University of Heidelberg

Department of Economics

Discussion Paper Series | No. 632

Murphy Diagrams: Forecast Evaluation of Expected Shortfall

Johanna F. Ziegel, Fabian Krüger,
Alexander Jordan, and Fernando Fasciati

Mai 2017

Murphy Diagrams: Forecast Evaluation of Expected Shortfall*

Johanna F. Ziegel
University of Bern

Fabian Krüger
Heidelberg University

Alexander Jordan
Heidelberg Institute for Theoretical Studies

Fernando Fasciati
Raiffeisen Schweiz

May 12, 2017

Abstract

Motivated by the Basel 3 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 (Murphy diagrams) 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 3 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 is challenging to select a specific member function from the FZ family on either economic or statistical grounds. 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

*We thank seminar and conference participants in Heidelberg and Augsburg (Statistische Woche 2016) for helpful comments. Johanna 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.

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

another under a relevant class of scoring functions.[2] While this class could be the entire FZ family, we argue that a subfamily which emphasizes ES – rather than VaR – is economically more plausible in the light of the Basel 3 standard. Analyzing the robustness 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 designed to detect significant deviations of the observed performance 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. on a grid of parameters. Following a suggestion by Ehm et al. (2016, Section 3.4), we discuss a permutation test which accounts for this circumstance via multiple testing corrections. The labels of two forecasting methods are randomly switched to enforce the null hypothesis of equal forecast performance, allowing the computation of p-values by Monte Carlo simulation. While a formal investigation of this test is left for future research, simulation evidence points to satisfactory size and power properties.

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 the HEAVY model tends to outperform its competitors, as indicated by Murphy diagrams and tests of forecast dominance.

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 2 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 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 permutation test for the hypothesis of forecast dominance in Section 3. We identify a class of scoring functions primarily suited for the evaluation of the expected shortfall component in a forecast of type (VaR, ES), illustrate its use in an empirical case study in Section 4, and draw a link to European put options in Section 5. A discussion in Section 6 concludes.

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

[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).

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

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

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

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 we consider elements of the action domain $\mathsf{A}_0 = \{x \in \mathbb{R}^2 : x_1 \geq x_2\}$, 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 $\int \mathrm{S}(x,y)\,\mathrm{d}F(y)$ exists for all $F \in \mathcal{F}_1$, $x \in \mathsf{A}_0$. The scoring function S is called *consistent* for T_α if

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

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

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 $x \in \mathsf{A}_0$. 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] As shown by Fissler and Ziegel (2016, Section 5), consistent scoring functions for T_α take the form $\mathrm{S}(x_1, x_2, y)$, where x_1 is a forecast of VaR_α, x_2 is a forecast of ES_α, and y is the realization. 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

$$\begin{aligned}\mathrm{S}(x_1,x_2,y) = {} & \big(\mathbb{1}\{y \leq x_1\} - \alpha\big)\big(G_1(x_1) - G_1(y)\big) \\ & + G_2(x_2)\Big(\frac{1}{\alpha}\mathbb{1}\{y \leq x_1\}(x_1 - y) - (x_1 - x_2)\Big) \\ & - \big(\mathcal{G}_2(x_2) - \mathcal{G}_2(y)\big),\end{aligned} \tag{2}$$

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

[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.

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 (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. Here we seek to develop methods for comparing forecasts under a class of scoring functions, thus avoiding the need to select a single specific function. 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 (X_1^A, X_2^A) and (X_1^B, X_2^B) made by methods A and B, respectively, we say that method A *weakly dominates* method *B with respect to* $\mathcal{S}$ if

$$\mathbb{E}\left(\mathrm{S}(X_1^A, X_2^A, Y)\right) \le \mathbb{E}\left(\mathrm{S}(X_1^B, X_2^B, Y)\right), \quad \text{for all } \mathrm{S} \in \mathcal{S},$$

where the expectations are with respect to the joint distribution of $(X_1^A, X_2^A, X_1^B, X_2^B, Y)$.

When $\mathcal{S}$ is a 'small' class it feasible to check an empirical version of dominance for all members. And importantly, once dominance has been established for a given class it can be translated to the extension including all mixtures, e.g. dominance with respect to $\{S_1, S_2\}$ implies dominance with respect to $\{aS_1 + bS_2 : a, b \ge 0\}$. 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). To this end, 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 T_α, 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 $v_1, v_2, y \in \mathbb{R}$, $(x_1, x_2) \in \mathsf{A}_0$, we define*

$$\mathrm{S}_{v_1}(x_1, y) = (\mathbb{1}\{y \le x_1\} - \alpha)\big(\mathbb{1}\{v_1 \le x_1\} - \mathbb{1}\{v_1 \le y\}\big)$$
$$\mathrm{S}_{v_2}(x_1, x_2, y) = \mathbb{1}\{v_2 \le x_2\}\left(\frac{1}{\alpha}\mathbb{1}\{y \le x_1\}(x_1 - y) - (x_1 - v_2)\right) + \mathbb{1}\{v_2 \le y\}(y - v_2).$$

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

$$\mathrm{S}(x_1, x_2, y) = \int \mathrm{S}_{v_1}(x_1, y)\,\mathrm{d}H_1(v_1) + \int \mathrm{S}_{v_2}(x_1, x_2, y)\,\mathrm{d}H_2(v_2). \tag{3}$$

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

The *elementary scores* S_{v_1} and S_{v_2} are themselves consistent scoring functions for T_α, which follows immediately by choosing Dirac-measures for H_1 or H_2 in (3). Note that S_{v_1} for T_α is also the elementary score in the class of consistent scoring functions for α-quantiles as identified by Ehm et al. (2016), an unsurprising result given that VaR_α is an α-quantile. The elementary score $\mathrm{S}_{v_1}(x_1, y)$ goes to zero as $v_1 \to \pm\infty$. The second elementary score for T_α, $\mathrm{S}_{v_2}(x_1, x_2, y)$, takes a more complex form in that it depends on the joint forecast $(x_1, x_2)'$ and the realization y. It goes to zero as $v_2 \to +\infty$, and converges to $(1/\alpha)(\mathbb{1}\{y \le x_1\} - \alpha)(x_1 - y)$ as $v_2 \to -\infty$. This explains the different restrictions on the corresponding mixing measures H_1 and H_2 in Proposition 2.1.

We now 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 a forecast of type T_α. 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.

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 3. Focusing on $\mathcal{S}_2$ also seems justified from a statistical perspective: First, $\mathcal{S}_2$ contains positively homogeneous scoring functions for T_α 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. Second, 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.

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

$$v_2 \mapsto \mathbb{E}(\mathrm{S}_{v_2}(X_1, X_2, Y)),$$

where the expectation is with respect to the joint distribution of (X_1, X_2, 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 4. 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

$$\mathbb{E}\left(\mathrm{S}_{v_2}(X_1^A, X_2^A, Y)\right) \leq \mathbb{E}\left(\mathrm{S}_{v_2}(X_1^B, X_2^B, Y)\right) \quad \text{for all } v_2 \in \mathbb{R};$$

compare Ehm et al. (2016, Corollary 1).

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, v_1 and v_2. Instead, when focusing on $\mathcal{S}_2$, it suffices to check inequalities along a single grid for v_2. We give results for all consistent scoring functions as a robustness check in Appendix D.

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 $X_t = (X_{t,1}, X_{t,2})' \in \mathsf{A}_0$ denote the $(\mathrm{VaR}_\alpha, \mathrm{ES}_\alpha)$ forecast of Y_t, with the understanding that X_t is 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 $Z_t = (X_t, Y_t)' \in \mathsf{A}_0 \times \mathbb{R}$ is stationary and ergodic, with stationary distribution F_Z.

This assumption rules out deterministic time trends, structural breaks and seasonalities, among others. At the same time, the forecasts and realizations are allowed to fluctuate over time, as long as the fluctuations ‘wash out’ eventually. In particular, 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 $\mathrm{S}(X_t, Y_t)$ does not depend on time, t. In particular, this holds when S equals an elementary score S_{v_2} from Proposition 2.1. We can thus define the notion of an expected elementary score, as follows. Consider a sequence of forecasts $\{X_t\}_{t\in\mathbb{N}}$ 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

$$\mathbb{E}\left(\mathrm{S}_{v_2}(X_{t,1}, X_{t,2}, Y_t)\right) = \int_{\mathsf{A}_0\times\mathbb{R}} \mathrm{S}_{v_2}(x_1, x_2, y)\ \mathrm{d}F_Z(x_1, x_2, y), \tag{4}$$

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 $\{X_t^A\}_{t\in\mathbb{N}}$ and $\{X_t^B\}_{t\in\mathbb{N}}$ denote two competing sequences of forecasts of T_α, and let $\{Y_t\}_{t\in\mathbb{N}}$ denote the corresponding realizations such that $\{(X_t^A, Y_t)\}_{t\in\mathbb{N}}$ and $\{(X_t^B, Y_t)\}_{t\in\mathbb{N}}$ 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

$$\mathbb{E}\left(\mathrm{S}_{v_2}(X_{t,1}^A, X_{t,2}^A, Y_t)\right) \leq \mathbb{E}\left(\mathrm{S}_{v_2}(X_{t,1}^B, X_{t,2}^B, Y_t)\right) \quad \text{for all } v_2 \in \mathbb{R},$$

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

Under standard regularity conditions, the expectations in Definition 3.1 can be consistently estimated by empirical averages over observed forecasts and realizations at dates $t = 1, \ldots, T$, e.g. as $T \to \infty$ it holds that

$$\frac{1}{T}\sum_{t=1}^{T} \mathrm{S}_{v_2}(X_{t,1}^A, X_{t,2}^A, Y_t) \overset{a.s.}{\to} \mathbb{E}\left(\mathrm{S}_{v_2}(X_{t,1}^A, X_{t,2}^A, Y_t)\right),$$

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. The test procedure, which we detail in Appendix B, can be summarized as follows:

- **Stage 1:** Test the null hypothesis that methods A and B perform equally well under a given elementary score, against the one-sided alternative that B performs strictly better, i.e. for a given v_2 the pointwise null and alternative hypotheses are

$$H0_{v_2} : \mathbb{E}\left(\mathrm{S}_{v_2}(X_{t,1}^A, X_{t,2}^A, Y_t)\right) = \mathbb{E}\left(\mathrm{S}_{v_2}(X_{t,1}^B, X_{t,2}^B, Y_t)\right),$$
$$H1_{v_2} : \mathbb{E}\left(\mathrm{S}_{v_2}(X_{t,1}^A, X_{t,2}^A, Y_t)\right) > \mathbb{E}\left(\mathrm{S}_{v_2}(X_{t,1}^B, X_{t,2}^B, Y_t)\right).$$

  This test is repeated for threshold values v_2 on a predefined grid, yielding a sequence of pointwise p-values.

- **Stage 2:** Compute corrected p-values which are designed to control the family-wise error rate (FWER) of the procedure. The FWER is defined as the probability of making at least one false rejection, i.e. rejecting $H0_{v_2}$ for at least one grid point v_2 at which A does not perform worse than B. Reject H0 if the minimum of the corrected p-values is below the chosen significance level.

The tests in the first stage are one-sided t-tests of the null hypothesis that the expected score difference between models A and B is zero. To implement the correction in the second stage, we apply the Westfall and Young (1993) algorithm to the pointwise p-values; see also Cox and Lee (2008) who investigate the properties of the algorithm in the context of functional data, i.e., the null hypotheses refer to a grid of values of some parameter as in our case. The most important implementation choice is how to simulate p-values under the null hypothesis that model A weakly dominates model B, as formulated in Definition 3.1. Our approach draws an i.i.d. sample where each realization of joint pointwise p-values is simulated by reassigning the forecasts' labels with equal probability at each time step. This enforces equality of the expected elementary scores of models A and B for all v_2, thus representing the boundary of the null hypothesis of weak dominance. We call the minimum of the corrected p-values the *minimal Westfall-Young* p-value.

Our results rest on the assumption that the score differences are independent over time (henceforth, IOT). Given that we consider one-day ahead forecasts, this assumption follows standard practice in the econometric forecasting literature, which is to consider autocorrelation up to lag $\tau - 1$, where τ is the forecast horizon; see e.g. Clark and McCracken (2013, Section 7). Consistent with the IOT assumption, the tests in the first stage do not account for possible autocorrelation, and the label switches in the second stage are performed independently over time.

It is possible to implement the test differently taking into account correlation of score differences over time. We consider this alternative implementation in Appendix D where we apply the resulting tests to the data example of Section 4. There, we also consider the use of both types of elementary scores for T_α, an extension where the pointwise null and alternative hypotheses are defined on two independent grids of values for v_1 and v_2, respectively.

We acknowledge that there are two open issues about the testing procedure just described. First, the relabeling step enforces *exchangeability* of the two models' scores. While exchangeability implies equality of expected scores, the converse is not true. It is unclear whether this imbalance vanishes or remains asymptotically. Second, our testing procedure controls its size α at the boundary of the null hypothesis, for equality of the expected scores of models A and B. Intuitively, and as conjectured by Ehm et al. (2016, p. 522), one would expect that the test's rejection probability is smaller than α in the interior of the null hypothesis, when A strictly dominates B at some grid points. However, a formal proof of this intuition is beyond the scope of this paper.

In view of these open issues, we investigate the testing procedure by simulation.[5] The data generating process is similar to the HEAVY forecasting model which we use in the empirical analysis of Section 4. To this end, we first create data from the deterministic process

$$\sigma_t^2 = 0.5 \ \mathrm{RK}_{t-1} + 0.7 \ \sigma_{t-1}^2,$$

where RK_t is a 'realized kernel' measure of intra-day volatility (Barndorff-Nielsen et al., 2008, 2009), i.e. between end-of-day time $t-1$ and t, and $\sigma_0^2 = 0.35$. We use the RK values as recorded in the S&P 500 data set in Section 4, 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 which are t-distributed with $\nu = 6$ degrees of freedom.[6] Given knowledge of the process and the sequence $\{\mathrm{RK}_j\}_{j \leq t-1}$, the perfect T_α forecast for R_t consists of

$$\mathrm{VaR}^*_{t|t-1,\alpha} = \sqrt{\frac{\nu - 2}{\nu}} \ \sigma_t \ q_{\alpha,\nu},$$
$$\mathrm{ES}^*_{t|t-1,\alpha} = \frac{1}{\alpha} \int_0^\alpha \mathrm{VaR}^*_{t|t-1,z} \ \mathrm{d}z,$$

[5] We use the R programming language (R Core Team, 2017) for all simulations and empirical analyses in this paper.

[6] 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.

where $q_{\alpha,\nu}$ is the α-quantile of the t-distribution with ν degrees of freedom. We fix $\alpha = 0.025$ and consider two forecasting models $m \in \{1, 2\}$, with forecasts given as follows:

$$\mathrm{VaR}_{t|t-1,m} = \mathrm{VaR}^*_{t|t-1} + \varepsilon_{t,m},$$
$$\mathrm{ES}_{t|t-1,m} = \mathrm{ES}^*_{t|t-1} + \varepsilon_{t,m},$$

where $\varepsilon_{t,m} \sim \mathcal{N}(0, \zeta_m)$, independently of t and m. The variance term $\zeta_m \geq 0$ 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. To investigate the size and power of the proposed testing procedure, we experiment with different choices of ζ_1 and ζ_2.[7]

First consider the case $\zeta_1 = \zeta_2 = 1$. This means that both models have equal expected scores, which is consistent with weak dominance. To investigate the size of our procedure, we simulate 1000 data sets comprising forecasts and realizations for $T = 500$ time periods, yielding a sample of 1000 minimal Westfall-Young p-values. The left panel of Figure 1 illustrates the results for nominal levels ranging from 0 to 15 percent. We observe a conservative behavior of the test with empirical size slightly below the nominal level. These results seem satisfactory, and suggest that our approach of controlling the test's size at the boundary of the null hypothesis is sufficient in the present context.

We further investigate the power behavior at the 5 percent level. Consider the case of $\zeta_1 > 0$ 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 implies a dominance relationship in favor of the second model following results from Tsyplakov (2014).[8] In the simulation, we vary ζ_1 from 0.05 to 0.5, consider $T = 200, 500$, and generate a sample of 500 minimal Westfall-Young p-values for each combination of ζ_1 and T. The right panel of Figure 1 suggests a monotonic power increase in both variables with convergence to 1. These qualitative features are in line with common sense, and thus provide a sanity check for the testing procedure.

4 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 $(\mathrm{VaR}_\alpha, \mathrm{ES}_\alpha)$ forecasts at level $\alpha = 0.025$:

- The HEAVY model (Shephard and Sheppard, 2010) which uses intra-daily realized measures to model the time-varying variance of financial returns. The model posits that
$$\mathbb{V}(R_t|\mathcal{W}_{t-1}) = \sigma_t^2 = \omega + \gamma\ \mathrm{RK}_{t-1} + \beta\ \sigma_{t-1}^2, \tag{5}$$
where $\mathbb{V}$ denotes variance, and RK_{t-1} is the realized kernel measure computed from intra-daily price movements at day $t-1$. The quantities $\omega > 0, \gamma$ and β are model parameters which we estimate via the quasi-likelihood method described in Shephard and Sheppard (2010, Section 2.4.1). 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

[7] For all investigations, the testing procedure uses 50 equally spaced grid points for v_2, whose range is determined from the empirical range of all forecasts and realizations. Furthermore, we use 500 iterations of the relabeling procedure in Stage 2 of each test.

[8] 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.

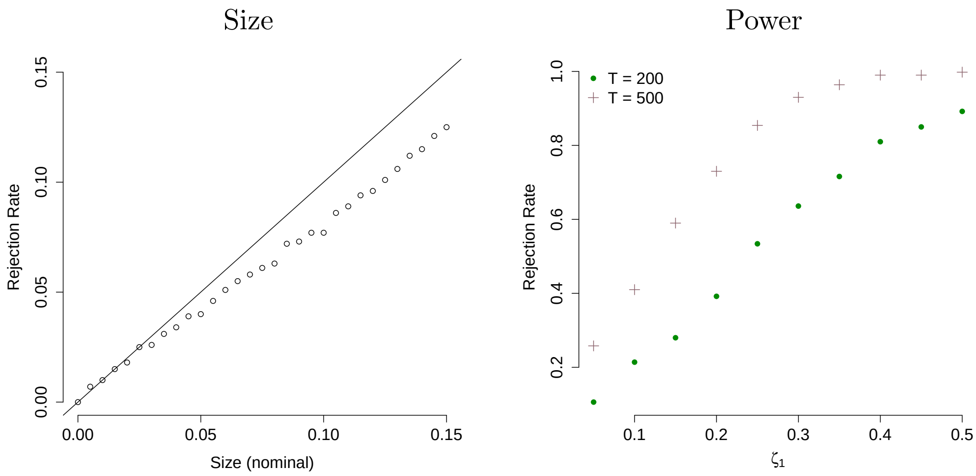

Figure 1: **Monte Carlo analysis of test for forecast dominance**. Left - Westfall-Young rejection rates plotted against nominal test size; data simulated under H0 ($\zeta_1 = \zeta_2 = 1$). Each simulated data set comprises $T = 500$ time periods. The results are based on 1000 Monte Carlo iterations. Right - Rejection rates of Westfall-Young tests at level 5 percent; data simulated under the alternative ($\zeta_1 > 0, \zeta_2 = 0$) for two sample sizes ($T = 200$ and $T = 500$). The results are based on 500 Monte Carlo iterations for each combination of ζ_1 and T.

		Avg. VaR_α	Avg. ES_α	VaR_α 'violation' rate
S&P 500	HEAVY	-2.056	-2.736	0.042
	GARCH(1,1)	-2.184	-2.906	0.040
	HS	-2.761	-4.028	0.029
DAX	HEAVY	-2.500	-3.326	0.037
	GARCH(1,1)	-2.607	-3.469	0.036
	HS	-3.130	-4.493	0.025

Table 1: **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 and should not exceed $\alpha = 0.025$.

data. 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 six degrees of freedom. This set of assumptions yields an estimate of VaR_α and ES_α of R_t, conditional on $\mathcal{W}_{t-1}$.

- A GARCH(1,1) model as proposed by Bollerslev (1986). The variance specification coincides with Equation (5), 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 with six degrees of freedom for the scaled conditional return distribution.
- 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.[9] The entire analysis is out-of-sample, i.e. we evaluate the forecasts against realizations which were not used for model fitting.

[9] 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.

Table 1 presents summary statistics on the forecasts; Figure 3 in Appendix C presents corresponding time series plots. On average, the HS model produces lower forecasts than those by the other two methods. For the S&P 500 data set, the average VaR_α forecast is -2.056 for HEAVY, compared to -2.184 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 returns surrounding the 2007-09 financial crisis. Figure 3 in Appendix C shows that the HEAVY and GARCH forecasts are highly correlated, and display much more time variation than the forecasts of the simple HS method. The latter observation shows that the HEAVY and GARCH models are much quicker to react to changes in the market environment than the HS method.

Figure 2 and Table 2 contain the main forecast evaluation results for the S&P 500 and DAX data sets. We perform forecast evaluation in three steps:

- The top row of Figure 2 presents Murphy diagrams for all three methods with the display for the S&P 500 data set 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 v_2. Forecasts based on the GARCH(1,1) model perform slightly worse, and the HS method's performance trails by a considerable margin.
- This is emphasized in the bottom row of Figure 2, where the method based on the HEAVY model is compared directly against GARCH(1,1) and HS, respectively. Examining the difference in elementary scores improves our ability 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 (Stage 1 in the dominance test) deliver an impression for the significance of the outperformance exhibited by the HEAVY model. It seems that most examiners would question a significant result only for the comparison of HEAVY to GARCH(1,1) in the DAX data example. However, the final significance decision depends mostly on the way in which the pointwise results are combined.
- Table 2 reports the minimal Westfall-Young p-value of the dominance test: There is ample support against the null hypothesis that HS dominates HEAVY, but no evidence against dominance of HEAVY over HS. These results are found for both the S&P 500 and the DAX data. In the comparison of HEAVY and GARCH(1,1) for S&P 500, we similarly find evidence against HEAVY dominating GARCH, but not vice versa. As the previous visual inspection of Figure 2 suggests, the HEAVY/GARCH comparison yields different results for the DAX data: At conventional significance levels, we do not find enough evidence to reject either direction of weak dominance.

The fact that the HEAVY model tends to outperform its competitors can perhaps be explained by its larger information set, incorporating intra-daily data in addition to daily returns. From Holzmann and Eulert (2014), we know 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.

In Appendix D, we analyze the robustness of our permutation test along two dimensions. First, we compare two different assumptions on the temporal dependence of the elementary scores. Second, we consider using both types of elementary scores S_{v_1} and S_{v_2} for the test. The results are generally similar to the ones reported here, with one exception: When considering both elementary scores without accounting for autocorrelation, the minimal Westfall-Young p-values tend to be small for all considered null hypotheses. We conjecture that these results are largely due to non-standard temporal dependence in the first extremal score. However, once one accounts for serial correlation, the results based on both elementary scores are similar to the ones based on S_{v_2} only.

5 Relationship to Option Pricing

In Section 2, we have provided a statistical justification for the class $\mathcal{S}_2$ of scoring functions. We next show that the elementary scores of $\mathcal{S}_2$ also bear an economic interpretation,

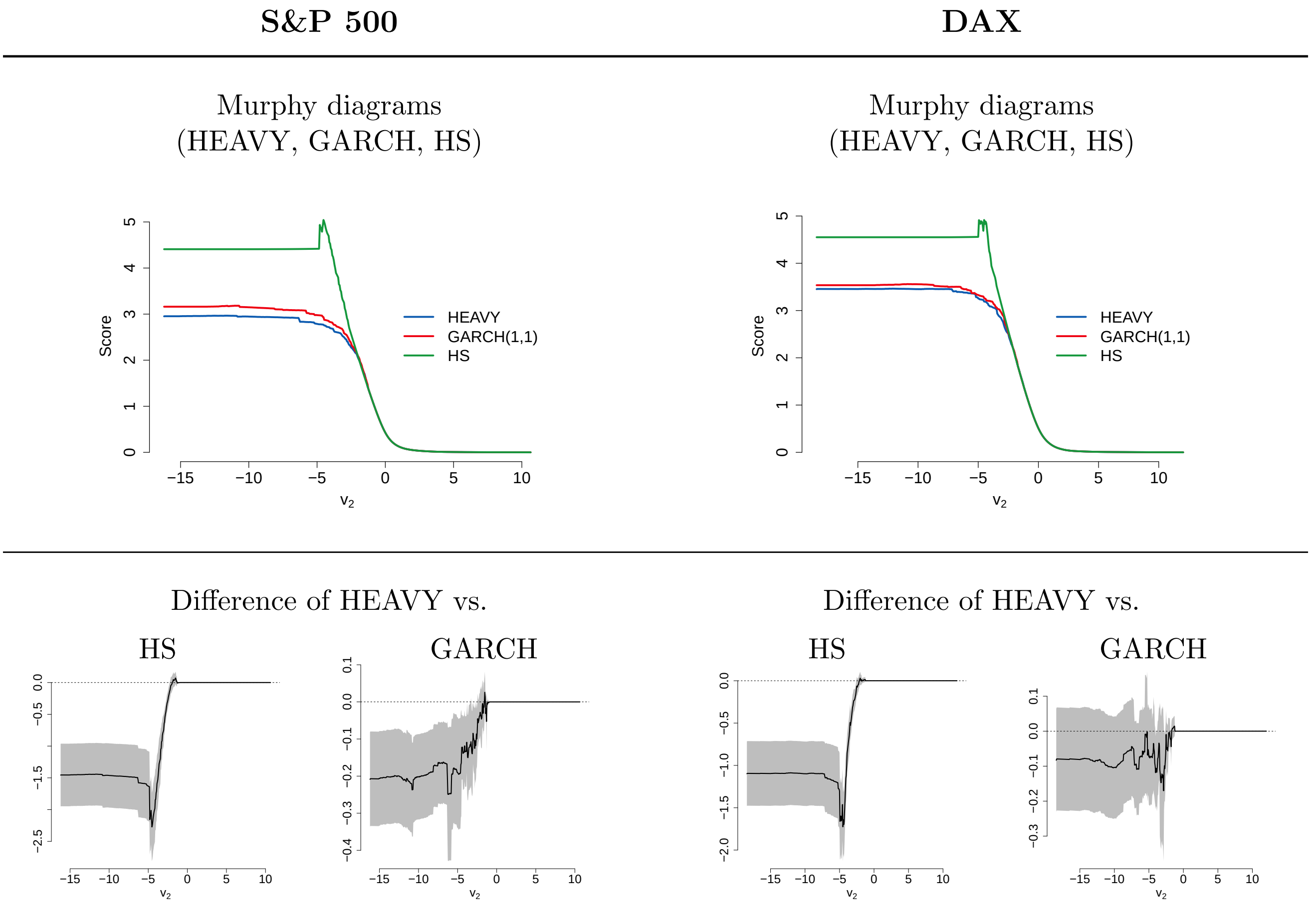

Figure 2: **Murphy diagrams for empirical forecasts**. Top panels: Smaller scores are better. Bottom panels: Negative difference means that HEAVY outperforms its competitor. Confidence intervals are pointwise at 95% level.

S&P 500		**DAX**	
Hypothesis	P-value	Hypothesis	P-value
HS weakly dominates HEAVY	0.000	HS weakly dominates HEAVY	0.000
HEAVY weakly dominates HS	0.772	HEAVY weakly dominates HS	0.876
GARCH weakly dominates HEAVY	0.000	GARCH weakly dominates HEAVY	0.174
HEAVY weakly dominates GARCH	0.998	HEAVY weakly dominates GARCH	0.924

Table 2: **Test results for empirical forecasts**. The table presents p-values for several hypotheses related to forecast dominance (see Definition 3.1). Results are based on the IOT assumption and the class $\mathcal{S}_2$. See Appendix D for additional results.

which resembles connections between VaR, ES and option prices drawn by Mitra (2015) and Barone Adesi (2016). Specifically, our elementary score S_{v_2} is equivalent in decision-theoretic terms to a short position in a European put option with its profit described by

$$\pi = P - \mathbb{1}\{S \leq K\}(K - S),$$

where P is the put option's price, K is the strike price, and S is the spot price. We draw the elementary scores' relation to π by identifying the spot price S with y, the strike price K with x_1, and by imposing $\alpha(x_1 - v_2)$ as the premium P's structure, such that

$$\begin{aligned} \pi &= \alpha(x_1 - v_2) - \mathbb{1}\{y \leq x_1\}(x_1 - y) \\ &= \alpha\mathbb{1}\{v_2 \leq y\}(y - v_2) - \alpha\mathrm{S}_{v_2}(x_1, x_2, y), \end{aligned} \tag{6}$$

conditional on a positive writing decision $x_2 \geq v_2$. Actions are limited to the choice of x_1 and x_2, corresponding to the strike price and the writing decision, respectively. The first term, $\alpha\mathbb{1}\{v_2 \leq y\}(y - v_2)$, describes the best case scenario without playing a role in the decision-making problem, while the second term can be interpreted as the regret, solely determining the best course of action.

Let F denote the distribution of the spot price at maturity of a given asset. From Proposition 2.1 and Equation (6), the expected profit is then

$$\mathbb{E}(\pi) = \mathbb{1}\{v_2 \leq x_2\}\left(x_1(\alpha - F(x_1)) + \int_{-\infty}^{x_1} y \,\mathrm{d}F(y) - \alpha v_2\right).$$

The expression in round brackets as a function of x_1 is concave with a maximum at $x_1 = \mathrm{VaR}_\alpha(F)$ taking the value $\alpha(\mathrm{ES}_\alpha(F) - v_2)$. Therefore, choosing $x_1 = \mathrm{VaR}_\alpha(F)$ is the optimal choice for x_1 given a positive writing decision $x_2 \geq v_2$. For x_2, any choice such that $(x_2 - v_2)(\mathrm{ES}_\alpha(F) - v_2) \geq 0$ is optimal.

Assuming that all market participants take only optimal actions, no options with non-zero expected profit will be traded. This implies that the option will only be traded if $v_2 = \mathrm{ES}_\alpha(F)$. This implies for the price of the option that

$$P = \alpha(\mathrm{VaR}_\alpha(F) - \mathrm{ES}_\alpha(F)). \tag{7}$$

Interestingly, Equation (7) coincides with the famous Black and Scholes (1973) pricing formula in one particular case. Specifically, assume that the asset price follows a geometric Brownian motion without trend, such that F is a log-normal distribution with parameters $\mu = \log(y_0) - 0.5\ \tau^2\ t$ and $\sigma = \tau\sqrt{t}$, where y_0 denotes the spot price at present, τ is the annual volatility, and t is the time to maturity, where $t = 1$ corresponds to one year. Under this form of F, Equation (7) recovers the Black-Scholes formula for the price of a European put option under the additional assumption that the risk-free interest rate is zero (see Hull, 2008, Chapter 13). The calculations that establish the equivalence are presented in Appendix E. Of course, Equation (7) may yield different prices than Black-Scholes under other forms for F. While we do not take a stance on which form for F – or, more generally, which option pricing scenario – is most appropriate, the similarity between statistical incentives (represented by elementary scores) and economic incentives (represented by option payoffs) seems intriguing.

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 forecasts, we focus on a class that puts as much emphasis on ES as possible. We also demonstrate a general principle for the construction of formal statistical tests for forecast dominance. While the test appears to work well in the simulation and data example, a detailed investigation of its theoretical properties is left for future work.

When using Murphy diagrams for comparing forecast performance, 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 deeper 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 in case that 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. *The Econometrics Journal*, 12:C1–C32, 2009.

G. Barone Adesi. VaR and CVaR implied in option prices. *Journal of Risk and Financial Management*, 9:2., 2016.

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

F. Black and M. Scholes. The pricing of options and corporate liabilities. *Journal of Political Economy*, 81:637–654, 1973.

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, page 1107. 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.

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.

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.

A. Harvey, E. Ruiz, and N. Shephard. Multivariate stochastic variance models. *The 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. *The Annals of Applied Statistics*, 8:595–621, 2014.

J. C. Hull. *Options, Futures and other Derivatives.* Pearson Prentice Hall, 7 edition, 2008.

H. Lütkepohl. *New Introduction to Multiple Time Series Analysis.* Springer Science & Business Media, 2005.

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

S. Mitra. The relationship between conditional Value at Risk and option prices with a closed-form solution. *The European Journal of Finance*, 21:400–425, 2015.

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

W. K. Newey and K. D. West. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55:703–708, 1987.

N. Nolde and J. F. Ziegel. Elicitability and backtesting: Perspectives for banking regulation. *Annals of Applied Statistics*, 2017. To appear.

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

R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2017. URL `https://www.R-project.org/`.

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.

C. Strähl and J. F. Ziegel. Cross-calibration of probabilistic forecasts. *Electronic Journal of Statistics*, 11:608–639, 2017.

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.

A. Zeileis. Econometric computing with HC and HAC covariance matrix estimators. *Journal of Statistical Software*, 11:1–17, 2004.

Appendix

A Proof of Proposition 2.1

Proof. The $\mathcal{F}_1$-consistency of S_{v_1} and S_{v_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

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

where H is a locally finite measure and $z \in \mathbb{R}$. As $G_2 \ge 0$, 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. $\square$

B Details on the permutation test

Here we provide implementation details for the permutation test introduced in Section 3.2. In Appendix D, we also apply the test for both types of elementary scoring functions. Therefore, we next describe the most general procedure which involves both elementary scores and thus two grids of parameters (for both v_1 and v_2). The simpler procedure considered Section 3.2 follows easily from the more general variant, by omitting the grid for v_1.

Stage 1

The pointwise tests are one-sided t-tests. The results in Sections 3.2 and 4 are based on the assumption that the score differences are independent over time (IOT); thus, the variance estimator entering the t-tests does not account for possible autocorrelation. In Appendix D, we present robustness checks using an autocorrelation-consistent Newey and West (1987) variance estimator, as implemented in the function `NeweyWest` of the R package `sandwich` (Zeileis, 2004), with a truncation lag of three.

Stage 2

Our method for correcting the pointwise p-values follows Westfall and Young (1993) and Cox and Lee (2008). We consider a total of $2M$ grid points for v_1 and v_2, leading to the grids $v_1^1, \ldots, v_1^M$ and $v_2^{M+1}, \ldots, v_2^{2M}$.[10] Let σ be the permutation of $\{1, \ldots, 2M\}$ such that $p(v_\cdot^{\sigma(1)}) \leq \ldots \leq p(v_\cdot^{\sigma(2M)})$; the subindex $\cdot$ equals either 1 or 2. Consider next two vectors of simulated p-values, $p^*(v_1^1), \ldots, p^*(v_1^M)$ and $p^*(v_2^{M+1}), \ldots, p^*(v_2^{2M})$, generated under the null hypothesis (see below). Define $q_m^* = \min\left\{p^*(v_\cdot^{\sigma(s)}) : s \geq m\right\}$. For example, q_1^* is the smallest of all simulated p-values, q_2^* is the smallest among the simulated p-values at grid points $v_\cdot^{\sigma(2)}, \ldots, v_\cdot^{\sigma(2M)}$, and so forth. We simulate L sets of p-values, obtaining values $q_{m,l}^*$ for $1 \leq m \leq 2M$ and $1 \leq l \leq L$. The adjusted p-values $r_1, \ldots, r_{2M}$ are finally obtained as

$$r_m = \frac{1}{L} \sum_{l=1}^{L} \mathbb{1}\left(q_{\sigma^{-1}(m),l}^* \leq p(v_\cdot^m)\right).$$

The minimal Westfall-Young p-value of the dominance test is given by $\min_{1 \leq m \leq 2M}\{r_m\}$. We reject the global null hypothesis if this minimum is smaller than α.

As suggested above, an important implementation aspect is how to enforce the null hypothesis when simulating the p-values. We do this by randomly permuting the labels of the forecasting methods A and B. Specifically, let

$$d_{v_1,t} \equiv \mathrm{S}_{v_1}(X_{t,1}^A, Y_t) - \mathrm{S}_{v_1}(X_{t,1}^B, Y_t)$$

denote the score difference between A and B, at time t, for the first elementary score, and

$$d_{v_2,t} \equiv \mathrm{S}_{v_2}(X_{t,1}^A, X_{t,2}^A, Y_t) - \mathrm{S}_{v_2}(X_{t,1}^B, X_{t,2}^B, Y_t)$$

denote the score difference between A and B, at time t, for the second elementary score. Under the H0 that A weakly dominates B, it holds that $\mathbb{E}(d_{v_1,t}) \leq 0$ and $\mathbb{E}(d_{v_2,t}) \leq 0$. At the boundary of the null hypothesis, it holds that $\mathbb{E}(d_{v_1,t}) = 0$ and $\mathbb{E}(d_{v_2,t}) = 0$. We enforce the latter equalities by simulating a sequence $s_t \in \{-1, +1\}, t = 1, \ldots, T$, and putting

$$\begin{aligned} d_{v_1,t}^* &= s_t \, d_{v_1,t}, \\ d_{v_2,t}^* &= s_t \, d_{v_2,t}; \end{aligned}$$

note that we use the same sign s_t for all values v_1, v_2, thus leaving the correlation structure of the grid points across v_1, v_2 intact. We then use the simulated time series $(d_{v_1,t}^*, d_{v_2,t}^*)$ to compute the pointwise p-values $p^*(v_1^1), \ldots, p^*(v_1^M)$ and $p^*(v_2^{M+1}), \ldots, p^*(v_2^{2M})$.

[10] The description in this paragraph loosely follows Strähl and Ziegel (2017, Section 6).

In our results in Sections 3.2 and 4, we draw the signs s_t independently across time t. This procedure is consistent with the IOT assumption that the score differences are not autocorrelated (see Section 3.2). In Appendix D, we present evidence that simulating the signs in blocks of length four (such that four consecutive periods t are multiplied with the same sign) yields similar test results.

C Additional figures

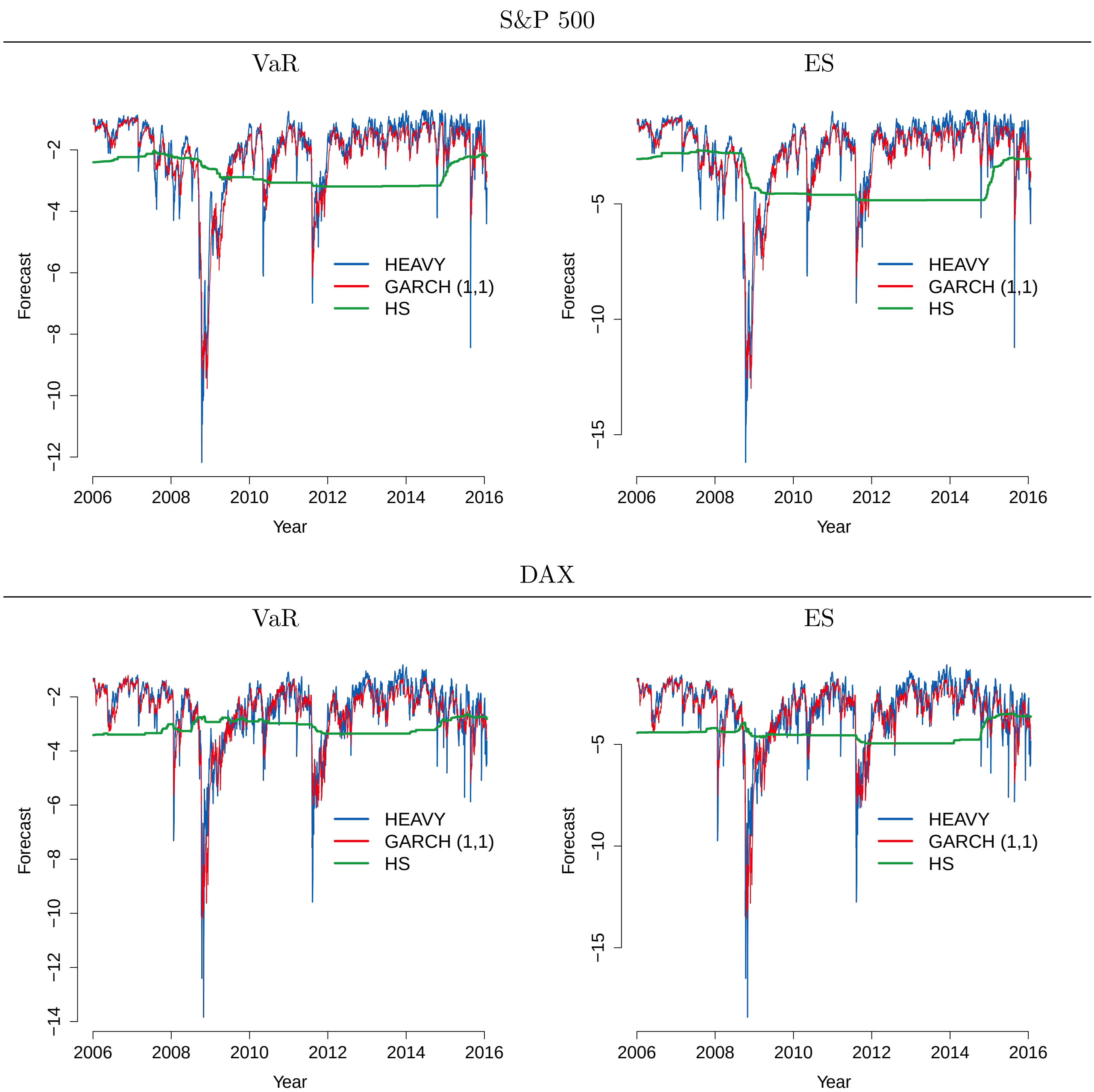

Figure 3: **Time series plots of empirical forecasts**. Left: Value-at-Risk, right: Expected Shortfall. The sample periods ranges from January 4, 2006 to January 25, 2016. See text for details.

D Robustness checks for the permutation test

Here we consider variations of the test described in Section 3.2 (and reported in Section 4), along two dimensions:

- *Assumptions on temporal dependence*
  - Option 1: Assume temporal independence of the score differences (as in the results in Section 4). Accordingly, we use independent sign permutations and do not account for possible autocorrelations in the pointwise t-tests.
  - Option 2: Allow for temporal independence of the score differences. That is, we use a fixed block length of four when drawing the sign permutations, and a corresponding lag length of three in the pointwise t-tests.
- *Set of elementary scores considered*
  - Option 1: Use all elementary scores.
  - Option 2: Use only the elementary scores corresponding to the second summand in Equation (3), i.e., only the scores involving both VaR and ES (as in the results in Section 4).

Independence, both elementary scores		*Independence, 2nd elementary score only*	
Hypothesis	P-value	Hypothesis	P-value
HS weakly dominates HEAVY	0.000	HS weakly dominates HEAVY	0.000
HEAVY weakly dominates HS	0.008	HEAVY weakly dominates HS	0.772
GARCH weakly dominates HEAVY	0.000	GARCH weakly dominates HEAVY	0.000
HEAVY weakly dominates GARCH	0.008	HEAVY weakly dominates GARCH	0.998

Dependence, both elementary scores		*Dependence, 2nd elementary score only*	
Hypothesis	P-value	Hypothesis	P-value
HS weakly dominates HEAVY	0.00	HS weakly dominates HEAVY	0.00
HEAVY weakly dominates HS	0.43	HEAVY weakly dominates HS	0.79
GARCH weakly dominates HEAVY	0.00	GARCH weakly dominates HEAVY	0.00
HEAVY weakly dominates GARCH	0.35	HEAVY weakly dominates GARCH	1.00

Table 3: **Test results for empirical forecasts (S&P 500)**. The table presents several variants of the permutation test, see text for details.

Independence, both elementary scores

Hypothesis	P-value
HS weakly dominates HEAVY	0.000
HEAVY weakly dominates HS	0.000
GARCH weakly dominates HEAVY	0.050
HEAVY weakly dominates GARCH	0.094

Independence, 2nd elementary score only

Hypothesis	P-value
HS weakly dominates HEAVY	0.000
HEAVY weakly dominates HS	0.876
GARCH weakly dominates HEAVY	0.174
HEAVY weakly dominates GARCH	0.924

Dependence, both elementary scores

Hypothesis	P-value
HS weakly dominates HEAVY	0.000
HEAVY weakly dominates HS	0.072
GARCH weakly dominates HEAVY	0.524
HEAVY weakly dominates GARCH	0.616

Dependence, 2nd elementary score only

Hypothesis	P-value
HS weakly dominates HEAVY	0.000
HEAVY weakly dominates HS	0.846
GARCH weakly dominates HEAVY	0.146
HEAVY weakly dominates GARCH	0.910

Table 4: **Test results for empirical forecasts (DAX)**. The table presents several variants of the permutation test, see text for details.

E Additional calculations for Section 5

Here we establish the equivalence between Equation (7) and the Black and Scholes (1973) pricing model as noted in Section 5. To this end, we express Equation (7) in terms of three factors: The strike price (x_1), the current spot price of the underlying asset (y_0), and the time to maturity (t). We proceed as follows:

- In Equation (7), set $\mathrm{VaR}_\alpha(F) = x_1$. This step follows from the optimality condition described in the text.
- Since F follows a lognormal distribution, we have that

$$\alpha = \int_{z=-\infty}^{x_1} d\ F(z) = \Phi\left(\frac{\ln x_1 - \ln y_0 + 0.5\ \tau^2\ t}{\tau\sqrt{t}}\right),$$

  where Φ is the cumulative distribution function of the standard normal distribution.
- Finally, compute $\mathrm{ES}_\alpha(F) = \mathbb{E}(Y|Y < x_1)$. To this end, note that if Y follows a log-normal distribution, then $(\ln Y|\ln Y < \ln x_1)$ follows a truncated normal distribution. Using the moment-generating function of the latter distribution, we obtain

$$\mathrm{ES}_\alpha(F) = \frac{y_0}{\alpha}\ \Phi\left(\frac{\ln x_1 - \ln y_0 - 0.5\ \tau^2\ t}{\tau\sqrt{t}}\right).$$

- Collecting terms, we find that

$$P = x_1\ \Phi\left(\frac{\ln x_1 - \ln y_0 + 0.5\ \tau^2\ t}{\tau\sqrt{t}}\right) - y_0\ \Phi\left(\frac{\ln x_1 - \ln y_0 - 0.5\ \tau^2\ t}{\tau\sqrt{t}}\right);$$

the latter formula is equal to the Black-Scholes put price in Equation (13.21) of Hull (2008) if the risk-free interest rate is zero. $\square$