0915	0.1445	0.9070	0.0835	0.0705	0.1325	0.1300	0.8640	0.1285	0.0755
250	0.1120	0.2500	1.0000	0.0910	0.0730	0.1890	0.2060	0.9990	0.1700	0.0805
500	0.1400	0.4030	1.0000	0.1060	0.0860	0.2435	0.2870	1.0000	0.1810	0.0940
1000	0.1540	0.6710	1.0000	0.1330	0.0775	0.3485	0.4425	1.0000	0.2795	0.0980
2000	0.1680	0.9050	1.0000	0.1400	0.0780	0.5345	0.6700	1.0000	0.4145	0.1235
Panel B: $\mathcal{U}(5\%)$										
90	0.0785	0.2225	0.9775	0.0825	0.0630	0.0960	0.1540	0.9605	0.1000	0.033
250	0.1370	0.3650	1.0000	0.1180	0.0945	0.1680	0.2320	1.0000	0.1560	0.0545
500	0.1510	0.5235	1.0000	0.1265	0.1075	0.2010	0.2970	1.0000	0.1710	0.0645
1000	0.1540	0.7820	1.0000	0.1350	0.0965	0.2065	0.4500	1.0000	0.1980	0.0635
2000	0.1700	0.9570	1.0000	0.1385	0.1155	0.2065	0.6765	1.0000	0.1865	0.0795
Panel C: $\mathcal{C}(5, 5\%)$										
90	0.1920	0.0945	0.0745	0.1680	0.1490	0.2940	0.1425	0.0670	0.2720	0.1945
250	0.2050	0.0900	0.2125	0.1640	0.1135	0.3990	0.1615	0.1890	0.3545	0.2020
500	0.2295	0.1050	0.5385	0.1650	0.1180	0.5390	0.1890	0.4840	0.4670	0.2670
1000	0.2665	0.1545	0.9180	0.1980	0.1225	0.7195	0.3330	0.8775	0.6680	0.3725
2000	0.3725	0.3080	1.0000	0.2590	0.1305	0.9075	0.5360	0.9985	0.8335	0.4880

For each panel, 2000 time series of in sample length $n = 2500$ and out of sample length $T - n \in \{90, 250, 500, 1000, 2000\}$ have been generated according to a GARCH(1,1) model with Standard Normal innovations with parameters $(\alpha_{DGP}, \beta_{DGP})$. Under the null hypothesis, the same model specification of the DGP is assumed, but using the specified (α, β) parameters.

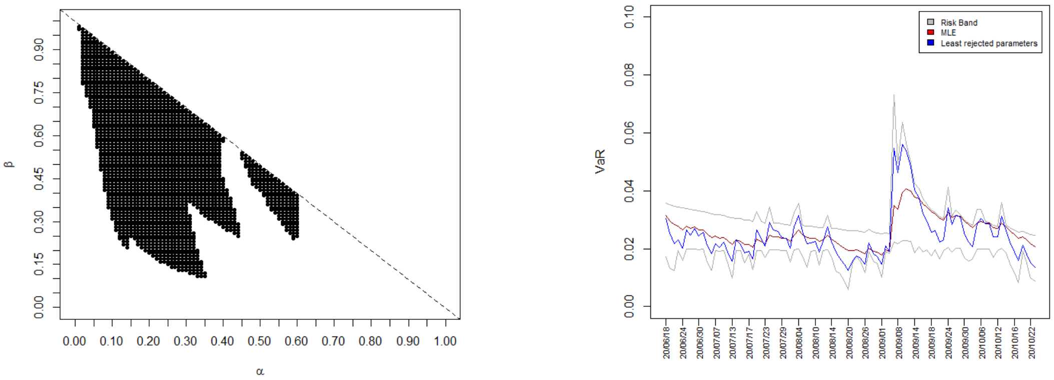
Table 7: Out-of-sample power 5% nominal level, standardized Student-t innovations.

	$\alpha_{DGP} = 0.05, \beta_{DGP} = 0.9$					$\alpha_{DGP} = 0.15, \beta_{DGP} = 0.75$				
	$\alpha = 0.05$	$\alpha = 0.45$	$\alpha = 0.9$	$\alpha = 0.05$	$\alpha = 0.05$	$\alpha = 0.05$	$\alpha = 0.45$	$\alpha = 0.9$	$\alpha = 0.05$	$\alpha = 0.05$
$T - n$	$\beta = 0.05$	$\beta = 0.05$	$\beta = 0.05$	$\beta = 0.45$	$\beta = 0.75$	$\beta = 0.05$	$\beta = 0.05$	$\beta = 0.05$	$\beta = 0.45$	$\beta = 0.90$
Panel A: $\mathcal{X}(5, q)$										
90	0.1120	0.1530	0.9155	0.1005	0.0720	0.1410	0.1360	0.8750	0.1245	0.0855
250	0.1435	0.2160	1.0000	0.1155	0.0755	0.1905	0.1875	0.9980	0.1845	0.0855
500	0.1370	0.3635	1.0000	0.1305	0.0855	0.2695	0.2500	1.000	0.2045	0.0980
1000	0.1645	0.5595	1.0000	0.1400	0.0800	0.3440	0.3675	1.0000	0.2910	0.0955
2000	0.1870	0.8155	1.0000	0.1310	0.0875	0.5025	0.5415	1.0000	0.4040	0.1080
Panel B: $\mathcal{U}(5\%)$										
90	0.0980	0.1820	0.9790	0.0840	0.0630	0.0900	0.1275	0.9630	0.0820	0.0485
250	0.1335	0.2885	1.0000	0.1225	0.0930	0.1540	0.1805	1.0000	0.1560	0.0485
500	0.1480	0.4370	1.0000	0.1215	0.1015	0.1650	0.2295	1.0000	0.1540	0.0570
1000	0.1380	0.6400	1.0000	0.1500	0.1005	0.1810	0.3255	1.0000	0.1725	0.0690
2000	0.1585	0.8835	1.0000	0.1400	0.1055	0.1845	0.5115	1.0000	0.1820	0.0830
Panel C: $\mathcal{C}(5, 5\%)$										
90	0.1870	0.0875	0.0755	0.1750	0.1580	0.2850	0.1410	0.0540	0.2640	0.1990
250	0.1950	0.0940	0.2280	0.1420	0.1150	0.3685	0.1735	0.1935	0.3545	0.2070
500	0.2185	0.1175	0.5515	0.1715	0.1105	0.5190	0.2210	0.4555	0.4560	0.2650
1000	0.2845	0.1575	0.9200	0.2150	0.1095	0.7160	0.3210	0.8525	0.6045	0.3165
2000	0.3595	0.2830	0.9995	0.2465	0.1265	0.8945	0.5035	0.9965	0.8300	0.4390

For each panel, 2000 time series of in-sample length $n = 2500$ and out of sample length $T - n \in \{90, 250, 500, 1000, 2000\}$ have been generated according to a GARCH(1,1) model with standardized Student-t innovations with parameters $(\alpha_{DGP}, \beta_{DGP})$. Under the null hypothesis, the same model specification of the DGP is assumed, but using the specified (α, β) parameters.

E Figures

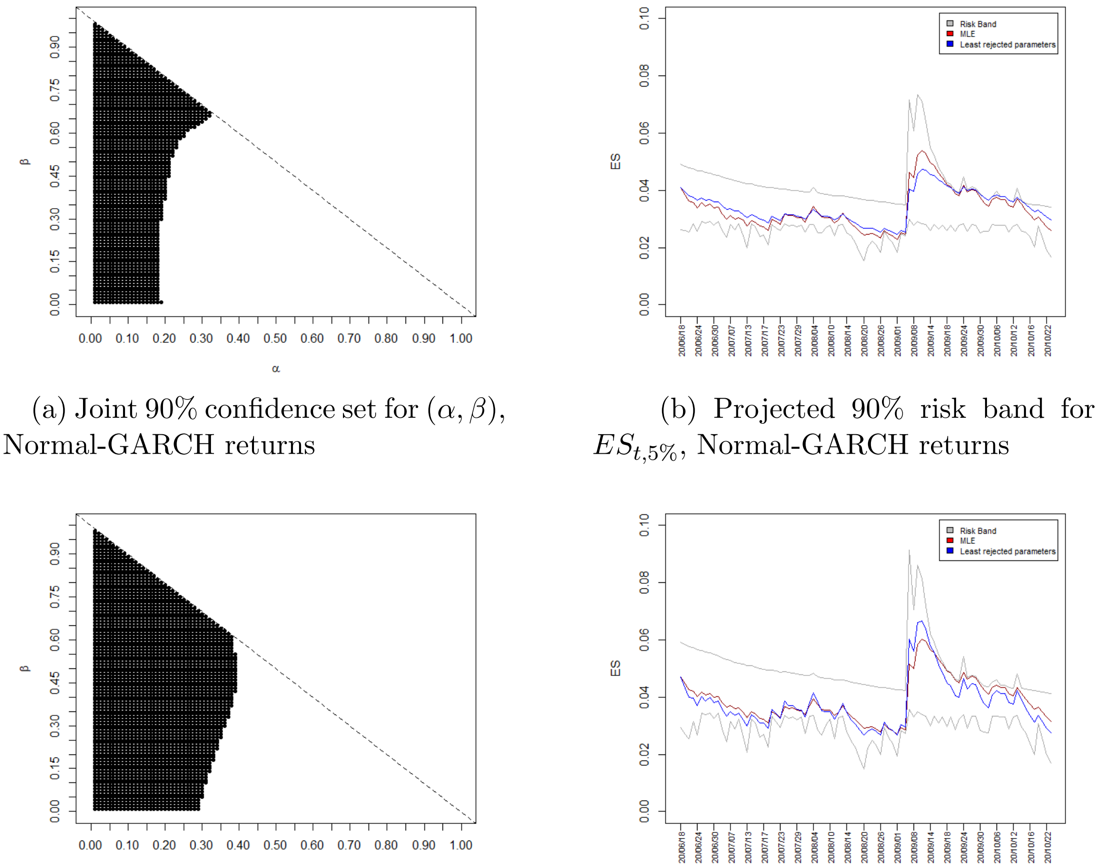
Figure 5: $\mathcal{X}$ test inversion on the Technology Select Sector SPDR Fund (XLK)



(a) Joint 90% confidence set for (α, β), Student t-GARCH returns

(b) Projected 90% risk band for $VaR_{t,5\%}$, Student t-GARCH returns

Figure 6: $\mathcal{U}$ test inversion on the Technology Select Sector SPDR Fund (XLK)



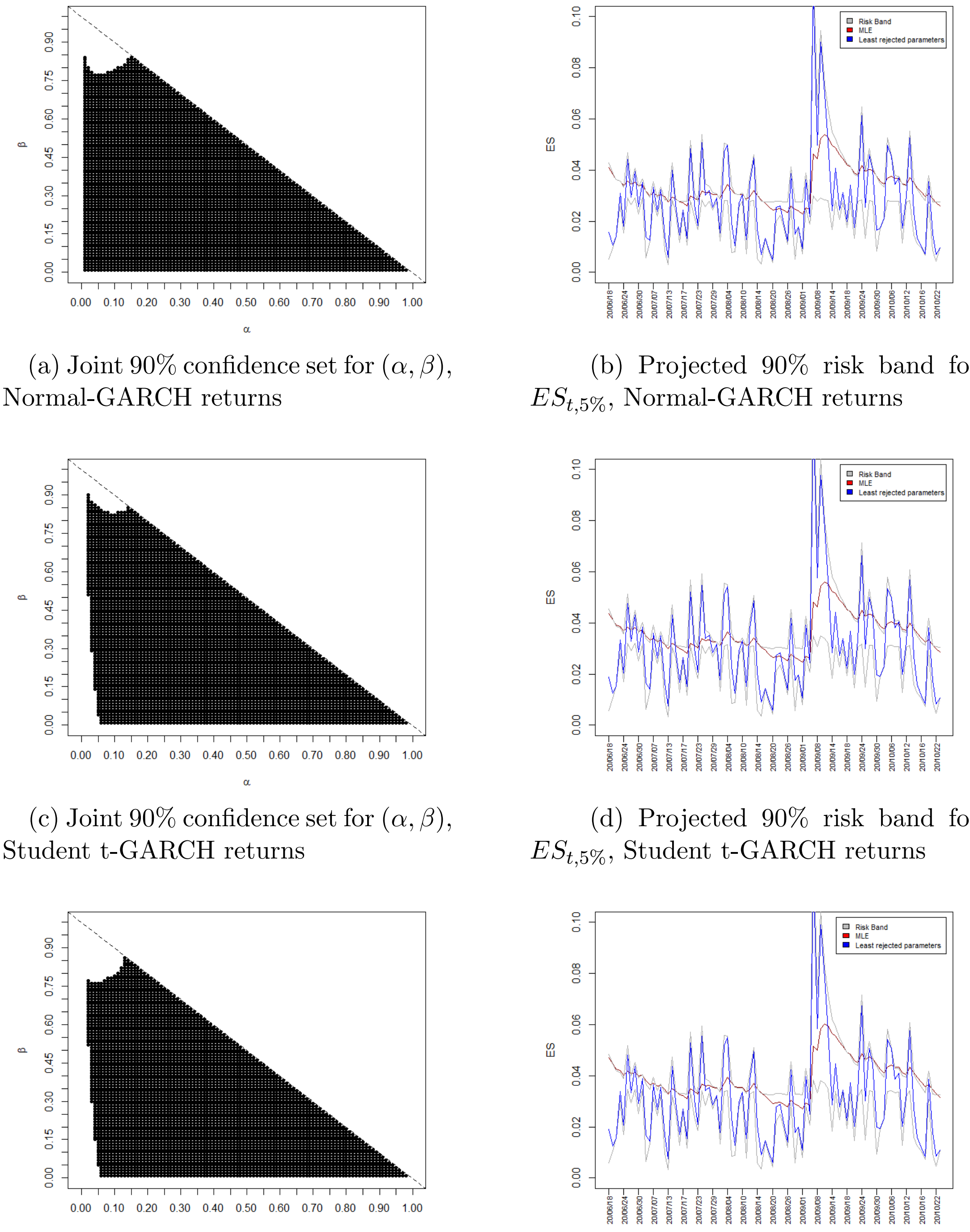
(a) Joint 90% confidence set for (α, β), Normal-GARCH returns

(b) Projected 90% risk band for $ES_{t,5\%}$, Normal-GARCH returns

(c) Joint 90% confidence set for (α, β), skewed Student t-GARCH returns

(d) Projected 90% risk band for $ES_{t,5\%}$, skewed Student t-GARCH returns

Figure 7: $\mathcal{C}$ test inversion on the Technology Select Sector SPDR Fund (XLK)



(a) Joint 90% confidence set for (α, β), Normal-GARCH returns

(b) Projected 90% risk band for $ES_{t,5\%}$, Normal-GARCH returns

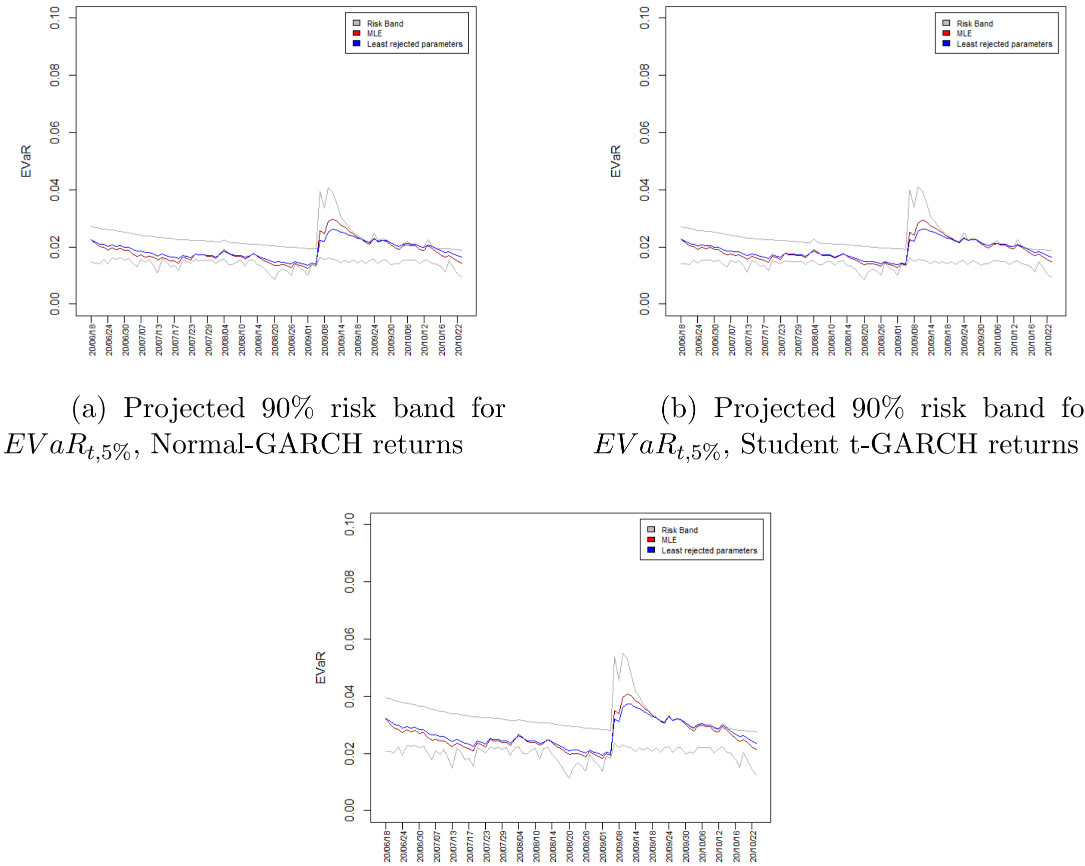
(c) Joint 90% confidence set for (α, β), Student t-GARCH returns

(d) Projected 90% risk band for $ES_{t,5\%}$, Student t-GARCH returns

(e) Joint 90% confidence set for (α, β), skewed Student t-GARCH returns

(f) Projected 90% risk band for $ES_{t,5\%}$, skewed Student t-GARCH returns

Figure 8: EVaR confidence intervals based on $\mathcal{U}$ test inversion for the Technology Select Sector SPDR Fund (XLK)

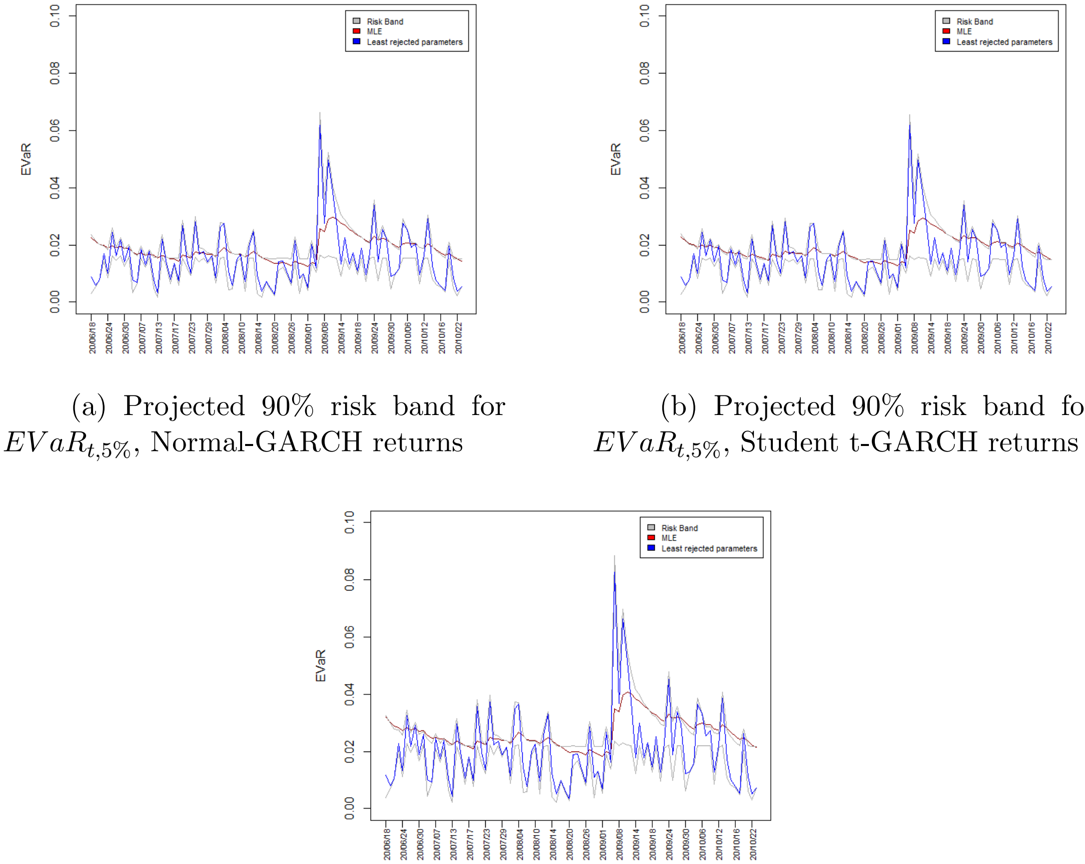


(a) Projected 90% risk band for $EVaR_{t,5\%}$, Normal-GARCH returns

(b) Projected 90% risk band for $EVaR_{t,5\%}$, Student t-GARCH returns

(c) Projected 90% risk band for $EVaR_{t,5\%}$, skewed Student t-GARCH returns

Figure 9: EVaR confidence intervals based on $\mathcal{C}$ test inversion for the Technology Select Sector SPDR Fund (XLK)



(a) Projected 90% risk band for $EVaR_{t,5\%}$, Normal-GARCH returns

(b) Projected 90% risk band for $EVaR_{t,5\%}$, Student t-GARCH returns

(c) Projected 90% risk band for $EVaR_{t,5\%}$, skewed Student t-GARCH returns