Robust Forecast Evaluation of Expected Shortfall[a]

Johanna F. Ziegel[b]
University of Bern

Fabian Krüger
Heidelberg University

Alexander Jordan
University of Bern

Fernando Fasciati
Raiffeisen Schweiz

October 25, 2018

[a]We thank seminar and conference participants in Heidelberg, Augsburg (Statistische Woche 2016), Karlsruhe (HeiKaMEtrics 2018) and Freiburg (GPSD 2018) for helpful comments. Johanna F. Ziegel gratefully acknowledges financial support of the Swiss National Science Foundation. The work of Fabian Krüger and Alexander Jordan has been funded by the European Union Seventh Framework Programme under grant agreement 290976. They also thank the Klaus Tschira Foundation for infrastructural support at the Heidelberg Institute for Theoretical Studies (HITS). The opinions expressed in this article are those of the authors do not necessarily reflect the views of Raiffeisen Schweiz. Calculations were performed on the HPC cluster at HITS, and UBELIX (http://www.id.unibe.ch/hpc), the HPC cluster at the University of Bern.

[b]Corresponding author. Affiliation: Institute of Mathematical Statistics and Actuarial Science, University of Bern, Switzerland. Postal Address: Alpeneggstrasse 22, CH-3012 Bern. Phone: +41 31 631 88 03. Email: johanna.ziegel@stat.unibe.ch.

Abstract

Motivated by the Basel III regulations, recent studies have considered joint forecasts of Value-at-Risk and Expected Shortfall. A large family of scoring functions can be used to evaluate forecast performance in this context. However, little intuitive or empirical guidance is currently available, which renders the choice of scoring function awkward in practice. We therefore develop graphical checks of whether one forecast method dominates another under a relevant class of scoring functions, and propose an associated hypothesis test. We illustrate these tools with simulation examples and an empirical analysis of S&P 500 and DAX returns.

Keywords: Forecasting, Expected Shortfall

JEL Classifications: C52, C53, G17

1 Introduction

The Basel III standard on minimum capital requirements for market risk (Basel Committe on Banking Supervision, 2016) uses Expected Shortfall (ES), rather than Value-at-Risk (VaR), to quantify the risk of a bank's portfolio. As described by McNeil et al. (2015, Chapter 8), ES possesses several desirable theoretical properties. However, it also has a major drawback: It is not elicitable, i.e. there is no scoring function that sets the incentive to report ES honestly, or that can be used to compare ES forecasts' accuracy.[1] As a partial remedy to this problem, Fissler and Ziegel (2016, henceforth FZ) show that ES is jointly elicitable with VaR and characterize the class of scoring functions that can be used to evaluate forecasts of type (VaR, ES). Fissler et al. (2016) provide a nontechnical introduction and discuss regulatory implications.

In applied work, it may be challenging to select a specific member function from the FZ family on either economic or statistical grounds. Furthermore, different choices of scoring functions may yield different forecast rankings in the case of imperfect forecasters or non-nested information sets, see for example Patton (2016). Motivated by this problem, we present a mixture representation using elementary members of the FZ family, which is mathematically similar to recent results by Ehm et al. (2016) for quantiles and expectiles. The mixture representation gives rise to Murphy diagrams which allow to check whether one forecast dominates another under a relevant class of scoring functions.[2] While this class could be the entire FZ family, we argue that a subfamily that emphasizes ES – rather than VaR – is economically more desirable in the light of the Basel III standard. Analyzing the robustness

[1] As detailed below, a scoring function (or loss function) assigns a real-valued score, given a forecast and a realizing observation.

[2] The name of the diagrams alludes to the meteorologist Allan H. Murphy (1931–1997) who pioneered similar diagrams in the context of a binary dependent variable (see Murphy 1977, as well as Ehm et al. 2016, p. 519).

of forecast rankings across this class of scoring functions is relevant both conceptually and practically, and referred to as *forecast dominance* in the following.

Forecast dominance holds at the population level - that is, it is defined in terms of expected performance, which is unobservable. Statistical tests are based on observable performance statistics, and are designed to detect significant deviations from hypotheses about expected performance; see e.g. Diebold and Mariano (1995) and Clark and McCracken (2013). In the present context, such tests are complicated by the fact that the null hypothesis refers to performance under all elementary members of the mixture representation, i.e. for all values of an auxiliary parameter. To tackle the resulting testing problem, we propose a variant of the test by Hansen (2005). The test was originally developed to conduct multiple comparisons among a finite set of forecast models. By contrast, our situation involves an infinite family of elementary functions which enter the comparison. We provide theoretical and simulation evidence that the test has good size and power properties in the present situation. Our test complements recent work by Ehm and Krüger (2018) and Yen and Yen (2018) who consider tests of forecast dominance for quantiles and expectiles. These papers differ from ours along several dimensions. First, they consider different forecast types (functionals) than we do. Second, their theoretical justification is based on Gaussian processes, whereas we use arguments from the multiple testing literature (Westfall and Young, 1993; Cox and Lee, 2008). Finally, Ehm and Krüger (2018) use independent permutation (rather than dependent bootstrap) methods to implement the null hypothesis.

In an empirical case study, we evaluate forecasts for daily log returns of the S&P 500 and DAX stock market indices. Three models with varying degree of sophistication are considered: the HEAVY model (Shephard and Sheppard, 2010) with access to the past's intra-daily data competes against two models using merely end-of-day data, a GARCH(1,1) model (Bollerslev, 1986) and a naive 'historical simulation' model. Our results suggest that both the HEAVY and the GARCH model dominate simple historical simulation; in contrast, we find no dominance relation between HEAVY and GARCH.

We emphasize that our interest lies in comparative forecast evaluation – that is, we seek to compare the (VaR, ES) forecasts of two competing methods.[3] Comparative evaluation is important to select a suitable forecasting method in practice, especially given the wealth of data sources and statistical techniques that could plausibly be used to generate forecasts. Comparative forecast evaluation is different from absolute evaluation which aims to determine whether a given forecast method possesses certain desirable optimality properties. The Basel II procedure of counting VaR 'violations', i.e. the number of times the actual return fell below the VaR forecast, is an example of absolute forecast evaluation. See Nolde and

[3]In financial jargon, the word 'backtesting' is sometimes used as a synonym for 'forecast evaluation'.

Ziegel (2017) for a detailed discussion of comparative versus absolute evaluation of financial forecasts.

The contributions of the present paper include a mixture representation of the FZ family in Section 2, which yields the Murphy diagrams, and a test for the hypothesis of forecast dominance. Sections 3 introduces the test and provides a theoretical justification; Section 4 presents simulation evidence on the test's size and power. We illustrate the mixture representation and the test in an empirical case study in Section 5. A discussion in Section 6 concludes. Three appendices contain proofs and technical details.

2 Consistent Scoring Functions for VaR and Expected Shortfall

To keep notation light, we start with a single-period outcome and move on to time-series considerations in the next section. Let $Y \in \mathbb{R}$ be a random variable describing the single-period return of a financial asset, where a negative return, $Y < 0$, corresponds to a loss. Value-at-Risk (VaR) and Expected Shortfall (ES) are popular measures of tail risk. Let F denote the distribution of Y, and assume that Y has finite mean. Then for a given level $\alpha \in (0, 1)$, the VaR and ES are defined as

$$\mathrm{VaR}_\alpha(F) = \inf\{z \in \mathbb{R} : F(z) \geq \alpha\}$$

and

$$\mathrm{ES}_\alpha(F) = \frac{1}{\alpha} \int_0^\alpha \mathrm{VaR}_u(F)\, \mathrm{d}u.$$

If $\mathrm{VaR}_\alpha(F)$ is unique, then $\mathrm{ES}_\alpha(F)$ can also be written as

$$\mathrm{ES}_\alpha(F) = \mathbb{E}(Y | Y \leq \mathrm{VaR}_\alpha(F)).$$

The latter representation explains the name 'expected shortfall': $\mathrm{ES}_\alpha(F)$ is the expectation of Y, given that Y is below its VaR. We are interested in small values of α, in particular $\alpha = 0.025$ which is the level that the Basel Committe on Banking Supervision (2016) requests for ES predictions. Then, VaR_α and ES_α will typically have negative values. Our sign convention corresponds to the sign convention of utility functions as used in Delbaen (2012) and it implies that $\mathrm{VaR}_\alpha \geq \mathrm{ES}_\alpha$ always holds.

Following Gneiting (2011a), Ehm et al. (2016), Patton (2016) and others, it is now widely recognized that consistent scoring functions are essential for comparing point forecasts. Consistency implies that, on average, a misspecified model may not outperform a correct model.

As discussed in Fissler and Ziegel (2016), ES_α cannot be evaluated consistently without joint consideration of VaR_α, so we stack the two functionals to obtain the two-dimensional functional

$$\mathrm{T}_\alpha(F) = (\mathrm{VaR}_\alpha(F), \mathrm{ES}_\alpha(F)).$$

As return distributions we consider members of the class $\mathcal{F}_1$ of distributions with finite mean and unique quantiles. The latter assumption allows us to simplify our presentation and does not seem restrictive in the context of financial returns. For example, the HEAVY and GARCH models used in our case study (Section 4) clearly satisfy the assumption. As forecasts of type T_α, where v is a forecast of VaR_α and e is a forecast of ES_α, we consider elements of the action domain $\mathsf{A}_0 = \{(v, e) \in \mathbb{R}^2 : v \geq e\}$, thereby ruling out irrational forecasts that violate $\mathrm{VaR}_\alpha \geq \mathrm{ES}_\alpha$. The following definition formalizes the notion of a consistent scoring function for T_α.

Definition 2.1. A *scoring function* $\mathrm{S} : \mathsf{A}_0 \times \mathbb{R} \to \mathbb{R}$ is a function such that $\int \mathrm{S}(v, e, y)\, \mathrm{d}F(y)$ exists for all $F \in \mathcal{F}_1$, $(v, e) \in \mathsf{A}_0$. The scoring function S is called *consistent* for T_α if

$$\mathbb{E}(\mathrm{S}(\mathrm{T}_\alpha(F), Y)) \leq \mathbb{E}(\mathrm{S}(v, e, Y)) \tag{1}$$

for all $(v, e) \in \mathsf{A}_0$ and all random variables Y with distribution in $\mathcal{F}_1$. The scoring function S is *strictly consistent* if equality in (1) implies $(v, e) = \mathrm{T}_\alpha(F)$.

Equation (1) says that, in expectation, it is a forecaster's best possible action to state the forecast $\mathrm{T}_\alpha(F)$, rather than an arbitrary alternative $(v, e) \in \mathsf{A}_0$. In this sense, a consistent scoring function sets the incentive for honest and accurate forecasting of T_α. Importantly, there is not only one scoring function that is consistent for T_α. Instead, there is a whole family of scoring functions with this property.[4] Here, we consider normalized scores for which $\mathrm{S}(y, y, y) = 0$ holds true. This normalization is in line with much of the existing literature (e.g. Gneiting, 2011a); other normalizations can easily be accommodated. Corollary 5.5 of Fissler and Ziegel (2016) implies that all scoring functions S of the form

$$\begin{aligned} \mathrm{S}(v, e, y) = {} & \big(\mathbb{1}\{y \leq v\} - \alpha\big)\big(G_1(v) - G_1(y)\big) \\ & + G_2(e)\Big(\frac{1}{\alpha}\mathbb{1}\{y \leq v\}(v - y) - (v - e)\Big) \\ & - \big(\mathcal{G}_2(e) - \mathcal{G}_2(y)\big), \end{aligned} \tag{2}$$

are consistent scoring functions for T_α, where G_1, G_2, and $\mathcal{G}_2$ are functions from $\mathbb{R}$ to

[4] The situation is similar for other functionals, i.e., there is typically a whole family of scoring functions that are consistent for a given functional. For example, Savage (1971) identifies a family of scoring functions that are consistent for the mean, and Gneiting (2011b) describes the family of scoring functions that are consistent for a quantile.

$\mathbb{R}$, $\mathcal{G}_2' = G_2$ (i.e., G_2 is the derivative of $\mathcal{G}_2$), G_1 and G_2 are increasing, $G_2 \geq 0$, and $\int G_1(y)\,\mathrm{d}F(y)$, $\int \mathcal{G}_2(y)\,\mathrm{d}F(y)$ exist and are finite for all $F \in \mathcal{F}_1$. If G_2 is strictly increasing, we obtain strict consistency. For example, the choice $G_1(z) = 0, G_2(z) = \exp(z)/(1+\exp(z))$ satisfies all of these requirements but there are many alternatives. Subject to regularity conditions, all normalized consistent scoring functions on the action domain A_0 are of the form (2).

Patton (2011), Nolde and Ziegel (2017) and Patton et al. (2018) have argued for the use of homogeneous scoring functions for forecast comparison. Such additional requirements on the scoring functions narrow down the possible choices for G_1, G_2 and $\mathcal{G}_2$ in (2). On a restricted action domain $\mathsf{A} = \mathbb{R} \times (-\infty, 0) \cap \mathsf{A}_0$, homogeneous scoring functions for T_α exist (Nolde and Ziegel, 2017, Theorem C.3), and under some additional assumptions there is even a unique zero-homogeneous choice (Patton et al., 2018, Proposition 1). Nevertheless, choosing one single scoring function for forecast evaluation implicitly imposes an order of preference on all sequences of forecasts which is usually hard and sometimes impossible to justify. Indeed, the results in Nolde and Ziegel (2017) show no clear preference between a zero-homogeneous choice for S or a (1/2)-homogeneous choice for S with respect to performance in forecast comparison. In their simulation study, these two different choices give emphasis to different aspects of model misspecification.

More generally, Patton (2016) and others have demonstrated that the choice of scoring function is relevant for the ranking of two competing forecasts in the presence of model misspecification and non-nested information sets, both of which are common in practice. The methods we consider in this paper are robust with respect to the choice of scoring function, in the sense that we compare forecasts under a class of scoring functions. We therefore make the following definition of forecast dominance which is analogous to Ehm et al. (2016, Definition 1).

Definition 2.2. Let $\alpha \in (0, 1)$ and let $\mathcal{S}$ be a class of consistent scoring functions for T_α. For two (possibly random) forecasts (V^A, E^A) and (V^B, E^B) made by methods A and B, respectively, we say that method A *weakly dominates* method B *with respect to* $\mathcal{S}$ if

$$\mathbb{E}\left(\mathrm{S}(V^A, E^A, Y)\right) \leq \mathbb{E}\left(\mathrm{S}(V^B, E^B, Y)\right), \quad \text{for all } \mathrm{S} \in \mathcal{S},$$

where the expectations are with respect to the joint distribution of (V^A, E^A, V^B, E^B, Y).

Once dominance has been established for a given class $\mathcal{S}$, it can be translated to the extension including all mixtures, e.g. dominance with respect to $\{\mathrm{S}_1, \mathrm{S}_2\}$ implies dominance with respect to $\{a\mathrm{S}_1 + b\mathrm{S}_2 : a, b \geq 0\}$. This simple observation is the basis for so-called Murphy diagrams which are graphical tools to check for forecast dominance empirically with

respect to all consistent scoring functions. Ehm et al. (2016) provide mixture representations of the families of consistent scoring functions for quantiles and expectiles. In order to derive similar methodology for $(\mathrm{VaR}_\alpha(F), \mathrm{ES}_\alpha(F))$, the following result presents a mixture representation for consistent scoring functions of the form given in (2).

Proposition 2.1. *Let $\alpha \in (0,1)$. For $\eta, y \in \mathbb{R}$, $(v,e) \in \mathsf{A}_0$, we define the elementary scores*

$$\begin{aligned}
\mathrm{S}_{\eta,1}(v,y) &= (\mathbb{1}\{y \le v\} - \alpha)\big(\mathbb{1}\{\eta \le v\} - \mathbb{1}\{\eta \le y\}\big) \\
\mathrm{S}_{\eta,2}(v,e,y) &= \mathbb{1}\{\eta \le e\}\left(\frac{1}{\alpha}\mathbb{1}\{y \le v\}(v-y) - (v-\eta)\right) + \mathbb{1}\{\eta \le y\}(y-\eta).
\end{aligned}$$

Let H_1 be a locally finite measure and H_2 a measure that is finite on all intervals of the form $(-\infty, x]$, $x \in \mathbb{R}$. Then all scoring functions $\mathrm{S} : \mathsf{A}_0 \times \mathbb{R} \to \mathbb{R}$ that are of the form (2) can be written as

$$\mathrm{S}(v,e,y) = \int \mathrm{S}_{\eta,1}(v,y)\,\mathrm{d}H_1(\eta) + \int \mathrm{S}_{\eta,2}(v,e,y)\,\mathrm{d}H_2(\eta). \tag{3}$$

The scores at (3) are consistent for T_α. They are strictly consistent if H_2 puts positive mass on all open intervals.

The first integral in Equation (3) represents the first line of Equation (2), whereas the second integral in (3) represents the second and third line of (2). In fact, for $x_1 \le x_2$, we have $H_1((x_1,x_2]) = G_1(x_2) - G_1(x_1)$ and for $x \in \mathbb{R}$, we have $H_2((-\infty,x]) = G_2(x)$. The scores $\mathrm{S}_{\eta,1}$ and $\mathrm{S}_{\eta,2}$ are themselves consistent scoring functions for T_α, which follows immediately by choosing Dirac-measures for H_1 or H_2 in (3). The score $\mathrm{S}_{\eta,1}(v,y)$ goes to zero as $\eta \to \pm\infty$, whereas the score $\mathrm{S}_{\eta,2}(v,e,y)$ goes to zero as $\eta \to +\infty$, and converges to $(1/\alpha)(\mathbb{1}\{y \le v\} - \alpha)(v-y)$ as $\eta \to -\infty$. This explains the different restrictions on the corresponding mixing measures H_1 and H_2 in Proposition 2.1.

We identify a subclass of consistent scoring functions for T_α whose members emphasize the evaluation of the ES_α component. The first integral in (3) corresponds to the mixture representation of consistent scoring functions for quantiles (Ehm et al., 2016, Theorem 1a), a class that in our context only evaluates the VaR_α forecast and ignores ES_α. Hence, choosing anything but a constant H_1 puts unnecessary emphasis on the VaR_α component of the forecast. The second integral corresponds to the evaluation of ES_α, conditional on VaR_α, where we cannot completely extinguish VaR_α in the evaluation due to the results on the (non-)elicitability of ES_α. Hence, we define $\mathcal{S}_2$ as the class of all consistent scoring functions for T_α as given at (3) with a constant H_1 (such that the first integral is zero), and focus on this class in the following.[5] In the context of the class $\mathcal{S}_2$, we denote the elementary scores $\mathrm{S}_{\eta,2}$ simply by S_η, since the scores $\mathrm{S}_{\eta,1}$ have been excluded.

[5] Another representation of the class $\mathcal{S}_2$, which does not make use of elementary scores, can be obtained by setting the function G_1 to zero in Equation (2).

Our focus on $\mathcal{S}_2$ is motivated by the aim to maximize the impact of the ES_α component in evaluation, which is in line with the emphasis set in Basel III. Focusing on $\mathcal{S}_2$ also seems justified from a statistical perspective: Dimitriadis and Bayer (2017) investigate several members of $\mathcal{S}_2$ in a regression framework. They argue that moving beyond $\mathcal{S}_2$ (i.e., considering non-constant choices of H_1 in Equation 2.1) does not improve the numerical performance of their estimators. Furthermore, $\mathcal{S}_2$ contains in particular positively homogeneous scoring functions for all possible degrees of homogeneity; see Nolde and Ziegel (2017, Section 2.3.1 and Theorem 6). As discussed there, positively homogeneous scoring functions enjoy a number of attractive properties.

The mixture representation at (3) allows graphical displays of the performance of T_α forecasts with respect to the elementary scores of $\mathcal{S}_2$,

$$\eta \mapsto \mathbb{E}(\mathrm{S}_\eta(V, E, Y)),$$

where the expectation is with respect to the joint distribution of (V, E, Y). In practice, the expectation is estimated by the average observed score. Examples of these displays, called Murphy diagrams (Ehm et al., 2016), are given in Figure 2 in Section 5. The diagrams provide simple graphical checks of whether one forecast dominates another under all scoring functions in $\mathcal{S}_2$. Specifically, Proposition 2.1 implies that the forecast of method A dominates that of method B with respect to $\mathcal{S}_2$ if and only if

$$\mathbb{E}\left(\mathrm{S}_\eta(V^A, E^A, Y)\right) \leq \mathbb{E}\left(\mathrm{S}_\eta(V^B, E^B, Y)\right) \quad \text{for all } \eta \in \mathbb{R};$$

this condition is analogous to the one in Ehm et al. (2016, Corollary 1) for quantiles and expectiles.

Clearly, one could also consider forecast dominance for T_α with respect to all consistent scoring functions. The procedures described in the following can be adapted to this case; an extension that is conceptually simple yet tedious in practice. This is because one needs to check inequalities across two grids of parameters, one for $\mathrm{S}_{\eta,1}$ and one for $\mathrm{S}_{\eta,2}$. Instead, when focusing on $\mathcal{S}_2$, it suffices to check inequalities along a single grid for η.

3 Testing forecast dominance

Here we first translate the methodology from Section 2 into a time series context, and then introduce a test of forecast dominance based on the elementary scores.

3.1 Comparing time series forecasts

So far, we have only considered a one-period forecasting problem. In most financial applications, however, the goal is to predict a time series $\{Y_t\}_{t\in\mathbb{N}}$, such as a sequence of asset returns observed at trading days $t = 1, 2, \ldots$. Furthermore, let $(V_t, E_t)' \in \mathsf{A}_0$ denote the $(\mathrm{VaR}_\alpha, \mathrm{ES}_\alpha)$ forecast of Y_t, based on an appropriate information set $\mathcal{W}_{t-1}$ generated by data available at time $t-1$. In applications, we seek to make forecasts and realizations comparable across time. We therefore require the following assumption.

Assumption 3.1. The time series $\{Z_t\}_{t\in\mathbb{N}}$ with $Z_t = (V_t, E_t, Y_t) \in \mathsf{A}_0 \times \mathbb{R}$ is stationary with distribution F_Z.

This assumption rules out deterministic time trends, structural breaks and seasonalities, among others. Nevertheless, many multivariate autoregressive models (e.g. Lütkepohl, 2005) or stochastic volatility models (e.g. Harvey et al., 1994) are stationary.

Consider any consistent scoring function S for T_α. Assumption 3.1 implies that the distribution of the random variable $\mathrm{S}(V_t, E_t, Y_t)$ does not depend on time, t. In particular, this holds when S equals an elementary score S_η in the class $\mathcal{S}_2$. We can thus define the notion of an expected elementary score, as follows. Consider a sequence of forecasts $\{V_t, E_t\}_{t\in\mathbb{N}}$ and corresponding realizations $\{Y_t\}_{t\in\mathbb{N}}$ which jointly define a stationary time series as in Assumption 3.1. The expected elementary scores for this process are given by

$$\mathbb{E}\left(\mathrm{S}_\eta(V_t, E_t, Y_t)\right) = \int_{\mathsf{A}_0\times\mathbb{R}} \mathrm{S}_\eta(v, e, y)\ \mathrm{d}F_Z(v, e, y), \tag{4}$$

where F_Z is defined in Assumption 3.1. Based on this definition, a notion of forecast dominance 'on average over time' follows naturally:

Definition 3.1. Let $\{V_t^A, E_t^A\}_{t\in\mathbb{N}}$ and $\{V_t^B, E_t^B\}_{t\in\mathbb{N}}$ denote two competing sequences of forecasts of $(\mathrm{VaR}_\alpha, \mathrm{ES}_\alpha)$, and let $\{Y_t\}_{t\in\mathbb{N}}$ denote the corresponding realizations such that $\{(V_t^A, E_t^A, Y_t)\}_{t\in\mathbb{N}}$ and $\{(V_t^B, E_t^B, Y_t)\}_{t\in\mathbb{N}}$ both satisfy Assumption 3.1 with stationary distributions F_Z^A and F_Z^B, respectively. We say that method A *weakly dominates* method B with respect to $\mathcal{S}_2$ if

$$\mathbb{E}\left(\mathrm{S}_\eta(V_t^A, E_t^A, Y_t)\right) \le \mathbb{E}\left(\mathrm{S}_\eta(V_t^B, E_t^B, Y_t)\right) \quad \text{for all } \eta \in \mathbb{R},$$

where the expectations are as at (4) with respect to the corresponding stationary distribution.

If $\{\mathrm{S}_\eta(V_t^A, E_t^A, Y_t)\}_{t\in\mathbb{N}}$ and $\{\mathrm{S}_\eta(V_t^B, E_t^B, Y_t)\}_{t\in\mathbb{N}}$ are ergodic, the expectations in Definition 3.1 can be consistently estimated by empirical averages over observed forecasts and

realizations at dates $t = 1, \ldots, n$, e.g. as $n \to \infty$ it holds that

$$\frac{1}{n} \sum_{t=1}^{n} \mathrm{S}_\eta(V_t^A, E_t^A, Y_t) \xrightarrow{p} \mathbb{E}\left(\mathrm{S}_\eta(V_t^A, E_t^A, Y_t)\right),$$

and analogously for method B.

3.2 Testing for forecast dominance

We are interested in the following null hypothesis:

H0: Method A weakly dominates method B;

Definition 3.1 gives a formal statement of the hypothesis. Importantly, we consider tests of finite-sample predictive ability (Giacomini and White, 2006; Clark and McCracken, 2013, Section 3.2) throughout this paper. That is, we ask whether method A outperforms method B, taking into account that both methods are based on imperfect parameter estimates. This setup is closely aligned with practical forecast comparisons where parameter uncertainty is a relevant concern.[6] Our proposed testing procedure employs the test by Hansen (2005). Whereas the test was originally designed to handle a finite number of comparisons, we adapt it to our problem which involves an infinite number of comparisons. We describe the procedure in the present section. In Sections 3.3 and 3.4, we present theory and connections to the multiple testing literature.

For each $\eta \in \mathbb{R}$, $t \in \mathbb{N}$, consider the score difference between methods A and B,

$$\delta_t(\eta) := S_\eta(V_t^A, E_t^A, Y_t) - S_\eta(V_t^B, E_t^B, Y_t),$$

with expectation $\mu(\eta)$ and standard deviation $\sigma(\eta)$. This translates to a null hypothesis of

$$\mu(\eta) \leq 0 \quad \text{for all } \eta \in \mathbb{R}.$$

[6] By contrast, comparisons of population-level predictive ability (Clark and McCracken, 2013, Section 3.1) ask whether model A would outperform model B if both models were estimated without error. They are useful to discriminate between alternative theories or assess the possible impact of a certain regressor, but are less in line with practical forecast situations which we consider here.

To construct test statistics, we use the finite sample estimates of $\mu(\eta)$ and $\sigma(\eta)$,

$$\begin{aligned}
\mu_n(\eta) &= \frac{1}{n}\sum_{t=1}^{n}\delta_t(\eta), \\
\sigma_n(\eta) &= \sqrt{\gamma_0(\eta) + 2\sum_{i=1}^{n-1}\kappa(n,i)\ \gamma_i(\eta)}, \\
\gamma_i(\eta) &= \frac{1}{n}\sum_{t=1}^{n-i}\left(\delta_t(\eta) - \mu_n(\eta)\right)\left(\delta_{t+i}(\eta) - \mu_n(\eta)\right), \\
\kappa(n,i) &= \frac{n-i}{n}(1-q_n)^i + \frac{i}{n}(1-q_n)^{n-i},
\end{aligned}$$

where $\mu_n(\eta)$ is the average of the empirical score differences, and $\sigma_n(\eta)$ is an autocorrelation-consistent estimator based on the empirical autocovariances $\gamma_i(\eta)$ and kernel weights $\kappa(n,i)$. Politis and Romano (1994) derive the estimator $\sigma_n(\eta)$ in the context of the stationary bootstrap under the requirement that $q_n \to 0$ as $n \to \infty$, with optimality considerations leading to $q_n = c\ n^{-1/3}$. We choose $c \approx 1.36$ as explained below, and compute the usual t-statistic given by

$$T(\eta) := \frac{\sqrt{n}\ \mu_n(\eta)}{\sigma_n(\eta)}.$$

Politis and Romano (1994) and Hansen (2005) justify the studentized bootstrap test statistic

$$T_0^*(\eta) := \frac{\sqrt{n}(\mu_n^*(\eta) - \mu_n(\eta))}{\sigma_n(\eta)},$$

where $\mu_n^*(\eta) := (1/n)\sum_{t=1}^{n}\delta_t^*(\eta)$ is the average of a bootstrap sample of score differences $\{(\delta_t^*(\eta))_{\eta\in\mathbb{R}} \colon t = 1,\ldots,n\}$. We use the stationary bootstrap of Politis and Romano (1994), where blocks of data are sampled at random. The block length is also random, following a geometric distribution with parameter q_n, leading to an average block length of $1/q_n$. As mentioned above, we let $q_n = c\ n^{-1/3}$, where we set the constant $c \approx 1.36$, so that we obtain an average block length of ten in our empirical study (for which n is around 2500). Choosing $\mu_n(\eta)$ to center the distribution of $\mu_n^*(\eta)$ corresponds to the conservative estimator $\hat{\mu}^u$ in Hansen (2005, p. 372). It is a convenient property of the stationary bootstrap that the formula for the studentizing factor $\sigma_n(\eta)$ does not depend on the specific observations drawn in a bootstrap iteration, and hence must be computed only once. Following Hansen (2005), we use the same variance estimator for the bootstrap iterations and the original sample.

We then compute the supremal test statistics on $\mathbb{R}$ for the original data and all bootstrap iterations $b = 1,\ldots,B$,

$$T^{\max} = \sup_{\eta\in\mathbb{R}} T(\eta),$$

$$T_{0,b}^{*,\max} = \sup_{\eta\in\mathbb{R}} T_{0,b}^*(\eta).$$

Due to the test statistics' structure these suprema can be computed exactly, as explained in Appendix B, but the computational cost increases quickly in sample size. The p-value for the joint null hypothesis that $\mu(\eta) \leq 0 \; \forall \; \eta \in \mathbb{R}$ is given by

$$p_H = \mathbf{P}_B \left(T_0^{*,\max} > T^{\max} \right) = \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}(T_{0,b}^{*,\max} > T^{\max}), \tag{5}$$

where $\mathbf{P}_B$ denotes the probability measure induced by bootstrap sampling.

In Appendix B, we show that it is sufficient to consider the test statistics as piece-wise functions with break points for all $\eta \in \{E_t^A, E_t^B\}_{t=1}^n$, i.e., the set of ES forecasts. The supremal test statistics are computed as the maximum of the left-sided and right-sided limits in all break points, and in all remaining critical points that fall in their respective interval of the piece-wise function partition.

As the sample size and the number of break points grow, computations become increasingly expensive so that we also explore whether computational shortcuts are attainable without compromising the test's size and power properties. To that end, we evaluate the test statistics at a grid of values $G = \{\eta^{[1]}, \eta^{[2]}, \ldots, \eta^{[m]}\} \subset \mathbb{R}$, with $\eta^{[1]} < \eta^{[2]} < \ldots < \eta^{[m]}$. We consider the following choices for G:

1. The grid G_n that consists of the (ordered) elements of the set $\{E_t^A, E_t^B\}_{t=1}^n$. This set is a natural choice in that it coincides with the jump points of the elementary scores S_η, see Proposition 2.1. We refer to this choice as 'jumps' in the following.
2. A thinned version of G_n, considering only every tenth element ('jumps/10').
3. An equally spaced grid ranging from the minimum of G_n to the maximum of G_n, containing as many elements as the thinned version in 2 ('equidistant').

See Sections 3.3 and 4 for theory and simulation evidence on the choice of grid points. We compute the maximal test statistics (original and bootstrap iteration) across G,

$$\tilde{T}^{\max} = \max_{\eta \in G} \; T(\eta),$$
$$\tilde{T}_0^{*,\max} = \max_{\eta \in G} \; T_0^*(\eta),$$

leading to a p-value for the joint null hypothesis that $\mu(\eta) \leq 0 \; \forall \; \eta \in G$,

$$\tilde{p}_H = \mathbf{P}_B \left(\tilde{T}_0^{*,\max} > \tilde{T}^{\max} \right), \tag{6}$$

the empirical probability that the bootstrap test statistic exceeds its sample counterpart.

3.3 Theoretical justifications

Following Cox and Lee (2008, Section 3.1), this section provides a theoretical justification for using the Hansen (2005) test in the present context. We first define the theoretical, supremal test statistic

$$T_0^{\max} = \sup_{\eta \in \mathbb{R}} \frac{\sqrt{n}(\mu_n(\eta) - \mu(\eta))}{\sigma_n(\eta)}.$$

This quantity can not be computed in practice since the expected score difference $\mu(\eta)$ is unknown even under the null hypothesis (the latter imposes no specific value but only that $\mu(\eta) \leq 0$). Nevertheless, $T_0^{\max}$ is useful for the theoretical justification of the testing procedure described in Section 3.2. Let F_* denote the bootstrap distribution function of $T_0^{*,\max}$ and F_*^{-1} its quantile function. The following high-level assumption states that the bootstrap procedure is consistent for the supremum $T_0^{\max}$.

Assumption 3.2. Let $\alpha \in (0,1)$. The bootstrap procedure is such that

$$\mathbb{P}\left(T_0^{\max} \geq F_*^{-1}(\alpha)\right) = 1 - \alpha + a_n$$

for some sequence $a_n \to 0$ as $n \to \infty$.

White (2000, Proposition 2.2 and Corollary 2.7) and Hansen (2005, Corollary 3) establish results similar to our Assumption 3.2 in the context of comparing multiple forecasting methods. In both of these studies, the test statistic of interest is the maximal element of a finite-dimensional vector. By contrast, our procedure is based on functional data, in that our test statistics are suprema over an uncountable set. While Assumption 3.2 seems plausible, we are not aware of a formal justification, but this issue is beyond the scope of the present paper.

Theorem 3.1. *Suppose that Assumption 3.2 holds and that $\mu(\eta) \leq 0$ for all $\eta \in \mathbb{R}$. Then, for any $\alpha \in (0,1)$,*

$$\mathbb{P}(p_H \leq \alpha) \leq \alpha + a_n,$$

where the sequence $(a_n)_{n \in \mathbb{N}}$ is as in Assumption 3.2.

Theorem 3.1 shows that the size of the test is under control for all elements of the null hypothesis (i.e., both on its boundary and in its interior, corresponding to equal performance and strict dominance respectively). This type of control is often hard to achieve; c.f. the comments and references in Ehm and Krüger (2018, Section 7).

Note that Theorem 3.1 refers to the p-value p_H in (5) that is based on analytical calculations for the relevant suprema. For computational reasons, we might not want to compute the p-value p_H but only the grid-based approximation $\tilde{p}_H$ given at (6). The

following result states conditions under which this approximation is justified. For a grid $G = \{\eta^{[1]}, \eta^{[2]}, \ldots, \eta^{[m]}\} \subset \mathbb{R}$, with $\eta^{[1]} < \eta^{[2]} < \ldots < \eta^{[m]}$ as in Section 3.2, we define for $i = 1, \ldots, m+1$

$$I_i = \begin{cases} (\eta^{[i-1]}, \eta^{[i]}] & \text{if } i = 2, \ldots, m, \\ (-\infty, \eta^{[1]}] & \text{if } i = 1, \\ (\eta^{[m]}, \infty) & \text{if } i = m+1. \end{cases}$$

We next assume that there are upper bounds on the variation of the test statistics within each interval:

Assumption 3.3. There exist constants $\tau_G \geq 0$, $\tau_G^* \geq 0$ such that

$$\sup_{\eta,\nu \in I_i} |T(\eta) - T(\nu)| \leq \tau_G, \quad \text{for all } i = 1, \ldots, m+1, \tag{7}$$

$$\sup_{\eta,\nu \in I_i} |T_0^*(\eta) - T_0^*(\nu)| \leq \tau_G^*, \quad \text{for all } i = 1, \ldots, m+1. \tag{8}$$

Proposition 3.2. *Suppose that Assumption 3.3 holds. Then,*

$$\tilde{p}_H(\tau_G) \leq p_H \leq \tilde{p}_H(-\tau_G^*),$$

where $\tilde{p}_H(\epsilon) = \mathbf{P}_B\left(\tilde{T}_0^{*,\max} > \tilde{T}^{\max} + \epsilon\right)$.

The proposition derives lower and upper bounds on the analytical p-value p_H. In particular, it states that the impact of the grid approximation is small if the test statistics display little variation within the intervals I_i. In the Monte Carlo study of Section 4, we follow the recommendation of Cox and Lee (2008, p. 626) and use the 'raw' p-values resulting from the grid approximation. In doing so, we essentially assume that $\tau_G = \tau_G^* = 0$ which seems reasonable when the grid is sufficiently dense, e.g. for large sample sizes and a continuous population distribution for the ES forecasts.[7]

Taking a slightly different perspective, the grid-based approximation can be seen as testing the following, restricted notion of forecast dominance:

Definition 3.1'. Method A *weakly dominates* method B with respect to $\mathcal{S}_2$ on the finite grid $G \subset \mathbb{R}$ if

$$\mathbb{E}\left(\mathrm{S}_\eta(V_t^A, E_t^A, Y_t)\right) \leq \mathbb{E}\left(\mathrm{S}_\eta(V_t^B, E_t^B, Y_t)\right) \quad \text{for all } \eta \in G,$$

where G has been defined before Assumption 3.3, and all other objects are as in Definition 3.1.

[7] In principle, one might try to estimate τ_G and τ_G^* in order to arrive at bounds for p_H. However, this procedure is likely to be computationally demanding, which contradicts the original motivation for using the grid approximation. It hence seems preferable to either set $\tau_G = \tau_G^* = 0$, or to compute p_H via the analytical supremum calculations detailed in Appendix B. We provide simulation results on both approaches in Section 4.

Definition 3.1' is a necessary condition for Definition 3.1. Furthermore, if the grid G is deterministic, then the Hansen (2005) test is valid for Definition 3.1' without further adjustment.[8] Proposition 3.2 quantifies the difference between p_H (the p-value for Definition 3.1) and $\tilde{p}_H$ (the p-value for Definition 3.1'). Its result is in line with the intuition that both p-values are similar if the grid is dense enough. Our simulation results in Table 1 provide direct numerical evidence on the quality of the grid-based approximation.

3.4 Related procedures

The testing procedure described in Sections 3.2 and 3.3 can be viewed as conducting pointwise tests for each η, adjusting the resulting test statistics or p-values for multiple testing and then taking the minimal adjusted p-value as a p-value for the joint hypothesis $\mu(\eta) \leq 0$ for all $\eta \in \mathbb{R}$. More specifically, the p-value adjustment implicit in the method of Hansen (2005) is a simplified ('one-step' or 'single-step') variant of the Westfall and Young (1993) step-down procedure for multiple testing; see Cox and Lee (2008, Section 3.2) and Meinshausen et al. (2011). Cox and Lee (2008) analyze the properties of applying Westfall and Young (1993) to functional data. In Appendix C, we describe the Westfall-Young procedure for our testing problem. The resulting p-value p_{WY} for the null hypothesis $\mu(\eta) \leq 0$ for all $\eta \in \mathbb{R}$ always fulfills $p_{WY} \leq p_H$, implying that the Westfall-Young procedure is more powerful. There are situations where the difference between both procedures is noticeable; see Cox and Lee (2008, Section 3.2). However, in our Monte Carlo study both approaches typically imply the same test decisions at conventional levels (see Appendix C), such that the difference between the two procedures is negligible in practice. We therefore focus on the simpler approach of Hansen (2005).

4 Monte Carlo Evidence on the Dominance Test

While we have provided a partial justification for the dominance test in Section 3.3, two important issues remain. First, does Assumption 3.2 provide a realistic description of the bootstrap in the present situation? Second, in case we substitute the analytical p-value p_H at (5) by a computational shortcut $\tilde{p}_H$ at (6), does the choice of grid points for η matter in practice? In view of these open issues, we next investigate the testing procedure by simulation. The data generating process (DGP) intends to be similar to the HEAVY forecasting

[8] Hansen's test conducts a comparison between a benchmark method and finitely many competitors. In the case of Definition 3.1', the comparison is between two methods at a finite number of fixed grid points. From a technical perspective, both of these comparisons boil down to testing whether all elements of a random vector have nonnegative expectation. Note that Hansen (2005) allows for cross-sectional dependence among the vector elements, as well as for certain forms of time series dependence, both of which are likely present in our setup.

model which we use in the empirical analysis of Section 5. To this end, we first create data from the following process:

$$\sigma_t^2 = 0.5\ \widetilde{\mathrm{RK}}_{t-1} + \beta\ \sigma_{t-1}^2,$$

where the series $\widetilde{\mathrm{RK}}_t$ mimics the 'realized kernel' measure of intra-day volatility (Barndorff-Nielsen et al., 2008, 2009), and the coefficient β determines the persistence of the series σ_t^2. To generate the series $\widetilde{\mathrm{RK}}_{t-1}$, we simulate from a Gaussian AR(1) model that we fit to the logarithmic values of the empirical series RK_t corresponding to the S&P 500 data used in our empirical analysis.[9] We further set $\sigma_0^2 = 0.35$, and assume that the return at day t is given by

$$R_t = \sqrt{\frac{\nu-2}{\nu}}\ \sigma_t\ X_t,$$

where $\{X_t\}_{t\in\mathbb{N}}$ is a sequence of i.i.d. random variables that are t-distributed with ν degrees of freedom. The factor $\sqrt{(\nu-2)/\nu}$ accounts for the fact that the variance of a t-distributed variable equals $\nu/(\nu-2)$. Hence, the factor ensures that the conditional variance of R_t is given by σ_t^2. Given knowledge of the process and the sequence $\{\mathrm{RK}_j\}_{j\le t-1}$, the perfect T_α forecast for R_t consists of

$$\begin{aligned}\mathrm{VaR}_{t|t-1,\alpha} &= \sqrt{\frac{\nu-2}{\nu}}\ \sigma_t\ Q_{\alpha,\nu},\\ \mathrm{ES}_{t|t-1,\alpha} &= \frac{1}{\alpha}\int_0^\alpha \mathrm{VaR}_{t|t-1,z}\ \mathrm{d}z,\end{aligned}$$

where $Q_{\alpha,\nu}$ is the α-quantile of the t-distribution with ν degrees of freedom. We consider two forecasting models $m \in \{1,2\}$ with perturbed ideal forecasts:

$$\begin{aligned}q_{t,\alpha,m} &= \mathrm{VaR}_{t|t-1,\alpha} + \varepsilon_{t,m},\\ e_{t,\alpha,m} &= \mathrm{ES}_{t|t-1,\alpha} + \varepsilon_{t,m},\end{aligned}$$

where $\varepsilon_{t,m} \sim \mathcal{N}(0,\zeta_m)$, independently of t and m.[10] The variance term $\zeta_m \ge 0$ is a measure of expected deviation from optimality. The limiting case $\zeta_m = 0$ means that $\varepsilon_{t,m} = 0$ almost surely, and corresponds to perfect forecasts. Note that model m incurs the same error in both components of its T_α forecast; this ensures that $q_{t,\alpha,m} \ge e_{t,\alpha,m}$ always holds as required. When $\zeta_1 = \zeta_2 > 0$, then both models have equal expected scores, which is consistent with weak dominance. To generate a difference in forecast quality, we consider the case of $\zeta_1 > 0$

[9] The resulting AR(1) model implies that the log of RK_t has a mean of -0.62, a first-order autoregressive coefficient of 0.83, and a residual variance of 0.38.

[10] We evaluate $\mathrm{ES}_{t|t-1,\alpha}$ using the function esT of the R package VaRES (Nadarajah et al., 2013), which employs numerical integration. The values for $\mathrm{ES}_{t|t-1,\alpha}$ thus obtained are very similar to an analytical expression (see Dobrev et al., 2017, Section 4, and the references therein) of which we became aware after completing this work. For example, for $\alpha \in \{0.01, 0.025, 0.05\}$ and a t-distribution with $\nu \in \{4, 6, 10\}$, the absolute difference is smaller than $2\ \times 10^{-9}$.

and $\zeta_2 = 0$, which means that the second model issues perfect forecasts while the first model deviates from optimality to a degree measured by ζ_1. This setup implies a dominance relationship in favor of the second model following results from Tsyplakov (2014),[11] and a violation of the null hypothesis that $m = 1$ weakly dominates $m = 2$. We note that a reversed dominance relation is the strongest type of null hypothesis violation: Reversed dominance implies that $\mu(\eta) \geq 0$ for all $\eta \in \mathbb{R}$, whereas $\mu(\eta) > 0$ for a small range of η would be sufficient to violate the null.

In the following investigation, we consider various values for the two DGP parameters (i.e., the persistence parameter β and the degrees of freedom ν), for the three hyperparameters (i.e., the functional level α, the type of grid for η, the number of observations n), and for the two parameters controlling forecast quality (i.e., the perturbation parameters ζ_1 and ζ_2). For the entire investigation we choose a significance level of five percent. The calculation of a single p-value uses 500 bootstrap iterations, and we draw 1000 p-values per scenario.

Tables 1 and 2 both show Monte Carlo simulation results for size and power. Table 1 adresses the question whether the grid specification for η is important at a sample size of 500. In Section 3.2, we discussed exact computation of the supremal test statistics and three variants of grid approximation ('jumps', 'jumps/10', and 'equidistant'). We can observe no exceedance of the nominal five percent level for the scenarios with equal predictive quality, regardless of whether we use the exact computation or any of the three grid approximations. For the power scenarios, we observe a nice grouping by parameter combination with evidence for a generally minor loss of power when using any of the grid approximation options. As the computational cost increases noticeably beyond a sample size of 500 and power properties are similar, our further simulation results are based on the thinned grid of ES forecasts ('jumps/10').

Table 2 gives a more comprehensive summary of the effects that different parameter values have on size and power. Again, while keeping the size controlled below the 5% level, we can observe that both higher persistence and heavier tails lead to a decrease in power. Similarly, forecasts at a functional level of 0.01 are much harder to evaluate than forecasts at a level of 0.05. In combination, the presence of high persistence while evaluating low-level ES forecasts can make it impossible to reach a power higher than the nominal size even for sample sizes of 2500. However, for moderate values of persistence and ES forecast level, forecast dominance can be rejected reliably.

[11] Both models have access to the same information base, which is used optimally by the second model, but suboptimally by the first model. Tsyplakov (2014) shows that this setup implies dominance of the second model under all proper scoring rules.

ES level	Grid	DGP parameters											
		$\beta = 0.0$			0.5			0.7			0.9		
		$\nu =$ 10	6	4	10	6	4	10	6	4	10	6	4
Size scenarios		$\zeta_1 = \zeta_2 = 1$											
$\alpha = 0.010$	exact	3.9	4.2	4.1	3.4	2.9	3.3	1.4	2.3	2.4	1.1	0.4	0.4
	jumps	2.0	2.4	2.4	2.1	2.4	1.2	1.9	1.2	1.0	0.4	0.5	0.1
	jumps/10	3.1	3.0	3.8	2.4	2.3	2.3	2.1	1.5	2.1	0.6	0.5	0.1
	equidist.	3.4	4.2	4.4	3.0	2.2	2.6	2.6	2.4	1.8	0.6	0.3	0.3
0.025	exact	3.7	4.0	4.4	3.1	3.4	3.3	3.7	3.4	2.7	2.1	1.0	1.3
	jumps	3.2	3.2	3.2	3.0	2.5	2.1	1.9	1.8	1.8	0.8	0.7	0.5
	jumps/10	2.9	2.9	4.1	2.3	2.3	2.8	2.6	2.0	1.5	1.0	0.8	0.9
	equidist.	3.2	3.9	2.9	2.7	3.5	2.1	3.0	3.2	1.9	1.5	1.0	0.6
0.050	exact	4.8	4.9	4.4	4.3	3.2	4.5	3.6	3.4	3.4	1.5	1.8	2.1
	jumps	3.6	3.6	4.7	3.6	3.4	2.7	1.1	2.7	2.6	1.3	1.2	1.1
	jumps/10	3.7	4.3	3.0	2.6	2.9	2.9	2.1	2.5	3.1	0.9	1.2	0.7
	equidist.	3.8	3.6	3.2	2.8	4.4	4.0	2.0	2.9	2.9	2.0	1.7	1.0
Power scenarios		$\zeta_1 = 0.1$	$\zeta_2 = 0$										
$\alpha = 0.010$	exact	64.2	45.8	29.8	8.1	4.5	2.7	1.4	1.2	0.5	0.1	0.0	0.1
	jumps	55.8	38.1	20.4	6.6	2.2	1.2	0.8	0.3	0.4	0.0	0.0	0.0
	jumps/10	55.7	37.4	22.4	4.9	1.6	1.4	1.1	0.1	0.3	0.1	0.1	0.1
	equidist.	57.5	38.3	25.3	5.1	3.1	1.2	1.0	0.4	0.2	0.2	0.3	0.0
0.025	exact	92.6	85.2	81.7	33.2	25.0	21.7	11.8	8.1	4.6	1.2	1.1	1.0
	jumps	85.7	80.9	73.6	27.5	21.1	15.2	7.6	6.3	3.5	0.6	0.7	0.5
	jumps/10	86.3	82.2	78.2	25.6	18.3	13.7	7.8	5.3	3.1	1.0	0.6	0.5
	equidist.	87.1	84.1	78.4	26.3	20.5	15.8	6.4	5.2	4.5	0.7	0.2	0.2
0.050	exact	98.7	98.2	97.6	58.9	57.4	53.1	26.6	21.6	18.6	4.5	3.0	3.0
	jumps	96.7	96.0	96.9	56.0	48.1	46.4	21.8	20.4	16.4	2.2	2.9	2.5
	jumps/10	96.9	95.6	96.4	56.6	49.5	48.8	22.2	17.5	16.3	4.3	2.7	2.9
	equidist.	97.8	96.8	96.9	54.6	50.6	49.7	23.2	20.7	15.6	2.7	3.6	2.6

Table 1: **Simulation results on size and power (grid-focused).** Size and power results (in percentage points) of the Monte Carlo investigation of the dominance test for a 5% significance level, with focus on the effect of the grid type. 'exact' denotes exact analytical computation of the test statistic; the three grid types ('jumps', 'jumps/10' and 'equidistant') are introduced at the end of Section 3.2. The sample size is fixed at 500 observations, p-values are generated using 500 bootstrap iterations, and 1000 p-values are drawn.

ES level	**Obs.**	**DGP parameters**								
		$\beta = 0.5$			0.7			0.9		
		$\nu =$ 10	6	4	10	6	4	10	6	4
Size scenarios		$\zeta_1 = \zeta_2 = 1$								
$\alpha = 0.010$	$n = 500$	2.4	2.3	2.3	2.1	1.5	2.1	0.6	0.5	0.1
	1000	2.7	2.6	2.7	2.7	2.1	2.1	0.6	0.6	0.6
	2500	4.2	3.9	2.4	1.5	2.8	1.6	1.0	1.0	0.6
0.025	500	2.3	2.3	2.8	2.6	2.0	1.5	1.0	0.8	0.9
	1000	2.4	3.3	3.9	2.4	2.8	1.9	1.3	1.1	1.1
	2500	4.0	4.1	2.8	2.8	3.5	3.3	2.3	1.7	0.9
0.050	500	2.6	2.9	2.9	2.1	2.5	3.1	0.9	1.2	0.7
	1000	2.6	2.7	2.9	2.8	2.6	2.0	1.8	2.1	1.1
	2500	2.7	2.5	4.5	2.9	2.5	2.9	3.2	2.8	1.7
Power scenarios		$\zeta_1 = 0.1$	$\zeta_2 = 0$							
$\alpha = 0.010$	$n = 500$	4.9	1.6	1.4	1.1	0.1	0.3	0.1	0.1	0.1
	1000	19.2	9.3	4.2	3.5	1.8	0.3	0.4	0.1	0.3
	2500	74.1	49.0	23.6	20.7	9.1	5.8	1.1	0.8	0.5
0.025	500	25.6	18.3	13.7	7.8	5.3	3.1	1.0	0.6	0.5
	1000	63.8	51.1	40.7	19.3	17.8	10.8	2.3	1.0	1.7
	2500	99.1	95.6	90.1	64.8	51.8	35.1	6.8	3.6	2.7
0.050	500	56.6	49.5	48.8	22.2	17.5	16.3	4.3	2.7	2.9
	1000	89.0	87.0	84.4	49.1	42.2	36.9	6.9	4.1	4.7
	2500	100.0	100.0	100.0	94.3	89.8	84.6	19.0	12.5	12.0
Power scenarios		$\zeta_1 = 0.2$	$\zeta_2 = 0$							
$\alpha = 0.010$	$n = 500$	31.7	18.6	10.2	6.3	2.4	1.9	0.3	0.2	0.1
	1000	75.9	56.3	33.1	18.2	10.9	4.4	0.3	0.3	0.4
	2500	99.8	97.8	86.9	72.7	46.6	20.0	4.7	2.0	1.3
0.025	500	72.7	64.3	56.1	25.2	20.4	12.4	2.1	1.7	0.6
	1000	98.3	95.8	92.3	65.1	52.9	39.3	5.4	3.8	3.3
	2500	100.0	100.0	100.0	98.8	94.3	87.9	24.1	16.2	9.8
0.050	500	92.5	92.3	90.6	56.8	50.5	49.6	7.7	8.3	6.5
	1000	100.0	99.9	99.7	90.0	86.7	86.0	17.8	13.1	14.0
	2500	100.0	100.0	100.0	99.9	100.0	100.0	53.5	44.7	41.1

Table 2: **Simulation results on size and power (focus on sample size).** Size and power results (in percentage points) of the Monte Carlo investigation of the dominance test for a 5% significance level, with focus on the effect of the number of observations. The grid of points η is thinned by a factor of 10 ('jumps/10', see end of Section 3.2), p-values are generated using 500 bootstrap iterations, and 1000 p-values are drawn.

5 Empirical Results for S&P 500 and DAX Returns

In this section, we apply our methodology to compare forecasts for the returns of two stock indices, the S&P 500 and the DAX. The return of the index (S&P 500 or DAX) is defined as

$$R_t = 100 \times (\log P_t - \log P_{t-1}),$$

where P_t is the level of the index at the end of trading day t. As before, let $\mathcal{W}_{t-1}$ denote the information set generated by data up to day $t-1$. We consider three models for daily log returns with corresponding $(\text{VaR}_\alpha, \text{ES}_\alpha)$ forecasts at level $\alpha = 0.025$.

The first specification is a HEAVY model (Shephard and Sheppard, 2010) that uses intra-daily realized measures to model the time-varying variance of financial returns. The model posits that

$$\mathbb{V}(R_t|\mathcal{W}_{t-1}) = \sigma_t^2 = \omega + \gamma \text{ RK}_{t-1} + \beta \sigma_{t-1}^2, \tag{9}$$

where $\mathbb{V}$ denotes variance, and RK_{t-1} is the realized kernel measure computed from intra-daily price movements at day $t-1$. We further assume that, conditional on $\mathcal{W}_{t-1}$, $(\sqrt{(\nu-2)/\nu}\ \sigma_{t-1})^{-1}\ R_t$ follows a t-distribution with ν degrees of freedom. We jointly estimate the model parameters (ω, γ, β and ν) via maximum likelihood. We re-fit the model only on the first trading day of each month using a rolling window of 1500 observations, i.e. roughly six years of daily data. The model yields an estimate of VaR_α and ES_α of R_t, conditional on $\mathcal{W}_{t-1}$. Second, we consider a GARCH(1,1) model as proposed by Bollerslev (1986). The variance specification coincides with Equation (9), except that the squared daily return, R_{t-1}^2, is used in place of RK_{t-1}. As for the HEAVY model, we assume a t-distribution and jointly estimate all parameters via maximum likelihood. As a third, simplistic benchmark we also consider the empirical unconditional VaR_α and ES_α computed from the returns in the 1500 observations up until day $t-1$. This approach resembles 'historical simulation' (HS) methods which are popular in practice (see e.g. McNeil et al., 2015, Section 9.2.3).

Our analysis is based on data from `http://realized.oxford-man.ox.ac.uk/`; this source covers both daily closing prices and realized measures computed from intra-daily data. We construct forecasts for the period from January 2006 to January 2016.[12] The entire analysis is out-of-sample, i.e. we evaluate the forecasts against realizations that were not used for model fitting. Table 3 gives a brief summary of the parameter estimates. For both the S&P500 and DAX series, the HEAVY model features a larger γ parameter (weight on realized measure) than does GARCH. Furthermore, HEAVY is less persistent than GARCH, as reflected in a smaller estimate of β. These findings are qualitatively in line with empir-

[12] More precisely, the S&P 500 sample comprises 2420 observations from January 6, 2006 to January 25, 2016; the DAX sample comprises 2494 observations from January 4, 2006 to January 25, 2016.

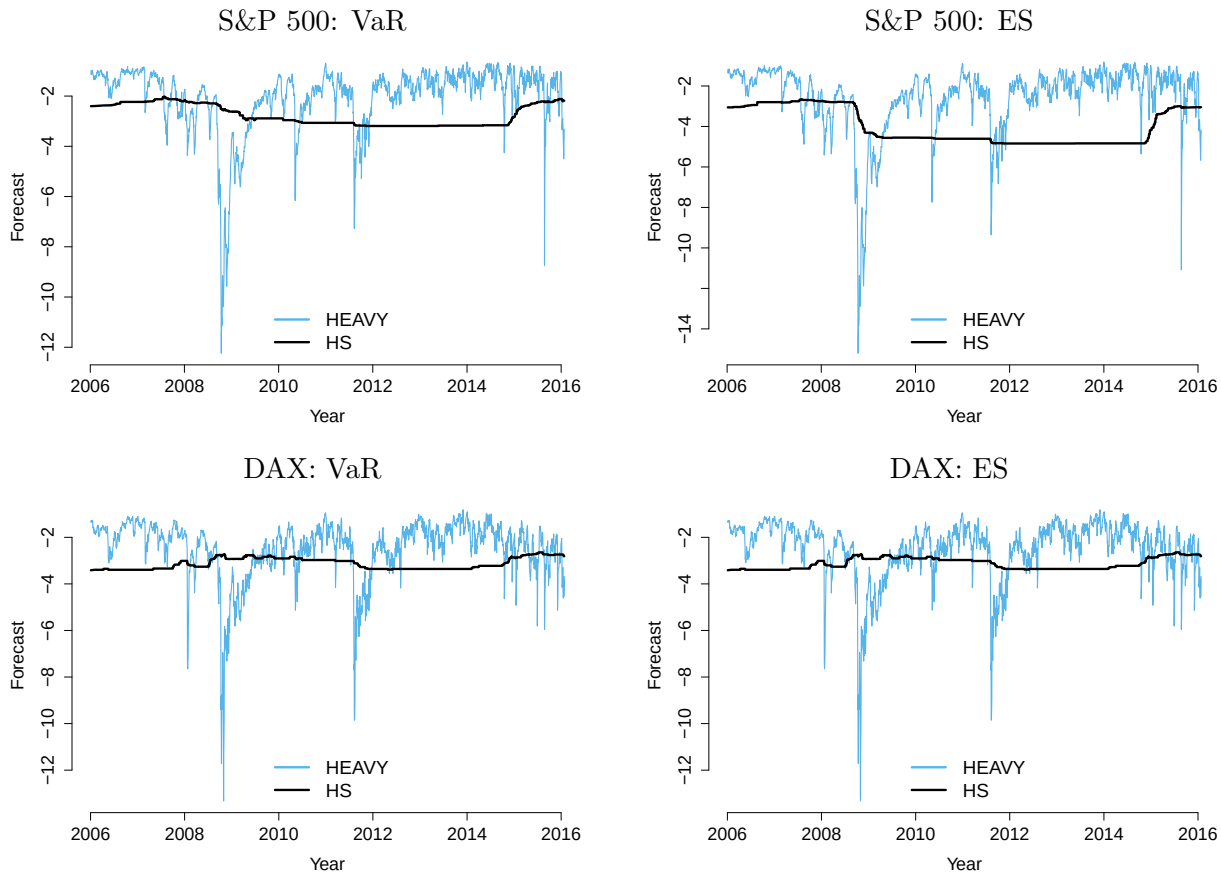

Figure 1: **Time series plots of empirical forecasts** for Value-at-Risk (VaR) and Expected Shortfall (ES). The sample periods ranges from January 2006 to January 2016. See text for details.

ical results by Shephard and Sheppard (2010). Finally, the estimated degrees of freedom are larger (i.e. closer to normality) for HEAVY than for GARCH. This suggests that the realized kernel measure may be more informative as a volatility proxy than squared returns, in the sense that conditioning on realized kernel leads to a lighter-tailed return distribution than does conditioning on squared returns.

Figure 1 presents time series plots of the HEAVY and HS forecasts (the GARCH forecasts are visually similar to the HEAVY ones, and are thus omitted for better display). The figure shows that the HEAVY forecasts display much more time variation than the forecasts of the simple HS method, suggesting that the HEAVY model is much quicker to react to changes in the market environment than the HS method. Table 4 presents some summary statistics on the forecasts. On average, the HS model produces lower forecasts than the other two methods. For the S&P 500 data set, the average VaR_α forecast is -2.065 for HEAVY, compared to -2.213 for GARCH and -2.761 for HS. The violation rates of the VaR_α forecasts are 4.2 percent (HEAVY), 4 percent (GARCH) and 2.9 percent (HS), with all three methods exceeding the nominal level of 2.5 percent, partially due to the negative

	S&P 500			
	ω	γ	β	ν
GARCH	0.009	0.085	0.912	6.984
HEAVY	0.000	0.344	0.742	10.684
	DAX			
	ω	γ	β	ν
GARCH	0.018	0.083	0.910	8.256
HEAVY	0.000	0.606	0.558	13.712

Table 3: **Parameter estimates for HEAVY and GARCH models** as presented in Equation (9). All parameters are re-estimated each month using rolling windows. Numbers in the table are medians across rolling windows.

	Avg. VaR_α	Avg. ES_α	VaR_α 'violation' rate
	S&P 500		
HEAVY	-2.065	-2.594	0.042
GARCH	-2.213	-2.852	0.040
HS	-2.761	-4.028	0.029
	DAX		
HEAVY	-2.499	-3.095	0.038
GARCH	-2.606	-3.322	0.038
HS	-3.130	-4.493	0.025

Table 4: **Summary statistics for empirical forecasts**. Sample period ranges from January 2006 to January 2016 (daily data). The VaR_α 'violation' rate is the fraction of days for which the actual returns falls below the VaR_α forecast (nominal rate: $\alpha = 0.025$).

returns around the 2007-09 financial crisis.

Figures 2 and 3 and Table 5 contain our main forecast evaluation results. Figure 2 presents Murphy diagrams for all three methods with the display for S&P 500 at left and the DAX results at right. For both data sets, the HEAVY model seems to attain the lowest average elementary score for the vast majority of thresholds η. Forecasts based on the GARCH(1,1) model perform slightly worse, and the HS method's performance trails by a considerable margin. This pattern is emphasized in Figure 3, where the HEAVY forecasts are compared directly against GARCH(1,1) and HS, respectively. Examining the difference in elementary scores makes it easier to detect which of two models is better at a certain threshold, especially when the difference is small. Pointwise confidence intervals at the 95 percent level deliver an impression for the significance of the outperformance exhibited by the HEAVY model. For the S&P 500 data, HEAVY seems to perform significantly better than GARCH for a majority of thresholds η; by contrast, the visual comparison for DAX returns does not indicate any clear dominance relation. Table 5 reports the p-value of the formal dominance test presented in Section 3: There is ample support against the null hypothesis that HS dominates HEAVY, but no evidence against dominance of HEAVY over HS. In the comparison of HEAVY and GARCH(1,1), we do not find enough evidence to reject either direction of weak dominance, at least at the five percent level. These results are found for both the S&P 500 and the DAX data.[13] Note that the results in Table 5 are based on a mean block length of ten in the block bootstrap implementation. Using a mean block length of twenty leads to the same test decisions at the five percent level. Furthermore, the results in Table 5 are based on exact calculation of the supremal test statistic; grid-based approximations yield very similar p-values.

The fact that both HEAVY and GARCH dominate HS can perhaps be explained by their use of conditioning information, in contrast to the unconditional distribution estimate implicit in HS. Holzmann and Eulert (2014) show that larger information sets lead to better scores under correct specification. While the latter assumption is unlikely to be satisfied in practice, one might expect similar results to hold under moderate degrees of misspecification.

6 Discussion

In this paper, we provide a mixture representation for the consistent scoring functions for the pair $(\mathrm{VaR}_\alpha, \mathrm{ES}_\alpha)$. This mixture representation facilitates assessments of whether one sequence of predictions for $(\mathrm{VaR}_\alpha, \mathrm{ES}_\alpha)$ dominates another across a suitable, user-specified class of scoring functions. As we are primarily interested in the comparison of the ES

[13] At the ten percent level, the null that GARCH dominates HEAVY is rejected for the S&P 500 data, in line with the visual impression conveyed by Figure 3.

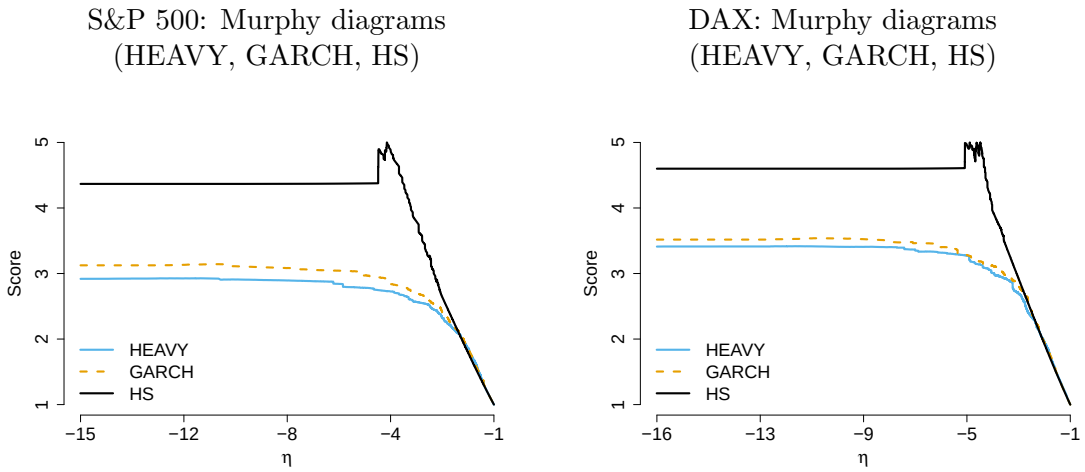

Figure 2: **Murphy diagrams for empirical forecasts**. Smaller scores are better.

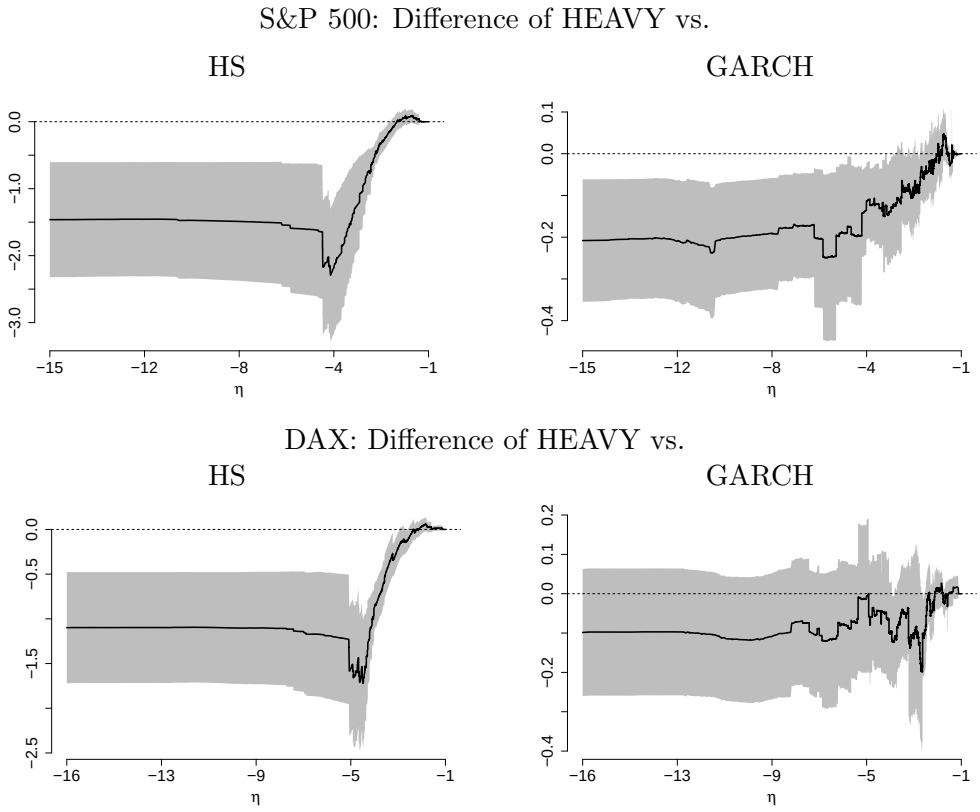

Figure 3: **Score differences for empirical forecasts**. Negative difference means that HEAVY outperforms its competitor. Confidence intervals are pointwise at 95% level.

S&P 500

Hypothesis	p-value
HS weakly dominates HEAVY	0.000
HEAVY weakly dominates HS	0.386
GARCH weakly dominates HEAVY	0.082
HEAVY weakly dominates GARCH	0.554

DAX

Hypothesis	p-value
HS weakly dominates HEAVY	0.000
HEAVY weakly dominates HS	0.488
GARCH weakly dominates HEAVY	0.172
HEAVY weakly dominates GARCH	0.732

Table 5: **Test results for empirical forecasts**. The table presents p-values for several hypotheses related to forecast dominance (see Definition 3.1). The results are based on exact calculation of the supremal test statistic (see Section 3.2), and the bootstrap implementation is based on a mean block length of ten.

forecasts, we focus on a class that puts as much emphasis on ES as possible.

We also propose a formal statistical test for forecast dominance in this context. Theoretical arguments are provided to show that the size of the test is controlled asymptotically, which is supported by the results of a detailed simulation study. This study also investigates the power properties of the test for a broad range of parameter choices in a practically relevant model for the data generating process. For the ES level of 0.025 recommended in the Basel III standard and a degree of volatility persistence that is similar to our empirical estimates, we observe good power properties for reasonably large sample sizes.

When comparing forecast performance in terms of forecast dominance, it is not necessary to select a specific scoring function prior to forecast evaluation. In the presence of possibly misspecified forecasts and non-nested information sets, this is an advantage as any choice of a particular consistent scoring function induces a preference ordering on all possible sequences of forecasts which is usually difficult or impossible to justify, or, even to describe; see Patton (2016). On the other hand, Murphy diagrams may lead to inconclusive situations in which neither of the two forecast methods dominates the other. This may be undesirable in contexts of decision making. Ideally, future work should develop a better understanding of Murphy diagrams, so that they can not only be used to check for forecast dominance but also guide the decision for a consistent scoring function appropriate for a specific application if a total order on forecasting methods is needed.

References

O. E. Barndorff-Nielsen, P. R. Hansen, A. Lunde, and N. Shephard. Designing realized kernels to measure the ex post variation of equity prices in the presence of noise. *Econometrica*, 76:1481–1536, 2008.

O. E. Barndorff-Nielsen, P. R. Hansen, A. Lunde, and N. Shephard. Realized kernels in practice: Trades and quotes. *Econometrics Journal*, 12:C1–C32, 2009.

Basel Committe on Banking Supervision. Minimum capital requirements for market risk. Available from `http://www.bis.org/bcbs/publ/d352.htm`, January 2016.

T. Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31:307–327, 1986.

T. Clark and M. McCracken. Advances in forecast evaluation. In G. Elliott and A. Timmermann, editors, *Handbook of Economic Forecasting*, volume 2, pages 1107–1201. Elsevier, 2013.

D. D. Cox and J. S. Lee. Pointwise testing with functional data using the Westfall-Young randomization method. *Biometrika*, 95:621–634, 2008.

F. Delbaen. *Monetary Utility Functions*. Osaka University Press, 2012.

F. X. Diebold and R. S. Mariano. Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13:253–263, 1995.

T. Dimitriadis and S. Bayer. A joint quantile and Expected Shortfall regression framework. Preprint, `arXiv:1704.02213`, 2017.

D. Dobrev, T. D. Nesmith, and D. H. Oh. Accurate evaluation of expected shortfall for linear portfolios with elliptically distributed risk factors. *Journal of Risk and Financial Management*, 10:5, 2017.

W. Ehm and F. Krüger. Forecast dominance testing via sign randomization. *Electronic Journal of Statistics*, forthcoming, 2018.

W. Ehm, T. Gneiting, A. Jordan, and F. Krüger. Of quantiles and expectiles: Consistent scoring functions, Choquet representations, and forecast rankings. *Journal of the Royal Statistical Society, Series B*, 78:505–562, 2016.

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

T. Fissler, J. F. Ziegel, and T. Gneiting. Expected Shortfall is jointly elicitable with Value at Risk - implications for backtesting. *Risk Magazine*, 2016. January issue.

R. Giacomini and H. White. Tests of conditional predictive ability. *Econometrica*, 74:1545–1578, 2006.

T. Gneiting. Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106:746–762, 2011a.

T. Gneiting. Quantiles as optimal point forecasts. *International Journal of Forecasting*, 27: 197–207, 2011b.

P. R. Hansen. A test for superior predictive ability. *Journal of Business & Economic Statistics*, 23:365–380, 2005.

A. Harvey, E. Ruiz, and N. Shephard. Multivariate stochastic variance models. *Review of Economic Studies*, 61:247–264, 1994.

H. Holzmann and M. Eulert. The role of the information set for forecasting - with applications to risk management. *Annals of Applied Statistics*, 8:595–621, 2014.

H. Lütkepohl. *New Introduction to Multiple Time Series Analysis*. Springer, 2005.

A. J. McNeil, R. Frey, and P. Embrechts. *Quantitative Risk Management: Concepts, Techniques and Tools*. Princeton University Press, 2 edition, 2015.

N. Meinshausen, M. H. Maathuis, and P. Bühlmann. Asymptotic optimality of the Westfall-Young permutation procedure for multiple testing under dependence. *Annals of Statistics*, 39:3369–3391, 2011.

A. H. Murphy. The value of climatological, categorical and probabilistic forecasts in the cost-loss ratio situation. *Monthly Weather Review*, 105:803–816, 1977.

S. Nadarajah, S. Chan, and E. Afuecheta. *VaRES: Computes value at risk and expected shortfall for over 100 parametric distributions*, 2013. URL `https://CRAN.R-project.org/package=VaRES`. R package version 1.0.

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

A. J. Patton. Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics*, 160:246–256, 2011.

A. J. Patton. Comparing possibly misspecified forecasts. Working paper, Duke University, 2016.

A. J. Patton, J. F. Ziegel, and R. Chen. Dynamic semiparametric models for Expected Shortfall (and Value-at-Risk). *Journal of Econometrics*, forthcoming, 2018.

D. N. Politis and J. P. Romano. The stationary bootstrap. *Journal of the American Statistical Association*, 89:1303–1313, 1994.

L. J. Savage. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66:783–801, 1971.

N. Shephard and K. Sheppard. Realising the future: Forecasting with high-frequency-based volatility (HEAVY) models. *Journal of Applied Econometrics*, 25:197–231, 2010.

A. Tsyplakov. Theoretical guidelines for a partially informed forecast examiner. Working Paper, Munich Personal RePec Archive, 2014.

P. Westfall and S. S. Young. *Resampling-Based Multiple Testing: Examples and Methods for P-Value Adjustment*. Wiley, 1993.

H. White. A reality check for data snooping. *Econometrica*, 68:1097–1126, 2000.

T.-J. Yen and Y.-M. Yen. Testing forecast accuracy of expectiles and quantiles with the extremal consistent loss functions. Preprint, `arXiv:1707.02048v3`, 2018.

Appendix

A Proofs

Proposition 2.1

Proof. The $\mathcal{F}_1$-consistency of S_{η_1} and S_{η_2} follows directly from Fissler and Ziegel (2016, Corollary 5.5). This implies the $\mathcal{F}_1$-consistency of S at (3) by a small modification of Gneiting (2011a, Theorem 2). To see that all scoring functions at (2) can be written as at (3), observe that an increasing function G can always be written as

$$G(x) = \int (\mathbb{1}\{v \leq x\} - \mathbb{1}\{v \leq z\}) \, \mathrm{d}H(v),$$

where H is a locally finite measure and $z \in \mathbb{R}$. As $G_2 \geq 0$, we can assume that the measure H_2 puts finite mass on all intervals of the form $(-\infty, x]$ and choose $z = -\infty$. Finally, G_2 is strictly increasing if and only if H_2 puts positive mass on all open intervals. □

Theorem 3.1

Proof. We have $T^{\max} \leq T_0^{\max}$, hence

$$p_H = \mathbf{P}_B\left(T_0^{*,\max} > T^{\max}\right) = 1 - F_*(T^{\max}) \geq 1 - F_*(T_0^{\max}).$$

Therefore, using Assumption 3.2, we obtain

$$\mathbb{P}\left(p_H \leq \alpha\right) \leq \mathbb{P}\left(1 - F_*(T_0^{\max}) \leq \alpha\right) = \mathbb{P}\left(T_0^{\max} \geq F_*^{-1}(1-\alpha)\right) = \alpha + a_n.$$

□

Proposition 3.2

Proof. By Assumption 3.3, we obtain

$$T^{\max} = \sup_{\eta\in\mathbb{R}} T(\eta) \geq \tilde{T}^{\max} = \max_{\eta\in G} T(\eta) \geq \sup_{\eta\in\mathbb{R}} \left(T(\eta) - \tau_G\right) = T^{\max} - \tau_G,$$
$$T_0^{*,\max} = \sup_{\eta\in\mathbb{R}} T_0^*(\eta) \geq \tilde{T}_0^{*,\max} = \max_{\eta\in G} T_0^*(\eta) \geq \sup_{\eta\in\mathbb{R}} \left(T_0^*(\eta) - \tau_G^*\right) = T_0^{*,\max} - \tau_G^*.$$

Therefore,

$$\mathbf{P}_B\left(T_0^{*,\max} > T^{\max}\right) \geq \mathbf{P}_B\left(T_0^{*,\max} > \tilde{T}^{\max} + \tau_G\right) \geq \mathbf{P}_B\left(\tilde{T}_0^{*,\max} > \tilde{T}^{\max} + \tau_G\right),$$
$$\mathbf{P}_B\left(T_0^{*,\max} > T^{\max}\right) \leq \mathbf{P}_B\left(T_0^{*,\max} > \tilde{T}^{\max}\right) \leq \mathbf{P}_B\left(\tilde{T}_0^{*,\max} > \tilde{T}^{\max} - \tau_G^*\right).$$

□

B Test statistic behavior

The elementary score difference $\delta(\eta) = \mathrm{S}_\eta(v^A, e^A, y) - \mathrm{S}_\eta(v^B, e^B, y)$ in $\mathcal{S}_2$ for a single case of forecasts and observation (v^A, e^A, v^B, e^B, y) is given by

$$\delta(\eta) = \begin{cases} 0, & \eta > \max(e^A, e^B), \\ \frac{1}{\alpha}\left(\mathbb{1}\left\{y \leq v^A\right\} - \alpha\right)\left(v^A - y\right) & \\ \quad -\frac{1}{\alpha}\left(\mathbb{1}\left\{y \leq v^B\right\} - \alpha\right)\left(v^B - y\right), & \eta \leq \min(e^A, e^B) \\ \frac{1}{\alpha}\left(\mathbb{1}\left\{y \leq v^A\right\} - \alpha\right)\left(v^A - y\right) - y + \eta, & e^B < \eta \leq e^A, \\ -\frac{1}{\alpha}\left(\mathbb{1}\left\{y \leq v^B\right\} - \alpha\right)\left(v^B - y\right) + y - \eta, & e^A < \eta \leq e^B. \end{cases}$$

As a function of η, the difference $\delta(\eta)$ is a discontinuous, piece-wise linear function with break points at e^A and e^B. Hence, the empirical mean difference $\mu_n(\eta)$ has breaks for all $\eta \in \{e_t^A, e_t^B\}_{t=1}^n$. The average $\mu_n^*(\eta)$ from a bootstrap sample exhibits only a subset of the breaks in $\mu_n(\eta)$.

We look at the structure of the test statistics T and T_0^*. As the numerator of T_0^* is the centered version $\mu_n^*(\eta) - \mu_n(\eta)$, and the denominator is the same for T_0^* and T, there is no difference in the break structure. Furthermore, both numerators are piece-wise linear functions, and the denominator is the square root of a quadratic function. Hence, both types of test statistics can be parameterized in the same way,

$$\begin{aligned} T(\eta) &= \frac{\sqrt{n}\mu_n(\eta)}{\sigma_n \eta} &&= \frac{\sqrt{n}(a+b\eta)}{\sqrt{c+2d\eta+e\eta^2}} \\ T_0^*(\eta) &= \frac{\sqrt{n}(\mu_n^*(\eta)-\mu_n(\eta))}{\sigma_n(\eta)} &&= \frac{\sqrt{n}(a_0^*+b_0^*\eta)}{\sqrt{c+2d\eta+e\eta^2}}, \end{aligned}$$

where $a, b, c, d, e, a_0^*, b_0^*$ are piece-wise constant functions in η, with the argument suppressed for readability. For the sake of completeness, the functions are given below, where $a_0^*(\eta)$ and $b_0^*(\eta)$ are computed analogously to $a(\eta)$ and $b(\eta)$ respectively,

$$\begin{aligned} a(\eta) &= \frac{1}{n}\sum_{t=1}^n a_t(\eta) = \frac{1}{n}\sum_{t=1}^n \left(\frac{1}{\alpha}\left(\mathbb{1}\left\{y_t \le v_t^A\right\} - \alpha\right)\left(v_t^A - y\right) - y\right)\mathbb{1}\left\{\eta \le e_t^A\right\} \\ &\qquad - \left(\frac{1}{\alpha}\left(\mathbb{1}\left\{y_t \le v_t^B\right\} - \alpha\right)\left(v_t^B - y\right) - y\right)\mathbb{1}\left\{\eta \le e_t^B\right\} \\ b(\eta) &= \frac{1}{n}\sum_{t=1}^n b_t(\eta) = \frac{1}{n}\sum_{t=1}^n \mathbb{1}\left\{\eta \le e_t^A\right\} - \mathbb{1}\left\{\eta \le e_t^B\right\} \\ c(\eta) &= \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n \kappa\left(n, |i-j|\right)\left(a_i(\eta) - a(\eta)\right)\left(a_j(\eta) - a(\eta)\right) \\ d(\eta) &= \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n \kappa\left(n, |i-j|\right)\left(a_i(\eta) - a(\eta)\right)\left(b_j(\eta) - b(\eta)\right) \\ e(\eta) &= \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n \kappa\left(n, |i-j|\right)\left(b_i(\eta) - b(\eta)\right)\left(b_j(\eta) - b(\eta)\right) \end{aligned}$$

On any of the intervals induced by the ordered set of break points, we can think of either test statistic as a function

$$f(\eta) = \frac{a+b\eta}{\sqrt{c+2d\eta+e\eta^2}},$$

with constant parameters $a, b, c, d, e \in \mathbb{R}$. If $f(\eta)$ is constant or linear only the endpoints of the interval are of interest. Otherwise, we check whether the only remaining critical point

for an extremum lies in the interval,

$$f'(\eta) = \frac{bc - ad + (bd - ae)\eta}{(c + 2d\eta + e\eta^2)^{3/2}}$$
$$f'(\eta_0) = 0 \Leftrightarrow \eta_0 = \frac{ad - bc}{bd - ae}.$$

The composite null hypothesis is checked on all critical points that fall in their respective interval, and the left-sided and right-sided limits of the break points.

Lastly, we can use these results to calculate $\sup_{\eta \in I} T(\eta)$ and $\sup_{\eta,\nu \in I} |T(\eta) - T(\nu)|$ for any interval I, and for either type of test statistic.

C Relation of Hansen's test to Westfall and Young (1993)

We also considered the method of Westfall and Young (1993) which controls the familywise error rate (i.e., the probability of making at least one false rejection) in multiple testing problems. To describe the procedure, let m be the number of points at which the test statistic is evaluated, and let π be the permutation of $\{1, \ldots, m\}$ such that $T(\eta^{[\pi(1)]}) \leq T(\eta^{[\pi(2)]}) \leq \ldots \leq T(\eta^{[\pi(m)]})$; that is, the permutation π arranges the sample t-statistics in ascending order. Define $U_k^* = \max\{T_0^*(\eta^{[\pi(s)]}) : s \leq k\}$. For example, U_m^* is the largest of all bootstrapped t-statistics, and U_5^* is the largest bootstrap t-statistic across the grid points $\eta^{[\pi(1)]}, \eta^{[\pi(2)]}, \ldots, \eta^{[\pi(5)]}$.

We generate B sets of bootstrap t-statistics, obtaining values $U_{k,b}^*$ for $1 \leq k \leq m$ and $1 \leq b \leq B$. The adjusted p-value for the pointwise test at $\eta = \eta^{[k]}$ is then given by

$$r_k = \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}\left(U_{\pi^{-1}(k),b}^* \geq T(\eta^{[k]})\right);$$

note that $\pi^{-1}(k)$ indicates the rank of $T(\eta^{[k]})$ among $\{T(\eta^{[1]}), \ldots, T(\eta^{[m]})\}$. Finally, the p-value for the joint hypothesis of interest is given by

$$p_{WY} = \min_{1 \leq k \leq m} r_k.$$

There is an interesting connection between Hansen's test and the Westfall-Young method (see Cox and Lee, 2008, Section 3.2): The Westfall-Young p-value is always weakly smaller than the p-value of Hansen's test (in the variant described in Section 3.2 above), such that the Westfall-Young method is potentially more powerful. To see this, let η_{k^*} be the grid point associated with the largest t-statistic, i.e. $\pi^{-1}(k^*) = m$. Then, Hansen's p-value is given by $p_H = r_{k^*}$, and it holds that $r_{k^*} \geq p_{WY}$ by construction. In some cases, the difference can be practically relevant, see Cox and Lee (2008, Section 4, Figure 4).

Observations	**Significance levels**									
	1%	2%	3%	4%	5%	6%	7%	8%	9%	10%
All scenarios										
$n = 500$	5	3	6	6	5	8	7	6	8	10
1000	1	2	1	0	3	4	3	4	3	3
2500	2	1	1	0	2	1	1	2	0	1
Size scenarios	$\zeta_1 = \zeta_2 = 0$									
$n = 500$	1	1	1	2	4	4	6	4	8	9
1000	0	0	0	0	1	1	1	1	1	1
2500	1	0	0	0	0	0	0	1	0	0

Table 6: **Maximum number of disagreements (per 1000).** Maximum number of disagreements (per 1000 p-value replications) between Hansen's test and the Westfall and Young correction among all simulated parameter combinations (scenarios) for various levels of significance. We consider 432 distinct parameter combinations for $n = 500$ and 81 distinct parameter combinations for $n = 1000$ and $n = 2500$, with one third of each category being size scenarios.

However, in our setup, the difference between the two procedures seems to be unimportant. Table 6 summarizes the maximum number of disagreements between the two procedures at significance levels of integer-valued percentage points from 1% to 10%. We observe that across all 594 parameter combinations in our simulation study from Section 4, the largest power difference lies at 1% and the largest size difference lies at a tenth of the respective nominal level. Additionally, it seems that these differences decrease as the number of observations grows. These findings suggests that the results of Meinshausen et al. (2011) showing a certain asymptotic optimality property of Hansen's procedure in some settings, may hold more generally.